

Paul Shapshak · Seetharaman Balaji  
Pandjassaram Kangueane  
Francesco Chiappelli · Charurut Somboonwit  
Lynette J. Menezes · John T. Sinnott *Editors*

# Global Virology III: Virology in the 21st Century

# Global Virology III: Virology in the 21st Century

Paul Shapshak • Seetharaman Balaji  
Pandjassaram Kangueane • Francesco Chiappelli  
Charurut Somboonwit • Lynette J. Menezes  
John T. Sinnott  
Editors

# Global Virology III: Virology in the 21st Century



*Editors*

Paul Shapshak

Department of Internal Medicine

University of South Florida

Tampa, FL, USA

Pandjassarame Kangueane

Biomedical Informatics 17,

Irlan Sandy Annex

Pondicherry, Pondicherry, India

Charurut Somboonwit

Department of Internal Medicine

University of South Florida

Tampa, FL, USA

John T. Sinnott

Department of Internal Medicine

University of South Florida

Tampa, FL, USA

Seetharaman Balaji

Department of Biotechnology

Manipal Institute of Technology,

Manipal Academy of Higher Education

Manipal, Karnataka, India

Francesco Chiappelli

Oral Biology and Medicine, CHS 63-090

UCLA School of Dentistry Oral Biology

and Medicine, CHS 63-090

Los Angeles, CA, USA

Lynette J. Menezes

Department of Internal Medicine

University of South Florida

Tampa, FL, USA

ISBN 978-3-030-29021-4

ISBN 978-3-030-29022-1 (eBook)

<https://doi.org/10.1007/978-3-030-29022-1>

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG

The registered company address is: Gwerbestrasse 11, 6330 Cham, Switzerland

# Preface

Viral diseases persist and develop with global increased risks for morbidity and mortality, in addition to social and financial disruption. Envelopment by viral disease spread due to global warming is a well-established contributor to these dire straits through increased vector range with multiple viral/microbial spread in addition to social disharmony and concomitant reduction of standards of living.

Answering the need for enhanced methods of study assists research establishments to accelerate scientific progress. This book provides readers with snapshots of where various fields are, so that they may be assisted as need be, to join this progress into the twenty-first century. The book is hopefully of help for professionals, students, and faculty, as well as for the interested reader.

We acknowledge and thank Alison Ball and Deepak Ravi of Springer Publishers for their help and guidance through the steps leading to the production of this book.

Tampa, FL, USA

Paul Shapshak

Manipal, Karnataka, India

Seetharaman Balaji

Pondicherry, Pondicherry, India

Pandjassarame Kangueane

Los Angeles, CA, USA

Francesco Chiappelli

Tampa, FL, USA

Charurut Somboonwit

Tampa, FL, USA

Lynette J. Menezes

Tampa, FL, USA

John T. Sinnott

# Introduction

Since the publication of *Global Virology* Volumes I and II, the need for *Global Virology* III became apparent because of the increased use and need of novel and forward-looking methods and techniques to accelerate virology research achievements around the globe [1–3].

The use of advanced methods for virology are accomplished, *ab initio*, as well as by using techniques agglomerated from many different fields, and are thereby used to accelerate application of what is relevant and useful for virology and human health [4–6].

This book provides views of work that has been undertaken and is planned in several fields of virology and is meant to promote current and future work, research, and health. Various fields and methods include virology, immunology, space research, astrovirology/astrobiology, plasmids, swarm intelligence, bioinformatics, data mining, machine learning, neural networks, critical equations, and advances in biohazard biocontainment. The use of novel and forward-looking methods, techniques, and approaches in research and development is promoted in this new book.

## References

1. Shapshak P, Somboonwit C, Kuhn J, Sinnott JT, editors. *Global virology I. Identifying and investigating viral diseases*. New York: Springer; 2015.
2. Shapshak P, Levine AJ, Somboonwit C, Foley BT, Singer E, Chiappelli F, Sinnott JT, editors. *Global virology II. HIV and NeuroAIDS*. New York: Springer; 2017.
3. Shapshak P, Balaji S, Kangueane P, Somboonwit C, Menezes L, Sinnott JT, Chiappelli F, editors. *Global virology III. Virology in the 21st century*. New York: Springer; 2019.
4. Girimonte D, Izzo D. Chapter 12: Artificial intelligence for space applications. In: Schuster AJ, editor. *Intelligent computing everywhere*. London: Springer; 2007. p. 235–53.
5. <https://www.esa.int/gsp/ACT/doc/AI/pub/ACT-RPR-AI-2007-ArtificialIntelligenceForSpaceApplications.pdf>.
6. Narayanan A, Keedwell EC, Olsson B. Artificial intelligence techniques for bioinformatics. *Appl Bioinform.* 2002. <https://pdfs.semanticscholar.org/bf9b/799a81e51a5cfa683f446cc5bca59e02d1e.pdf>.

# Contents

<b>Applications of Artificial Intelligence and Machine Learning in Viral Biology .....</b>	1
Sonal Modak, Deepak Sehgal, and Jayaraman Valadi	
<b>Non-immune Modulators of Cellular Immune Surveillance to HIV-1 and Other Retroviruses: Future Artificial Intelligence-Driven Goals and Directions.....</b>	41
Francesco Chiappelli, Allen Khakshooy, and Nicole Balenton	
<b>Emerging Technologies for Antiviral Drug Discovery.....</b>	59
Badireddi Subathra Lakshmi, Mohan Latha Abillasha, and Pandjassarame Kangueane	
<b>Wavelet-based Multifractal Spectrum Estimation in Hepatitis Virus Classification Models by Using Artificial Neural Network Approach .....</b>	73
Yeliz Karaca	
<b>Computational Coarse Protein Modeling of HIV-1 Sequences Using Evolutionary Search Algorithm .....</b>	97
Sandhya Parasnath Dubey and Seetharaman Balaji	
<b>Drug Development for Hepatitis C Virus Infection: Machine Learning Applications.....</b>	117
Sajitha Lulu Sudhakaran, Deepa Madathil, Mohanapriya Arumugam, and Vino Sundararajan	
<b>Modern Developments in Short Peptide Viral Vaccine Design.....</b>	131
Christina Nilofer, Mohanapriya Arumugam, and Pandjassarame Kangueane	
<b>Artificial Life and Therapeutic Vaccines Against Cancers that Originate in Viruses .....</b>	149
María Elena Escobar-Ospina and Jonatan Gómez	

<b>Mystery of HIV Drug Resistance: A Machine Learning Perspective.....</b>	307
Mohanapriya Arumugam, Nirmaladevi Ponnusamy, Sajitha Lulu Sudhakaran, Vino Sundararajan, and Pandjassarame Kangueane	
<b>Swarm Intelligence in Cell Entry Exclusion Phenomena in Viruses and Plasmids: How to Exploit Intelligent Gene Vector Self-Scattering in Therapeutic Gene Delivery .....</b>	325
Oleg E. Tolmachov	
<b>A Combinatorial Computational Approach for Drug Discovery Against AIDS: Machine Learning and Proteochemometrics .....</b>	345
Sofia D'souza, Prema K. V., and Seetharaman Balaji	
<b>Application of Support Vector Machines in Viral Biology .....</b>	361
Sonal Modak, Swati Mehta, Deepak Sehgal, and Jayaraman Valadi	
<b>Eliminating Cervical Cancer: A Role for Artificial Intelligence.....</b>	405
Lynette J. Menezes, Lianet Vazquez, Chilukuri K. Mohan, and Charurut Somboonwit	
<b>HIV and Injection Drug Use: New Approaches to HIV Prevention.....</b>	423
Charurut Somboonwit, Lianet Vazquez, and Lynette J. Menezes	
<b>Innovative Technologies for Advancement of WHO Risk Group 4 Pathogens Research.....</b>	437
James Logue, Jeffrey Solomon, Brian F. Niemeyer, Kambez H. Benam, Aaron E. Lin, Zach Bjornson, Sizun Jiang, David R. McIlwain, Garry P. Nolan, Gustavo Palacios, and Jens H. Kuhn	
<b>Space Exploration and Travel, Future Technologies for Inflight Monitoring and Diagnostics.....</b>	471
Jean-Pol Frippiat	
<b>Futuristic Methods in Virus Genome Evolution Using the Third-Generation DNA Sequencing and Artificial Neural Networks .....</b>	485
Hyunjin Shim	
<b>Futuristic Methods for Treatment of HIV in the Nervous System.....</b>	515
Allison Navis and Jessica Robinson-Papp	
<b>Tuberculosis: Advances in Diagnostics and Treatment .....</b>	529
Ju Hee Katzman, Mindy Sampson, and Beata Casañas	
<b>Astrovirology, Astrobiology, Artificial Intelligence: Extra-Solar System Investigations .....</b>	541
Paul Shapshak	

Contents	xi
<b>Climate Crisis Impact on AIDS, IRIS and Neuro-AIDS . . . . .</b>	<b>575</b>
Francesco Chiappelli, Emma Reyes, and Ruth Toruño	
<b>21st Century Virology: Critical Steps . . . . .</b>	<b>605</b>
Paul Shapshak	
<b>Futuristic Methods for Determining HIV Co-receptor Use . . . . .</b>	<b>625</b>
Jacqueline K. Flynn, Matthew Gartner, Annamarie Laumaea, and Paul R. Gorry	
<b>Index . . . . .</b>	<b>665</b>

# Applications of Artificial Intelligence and Machine Learning in Viral Biology



Sonal Modak, Deepak Sehgal, and Jayaraman Valadi

**Abstract** Present research efforts coupled with improved experimental techniques have provided voluminous genomic data. To convert this data into useful knowledge, novel tools for phenomenological and data driven modelling approaches are needed. This need has spurred initiation of a lot of rigorous efforts and has resulted in development of robust artificial intelligence (AI) and machine learning (ML) based models. While these paradigms individually and in synergistic combinations have been employed in various bioinformatics applications, the viral biology discipline has particularly benefitted most. These methodologies can efficiently handle single dimensional sequence to higher dimensional protein structures, microarray data, image and text data, experimental data emanating from spectroscopy, etc. Our analysis deals with ML tools like support vector machines (SVM), neural networks, deep neural networks, random forest, and decision tree. Analysis and interpretations are provided along with ample illustrations of their relevance to real-life applications. AI and evolutionary computing based tools like Genetic Algorithms, Ant Colony optimization, Particle swarm optimization and their applicability to viral biology problems are also discussed. Hybrid combination of these tools with ML techniques have resulted in simultaneous selection of informative attributes and high performance classification. This hybrid methodology has been discussed in detail.

In this chapter we describe the applications artificial intelligence and machine learning in virology. While there are AI has a multitude of tools, the focus would be

---

S. Modak

Life Sciences and Healthcare Unit, Persistent Systems Inc., Santa Clara, California, United States

D. Sehgal

Life Sciences Department, School of Natural Sciences (SoNS), Shiv Nadar University,  
Greater Noida, Uttar Pradesh, India

e-mail: [deepak.sehgal@snu.edu.in](mailto:deepak.sehgal@snu.edu.in)

J. Valadi (✉)

Center for Informatics, School of Natural Sciences (SoNS), Shiv Nadar University,  
Gautham Buddha Nagar, Uttar Pradesh, India

Centre for Modelling and Simulation, Savitri Bai Phule Pune University, Pune, India  
e-mail: [jayaraman.valadi@snu.edu.in](mailto:jayaraman.valadi@snu.edu.in); [jayaraman@cms.unipune.ac.in](mailto:jayaraman@cms.unipune.ac.in)

on a specific aspect of AI, known as evolutionary and heuristic computing. These are mainly employed as an alternative paradigm of optimization. They are mainly nature inspired algorithms. Although very simple and straightforward to use they have been deputed to solve several problems successfully in different domains of science and engineering. Machine learning on the other hand deals with a mountain of available data, recognize hidden patterns useful and interesting to upgrade it to structure and knowledge. We provide examples of the power of AI and Machine learning with the illustration of several examples from different subdomains of viral biology. We will also provide examples where the synergistic combination of AI and ML has been found to be a very potent tool for solving several important problems in viral biology.

**Keywords** Decision trees · Random Forest algorithm · Neural networks  
Activation functions · Convolutional neural networks · Genetic algorithms  
Ant Colony optimization · Particle swarm optimization · Attribute selection viral  
biology

## 1 Introduction

Machine learning has a rich collection of ever-increasing algorithms. While we have explained the use of Support Vector Machine (SVM) in virology in another chapter in detail, in this chapter we elucidate the desirable properties of three high performance algorithms, viz., Decision tree, Random forest (RF), Neural networks including deep architecture. Decision tree repeatedly splits attributes starting from a head node to the decision nodes known as leaf nodes. The results can be interpreted in terms of easily explainable form with domain attributes. Random forest is a collection of large number of decision tree algorithms. Randomness is introduced in random forests in two ways; (1) in each tree, bootstrap sampled examples form the input and (2) in every tree only a subset of randomly selected attributes are used . The final decision is based on majority decision of individual trees. Random forest reduce the variance of performance measures while maintaining the desirable low bias of decision trees. Neural Networks are connected by the information flow through a network of neurons. They mimic the combined action of neurons in the brain. Conventional architecture contains a layer of hidden neurons connecting the input. Recently deep neural networks with large number of hidden layers have been proposed to solve problems with huge amounts of text and image data. Several configurations have been proposed and Convolution neural networks (CNN) are most widely used.

Evolutionary and heuristic methods form a subset of AI methods. These methods have been successfully employed in biological domain with great success. These methods are employed as optimization tools which differ from conventional mathematical programming methods. While conventional methods are mainly gradient

based methods, evolutionary and heuristic computational methods do not require the evaluation of derivatives. They are simple to use but have rigorous basis and produce reasonably good solutions without having to formulate difficult model equations. In this work we have described mainly three methods, viz., genetic algorithms, Ant Colony Optimization and particle swarm optimization. All these methods are population based and provide several equally good solutions and allow the user to choose the solutions most useful. GA is inspired by natural evolution and uses the selection, crossover and mutation mechanisms to iteratively update and arrive at the best possible solution(s). Ant colony optimization is inspired by the cooperative search behaviour of real life ants. Almost blind ants are able to cooperatively carry out several tasks including optimizing their route to food source and back is due to their capabilities to deposit a chemical known as pheromone . They also get attracted to the pheromone rich trail and enhance the shortest trail in an auto catalytic feedback manner. The swarm behaviour is differently portrayed in Particle swarm optimization where the artificial swarm particles mimic the way in which the real life birds cooperatively synergise their movement adjusting their speed with the swarm. We have elaborated the algorithms of each method, both machine learning and Artificial Intelligence. We have also provided examples to illustrate the use of these algorithms in biology.

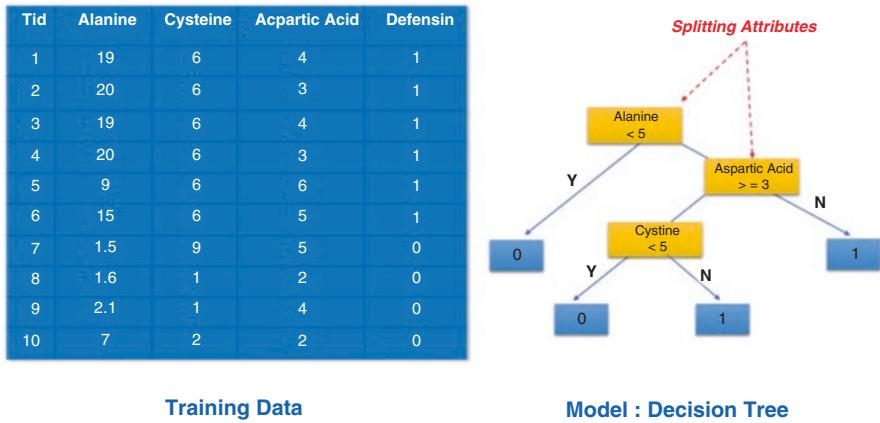
## 2 Decision Tree Algorithm

Decision trees are a class of learning algorithms employed for classification and regression [1, 2]. Starting with a given data set it breaks the set into smaller and smaller subsets simultaneously growing the decision tree. The final tree consists of a head node, intermediate decision nodes and leaf nodes. The leaf nodes provide final outputs and for a classification problem it is the predicted class of any given example. Each example is sent through a tree starting from the head node until the final leaf node following the appropriate branch as per the condition satisfied. For regression problem it is the predicted real value for any given example A two way split of a node results in two children nodes while a three-way split result in three children nodes. Multiway splits are also possible. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor attribute is called the root node. Decision trees can handle both categorical and numerical data.

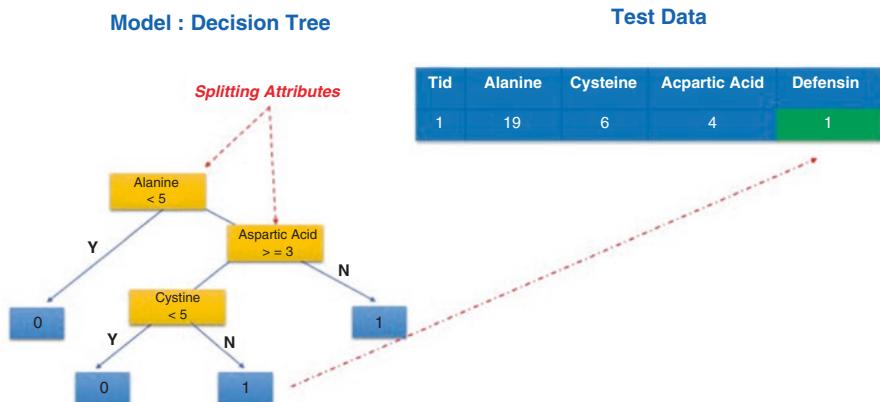
Decision Trees repeatedly split attributes starting from the head node until a leaf node is obtained. For the head node the most informative attribute and the split position is obtained by using different performance measures like Gini index, mutual information and misclassification error. For example, if attribute values lie between 0–10 the attribute is split at different split points 2, 4, 6 and 8 one by one. Using appropriate logical conditions the goodness of splits are evaluated using different performance measures and the split point with best performance measure and best informative attribute is used to split the head node. At every intermediate node posi-

tion the split point and most informative attribute are found in a similar fashion. The leaf node and stopping of splitting are ascertained by different stopping conditions. For example, if all the attributes have similar values or number of examples coming to particular node is less than predetermined value, then splitting is stopped. In this way a decision tree is built.

Using a function annotation problem in viral biology We can illustrate the working of a decision tree. We are given a data set of defensin peptides (denoted by zero class) and non-defensin peptides (denoted by one class). We are provided the alanine, cysteine and aspartic acid concentrations as attributes for each peptide. The final grown tree model is shown in Fig. 1. While, Fig. 2 shows how a test example can be sent through the tree down until the leaf node to determine the functional class of a test peptide.



**Fig. 1** Decision Tree example for functional annotation



**Fig. 2** Determination of functional class of test data by Decision Tree

## 2.1 Applications of Decision Trees in Virology

A decision tree is used as a classifier for determining an appropriate action (among a predetermined set of actions), which makes it preferred method since it's a common scenario in various problem statements in computational biology [1]. A case related to virology which can be considered as an example is to predict diagnosis and outcome of an illness caused by viral infection. Dengue viruses are responsible for causing dengue fever/dengue haemorrhagic fever (DF/DHF), transmitted by a Aedes aegypti mosquitos as vectors [3]. In the early phase of Dengue illness is often confused with febrile illnesses due to its nonspecific clinical symptoms, but the symptoms in later stage of illness are more definitive. Correct diagnosis of dengue in early phase requires laboratory tests which are costly [4, 5]. There are studies attempting diagnosis of dengue disease univariate or multivariate analysis of clinical symptoms and signs, haematological or biochemical parameters [6, 7]. Lukas Tanner et al. worked on developing an algorithm with decision tree approach which can efficiently diagnose dengue in early hours of illness [8]. Clinical data from different age groups and various time points of infection was collected. C4.5 decision tree classifier [9] was used by the authors and pruning confidence of 25% was used to remove branches. To overcome data over-fitting, the algorithms were validated using the k-fold cross validation approach [10] where fold value was set to 10 ( $k = 10$ ). Receiver-operating characteristic (ROC) curve was constructed to quantify the sensitivity and specificity of the decision algorithm. The overall error rate estimated after k-fold cross validation was 15.7%, with a sensitivity and specificity of 71.2% and 90.1%, respectively. In summary, diagnostic algorithm was able to differentiates dengue from non-dengue febrile illness with an accuracy of 84.7%.

Another dreadful virus is West Nile Virus (WNV) which can cause chronic medical conditions and even death after severe infection [11]. Similar to Dengue virus, mosquitos are vectors of disease transmission for WNV in humans, but other known modes for this virus is through blood transfusion, breastfeeding, transplacental transmission, occupational exposure in laboratory workers and stem cell and solid organ transplantation [12]. In January 2004, the Organ Procurement and Transplantation Network (OPTN) and the Health Resources and Services Administration (HRSA) released their recommendations on the role of deceased donor screening in [13]. They recommended to reject donor from geographic areas affected by WNV infections. Thus, Bryce A. Kiberd et al. demonstrated use of medical decision analysis to decide whether or not to implement deceased donor WNV screening by integrating differences in the type of organ transplanted, WNV disease prevalence, test characteristics and survival on the wait list [14]. The results of their analysis showed the potential loss of 452.4 life years (cumulative for heart, liver and kidney) due to screening annually, since most positive test results would be false-positive.

In another example decision tree was used to evaluate the performance of commercial software used for clinical diagnosis. SELDI (surface-enhanced laser desorption/ionization) is mass spectrometry proteomic approach developed recently which

potentially can help in biomarker discovery [15, 16]. Attempts have been made for associating such biomarkers with various types of cancers [17–20]. Recent in many studies SELDI data has been used along with machine learning algorithms in identifying protein fingerprints specific for particular cancer which can be effectively used to accurately differentiate cancer from the noncancer groups [21–24]. Antonia Vlahou et al. attempted to evaluate the classification algorithm called biomarker pattern software [BPS], which is commercially available for analysis of the SELDI serum protein profiling data [25]. Total 139 serum sample, 124 were considered for this study out of which 85 were controls and 39 were cancer samples. Randomly set of 15 was selected as learning set out of which 10 were controls and 5 were cancers to form test set for the algorithm. Decision tree that was generated from the learning set to classify the two groups. For evaluation the accuracy of the algorithm in predicting ovarian cancer, ten-fold cross-validation analysis was performed. It yielded 80% of specificity and 84.6% of sensitivity. When test set was processed by the algorithm, 80% of sensitivity and specificity was obtained. In conclusion, this study highlighted some advantages of BPS software and also pointed out some drawbacks like it is prone to data overfitting.

As in many other areas, decisions play an important role also in medicine, especially in medical diagnostic processes. Decision support systems helping physicians are becoming a very important part in medical decision making, particularly in those situations where decision must be made effectively and reliably. Since conceptual simple decision making models with the possibility of automatic learning should be considered for performing such tasks, decision trees are a very suitable candidate. They have been already successfully used for many decision-making purposes. As in many other areas, decisions play an important role also in medicine, especially in medical diagnostic processes. Decision support systems helping physicians are becoming a very important part in medical decision making, particularly in those situations where decision must be made effectively and reliably. Since conceptual simple decision making models with the possibility of automatic learning should be considered for performing such tasks, decision trees are a very suitable candidate. They have been already successfully used for many decision making purposes.

### 3 Random Forest Algorithm

Random Forest (RF) is an ensemble of randomly constructed independent (and unpruned i.e. fully grown) decision trees [26–28]. It uses bootstrap sampling technique, which is an improved version of bagging. Each tree differs from all others owing to the randomness introduced in RF algorithm in two ways: one in the sample dataset for growing the tree and the other in the choice of the subset of attributes for node splitting while growing each tree. Such a RF is grown in the following manner:

1. From the training data of  $n$  examples, draw a bootstrap sample (i.e., randomly sample, with replacement, ' $n$ ' examples).

2. For each bootstrap sample, grow a regression tree with the following modification: at each node, choose the best split among a randomly selected subset of m (rather than all) features. Each tree is grown to the maximum size.
3. Repeat the above steps until (a sufficiently large number) N such trees are grown.

For each tree, a bootstrap sample (with replacement) is drawn from the original training data set, i.e. a sample is taken from the training data set and is then replaced again in the data set before drawing the next sample. Likewise, ‘n’ numbers of samples are taken to form ‘In-Bag’ data for a particular tree, where ‘n’ is the size of the training data set. The main advantage of bootstrap sampling is to avoid over fitting the training data. In each of the Bootstrap training sets, about one-third of the instances are unused for making the ‘In Bag’ data on an average and these are called the Out-Of-Bag (OOB) data for that particular tree. The decision tree is induced using this ‘In-Bag’ data using the CART (Classification and Regression Trees) algorithm [2].

Pruning is not necessary in RF, since bootstrap sampling takes care of the over fitting problem. This further reduces the computational load of the RF algorithm. There is no need for a separate test data in RF for checking the overall accuracy of the forest. It uses the OOB data for cross validation. After all the trees are grown, the  $k^{\text{th}}$  tree classifies the instances that are OOB for that tree (left out by the  $k^{\text{th}}$  tree). In this manner, each case is classified by about one third of the trees. A majority voting strategy is then employed to decide on the class affiliation of each case. The proportion of times that the voted class is not equal to the true class of case-‘n’, averaged over all the cases in the training data set is called as the OOB error estimate. Now after growing the forest, if an unseen validation test dataset is given for regression, each tree in the Random Forest contributes a unit vote. The output of the classifier is determined by a majority vote of the trees. The prediction error rate of the forest, depends on the strength of each tree and the correlation between any two trees in the forest. The key to higher prediction accuracy is to keep low bias and low correlation among the trees. This may be done by adjusting the number of variables randomly selected for each tree (mtry). If the value of ‘mtry’ is decreased, the strength of each tree decreases, but with increase in ‘mtry’ the correlation among the trees increases and the computational load may also increase. The default value of ‘mtry’ is chosen as  $M/3$  for regression problems and  $\sqrt{M}$  for classification problems, where ‘M’ is the total number of attributes.

The important features of Random Forests are that they can handle most high dimensional and multi-class data easily and the threshold noise limit is more for Random Forest compared to the other algorithms. It can be used even if the number of attributes is more than the number of examples.

### **3.1 Variable Selection Using Random Forests**

Random Forest can also be used to get an estimate of the variables that are less important for prediction. All the cases that are OOB for a particular tree are put down the tree to get a prediction with some votes. Now to get an estimate of vari-

able importance, the value of each of the attributes is randomly permuted in the OOB cases of a particular tree and the decrease in the number of votes for the majority voted class is calculated. This decrease in the number of votes, when averaged over all the trees in the forest, gives the raw importance score for that variable. So, higher the raw importance score, greater is the importance of that variable in classification. Thus the raw importance score can be employed for feature ranking.

### 3.2 Applications of Random Forests in Virology

With the knowledge of all aspects of Random Forest technique, it can effectively use as a tool to construct prediction models for problems in virology domain. One of the most notorious viral strains in influenza A, responsible for at least one major episode of global health threat in a decade. It occasionally breaks the restriction barrier of the primary host, which is mostly animal populations, and infect humans leading to potential pandemic.

Host tropism is a property of viruses which defines its infection specificity to particular hosts and host tissues. Thus, it explains why viruses are only capable of infecting a limited range of host organisms. To greater extent the species barrier restricts influenza strains to infect other hosts since new viral stains needs to overcome host range restriction to adapt to a new host. Most important determinant of tropism is hemagglutinin protein (HA) receptor specificity on host cells. Studies have already revealed the preference of stains affecting humans recognizes a2,6-sialic acid linkage while avian strains preferentially bind receptors of a2,3-sialic acid linkages [29–31]. The second most crucial determinant is PB2 subunit of viral polymerase complex. Host range of influenza viruses can be efficiently determined by the amino acid residue residing at position 627 in PB2 [32–34]. Apart from these important factors, comparison of genomic signatures of the hosts [35] and position specific mutations might be explored to evaluate the capability of avian stains infecting humans. Christine LP Eng et al. studied host tropism of influenza A virus proteins using random forest [36]. A combined prediction model was trained using 3272 positive human samples and 3923 negative avian samples, while 799 positive samples and 989 negative samples used as external testing dataset. These proteins sequences were transformed into feature vectors extraction their physicochemical properties. Twenty feature vectors were derived from composition of each of the 20 standard amino acids. The next step of transformation was performed using a method developed by Dubchak et al., in which three descriptors: composition, transition, and distribution, were calculated to globally describe amino acid properties [37, 38]. Training of Random Forest prediction models were conducted using ten-fold cross-validation, where entire dataset is divided into 9 training subsets and 1 testing subset. Grid search approach was employed to fine tune the parameters for best performance. In this comparative study, Random Forest outperformed over Naïve Bayes, k-Nearest

Neighbours algorithm (kNN), SVM and Artificial Neural Network (ANN) classifiers, yielding 98.58% prediction accuracy (AUC = 0.996; MCC = 0.972), and hence was chosen as the classifier to train the remaining prediction models for individual proteins.

In another study Yu Wei et al. demonstrated effective use of Random Forest technique in discovery of novel potent targets for developing new drugs to block virus infection [39]. The viral species targeted for this study was hepatitis C virus (HCV), because its chronic infection can result in chronic liver disease, progressing to cirrhosis and hepatocellular carcinoma [40]. There is urgent need to develop new anti-HCV drugs because of several critical issues with current HCV therapies, which includes side effects and drug resistances [41]. HCV NS5B polymerase is an RNA-dependent RNA polymerase which plays an important role in replication process of genomic RNA of HCV [42, 43]. Current studies based on X-ray structures of inhibitor-bound HCV NS5B polymerase [44, 45] is proving extremely informative in discovering and developing of new structure-based NS5B polymerase inhibitors. Authors developed a virtual screening workflow that includes random forest, e-pharmacophore, and molecular docking methods to discover a series of novel small molecule NS5B polymerase inhibitor leads. Random Forest method was first used to build the predictive models of the NS5B polymerase inhibitors. Sixteen descriptors were selected, and the overall classification accuracy of the model was 84.4%. The outcome of this study was 5 compounds which showed inhibitory potency against NS5B polymerase with IC<sub>50</sub> value of 2.01–23.84 μM. Furthermore these compounds further optimized and developed to be potent and highly active NS5B polymerase inhibitors.

Some studies correlated the increase in incidences of hepatocellular carcinoma (HCC) with increased prevalence of HCV infection [46–48]. The significance of HCV viral infection in the pathogenesis of HCC can be validated by understanding the transition of liver tissues from benign to malignant. Valeria R Mas et al. studied the gene expression patterns of 108 liver tissue samples at different stages, including normal, cirrhosis, and different HCC stages [49]. For 58 HCV cirrhotic tissues, 863 differentially expressed probe sets were yielded by comparing cirrhotic tissues with (n = 17) and without (n = 41) HCC. There was a need of a classifier to predict whether the HCV cirrhotic tissue was from a patient without HCC versus cirrhotic tissue with HCC. Fifteen probe sets were consistently identified among the random forest classifiers, which helped authors to identify gene signatures that distinguish the pathological stages of HCC and potential molecular markers for early HCC diagnosis in high risk cirrhotic HCV patients.

The predictive power of Random Forest method can also be employed in development of time series models in disease prediction. A comparative analysis of viral outbreak data was performed by Michael J Kane et al. between an autoregressive integrated moving average (ARIMA) model and Random Forests model [50]. Time series models of both the methods was applied to outbreaks incidence data of avian influenza (H5N1) virus in Egypt. Authors not only found Random Forest model outperforming the ARIMA model in predictive ability, but also inferred that it effective for predicting outbreaks of H5N1 in Egypt.

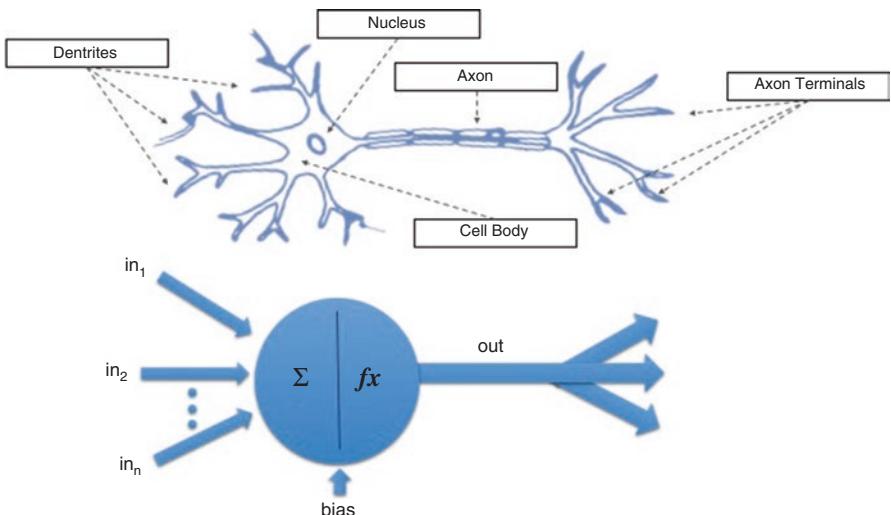
## 4 Neural Network Algorithm

A Neural Network is an artificial intelligence tool which mimics human brain for carrying out useful tasks rapidly [51]. More specifically, ANN is inspired by human information processing through the interaction of many billions of neurons connected to each other. Figure 3 illustrates how the neural network algorithm is inspired by the properties of brain cells and its analogy with the actual functioning of the neurons. A typical dendrite in the human brain receives signals from other neurons and cell body sums the incoming signals and when the sum exceeds a threshold value, neuron fires and the signal is transmitted through axons to other neurons. The signal quantity is proportional to the strength of the connections which can be inhibitory or excitatory.

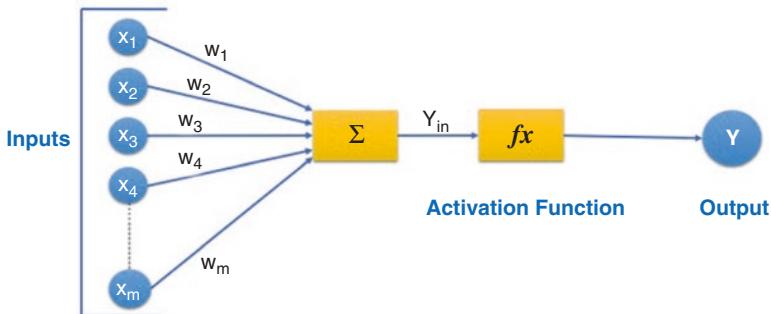
ANNs mimic this cooperative functioning of the neurons by connecting the inputs of a given data (input neurons) to the required outputs of a specific task through a series of layers of neurons. The structure of a standard neural network architecture consists of input, weights, activation function hidden layers of neurons and outputs.

### 4.1 Model of Artificial Neural Network

A general model of ANN is schematically represented in Fig. 4, followed by its processing. For the above general model of artificial neural network, the net input can be calculated as follows:



**Fig. 3** Diagrammatic representation of neural network



**Fig. 4** General model of an Artificial Neural Network

$$y_{in} = x_1 \cdot w_1 + x_2 \cdot w_2 + x_3 \cdot w_3 + \dots + x_m \cdot w_m$$

i.e., Net input is:

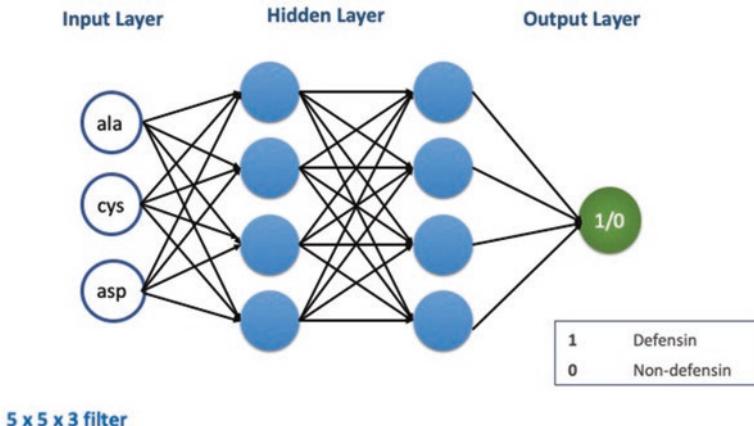
$$y_{in} = \sum_i^m x_i \cdot w_i$$

The output can be calculated by applying the activation function over the net input. Hence, output is the net function of the net input. Activation functions are used to achieve non-linear functional mapping. Such non-linear mapping is necessary for handling data which are not linearly classifiable.

Some commonly used activation functions are:

- (a) Sigmoid or Logistic
- (b) Tanh (Hyperbolic tangent)
- (c) ReLu (Rectified linear units)

A typical ANN architecture consists of an Input layers, 1 or 2 Hidden layers and 1 or multiple Output layers. For the Defensin classification problem illustrated in Fig. 2 the input layer denotes the concentrations of Alanine, Cysteine and Aspartic Acid amino acids. In Fig. 5, there are 4 hidden neurons in each of the two layers. The inputs are weighted and then sent to each of the neurons in the first hidden layer. These are summed, squashed (non-linearly mapped), weighted and then sent to the next hidden layer of neurons. These are summed and further squashed by activation functions, summed up and sent to the output layer. Every input example is sent through the layers, following the same procedure. The network output is compared with the actual output and overall error is computed. The weights are revised using a gradient decent algorithm known as Back Propagation algorithm. The procedure is repeated until the total error is minimised.



**Fig. 5** Schematic block diagram of an Artificial Neural Network

#### 4.2 Applications of Neural Networks in Virology

Viral epidemics are caused because of outbreak of a viral infection which can readily transmitted to other targets. One of the notorious example is Zika virus disease, which is caused by a virus transmitted primarily by Aedes mosquitoes [52]. In early period of infection Zika virus infection symptoms might not visible in most of the patients, but consequences of severe cases are very frightening like innate microcephaly in new-borns, preterm birth and miscarriage if infected during pregnancy, congenital malformations, etc. [53–55]. Even after advancements in several fields of computational biology, there is lack of reliable approach to correctly predict an outbreak and expected geographic scale. Mahmood Akhtar et al. attempted to build a dynamic neural network model to predict the geographic spread of outbreaks in real-time [56]. Most important part was gathering data for model building from diverse source that must include socioeconomic, population, epidemiological, travel and mosquito vector suitability data. For this problem Nonlinear AutoRegressive models based neural network was employed with exogenous inputs known as NARX neural networks [57–59]. For identifying top 10% of at-risk regions, the average accuracy of the model remains above 87% for prediction up to 12-weeks in advance. Further, the model is almost 80% accurate for 4-week ahead prediction for all classification schemes, and almost 90% accurate for all 2-week ahead prediction scenarios, i.e., the correct risk category of 9 out of 10 locations can always be predicted. There were several other important finding of this study, indicating the efficiency of neural networks is solving such prediction problems.

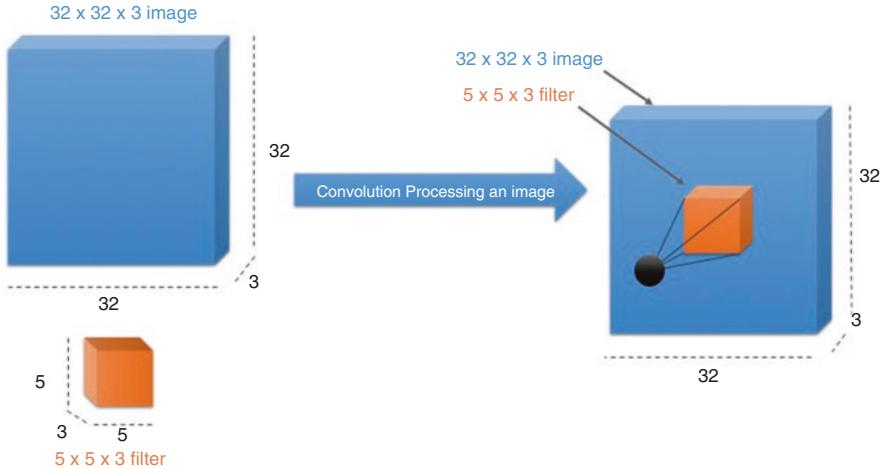
Certain properties of HIV-1 isolates can be helpful in classifying the viruses phenotypically. One such properties includes ability to replicate form multinucleated cell fusion with MT-2 cell, which is transformed T-cell line [60]. Another property is based on use of primary coreceptor to enter cells [61–66]. In recent studies,

the V3 region of HIV-1 envelope protein has been identified as a major determinant of coreceptor usage [67–70]. Wolfgang Resch et al. generated neural networks to predict coreceptor usage or MT-2 cell tropism from the amino acid sequence using a subset of positions in V3 [71]. For evaluating existing methods and by implementing neural network, set of MT-2 cell tropism (NSI/SI set), and set of known coreceptor usage (R5/X4 set) was assembled. Additional features included in this set was the epidemiologic relatedness, which was never considered before in sequence sets used in earlier studies. Neural networks were fully connected feed-forward networks with 16 sigmoidal input nodes, three hidden sigmoidal nodes, and one linear output node. Amino acids and gaps were encoded numerically by consecutive numbers from 1 to 21. Training was done using a Bayesian regulation modification of backpropagation and started with random weights [72]. The training target used values of 0 for R5/NSI and 1 for X4/SI. In summary, The mean reliability for X4 prediction of the R5/X4 neural network was 0.69 for 100 subsets of unrelated sequences, a considerable improvement over the reliability of 0.48 achieved by method.

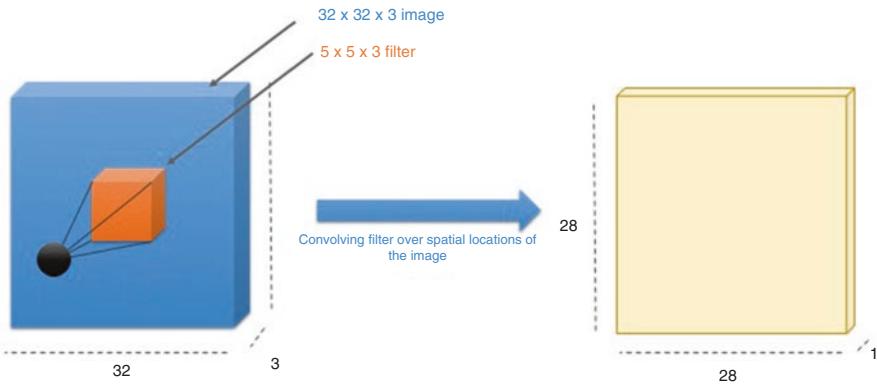
## 5 Deep Neural Networks

Deep-learning networks differ from conventional neural networks by their depth. Most of the earlier versions were shallow consisting of one input and one output layer, and at most one hidden layer in between. Deep neural networks on the other hand consist of several hidden layers between input and output. Additionally, in deep-learning networks, each layer of nodes trains on a distinct set of features based on the previous layer's output. With further advancement in layers nodes recognize higher level features. As the further you advance into the neural net, the more complex the features your nodes can recognize, since they aggregate and recombine features from the previous layer. CNNs, like conventional neural networks, consists of layers of neurons which receive input data, take a weighted sum and propagates through an activation function. The outputs received from the last layer of hidden neurons is compared with the actual output and the weights are corrected using back propagation algorithm.

Unlike neural networks, where the input is a vector, here the input is a multi-channelled image. For an RGB image let us assume CNN receives an image of size 32X32X3. This input undergoes a series of convolution operations in CNN. For this operation several filters each having random weights are used and they convolve over the image, shown in Fig. 6. Let us assume we take the 5\*5\*3 filter and slide it over the complete image covering all possible unique 5X5X3 subsets of the image. On every convolution operation we obtain a dot product between the image and the filter and the output ( $WT \cdot X + B$ ) is a scalar(one number) Similarly for every other dot product taken, the result is a scalar. It is easy to arrive at the figure of 28×28 unique image subsets are to be convolved and a complete convolution operation with a single filter yields an output of size 28X28X1, shown in Fig. 7. The convolu-



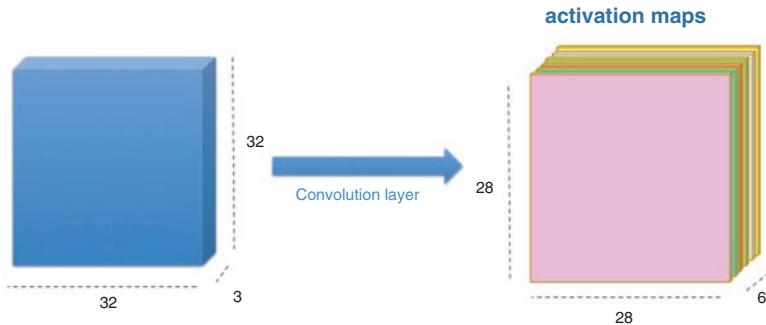
**Fig. 6** Example of multi-channelled image as input for Convolutional Neural Network



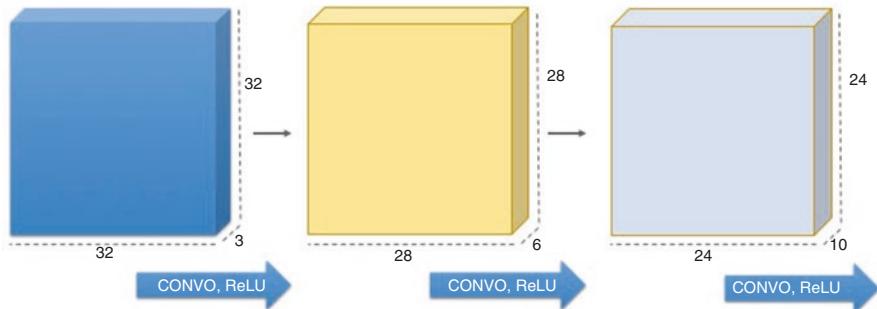
**Fig. 7** Convolution operation with a single filter

tion layer normally consists of several filters and if we assume six filters are taken each of the six independent layers convolve and the total output will be six feature maps and the combined size will be  $28 \times 28 \times 6$ . Each filter is independently convolved with the image and we end up with 6 feature maps of shape  $28 \times 28 \times 1$ , which is diagrammatically represented in Fig. 8. The architecture consisting of several convolution layers in sequence will look like Fig. 9.

So with each layer there is a thickening of the width and thinning of the breadth. If the finalized filters with random weights learn at the entire set of layers through back propagation each successive layers will learn higher and higher levels of features. Another building block of CNN is the pooling layer. This layer down samples the image and progressively reduces the size and the parameters to learn. This pool-



**Fig. 8** Output of multiple feature maps in a Convolutional Neural Network



**Fig. 9** Convolution layers in a Convolutional Neural Network

ing layer operates on each feature map. CNN like the conventional network consists of activation blocks and the most commonly used in CNN is the ReLU activation function. The fully connected layer of neurons converts the image to a linear structure like the ones in regular neural networks.

## 5.1 Applications of Deep Neural Networks in Virology

Well accepted applications of deep neural networks algorithms includes image processing and face recognition [73]. The most commonly employed deep learning network architecture for image analysis is the convolutional neural network (CNN). The basic cores of CNN are Pattern matching (convolution) and aggregation (pooling) operations [74]. Reza Ahsan et al. attempted a novel approach of developing, training and validating image processing convolution neural network algorithms for prediction of influenza proteins [75]. The method used was conversion of two important influenza virus A subtypes protein sequences (HA and NA) into

binary images. Sequences of five hemagglutinin (HA) proteins (H1, H3, H4, H5, H9) and four neuraminidase (NA) proteins (N1, N2, N6, and N8) extracted from UniProt Protein database [76]. HA Polynomial Dataset (HAPD) and NA Polynomial Dataset (NAPD) are created by converting each amino acid position into one feature or variable. Thus number of features (or column) for each sample will be equal to length of the longest protein sequence. While the Binary Image Datasets, viz. HA Binary Image Dataset (HABID) and NA Binary Image Dataset (NABID), for these proteins are created by converting sequence character of single-letter codes of an amino acid to an integer. Then numeric data of HA sequences converted to the binary image, composed of nineteen 0 and 1. For an example authors assigned amino acid Arginine (R) number 2, to get the binary numbers 010000000000000000000000. Likewise, image of the binary matrix for 20\*(number of protein sequences) was created. The polynomial datasets of HA and NA amino acids sequences was created which was later used for constructing a binary image datasets of the amino acids sequences. Conventional predictive models were trained and tested using the polynomial datasets. Finally the prediction model for the virus subtypes based on images of protein sequences was developed, trained and validated using CNN, followed by its comparison with conventional predictive models. The performances of conventional predictive models varied, from 35% to 99%, while authors were able to reach 99% accuracy with Naïve Bayes model in predicting the HA subtype, that dataset created based on thousands of physicochemical features of proteins, not protein sequence. While the image processing models using CNN yielded performance upto100%. The main outcome of this work was highlighting that raw amino acid sequences can be directly fed into the prediction model, and extraction of physicochemical properties as features can be skipped.

Similar work was done by Youngmahn Han et al. where they developed an approach for computationally scanning the peptide candidates that bind to a specific major histocompatibility complex (MHC) to speed up the peptide-based vaccine development process [77]. For this problem Deep convolutional neural network (DCNN) was employed. The peptide-MHC interactions were encoded into image-like array(ILA) data. The dataset used for this work was nonapeptide i.e. 9 physicochemical scores [78], binding data for HLA-A and -B. For the binary classification of peptide binding affinities, peptides with a halfmaximal inhibitory concentration ( $IC_{50}$ ) value of less than 500 nM were designated as binders. The contact site between the peptide and MHC molecule is corresponded to a “pixel” of the ILA. For each “pixel”, physicochemical property values of the amino acid pair at the contact site are assigned to its channels. The predictive performance DCNN was evaluated with leave-one-out and five-fold cross-validation approaches. The mean validation losses were 0.318 in leave one-out and 0.254 in five-fold cross-validation, and the mean validation accuracies were 0.855 and 0.892, respectively, and this indicate that our DCNN was able to be generally trained on the ILA data without much overfitting problems. The DCNN showed a reliable performance for the independent benchmark datasets. DCNN significantly outperformed other tools in peptide binding predictions for alleles belonging to the HLA-A3 supertype.

**Table 1** Deep meta-architectures for object detection

Architecture	Title	Description
Faster R-CNN	Faster region-based convolutional neural Network	Region proposal Network (RPN) takes an image as input and processes it by a feature extractor and features are used to predict objects [154].
SSD	Single shot multibox detector	Object recognition in a fixed-size collection of bounding boxes, which are produced by feed-forward convolutional network [155].
R-FCN	Region-based fully convolutional networks	It uses position-sensitive maps to address the problem of translation invariance [156].

Virus causing infections in plants is another concerning area that can severely affect economy of a country when case of an viral outbreak. Usually climate change in a region affects ecological variable like precipitation humidity and temperature, which consequently serve as a vector in which viruses to spread if changes are favourable [79]. Alvaro Fuentes et al. worked on developing an approach to identify and recognize of diseases affects tomato plants using deep neural network algorithm [80]. Dataset used in this approach was images affected by several diseases and pests in tomato plants. Additional important data used annotations, which were added manually by experts by creating the bound box around the anomaly in the image and assign the class to define the impact.

Input Images are passed through CNN meta-architectures mentioned in Table 1. The output of the CNN architecture is passed through a fully connected layer (feature extractor). Finally SoftMax layer is used to produce the output. The fully connected layer used in this work employs different standard feature extractors, already available in the literature. These are AlexNet [81], VGG-16 [82], GoogLeNet [83], ResNet-50 [84], ResNet-101 [84], ResNetXt-101 [85] etc. While the performance of all the architecture is generally very good, due to the small number of samples in few classes, these examples were predicted poorly. Resulting in false positive and lower average precision. The input image with different resolutions and scales was feed into the system. These images were first pre-processed and later used for extracting features for deep neural networks. The outcome of the pipeline was class disease and localization of the infected area of the plant in the mage. In this study, authors demonstrated a non-destructive local solution in identification of plant disease of pest infection. This approach can be proved extremely helpful in making correct remedial approach, avoid the disease expansion to the whole crop and reduce the excessive use of chemical solutions.

## 6 Genetic Algorithms

Genetic algorithms belong to a family of computational models, which has been inspired by evolution [86–88]. They are immensely popular because they are simple to implement and have widespread applications. Genetic algorithms are population-based, stochastic algorithms and are popularly used as optimization tools. GA for

most optimization problems, starts with a randomly generated initial population, where each individual of the population represents a possible solution, and is encoded into a string. There are different encoding techniques like binary encoding where each solution is converted into a string of a given size consisting of zeros and ones and real encoding where each solution is represented by a real number. The encoding technique must be clearly defined in advance. Each individual is evaluated for its fitness. The fitness of a solution is either the value of the objective function which we want to optimize, or a function of the objective function. The function that defines the fitness has to be specified distinctly for each problem. Generally, a fitter individual has a better probability to be selected for further operations to evolve newer solutions with better fitness. In most GAs there are three primary genetic operations, which are applied to the population members repeatedly until the solution has converged.

### 1. Selection

This operation involves the selection of individuals from the current population, to create a mating pool for the next generation. Individuals with higher fitness values have a greater chance of being selected. Tournament and Roulette wheel selection are the most popular selection schemes.

### 2. Crossover

Where (randomly selected) elements or chunks of elements are swapped (with a probability known as crossover probability) between individuals, to create population members of a new generation.

### 3. Mutation

Where (randomly chosen) elements are modified.

As can be seen from the above description, the encoding and the fitness evaluation are defined specifically for each problem whereas the implementation of the genetic operators is a common one.

## 6.1 GA for Attribute Selection

Selection of the most informative attributes is an important pre-processing steps involved in a function annotation problem in viral biology. GA employing the three genetic operators (selection, crossover and mutation) iteratively evolves the best attributes from a set of attributes in a given data set. The size of an individual is the size of the total number of attributes. As an example, if the original set of descriptors are six in number each member will have a string length of six. The algorithm starts with random generation of a predefined number of solutions. For each solution, every bit is randomly filled with ones and zeros. Each bit represents one attribute and a value of one represents presence of an attribute in the solution and zero represents absence of a solution. Once the solutions are generated the attributes selected

in each solution are input to a classifier and the performance is measured in terms of suitable performance measures like Cross Validation (CV) accuracy. After evaluation the fitter solutions are selected by a selection process like tournament selection and the crossover process is carried out with a crossover probability on the selected solutions . After this the mutation step is conducted in which each of the bit is flipped (ones to zero and zero to one). This completes one generation and the next and subsequent generations the process of selection, crossover and mutation are conducted . This process is repeated until convergence and the best solution provides the most informative subset of descriptors.

## 6.2 *Generalized GA*

The algorithm consisting of generating random population, selection and mutation, is illustrated below for a representative data set with six features:

A population is randomly generated with each solution having number of bits equivalent to the total number of attributes. The attributes which are selected in each individual represented by ones are sent to a standard classifier to get the performance measure like CV accuracy.

**CV Accuracy = 78%**

1		1		0		0		1		1		0
---	--	---	--	---	--	---	--	---	--	---	--	---

**Accuracy = 82%**

1		1		1		1		1		1		1
---	--	---	--	---	--	---	--	---	--	---	--	---

**Accuracy = 74%**

1		1		1		0		0		0		0
---	--	---	--	---	--	---	--	---	--	---	--	---

**Accuracy = 75%**

0		1		0		1		1		0		0
---	--	---	--	---	--	---	--	---	--	---	--	---

**Accuracy = 73%**

1		1		0		1		0		0		0
---	--	---	--	---	--	---	--	---	--	---	--	---

**Accuracy = 81%**

1		1		1		0		1		1		1
---	--	---	--	---	--	---	--	---	--	---	--	---

**Accuracy = 71%**

0		1		1		0		0		0		0
---	--	---	--	---	--	---	--	---	--	---	--	---

**Accuracy = 72%**

1	0	1	1	0	1
---	---	---	---	---	---

**Accuracy = 78%**

0	1	1	0	0	1
---	---	---	---	---	---

**Accuracy = 74.5%**

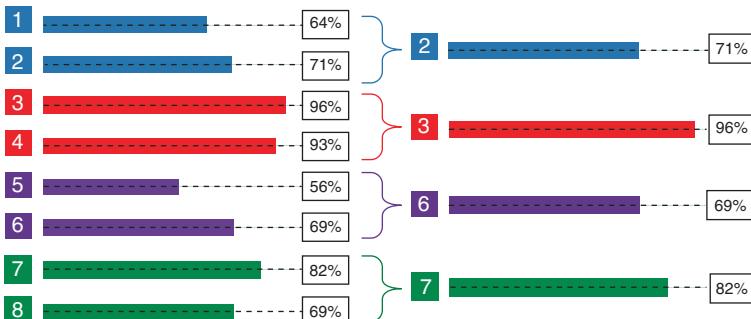
1	1	0	0	0	1
---	---	---	---	---	---

## 6.2.1 Selection

Accuracy is directly used as fitness measure and in the selection step solutions are selected based on the selection mechanism. Here we illustrate the process with tournament selection process.

### 6.2.1.1 Tournament Selection

From a given populations, two chromosomes are chosen at random, and the one with higher accuracy is selected for crossover. See Fig. 10 for diagrammatic representation of the Tournament Selection where the length represents the accuracies and longer chromosome means better accuracy. It can be seen in Fig. 10 that chromosomes 2,3,6 and 7 are selected, because their accuracies are better than the chromosomes they are compared with. This selection process is conducted twice so that number of chromosomes before selection and after selection remains same. In Tournament selection, it is guaranteed that worst solution will never chose for crossover.



**Fig. 10** Diagrammatic representation of the Tournament Selection

### 6.2.2 Crossover

In the crossover process new solutions are generated from an existing population stochastically. Solutions are chosen at random from population with a crossover probability. There are different types of crossover and the following three are most popular:

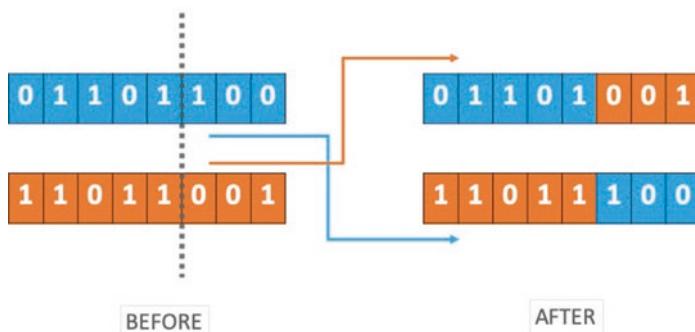
1. *Single point crossover*
2. *Multi point crossover*
3. *Uniform Crossover*

#### 6.2.2.1 Single Point Crossover

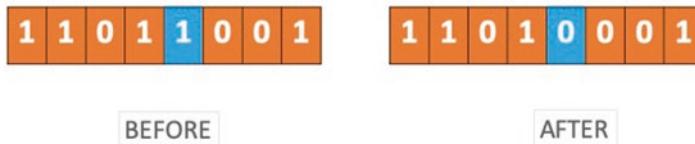
In this illustration we employ single point crossover in which two randomly chosen members are made to undergo the process of crossover with a predefined probability. A random intersection point is chosen and using this intersection point two new solutions are generated as shown in Fig. 11. This process is repeated until a new population is created after crossover with the same number of solutions originally present. After completion of crossover the solutions undergo the process of mutation with a small mutation probability.

### 6.2.3 Mutation

It is used to maintain genetic diversity of solutions from generation to generation. It is used to avoid problem of rapid convergence to a poor local optimum. The flip mutation operation flips one or more-bit values (from zero-to-one or from one-to-zero) from a crossover chromosome from its initial state stochastically. Mutation operation is done according to mutation probability, usually very small. Starting from the first offspring after crossover, each bit of the solution flipped (zero to one



**Fig. 11** Example of single point crossover in Genetic Algorithms



**Fig. 12** Example of Mutation in Genetic Algorithms

or one to Zero) with the predefined mutation probability. It can be seen in Fig. 12, the fifth bit got flipped. Similarly every solution is subjected to mutation operation and after mutation operation accuracy of each solution is estimated by sending to a classifier. These three operations, namely selection, crossover and mutation, complete in generation and the same steps of selection, crossover and mutation are carried out for a large number of generations until convergence.

### 6.3 Applications of GA in Virology

Applications of Genetic Algorithms are not only restricted primarily to solve optimization problems but also, they are frequently used in diverse areas like training Neural Networks, digital image processing, genetics-based machine learning, spectrometric data analysis, etc. Due to recent advancements in laboratory methodologies, there is a rapid increase in the amount of published and experimental data in several domains of Life Sciences. Virology is no exception and there is a recognized need for better optimization method to address problems like fitting a model to observed data generated by virology studies.

Viral genomes show great variation in nature and genome sequencing projects are uncovering many unique features of these that had been previously known. Human immunodeficiency virus type 1 (HIV-1) is one of the two types of HIV viruses that causes AIDS, which is the most advanced stage of HIV infection [89, 90]. Proivirus in retroviruses like HIV-1, is referred to the genomic unit formed when viral genetic material is translocated to the nucleus and integrated into the host-cell chromosomal DNA. Prior to provirus formation, a double-stranded molecule of DNA is generated by reverse transcribing two viral RNA copies. During metamorphosis of RNA into DNA, point mutations can occur. These mutations were in focus for understanding the viral biology with a view to identify drug targets for clinical intervention. However, recently it has been shown that the majority of HIV-1-infected cells *in vivo* can contain multiple proviruses [91]. The number of proviruses may vary from one to eight copies per infected splenocyte. This implies that recombination could also be playing major role in the intrapatient evolution of HIV. To analyze and understand HIV evolution in host, Gennady Bocharov et al. developed a stochastic model that reflects in some detail both the biology of HIV replication and the infection process within a host [92]. In this study, multiple fac-

tors impacting the viral evolution has to be considered to mimic real HIV infection. These factors include the extent of virus expansion and degradation, selection processes and the multiplicity of infected cells. Thus, genetic algorithms fits perfectly here to take into account all factors as variation operators and simulate viral evolution. Genetic algorithms proved effective in segregating the contribution of the inherently linked processes of multi-infection and recombination. The model developed by the authors in this work, provides a versatile platform for predicting the response of HIV towards therapeutic interventions.

Genomics studies have revealed the sequence of molecular events in the replication cycle of the HIV [93], including the following seven steps:

- (i) viral entry
- (ii) reverse transcription
- (iii) integration
- (iv) gene expression
- (v) assembly
- (vi) budding
- (vii) and maturation

To design strategies to inhibit the HIV replication or develop effective antiviral agents, each individual step within the HIV life cycle may be used as a potential target. Antiviral chemotherapy is effective in some extent to suppress the infection, but it comes with deleterious side effects. Styrylquinoline derivatives are class of compounds, which at non-toxic concentrations shown to inhibit integration activity in vitro and to block viral replication [94]. Nasser Goudarzi et al. used genetic algorithms for descriptor selection in quantitative structure–activity relationships (QSAR) based study to understand the pharmacophore properties of styrylquinoline derivatives and to design inhibitors of HIV-1 integrase [95]. Two factors which governs the predictive accuracy of QSAR models are: predictive model selected, and descriptor selection that sufficiently represent the structural information. Thus genetic algorithm–multiple linear regression (GA–MLR) was considered as best option for predicting the anti-HIV activity ( $\text{pIC}_{50}$ ) values of styrylquinoline derivatives. For this work  $\text{pIC}_{50}$  values of for 36 molecules of styrylquinoline derivatives from the literature [96] were taken. GA process first generated random feature subsets of the molecule, followed by subset-wise evaluation of selected descriptors for fitness to predict  $\text{pIC}_{50}$ . Based on the fitness GA operators of selection, crossover and mutation were repeatedly applied to get better subsets of descriptors, as iteration proceeded. After convergence, GA narrowed down the search from 302 descriptors to 7 best descriptors by iterating 100 generation of simulation, on population size 64, mutation rate 0.005, and cross-over 0.6. The correlation coefficients ( $R^2$ ) GA–MLR model for training set was 0.9519 while for test set it was 0.7977. The results of this study provided enough information related to different molecular properties, which can participate in the physicochemical process that affected the HIV inhibition activity of styrylquinoline derivatives.

Similar work has been done by Yong Cong et al., where another variant of GA with Partial Least Square (GA–PLS) was employed to select best descriptor subset

for QSAR modeling in a linear model to study influenza virus neuraminidase (H1N1) inhibitors [97]. In this work SVM (GA-SVM) was used to build regression model to evaluate structural and physicochemical features of compounds contributing to the influenza virus NA inhibitory activity. Data used by this group was 108 compounds with carbocyclic and flavonoid scaffolds, which have clear inhibitory activity against influenza virus strain A/PR/8/34 (H1N1) reported in the literature [98–105]. Further, these compounds were separated into the training set (80 compounds) and test set (28 compounds) based on their similarity and distribution in the chemical space. The chemical space here denotes the used structural and chemical descriptors [106]. GA generated random population to subsets of descriptors and these descriptors were evaluated by GA-PLS to calculate the fitness, fitness operator described before. After large number of iterations of subsequent evaluation, best top 9 descriptors were found to give the highest performance. These selected 9 descriptors were used by GA-SVM to create regression models. Here GA was used to select best set of kernel parameters, to provide the highest correlation coefficient (R) of 0.9189 for the training set. While the correlation coefficient values achieved for testing set was 0.9415. for the testing set. Thus, authors demonstrated how combinatory methods can be effectively used to address complex problems like investigating inhibitory activity of compounds against of viral proteins, which potentially can be used as base for receptor-based and ligand-based anti-influenza drug design.

There are other examples where GA was also used for applications which deals with handling genomic sequence data. Chunlin Wang et al. performed a benchmarking experiment where genetic algorithm was implemented in parallel mode to optimize multiple genomic sequence alignments initially generated by various alignment tools [107]. They developed a program, GenAlignRefine, which improves the overall quality of global multiple sequence alignments (MSA) by using a genetic algorithm to improve alignments in local regions. Addressing such a problem statement was a challenge since MSA can provide only approximate solutions to alignments except for the smallest alignments. Already a number of novel heuristic algorithms have been proposed [108]. Deciding factors of the effectiveness are: (a) choice of an objective function (OF) that assesses the quality of an alignment, (b) algorithm design to optimize the score from that objective function. Sum-of-pair (SP) function is frequently used OF [109], which is an extension of the scoring method used in pair-wise alignments. Alternatively, COFFEE (Consistency based Objective Function For alignmEnt Evaluation) [110] function can be used which assesses the evenness between a multiple alignment and libraries of optimal pair-wise alignments of the same sequences. Authors used the COFFEE OF as a measure of the optimization of the MSA, since other studies proved its robustness better alignments [111]. Genetic algorithm was employed to optimize an alignment by attempting to maximize its COFFEE score. The columns in an alignment that contain a gap adjacent to a gap-free region of at least 20 nucleotides as defined in this study as “fuzzy” regions. The starting point for the genetic algorithm in the method developed was the initial alignment produced by T-Coffee [111] alignment on fuzzy regions. GenAlignRefine then optimizes the application of the genetic operators by using a combination of only 3 operators rather than the full set by pre-aligning each

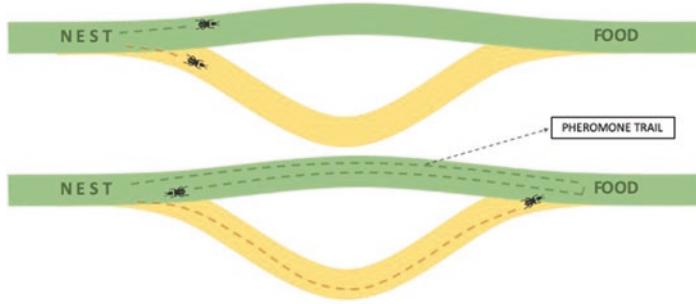
fuzzy region using T-Coffee, similar to studies done earlier [112]. Using these 3 genetic operators as genetic operators, authors effectively utilized genetic algorithms to efficiently improve MSA of whole genome sequences.

Totally other side of virology is the remedial approaches undertaken to either avoid, or reduce the dreadful consequences of viral infection. Worldwide, immunologists are working actively to develop preventive and therapeutic vaccines against cancer. There are several challenges related to this work, out of which most critical is translating positive immunoprevention from animal models to human situations. Thus, a successful experiment confirming effectiveness of vaccine on a particular cancer, seeks devising an optimal vaccination schedule that maximizes chances of demonstrating best effects. Cristiano Calonaci et al. [113] developed an agent-based model (ABM) [114] to summarize outcome of vaccination experiments for mammary carcinoma [115–119]. Genetic algorithms in this case was employed to deduce optimal vaccination schedule. To make this process more robust and effective, genetic algorithm was parallelized using Message Passing Interface (MPI), where a simulator was used as a fitness evaluator. The suggested schedule was then tested *in vivo*, giving good results. Thus, successful application of drug optimization using parallel computing was demonstrated by authors, leading to the development of a real virtual lab to analyze and optimize vaccine protocol administrations.

## 7 Ant Colony Optimization

The Ant System (AS) was initially proposed as a metaheuristic for optimization problems, by Marco Dorigo in 1992 [120]. It constitutes a class of algorithms in the area of Swarm Intelligence. The first problem studied in AS was that of searching for the most optimal path in a graph popularly known as the Traveling Salesman Problem (TSP) [121]. Over a period of time, the Ant System branched into several variations, sometimes to give better results for benchmark problems and sometimes varying as per the requirements of a domain problem. Thus, Ant Colony Optimization algorithm (ACO) is a probabilistic algorithm aimed at solving computationally intensive problems by drawing on random ant system behavior, towards incrementally finding better solutions.

In addition, ACO displays a reinforcement learning behavior which gives it a remarkable capability to learn while building its solutions. As a result, owing to multiple important properties of the ACO algorithm, a majority of published papers have reported ACO performing very well in many problem domains in comparison to other metaheuristics. One such class where ACO has been known to perform very well is in the area of combinatorial bioinformatics optimization problems. In this context, the attribute selection problem is of extreme importance. As an example, Microarray datasets are composed of a huge number of gene expression profiles. These profiles from a computational perspective are extremely noisy and redundant. A model (predictive or otherwise) when derived out of this data, will therefore also



**Fig. 13** Pheromone trail for exploration by virtual ants

be inefficient and possibly misleading. As a result, pre-processing of these datasets is paramount. Collecting informative gene subsets, from this aspect, thus turns out to be very important. The reduced informative gene subsets thus obtained, help in building more expressive predictive models. At this time, popular classifiers like SVM, Random Forests etc. may take over. Sometimes, a feedback loop with the subset selection algorithm may help to improve the final model.

ACO has been motivated by the cooperative search behaviour of real life ants of a colony for finding food. As naturally observed, an ant wanders randomly and on finding food returns to its colony while laying down pheromone trails. Random ants on finding such trails follow the same with a very high probability and return to the nest by reinforcing the pheromone concentration on these trails. More and more ants follow the pheromone rich trail and the shortest route is established. Figure 13 illustrates this process. This probabilistic behaviour thus ensures that searching for food is not just in a local region and exploration thus continues. Once another new good path appears, ants start using that route. More information on this can be found in [122].

In terms of an optimization algorithm, ACO is fundamentally described by the algorithm mentioned next.

## 7.1 Generalized ACO Algorithm

### 1. Initialisation

Place ants at their initial positions;

Initialize a Pheromone matrix that records an initial pheromone value for all possibilities;

### 2. For 'itr' iterations -

For 'k' ants -

For 'n' moves towards building a complete solution-

Select a partial solution probabilistically using problem heuristic and transition function using pheromone values;

- Evaluate the k-th ant's solution and store it;
  - Store and Select the best solution/s;
  - Simulate reinforcement behaviour by increasing pheromone values for above selected solutions;
  - Simulate pheromone evaporation by decreasing other solution components not selected;
  - Repeat.*
3. Extract and report the final complete solution as the most optimal for the given parameters

While initialization of a generalized ACO, artificial ants may be placed on random positions (partial solution components). Next, in a pre-determined ‘itr’ set of iterations with a certain ‘k’ ants, a solution is explored considering there are ‘n’ partial components of the complete solution.

The problem heuristic is normally associated with the amount of information provided by the partial component of the complete solution.

In a later approach, Dorigo et al. [121] introduced the notions of exploration and exploitation to the ACO algorithm for the symmetric TSP problem. This process involved the generation of a random value called  $q$ , between 0 and 1, which was tested against a threshold  $q0$  (user defined). An exploitation, where the best available partial solution component would be chosen (the shortest edge with maximum pheromone concentration for TSP), constituted the next option if  $q$  was less than  $q0$ . Otherwise, exploration, where a random solution component according to a probability distribution, would be selected. Elitism has also been used to improve results frequently. Such exploration and exploitation based search measures thus overcame many problems which normally a greedy algorithm would suffer from, for example the solution search being stuck in local optima.

The set of complete solutions for one iteration are then evaluated and the best are selected for updating pheromone concentration corresponding to a global update. Other solutions go through a local updation with pheromone evaporation.

## 7.2 Applications of ACO in Virology

Several viral diseases and outbreaks not only cause threats to humans, but also adversely affects the plant agriculture and animal husbandry in worst possible ways. One such example is shrimp aquaculture which has been severely affected by White spot disease (WSD) [123–125] resulting in a huge economic burden to the industry. Researchers have been working on developing approaches to find potential antiviral agents which will be used in docking analysis. These drug-like molecule obtained from the docking experiments would be used to optimize to a candidate drug. The objective is to find the inhibitors that blocks the binding of the viral protein to the receptor, thus averting the viral infections.

Finding or developing new drug is a lengthy, complex, and costly process, with no assurance that the drug will actually be effective. There is a lack of validated diagnostic and therapeutic biomarkers to objectively detect and measure biological states. In-silico techniques have become important part of the drug discovery cycles because of their crucial role from hit identification to lead optimization. These methodologies are employed screen numerous molecules and narrow down the search to few potent candidates. One such widely used approach is ligand or structure based virtual screening [126]. Protein-ligand docking problem (PLDP) involves the calculation of approximate binding free energy of the complex formation, based on which ligands are ranked. To address this problem Oliver Korb et al. proposed new algorithm based on ant colony optimization (ACO), called Protein–Ligand ANT System (PLANTS) for sampling the search space [127]. In conclusion, different parameter settings were evaluated in this study to assure high success rates in pose prediction for different timings. Default docking settings were able to reproduce ligand geometries similar to the crystal geometry in about 72% of the cases at average docking times of 97 seconds.

HIV-1 and HIV-2 viral strains have different amino acid and nucleotide sequences. As discussed in genetic algorithm section, both of these viruses require a reverse transcriptase (RT) to convert viral RNA into proviral DNA that can then be inserted into the host DNA. Thus a lot of focus has been targeted on RT for drug discovery against HIV. 1-[(2-Hydroxyethoxy)methyl]-6-(phenylthio)-thymine (HEPT) acts as nonnucleoside inhibitor against HIV-1 [128, 129]. HEPT derivatives has been extensively exploited QSAR studies [130–142]. Vali Zare-Shahabadi et al. worked on developing QSAR model for a large set of HEPT derivatives to predict its anti-HIV1 activity, where ant colony system (ACS) was employed to select best descriptors [143]. Probability vectors were derived as colony of ants, where each ant is a bit string representation of all descriptors. That means that the elements in the bit string are set to zero for the nonselected descriptors, whereas the selected ones are set to one [144]. With randomly selected set of descriptors, a regression model was built, followed by assessment of each ant by fitness function, which in this case was cross-validation correlation coefficient. Outlier detection and regeneration of the linear model was employed to increase the quality of the linear model. The final model yielded RMSE values for the training and prediction sets 0.47 and 0.52, respectively. The  $R^2$  value for the training was 0.90 along with an F statistic value of 100.7. The RMSE values for the training and prediction sets were 0.56 ( $R^2 = 0.86$ ) and 0.58 ( $R^2 = 0.85$ ), respectively.

## 8 Particle Swarm Optimization

The particle swarm is a population-based stochastic algorithm for optimization which is based on social–psychological principles of bird flocking. The synchrony of flocking behaviour of a group of birds is believed to be a function of bird's efforts to maintain an ideal separation among themselves and their neighbours. Birds change

their movement and path to stay away from predators, look for maintaining their life existence, enhance survivability in different environmental parameters and so on.

PSO is similar to a genetic algorithm (GA), as PSO also is initialized with random population called particles. Unlike GA, in PSO all population members survive from the beginning of a trial until the end, each potential solution is also assigned a randomized velocity [145, 146]. Each particle keeps track of its coordinates in the search space associated with the best fitness achieved so far. At each time step (generations) the particle is updated by following two ‘best’ values:

- (a) Best solution obtained by a given particle so far. This values is called as *pBest*
- (b) Best value obtained so far by any particle in the swarm. This values is called as *gBest*

## 8.1 Generalized PSO Algorithm

1. Initialize a population of particles with random positions and velocities on d dimensional search space.
2. Each particle fitness is evaluated over a desired optimization function.
3. *pBest* and *gBest* values are computed.
4. Compare each particle fitness with, particle having best fitness (*gBest*). If current fitness of a particle is better than best particle, then replace current particle as best particle along with position and velocities.
5. Update the velocity and position of the particles according to following equations.

$$\begin{aligned} v[ ] &= v[ ] + c_1 * \text{rand}( ) * (p\text{Best}[ ] - \text{present}[ ]) \\ &\quad + c_2 * \text{rand}( ) * (g\text{Best}[ ] - \text{present}[ ]) \end{aligned}$$

$$\text{present}[ ] = \text{present}[ ] + v[ ]$$

6. Update *pBest* and *gBest*.
7. Repeat procedure from step 2 until convergence

### 8.1.1 Advantages and Disadvantages

Two notable advantages includes:

- (a) very few parameters to tune
- (b) slight variations works well in a wide variety of applications

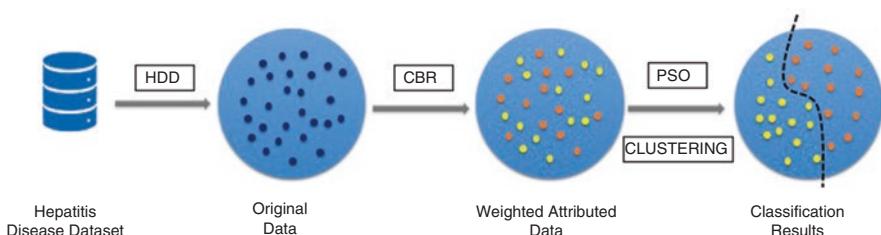
While downsides includes:

- (a) easy to fall into local optimum in high-dimensional space
- (b) low convergence rate in the iterative process

## 8.2 Applications of PSO in Virology

There are several noteworthy examples in problems in life sciences, where Particle Swarm Optimization was efficiently used to optimize the pool of candidate solution by iterative screening based on quality of measure. One excellent example in field of virology is work done by Mehdi Neshat et al. where PSO was used to diagnose hepatitis disease type [147]. Hepatitis literally means inflammation of the liver and it can be caused because of several factors. One of the major causative agents are viruses. Viral hepatitis is an infection that causes liver inflammation and damage. Treatment and medication of Hepatitis heavily depends on its correct diagnosis. Researchers have already started exploiting computational intelligence in diagnosing different diseases. The most frequently used method for this purpose is neural networks. Different kinds of neural networks with various specifications have been used in diagnosing diseases [148]. There are other studies employing neural networks and fuzzy system for diagnosis of B hepatitis disease [149, 150]. Mehdi Neshat et al. used combination of two methods of PSO and CBR (case-based reasoning). This is a classification problem of determining whether patients with hepatitis will live or die. Thus, dataset of 155 samples considered for this study has these two classes (32 “die” cases, 123 “live” cases). The database created in this study using the patient data contains 19 attributes. These attributes include details like physiology of the patient (age, sex, etc.), symptoms (Fatigue, Anorexia, etc.), treatments (Steroid, Antivirals, etc.) and clinical test results (Bilirubin, Alk phosphate, etc.). CBR generates weighted attributes for the original dataset. Centroids are randomly selected from the dataset, which acts as classes to which appropriate data points will be assigned. This is followed by calculation of accuracy for each cluster. Figure 14 is the diagrammatic representation of the methodology used by the authors. PSO performs these steps for each of particle and outcome of large number of iterations is the best accuracy. The accuracy of CBR-PSO method in diagnosing hepatitis disease was found to be 94.58%, far better compared to PSO method whose best accuracy was 89.46%.

Viral Load (VL) Test is a laboratory test that measures the amount of HIV in a blood sample. Results are reported as the number of copies of HIV RNA per milliliter of blood. HIV-1 infection cannot be effectively diagnosed without viral load testing [151, 152] thus routine use of this test is also recommended by World Health Organization (WHO). However, this implementation is subjected to cost, availability and accessibility of testing instruments. K. Kamalanand et al. worked on effi-



**Fig. 14** Diagrammatic representation of PSO-CBR methodology

ciently estimating HIV-1 viral load from CD4 cell count using a computational swarm intelligence (PSO) technique in conjunction with the three-dimensional HIV model [153]. In this work authors attempted to estimate the HIV-1 viral load from CD4 cell count in the acute and chronic phase of the HIV1 infection. For this purpose, below nonlinear differential equation was employed:

$$\begin{aligned}\frac{dx(t)}{dt} &= a(x_0 - x(t)) - bx(t)z(t) \\ \frac{dy(t)}{dt} &= c(y_0 - y(t)) + dy(t)z(t) \\ \frac{dz(t)}{dt} &= z(t)(ex(t) - fy(t))\end{aligned}$$

In these equations,  $x(t)$ ,  $y(t)$ ,  $z(t)$  are the concentrations of the CD4, CD8 lymphocyte population, and concentrations of the HIV-1 viral load respectively. While  $x_0$  and  $y_0$  are the normal unperturbed concentrations of the CD4 and CD8 lymphocyte population respectively. Here,  $a$ ,  $b$ ,  $c$ ,  $d$ ,  $e$  and  $f$  are the system parameters.

The objective function used in this study that needed to be minimized for estimation of HIV-1 viral load, can be given as:

$$J_\theta = \sum_{n=1}^N \frac{(\hat{x}_n - x_n)^2}{N \text{mean}(x_n)}$$

Where,  $\theta$  is the set of HIV parameters to be estimated;  $\hat{x}_n$  represents the CD4 cell population;  $N$  is the total number of samples available for CD4 data.

Thus, using the principles of PSO, newer parameters will be generated for all the samples, until best results are obtained as per the fitness function. Moving particles in PSO methodology, here is equivalent to trying random values for parameters of the differential equation to calculate the viral load. The average error in estimation of viral load was found to be 3.317%. Further, the maximum estimation error in the acute stage of the disease was found to be 14.19%, whereas, the maximum estimation error in the chronic phase of the disease was found to be 0.4399%. Hence it appears that the PSO algorithm for estimation of HIV-1 viral load is highly efficient during the chronic phase of the disease.

## 9 Concluding Remarks

In this review, we illustrated the use of Artificial Intelligence and Machine learning methods in viral biology. We have shown the power of machine learning to extract useful patterns from large biological data and convert to useful knowledge. Different machine learning algorithms including decision tree, random forest, neural net-

works and deep neural networks have been explained lucidly. We also dealt with the use of Artificial intelligence methods like genetic algorithms, ant colony optimization and particle swarm optimization methods in synergistic combination with machine learning methods to provide optimal solutions computationally faster and with increased accuracy and robustness. We have also listed large number of case studies and examples in different areas of viral biology where AI and ML tools have been beneficially employed for solving real life problems.

## References

1. Sousa MS, Mattoso ML, Ebecken NF. Data mining: a database perspective. *WIT Transactions on Information and Communication Technologies*; 1970 Jan 1;22.
2. Steinberg D, Colla P. CART: classification and regression trees. In Wu X, Kumar V, editors. *The top ten algorithms in data mining*. Knowledge and information systems (Boca Raton, FL). 2008 Jan 1;14(1):1-37. p. 179–201.
3. Gubler DJ. Dengue and dengue hemorrhagic fever. *Clin Microbiol Rev*. 1998;11(3):480–96.
4. Vaughn DW, Green S, Kalayanarooj S, Innis BL, Nimmannitya S, Suntayakorn S, Rothman AL, Ennis FA, Nisalak A. Dengue in the early febrile phase: viremia and antibody responses. *J Infect Dis*. 1997;176(2):322–30.
5. Halstead SB. Dengue. *Lancet*. 2007;370(9599):1644–52.
6. Kalayanarooj S, Vaughn DW, Nimmannitya S, Green S, Suntayakorn S, Kunentrasai N, Viramitracai W, Ratanachu-Eke S, Kiatpolpoj S, Innis BL, Rothman AL. Early clinical and laboratory indicators of acute dengue illness. *J Infect Dis*. 1997;176(2):313–21.
7. Chadwick D, Arch B, Wilder-Smith A, Paton N. Distinguishing dengue fever from other infections on the basis of simple clinical and laboratory features: application of logistic regression analysis. *J Clin Virol*. 2006;35(2):147–53.
8. Tanner L, Schreiber M, Low JG, Ong A, Tolfsenstam T, Lai YL, Ng LC, Leo YS, Puong LT, Vasudevan SG, Simmons CP. Decision tree algorithms predict the diagnosis and outcome of dengue fever in the early phase of illness. *PLoS Negl Trop Dis*. 2008;2(3):e196.
9. Quinlan JR. C4. 5: programs for machine learning. Elsevier (San Mateo, CA); 2014 Jun 28.
10. Kothari R, Dong M. Decision trees for classification: a review and some new results. In *Pattern recognition: from classical to modern approaches*. World Scientific (Singapore). 2001. pp. 169–84.
11. Solomon T, Ooi MH, Beasley DW, Mallewa M. West Nile encephalitis. *BMJ*. 2003;326(7394):865–9.
12. Sampathkumar P. West Nile virus: epidemiology, clinical presentation, diagnosis, and prevention. In *Mayo clinic proceedings* 2003 Sep 1, vol. 78, no. 9, p. 1137–44, Elsevier.
13. Organ Procurement and Transplantation Network. <http://www.optn.org/news/newsDetail.asp?id=303>. Accessed on-line February 24, 2004.
14. Kiberd BA, Forward K. Screening for West Nile virus in organ transplantation: a medical decision analysis. *Am J Transplant*. 2004;4(8):1296–301.
15. Hutchens TW, Yip TT. New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Commun Mass Spectrom*. 1993;7(7):576–80.
16. Merchant M, Weinberger SR. Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis*. 2000;21(6):1164–77.
17. Jones MB, Krutzsch H, Shu H, Zhao Y, Liotta LA, Kohn EC, Petricoin EF III. Proteomic analysis and identification of new biomarkers and therapeutic targets for invasive ovarian cancer. *Proteomics*. 2002;2(1):76–84.

18. Bergman AC, Benjamin T, Alaiya A, Waltham M, Sakaguchi K, Franzén B, Linder S, Bergman T, Auer G, Appella E, Wirth PJ. Identification of gel-separated tumor marker proteins by mass spectrometry. *Electrophoresis*. 2000;21(3):679–86.
19. Alaiya AA, Franzén B, Fujioka K, Moberger B, Schedvins K, Silfversvärd C, Linder S, Auer G. Phenotypic analysis of ovarian carcinoma: polypeptide expression in benign, borderline and malignant tumors. *Int J Cancer*. 1997;73(5):678–82.
20. Thompson S, Turner GA. Elevated levels of abnormally-fucosylated haptoglobins in cancer sera. *Br J Cancer*. 1987;56(5):605–10.
21. Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, Wright GL. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res*. 2002;62(13):3609–14.
22. Qu Y, Adam BL, Yasui Y, Ward MD, Cazares LH, Schellhammer PF, Feng Z, Semmes OJ, Wright GL. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin Chem*. 2002;48(10):1835–43.
23. Petricoin EF III, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*. 2002;359(9306):572–7.
24. Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem*. 2002;48(8):1296–304.
25. Vlahou A, Schorge JO, Gregory BW, Coleman RL. Diagnosis of ovarian cancer using decision tree classification of mass spectral data. *Biomed Res Int*. 2003;2003(5):308–14.
26. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
27. Liaw A, Wiener M. Classification and regression by random forest. *R News*. 2002;2(3):18–22.
28. Cutler DR, Edwards TC Jr, Beard KH, Cutler A, Hess KT, Gibson J, Lawler JJ. Random forests for classification in ecology. *Ecology*. 2007;88(11):2783–92.
29. Matrosovich MN, Gambaryan AS, Teneberg S, Piskarev VE, Yamnikova SS, Lvov DK, Robertson JS, Karlsson KA. Avian influenza A viruses differ from human viruses by recognition of sialyloligosaccharides and gangliosides and by a higher conservation of the HA receptor-binding site. *Virology*. 1997;233(1):224–34.
30. Rogers GN, Paulson JC. Receptor determinants of human and animal influenza virus isolates: differences in receptor specificity of the H3 hemagglutinin based on species of origin. *Virology*. 1983;127(2):361–73.
31. Suzuki Y. Gangliosides as influenza virus receptors. Variation of influenza viruses and their recognition of the receptor sialo-sugar chains. *Prog Lipid Res*. 1994;33(4):429–57.
32. Li OT, Chan MC, Leung CS, Chan RW, Guan Y, Nicholls JM, Poon LL. Full factorial analysis of mammalian and avian influenza polymerase subunits suggests a role of an efficient polymerase for virus adaptation. *PLoS One*. 2009;4(5):e5658.
33. Jagger BW, Memoli MJ, Sheng ZM, Qi L, Hrabal RJ, Allen GL, Dugan VG, Wang R, Digard P, Kash JC, Taubenberger JK. The PB2-E627K mutation attenuates viruses containing the 2009 H1N1 influenza pandemic polymerase. *MBio*. 2010;1(1):e00067–10.
34. Subbarao EK, London W, Murphy BR. A single amino acid in the PB2 gene of influenza A virus is a determinant of host range. *J Virol*. 1993;67(4):1761–4.
35. Chen GW, Chang SC, Mok CK, Lo YL, Kung YN, Huang JH, Shih YH, Wang JY, Chiang C, Chen CJ, Shih SR. Genomic signatures of human versus avian influenza A viruses. *Emerg Infect Dis*. 2006;12(9):1353–60.
36. Eng CL, Tong JC, Tan TW. Predicting host tropism of influenza A virus proteins using random forest. *BMC Med Genet*. 2014;7(3):S1.
37. Dubchak I, Muchnik I, Holbrook SR, Kim SH. Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci*. 1995;92(19):8700–4.

38. Dubchak I, Muchnik I, Mayor C, Dralyuk I, Kim SH. Recognition of a protein fold in the context of the SCOP classification. *Proteins Struct Function Bioinform.* 1999;35(4):401–7.
39. Wei Y, Li J, Qing J, Huang M, Wu M, Gao F, Li D, Hong Z, Kong L, Huang W, Lin J. Discovery of novel hepatitis C virus NS5B polymerase inhibitors by combining random forest, multiple e-pharmacophore modeling and docking. *PLoS One.* 2016;11(2):e0148181.
40. Lavanchy D. The global burden of hepatitis C. *Liver Int.* 2009;29:74–81.
41. Sarrazin C, Zeuzem S. Resistance to direct antiviral agents in patients with hepatitis C virus infection. *Gastroenterology.* 2010;138(2):447–62.
42. Behrens SE, Tomei L, De Francesco R. Identification and properties of the RNA-dependent RNA polymerase of hepatitis C virus. *EMBO J.* 1996;15(1):12–22.
43. Moradpour D, Brass V, Bieck E, Fribe P, Gosert R, Blum HE, Bartenschlager R, Penin F, Lohmann V. Membrane association of the RNA-dependent RNA polymerase is essential for hepatitis C virus RNA replication. *J Virol.* 2004;78(23):13278–84.
44. Ago H, Adachi T, Yoshida A, Yamamoto M, Habuka N, Yatsunami K, Miyano M. Crystal structure of the RNA-dependent RNA polymerase of hepatitis C virus. *Structure.* 1999;7(11):1417–26.
45. Lesburg CA, Cable MB, Ferrari E, Hong Z, Mannarino AF, Weber PC. Crystal structure of the RNA-dependent RNA polymerase from hepatitis C virus reveals a fully encircled active site. *Nat Struct Mol Biol.* 1999;6(10):937.
46. Tagger A, Donato F, Ribero ML, Chiesa R, Portera G, Gelatti U, Albertini A, Fasola M, Boffetta P, Nardi G. Case-control study on hepatitis C virus (HCV) as a risk factor for hepatocellular carcinoma: the role of HCV genotypes and the synergism with hepatitis B virus and alcohol. *Int J Cancer.* 1999;81(5):695–9.
47. Tsukuma H, Hiyama T, Tanaka S, Nakao M, Yabuuchi T, Kitamura T, Nakanishi K, Fujimoto I, Inoue A, Yamazaki H, Kawashima T. Risk factors for hepatocellular carcinoma among patients with chronic liver disease. *N Engl J Med.* 1993;328(25):1797–801.
48. El-Serag HB, Mason AC. Rising incidence of hepatocellular carcinoma in the United States. *N Engl J Med.* 1999;340(10):745–50.
49. Mas VR, Maluf DG, Archer KJ, Yanek K, Kong X, Kulik L, Freise CE, Olthoff KM, Ghobrial RM, McIver P, Fisher R. Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma. *Mol Med.* 2009;15(3–4):85–94.
50. Kane MJ, Price N, Scotch M, Rabinowitz P. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks. *BMC Bioinform.* 2014;15(1):276.
51. Zurada JM. Introduction to artificial neural systems. St. Paul: West Publishing Company; 1992.
52. Marcondes CB, Ximenes MD. Zika virus in Brazil and the danger of infestation by Aedes (Stegomyia) mosquitoes. *Rev Soc Bras Med Trop.* 2016;49(1):4–10.
53. Mlakar J, Korva M, Tul N, Popović M, Poljšak-Prijatelj M, Mraz J, Kolenc M, Resman Rus K, Vesnaver Vipotnik T, Fabjan Vodusek V, Vizjak A. Zika virus associated with microcephaly. *N Engl J Med.* 2016;374(10):951–8.
54. Driggers RW, Ho CY, Korhonen EM, Kuivanen S, Jääskeläinen AJ, Smura T, Rosenberg A, Hill DA, DeBiasi RL, Vezina G, Timofeev J. Zika virus infection with prolonged maternal viremia and fetal brain abnormalities. *N Engl J Med.* 2016;374(22):2142–51.
55. Brasil P, Pereira JP Jr, Moreira ME, Ribeiro Nogueira RM, Damasceno L, Wakimoto M, Rabello RS, Valderramos SG, Halai UA, Salles TS, Zin AA. Zika virus infection in pregnant women in Rio de Janeiro. *N Engl J Med.* 2016;375(24):2321–34.
56. Akhtar M, Kraemer MU, Gardner L. A dynamic neural network model for real-time prediction of the Zika epidemic in the Americas bioRxiv 2018 Jan 1:466581.
57. Leontaritis IJ, Billings SA. Input-output parametric models for non-linear systems part I: deterministic non-linear systems. *Int J Control.* 1985;41(2):303–28.
58. Narendra KS, Parthasarathy K. Identification and control of dynamical systems using neural networks. *IEEE Trans Neural Netw.* 1990;1(1):4–27.

59. Chen S, Billings SA, Grant PM. Non-linear system identification using neural networks. *Int J Control.* 1990;51(6):1191–214.
60. Tersmette MJ, De Goede RE, Al BJ, Winkel IN, Gruters RA, Cuypers HT, Huisman HG, Miedema F. Differential syncytium-inducing capacity of human immunodeficiency virus isolates: frequent detection of syncytium-inducing isolates in patients with acquired immunodeficiency syndrome (AIDS) and AIDS-related complex. *J Virol.* 1988;62(6):2026–32.
61. Alkhatib G, Combadiere C, Broder CC, Feng Y, Kennedy PE, Murphy PM, Berger EA. CC CKR5: a RANTES, MIP-1 $\alpha$ , MIP-1 $\beta$  receptor as a fusion cofactor for macrophage-tropic HIV-1. *Science.* 1996 Jun 28;272(5270):1955–8.
62. Liu R, Paxton WA, Choe S, Ceradini D, Martin SR, Horuk R, MacDonald ME, Stuhlmann H, Koup RA, Landau NR. Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell.* 1996;86(3):367–77.
63. Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhardt M, Marzio PD, Marmon S, Sutton RE, Hill CM, Davis CB. Identification of a major co-receptor for primary isolates of HIV-1. *Nature.* 1996;381(6584):661–6.
64. Doranz BJ, Rucker J, Yi Y, Smyth RJ, Samson M, Peiper SC, Parmentier M, Collman RG, Doms RW. A dual-tropic primary HIV-1 isolate that uses fusin and the  $\beta$ -chemokine receptors CKR-5, CKR-3, and CKR-2b as fusion cofactors. *Cell.* 1996;85(7):1149–58.
65. Dragic T, Litwin V, Allaway GP, Martin SR, Huang Y, Nagashima KA, Cayanan C, Maddon PJ, Koup RA, Moore JP, Paxton WA. HIV-1 entry into CD4+ cells is mediated by the chemokine receptor CC-CKR-5. *Nature.* 1996 Jun;381(6584):667–73.
66. Feng Y, Broder CC, Kennedy PE, Berger EA. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science.* 1996;272(5263):872–7.
67. Chesebro B, Wehrly K, Nishio J, Perryman S. Mapping of independent V3 envelope determinants of human immunodeficiency virus type 1 macrophage tropism and syncytium formation in lymphocytes. *J Virol.* 1996;70(12):9055–9.
68. Choe H, Farzan M, Sun Y, Sullivan N, Rollins B, Ponath PD, Wu L, Mackay CR, LaRosa G, Newman W, Gerard N. The  $\beta$ -chemokine receptors CCR3 and CCR5 facilitate infection by primary HIV-1 isolates. *Cell.* 1996;85(7):1135–48.
69. Cocchi F, DeVico AL, Garzino-Demo A, Cara A, Gallo RC, Lusso P. The V3 domain of the HIV-1 gp120 envelope glycoprotein is critical for chemokine-mediated blockade of infection. *Nat Med.* 1996;2(11):1244–7.
70. Hwang SS, Boyle TJ, Lyerly HK, Cullen BR. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science.* 1991;253(5015):71–4.
71. Resch W, Hoffman N, Swanstrom R. Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. *Virology.* 2001;288(1):51–62.
72. MacKay DJ. Bayesian interpolation. *Neural Comput.* 1992;4(3):415–47.
73. Rodellar J, Alférez S, Acevedo A, Molina A, Merino A. Image processing and machine learning in the morphological analysis of blood cells. *Int J Lab Hematol.* 2018;40:46–53.
74. Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol.* 2016;12(7):878.
75. Ebrahimi M, Ahsan R. The first implication of image processing techniques on influenza a virus sub-typing based on HA/NA protein sequences, using convolutional deep neural Network. *BioRxiv* 2018 Jan 1:448159.
76. <https://www.uniprot.org>.
77. Han Y, Kim D. Deep convolutional neural networks for pan-specific peptide-MHC class I binding prediction. *BMC Bioinform.* 2017;18(1):585.
78. Liu W, Meng X, Xu Q, Flower DR, Li T. Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinform.* 2006 Dec;7(1):182.

79. The World Bank. Reducing Climate-Sensitive Risks. 2014, Volume 1. Available online: <http://documents.worldbank.org/curated/en/486511468167944431/Reducing-climate-sensitive-disease-risks>. Accessed on 20 June 2017.
80. Fuentes A, Yoon S, Kim S, Park D. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *Sensors*. 2017;17(9):2022.
81. Krizhenivshky A, Sutskever I, Hinton G. Imagenet classification with deep convolutional networks. In Proceedings of the Conference Neural Information Processing Systems (NIPS). p. 1097–105.
82. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*. 2014 Sep 4.
83. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A. Going deeper with convolutions. In Proceedings of the IEEE conference on computer vision and pattern recognition 2015. p. 1–9.
84. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition 2016. p. 770–8.
85. Xie S, Girshick R, Dollár P, Tu Z, He K. Aggregated residual transformations for deep neural networks. In Proceedings of the IEEE conference on computer vision and pattern recognition 2017. p. 1492–500.
86. Holland JH. Genetic algorithms. *Sci Am*. 1992;267(1):66–73.
87. Goldberg DE, Holland JH. Genetic algorithms and machine learning. *Mach Learn*. 1988;3(2):95–9.
88. Michalewicz Z, Janikow CZ, Krawczyk JB. A modified genetic algorithm for optimal control problems. *Comput Math Appl*. 1992;23(12):83–94.
89. Gallo RC, Sarin PS, Gelmann EP, Robert-Guroff M, Richardson E, Kalyanaraman VS, Mann D, Sidhu GD, Stahl RE, Zolla-Pazner S, Leibowitch J. Isolation of human T-cell leukemia virus in acquired immune deficiency syndrome (AIDS). *Science*. 1983;220(4599):865–7.
90. Barré-Sinoussi F, Chermann JC, Rey F, Nugeyre MT, Chamaret S, Gruest J, Dauguet C, Axler-Blin C, Vézinet-Brun F, Rouzioux C, Rozenbaum W. Isolation of a T-lymphotropic retrovirus from a patient at risk for acquired immune deficiency syndrome (AIDS). *Science*. 1983;220(4599):868–71.
91. Jung A, Maier R, Vartanian JP, Bocharov G, Jung V, Fischer U, Meese E, Wain-Hobson S, Meyerhans A. Recombination: multiply infected spleen cells in HIV patients. *Nature*. 2002;418(6894):144.
92. Bocharov G, Ford NJ, Edwards J, Breinig T, Wain-Hobson S, Meyerhans A. A genetic-algorithm approach to simulating human immunodeficiency virus evolution reveals the strong impact of multiply infected cells and recombination. *J Gen Virol*. 2005;86(11):3109–18.
93. De Clercq E. Emerging anti-HIV drugs. *Expert Opin Emerg Drugs*. 2005;10(2):241–74.
94. Mekouar K, Mouscadet JF, Desmaële D, Subra F, Leh H, Savouré D, Auclair C, d'Angelo J. Styrylquinoline derivatives: a new class of potent HIV-1 integrase inhibitors that block HIV-1 replication in CEM cells. *J Med Chem*. 1998;41(15):2846–57.
95. Goudarzi N, Goodarzi M, Chen T. QSAR prediction of HIV inhibition activity of styrylquinoline derivatives by genetic algorithm coupled with multiple linear regressions. *Med Chem Res*. 2012;21(4):437–43.
96. Leonard JT, Roy K. Exploring molecular shape analysis of styrylquinoline derivatives as HIV-1 integrase inhibitors. *Eur J Med Chem*. 2008;43(1):81–92.
97. Cong Y, Li BK, Yang XG, Xue Y, Chen YZ, Zeng Y. Quantitative structure–activity relationship study of influenza virus neuraminidase A/PR/8/34 (H1N1) inhibitors by genetic algorithm feature selection and support vector regression. *Chemom Intell Lab Syst*. 2013;127:35–42.
98. Williams MA, Lew W, Mendel DB, Tai CY, Escarpe PA, Laver WG, Stevens RC, Kim CU. Structure–activity relationships of carbocyclic influenza neuraminidase inhibitors. *Bioorg Med Chem Lett*. 1997;7(14):1837–42.
99. Kim CU, Lew W, Williams MA, Wu H, Zhang L, Chen X, Escarpe PA, Mendel DB, Laver WG, Stevens RC. Structure–activity relationship studies of novel carbocyclic influenza neuraminidase inhibitors. *J Med Chem*. 1998;41(14):2451–60.

100. Lew W, Wu H, Mendel DB, Escarpe PA, Chen X, Laver WG, Graves BJ, Kim CU. A new series of C3-aza carbocyclic influenza neuraminidase inhibitors: synthesis and inhibitory activity. *Bioorg Med Chem Lett.* 1998;8(23):3321–4.
101. Lew W, Wu H, Chen X, Graves BJ, Escarpe PA, MacArthur HL, Mendel DB, Kim CU. Carbocyclic influenza neuraminidase inhibitors possessing a C3-cyclic amine side chain: synthesis and inhibitory activity. *Bioorg Med Chem Lett.* 2000;10(11):1257–60.
102. Lew W, Williams MA, Mendel DB, Escarpe PA, Kim CU. C3-Thia and C3-carba isosteres of a carbocyclic influenza neuraminidase inhibitor, (3R, 4R, 5S)-4-acetamido-5-amino-3-propoxyl-1-cyclohexene-1-carboxylic acid. *Bioorg Med Chem Lett.* 1997;7(14):1843–1846. C3-Thia and C3-carba isosteres of a carbocyclic influenza neuraminidase inhibitor, (3R,4R,5S)-4-acetamido-5-amino-3-propoxyl-1-cyclohexene-1-carboxylic acid.
103. Zhang L, Williams MA, Mendel DB, Escarpe PA, Kim CU. Synthesis and activity of C2-substituted analogs of influenza neuraminidase inhibitor GS 4071. *Bioorg Med Chem Lett.* 1997;7(14):1847–50.
104. Zhang L, Williams MA, Mendel DB, Escarpe PA, Chen X, Wang KY, Graves BJ, Lawton G, Kim CU. Synthesis and evaluation of 1, 4, 5, 6-tetrahydropyridazine derivatives as influenza neuraminidase inhibitors. *Bioorg Med Chem Lett.* 1999;9(13):1751–6.
105. Liu AL, Wang HD, Lee SM, Wang YT, Du GH. Structure–activity relationship of flavonoids as influenza virus neuraminidase inhibitors and their in vitro anti-viral activities. *Bioorg Med Chem.* 2008;16(15):7141–7.
106. Todeschini R, Consonni V. *Handbook of molecular descriptors*. Wiley (Weinheim, Germany); 2008.
107. Wang C, Lefkowitz EJ. Genomic multiple sequence alignments: refinement using a genetic algorithm. *BMC Bioinform.* 2005;6(1):200.
108. Hirosawa M, Totoki Y, Hoshida M, Ishikawa M. Comprehensive study on iterative algorithms of multiple sequence alignment. *Bioinformatics.* 1995;11(1):13–8.
109. Nicholas HB Jr, Ropelewski AJ, Deerfield DW. Strategies for multiple sequence alignment. *BioTechniques.* 2002;32(3):572–91.
110. Notredame C, Holm L, Higgins DG. COFFEE: an objective function for multiple sequence alignments. *Bioinformatics (Oxford, England).* 1998;14(5):407–22.
111. Notredame C, Higgins DG, Heringa J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol.* 2000;302(1):205–17.
112. Thomsen R, Boomsma W. Multiple sequence alignment using SAGA: investigating the effects of operator scheduling, population seeding, and crossover operators. In: *Workshops on applications of evolutionary computation 2004 Apr 5, p. 113–22.* Springer, Berlin, Heidelberg.
113. Calonaci C, Chiacchio F, Pappalardo F. Optimal vaccination schedule search using genetic algorithm over MPI technology. *BMC Med Inform Decis Mak.* 2012;12(1):129.
114. Pappalardo F, Lollini PL, Castiglione F, Motta S. Modeling and simulation of cancer immunoprevention vaccine. *Bioinformatics.* 2005;21(12):2891–7.
115. Pappalardo F, Forero IM, Pennisi M, Palazon A, Melero I, Motta S. SimB16: modeling induced immune system response against B16-melanoma. *PLoS One.* 2011;6(10):e26523.
116. Pappalardo F, Halling-Brown MD, Rapin N, Zhang P, Alemani D, Emerson A, Paci P, Duroux P, Pennisi M, Palladini A, Miotto O. ImmunoGrid, an integrative environment for large-scale simulation of the immune system for vaccine discovery, design and optimization. *Brief Bioinform.* 2009;10(3):330–40.
117. Pennisi M, Catanuto R, Pappalardo F, Motta S. Optimal vaccination schedules using simulated annealing. *Bioinformatics.* 2008;24(15):1740–2.
118. Halling-Brown M, Pappalardo F, Rapin N, Zhang P, Alemani D, Emerson A, Castiglione F, Duroux P, Pennisi M, Miotto O, Churchill D. ImmunoGrid: towards agent-based simulations of the human immune system at a natural scale. *Philos Trans A Math Phys Eng Sci.* 2010;368(1920):2799–815.
119. Pennisi M, Pappalardo F, Palladini A, Nicoletti G, Nanni P, Lollini PL, Motta S. Modeling the competition between lung metastases and the immune system using agents. *BMC Bioinform.* 2010;11(7):S13. BioMed Central.

120. Colorni A, Dorigo M, Maniezzo V. Distributed optimization by ant colonies. In Proceedings of the first European conference on artificial life 1992 Dec, vol. 142, p. 134–42.
121. Dorigo M, Gambardella LM. Ant colonies for the travelling salesman problem. *Biosystems*. 1997;43(2):73–81.
122. Bonabeau E, Marco DD, Dorigo M, Theraulaz G. Swarm intelligence: from natural to artificial systems: Oxford University Press (New York); 1999.
123. Pemula AK, Krishnan S. Temporal analysis of molecular changes in shrimp (*Penaeus vannamei*) tissues with respect to white spot disease. *J Food Sci Technol*. 2015;52(11):7236–44.
124. Verbruggen B, Bickley L, van Aerle R, Bateman K, Stentiford G, Santos E, Tyler C. Molecular mechanisms of white spot syndrome virus infection and perspectives on treatments. *Viruses*. 2016;8(1):23.
125. Solís-Lucero G, Manoutcharian K, Hernández-López J, Ascencio F. Injected phage-displayed-VP28 vaccine reduces shrimp *Litopenaeus vannamei* mortality by white spot syndrome virus infection. *Fish Shellfish Immunol*. 2016;55:401–6.
126. Oprea TI, Matter H. Integrating virtual screening in lead discovery. *Curr Opin Chem Biol*. 2004;8(4):349–58.
127. Korb O, Stützle T, Exner TE. PLANTS: application of ant colony optimization to structure-based drug design. In International workshop on ant colony optimization and swarm intelligence 2006 Sep 4, Springer, Berlin, Heidelberg, p. 247–58.
128. Tanaka H, Takashima H, Ubasawa M, Sekiya K, Nitta I, Baba M, Shigeta S, Walker RT, De Clercq E, Miyasaka T. Structure-activity relationships of 1-[(2-hydroxyethoxy) methyl]-6-(phenylthio) thymine analogs: effect of substitutions at the C-6 phenyl ring and at the C-5 position on anti-HIV-1 activity. *J Med Chem*. 1992;35(2):337–45.
129. Tanaka H, Takashima H, Ubasawa M, Sekiya K, Nitta I, Baba M, Shigeta S, Walker RT, De Clercq E, Miyasaka T. Synthesis and antiviral activity of deoxy analogs of 1-[(2-hydroxyethoxy) methyl]-6-(phenylthio) thymine (HEPT) as potent and selective anti-HIV-1 agents. *J Med Chem*. 1992;35(25):4713–9.
130. Gayen S, Debnath B, Samanta S, Jha T. QSAR study on some anti-HIV HEPT analogues using physicochemical and topological parameters. *Bioorg Med Chem*. 2004;12(6):1493–503.
131. De Clercq E. Perspectives of non-nucleoside reverse transcriptase inhibitors (NNRTIs) in the therapy of HIV-1 infection. *Farmaco*. 1999;54(1–2):26–45.
132. Hansch C, Zhang L. QSAR of HIV inhibitors. *Bioorg Med Chem Lett*. 1992;2(9):1165–9.
133. Bajaj S, Sambi SS, Madan AK. Topochemical model for prediction of anti-HIV activity of HEPT analogs. *Bioorg Med Chem Lett*. 2005;15(2):467–9.
134. Jalali-Heravi M, Parastar F. Use of artificial neural networks in a QSAR study of anti-HIV activity for a large group of HEPT derivatives. *J Chem Inf Comput Sci*. 2000;40(1):147–54.
135. Garg R, Gupta SP, Gao H, Babu MS, Debnath AK, Hansch C. Comparative quantitative structure–activity relationship studies on anti-HIV drugs. *Chem Rev*. 1999;99(12):3525–602.
136. Kireev DB, Chrétien JR, Grierson DS, Monneret C. A 3D QSAR study of a series of HEPT analogues: the influence of conformational mobility on HIV-1 reverse transcriptase inhibition. *J Med Chem*. 1997;40(26):4257–64.
137. Hannongbua S, Nivesanond K, Lawtrakul L, Pungpo P, Wolschann P. 3D-quantitative structure–activity relationships of HEPT derivatives as HIV-1 reverse transcriptase inhibitors, based on Ab initio calculations. *J Chem Inf Comput Sci*. 2001;41(3):848–55.
138. Bazouzi H, Zahouily M, Boulajaaj S, Sebti S, Zakarya D. QSAR for anti-HIV activity of HEPT derivatives. *SAR QSAR Environ Res*. 2002;13(6):567–77.
139. Douali L, Villemain D, Cherqaoui D. Neural networks: accurate nonlinear QSAR model for HEPT derivatives. *J Chem Inf Comput Sci*. 2003;43(4):1200–7.
140. Akhlaghi Y, Kompany-Zareh M. Application of radial basis function networks and successive projections algorithm in a QSAR study of anti-HIV activity for a large group of HEPT derivatives. *J Chemometr*. 2006;20(1–2):1–2.
141. Gaudio AC, Montanari CA. HEPT derivatives as non-nucleoside inhibitors of HIV-1 reverse transcriptase: QSAR studies agree with the crystal structures. *J Comput Aided Mol Des*. 2002;16(4):287–95.

142. Luco JM, Ferretti FH. QSAR based on multiple linear regression and PLS methods for the anti-HIV activity of a large group of HEPT derivatives. *J Chem Inf Comput Sci.* 1997;37(2):392–401.
143. Zare-shahabadi V, Abbasitalab F. Application of ant colony optimization in development of models for prediction of anti-HIV-1 activity of HEPT derivatives. *J Comput Chem.* 2010;31(12):2354–62.
144. Shamsipur M, Zare-Shahabadi V, Hemmateenejad B, Akhond M. Combination of ant colony optimization with various local search strategies. A novel method for variable selection in multivariate calibration and QSPR study. *QSAR Comb Sci.* 2009;28(11–12):1263–75.
145. Sammut C, Webb GI, editors. *Encyclopedia of machine learning*. Springer Science & Business Media (New York); 2011 Mar 28.
146. Eberhart RC, Hu X. Human tremor analysis using particle swarm optimization. In *Proceedings of the 1999 congress on evolutionary computation-CEC99* (Cat. No. 99TH8406) 1999, vol. 3, p. 1927–30. IEEE.
147. Neshat M, Sargolzaei M, Nadjaran Toosi A, Masoumi A. Hepatitis disease diagnosis using hybrid case based reasoning and particle swarm optimization. *ISRN Artificial Intelligence* 2012 Jul 8;2012.
148. Lisboa PJ, Ifeachor EC, Szczepaniak PS, editors. *Artificial neural networks in biomedicine*: Springer Science and Business Media (Tyne & Wear, England); 2000.
149. Neshat M, Yaghobi M. FESHDD: fuzzy expert system for hepatitis B diseases diagnosis. In *2009 Fifth International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control 2009 Sep.*
150. Neshat M, Yaghobi M. Designing a fuzzy expert system of diagnosing the hepatitis B intensity rate and comparing it with adaptive neural network fuzzy system. In *Proceedings of the World Congress on Engineering and Computer Science 2009 Oct*, vol. 2, p. 797–802.
151. Greig J, du Cros P, Klarkowski D, Mills C, Jørgensen S, Harrigan PR, O'Brien DP. Viral load testing in a resource-limited setting: quality control is critical. *J Int AIDS Soc.* 2011;14(1):23.
152. Calmy A, Ford N, Hirscher B, Reynolds SJ, Lynen L, Goemaere E, De La Vega FG, Perrin L, Rodriguez W. HIV viral load monitoring in resource-limited regions: optional or necessary? *Clin Infect Dis.* 2007;44(1):128–34.
153. Kamalan K, Jawahar PM. Particle swarm optimization based estimation of HIV-1 viral load in resource limited settings. *Afr J Microbiol Res.* 2013;7(20):2297–304.
154. Ren S, He K, Girshick R, Sun J. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems 2015*. p. 91–9.
155. Liu W, Anguelov D, Erhan D, Szegedy C, Reed S, Fu CY, Berg AC. Ssd: single shot multibox detector. In *European conference on computer vision 2016 Oct 8*. Springer, Cham, p. 21–37.
156. Dai J, Li Y, He K, Sun J. R-fcn: Object detection via region-based fully convolutional networks. In *Advances in neural information processing systems 2016*, p. 379–87.

# Non-immune Modulators of Cellular Immune Surveillance to HIV-1 and Other Retroviruses: Future Artificial Intelligence-Driven Goals and Directions



Francesco Chiappelli, Allen Khakshooy, and Nicole Balenton

**Abstract** Immune surveillance to viruses and other foreign pathogens involves a specific process, which is well-characterized in terms of the immune and non-immune cells and factors involved, and their specific timeline. For example, first exposure to a viral antigen involves major histocompatibility complex type I molecules to present the ‘self’ and ‘nonself’ antigen to CD3 + CD8 + CD45RA+ T cells. This presentation initiates the event of T cell activation, which results in the production of T cell growth factor (interleukin 2, IL-2), interferon-gamma (IFN- $\gamma$ ) and other immune soluble products, which together act to promote the clonal proliferative expansion of the responding cell population. Soluble immune factors produced during this initial phase of the immune response also initiate the maturation of the responding T cells into CD3 + CD8 + CD45R0+ memory cells against that specific viral antigen, and of related cell populations that act to control and dampen cellular immune reactivity. Contemporaneously, these soluble immune factors trigger non-immune cells to release of non-immune soluble factors, including pituitary and adrenocortical hormones (e.g., glucocorticoids), which finely modulate immune cell responses. Together, immune and non-immune cells and soluble factors act in concert to engender and sustain a finely tuned immune surveillance process, whose ultimate end ( $Y$ ) is to regain homeostatic balance of the organism – a healthy state, that is eradication of the immune-pathological signs and symptoms that derived from the viral infection, referring here to homeostatic balance  $Y$ .

In brief, immune and non-immune cells and soluble factors are concerted predicting factors for regaining  $Y$ , to the same extent as the initial viral challenge, and other factors related to the state of heterostasis of the organism. These states, which include unhealthy eating habits, sleep deprivation, stress, and the like, act in concert to delay, hamper and counter  $Y$ .

Therefore, from a biostatistical viewpoint, the problem becomes a relatively basic multiple regression, in which the outcome  $Y$ , the homeostatic state of health regained following a viral infection, is simply the sum of positive and negative factors and/or events. Positive factors and events ( $\Pi$ ) inherently push allostasis forward

---

F. Chiappelli (✉) · A. Khakshooy · N. Balenton  
UCLA School of Dentistry, Los Angeles, CA, USA  
e-mail: [fchiappelli@dentistry.ucla.edu](mailto:fchiappelli@dentistry.ucla.edu)

(i.e., the orderly process of immune activation and maturation), but the negative ( $N$ ) factors and events, allostatically speaking, interfere with attaining  $Y$ . Simplistically,  $Y$  is the product of the fine, coordinated and time-regulated interaction between all the interacting  $\Pi$ 's and  $N$ 's during the immune surveillance process ( $Y = \Sigma\Pi + \Sigma N$ ).

The question then becomes, knowing what we know today about the constituents of  $\Sigma\Pi$  and of  $\Sigma N$ , can we not design, by means of bioinformatics, artificial  $\Pi$ 's and  $N$ 's, that may push the organism's response more securely through all the allostatic phases to  $Y$ , the homeostatic state of health regained following a viral infection? Physiology has been able to a related feat by producing bioinformatics particles, which when injected in patients help regulate cholesterol levels. Future artificial intelligence (AI) advances will produce the artificial  $\Pi$ 's and  $N$ 's, which will aid regaining  $Y$ . In the meanwhile, as science continues to complete our knowledge of all the  $\Pi$ 's and of all the  $N$ 's involved, "tweening", the computerized process by which, knowing the end-product of a sequence, the steps in-between can be programmed, will be applied to our conceptualization of immune surveillance events.

In conclusion, the novel science of immune-tweening will help us understand and complete a set of immune and non-immune events that lead to  $Y$ , the homeostatic state of health regained following a viral infection. AI, on the other hand, holds strong promise to help us generate and produce bioinformatics or 'micro-adjuvants', as it were, of immune surveillance. We envisage that these giant steps in the future of viral immunity will first be achieved in the context of infection with the human immunodeficiency virus (HIV) because it has become the model for our understanding of anti-viral immune surveillance.

**Keywords** Viral immune surveillance · Neuroendocrine modulators of cellular immunity · Allostasis · Bayesian prediction model · Artificial intelligence · Immune-tweening

## 1 Cellular Immune Surveillance to Retroviruses

### 1.1 *The Immune System*

As the primary host defense system, the immune system is comprised of several soluble and insoluble components, which can be segregated functionally into an innate component of the immunity, and an antigen-dependent (or acquired) immune component [1–3]. Soluble immune factors include growth, migration and maturation factors, which are typically subsumed as cytokines under a few principal family groups, such as interleukins (ILs), interferons (INFs), complement factors, antimicrobial peptides (e.g.,  $\beta$ -defensins), enzymes (e.g., lysozyme, salivary phospholipase A2), and others. Soluble immune factors also include immunoglobulins (Ig) [3].

Insoluble components of immunity comprise a circulatory system, the lymphatic system, which is the structural anatomical bridge between cell-mediated immune

surveillance and the arterio-venous system. Anatomically, the lymphatic system is widespread within the body running from every major organ to the most distant and minute extremities. It connects the major lymphoid organs, such as the bone marrow, the thymus, and the spleen to other organs that can also play a significant role in immune surveillance, including the intestinal mucosa, the liver, the skin and others. The lymphatic system, like the venous and the arterial systems, consists of vessels, which structurally are juxtaposed endothelial cells by means of tight junctions, capillary beds, and smooth musculature intertwined by sympathetic innervation. The sympathetic nervous system plays an important role in the regulation of lymphatic system events. Lymphatic vessels are characterized by the presence of lymph nodes at roughly regular intervals. The lymph, or the fluid within the lymphatic vasculature, utilize this system and surveil the contents of this fluid that drain into and out of lymph nodes and back to the blood.

Immune cells that circulate within the lymphatic vessels can penetrate in the cortical region of the lymph nodes to engage specialized cells that present antigens, thus triggering the maturational events that will signify, within the span of 5–7 days, a concerted immune reaction specifically directed at the removal of the antigen and any cell in the organism that might have been infected by the antigen. The process of maturation of immune cells occurs as the challenged cells traverse the cortical region into the medullary region of the node. Once mature, they exit the lymph node and engage again in lymphatic circulation. In brief, the cells that circulate within the lymphatic vessels consist of a mixture of functionally mature and immature specialized immune cells. The healthier the organisms, the more naïve cells are found in the lymphatic vessels; the more widespread an infection, the more mature cells circulate in the lymphatic system. The lymphatic system has two principal ports of drainage in the venous circulation, which ensure that mature immune cells routinely enter the venous and arterial circulation, and therefore surveil the entire organism [4–8].

Besides the anatomical lymphatic structure, the insoluble component of immune surveillance entails three principal functionally distinct families of cells, which ontogenetically arise from the bone marrow, but terminate their maturational sequence elsewhere. One family has the primary function of destroying foreign pathogens, either chemically, by phagocytosis, or by perforating holes in their plasma membrane and killing them. Neutrophils are the most abundant type of phagocyte, normally representing 50–60% of the total circulating immune cells, rapidly migrate toward the site of inflammation, characterized by redness, swelling, heat, and pain, in a process called chemotaxis, and destroy their targets either by phagocytosis or chemically, which tends to aggravate inflammatory responses. Mast cells reside in connective tissues and mucous membranes, and regulate the inflammation. The process of phagocytosis cannot be dependent on specific recognition sites because each pathogen is different, eventually leading to a presentation of certain peptidic structures of the pathogen by means of the major histocompatibility complex (MHC) class II. In this case, the immune cells doing the presentation of foreign pathogen structures are called antigen-presenting cells (APC). There several sub-families of APC's, but the principal ones are either generally found in the tissues,

i.e., dendritic cells, or in the circulation, i.e., monocytes and macrophages M0, M1 or M2. It is the presence of these ‘nonsel’ structures that are presented within the ‘sel’ structure of the MHC that triggers an antigen-dependent immune response. When the foreign pathogen is internal to the cell, such as might be the case in a cancer cell or a virally-infected cell, then any cell in the body has the potential of turning in an APC, and it will utilize its MHC Class I, which all cells of the same organism share, to present peptidic aspects of the cancer or viral genome. The process of killing any cell recognized as foreign pathogen can also be independent of any specific site recognition, and manifests what is referred to, in those cases, as ‘natural’ killing. Thus the sub-family of cells specialized in this type of killing are the Natural Killer (NK) cells [2, 3, 9–11].

The second major group of immune cells involved in immune surveillance events are derived from a group of cells that leave the bone marrow as pre-thymocytes, are educated in the outer cortex of the thymus by epithelial cells mainly secreting IL-7. This education prepares the response to an immune challenge and, most importantly, to the fact that ‘sel’ – i.e., MHC class I or MHC Class II per se, is not an immune challenge. Now refined, they migrate toward the medulla and exit as leukocytes expressing the common leukocyte cluster of differentiation (CD), CD45+, expressing the T cell marker CD3, which is associated with the T cell receptor (TcR – predominantly alpha/beta chains,  $\alpha/\beta$ , but in the mucosal immune system gamma/delta chains,  $\gamma\delta$ ), and either expressing CD4 or CD8. These are naïve T cells, having never encountered a ‘nonsel’ antigen, and thus express the longest restriction fragment, A, of CD45, and are CD45RA+.

It is not clear where this education occurs after thymus atrophy in puberty. What is clear is that a large proportion of naïve T cells are found circulating throughout the lifespan into aging. Even centenarians have been reported to maintain a healthy 40–50% proportion of naïve T cell in the circulation. It has been hypothesized that, following thymic atrophy, the intestinal Peyer’s patches may take on the role of T cell education – improbable, yet possible. It has also been hypothesized that the process of thymic atrophy is not generalized, but, on the contrary, rather specific: the regions with limited sympathetic innervation atrophy, but the regions that are richly innervated condense and continue to sustain thymic function. As the organism ages and these regions become increasing rarified, then the proportion of circulating naïve cells begin to fall, a process that is now believed to be an important factor in immune senescence – possible, probable and indeed supported by an emerging body of data. Naïve T cells that circulate in the venous/arterial or the lymphatic systems are recognized as CD45RA+ CD3+ CD4+/CD8+. As these cells encounter an APC that presents ‘sel’ plus ‘nonsel’. CD4+ and CD8+ T cells go forward to progress along several lines of maturation, some of which are believed to be terminal and irreversible, whereas others are thought to be more plastic in nature allowing the maturing subpopulation to go back and forth between functional states, depending on the microenvironment of IL’s, INF’s and other factors. CD4+ T cells will respond to a foreign antigen challenge presented on MHC class II. CD8+ T cells will respond to a foreign antigen challenge presented on MHC class I. Both the CD4+ and CD8+ sub-population is endowed on naïve (CD45RA+) and memory

(CD45R0+), resting (CD25-) and activated (CD25+) cells that express the genomic marker FoxP3. These cells are regulatory T cells (TRegs) in that they attenuate T cell-mediated immune responses. By far the larger proportion of TRegs are recognized by the characteristic immunophenotypic profile of CD45RA + CD3 + CD4 + CD25 + Foxp3+ [2, 3, 12–25].

The third principal group of immune cells is indifferent to viral immune surveillance. The B cells, so called due to their first identification in birds, observed noted to become educated in the Bursa of Fabricius. In humans, B cells are educated in the bone marrow, and terminate their maturation ‘on site’, as it were. Upon foreign encounter, the clonal expansion, that is, the activation and proliferation of B cells produce several types and sub-types of immunoglobulins (Ig) or antibodies. The ubiquitous structural integrity of immunoglobulins allows the binding to specific epitopes on the pathogens and aid in their destruction either IG Fc-region dependent killing via a killer cell immunoglobulin receptors (KIR), or by emulsification [3, 26, 27].

In brief, immune surveillance is the process by which organisms are protected from infectious pathogens by means of layered defenses of increasing specificity. In the context of this chapter, we are concerned principally by immune surveillance events brought about and directed by CD8+ T cells [2, 3].

## 1.2 *Retroviral Immune Surveillance*

Viral cellular immune surveillance is a dynamic and fluid system that is driven by finely regulated cellular processes including cytokines and other factors locally in the microenvironment and systemically throughout the body [2, 3, 28–30]. The innate immune system plays an important role in viral immunity. Innate cells and humoral factors are essential for the initial detection of invading viruses and subsequent activation of acquired immunity [31]. Toll-like receptors (TLRs) are single membrane-spanning non-catalytic receptors expressed by myeloid and dendritic sentinel cells. TLRs act as pattern recognition receptors (PRR) that recognize pathogen-associated molecular patterns (PAMPs), specific structurally conserved molecules derived from invading pathogens, that activate innate immune cell responses and transduce the signal. TLRs initially detect viral invasion and activate transcription factors, including IRF3, leading to induction of IFN’s production. These events are facilitated by the membrane phospholipid phosphatidyl inositol-5-phosphate (PtdIns5P) by binding to IRF3 and its up-stream kinase TRAF family member-associated NF- $\kappa$ B activator (TANK)-binding kinase 1 (TBK1). This interaction promotes TBK1-mediated IRF3 phosphorylation and activation because TBK1 is a serine/threonine kinase that plays an essential role in regulating inflammatory responses to viral pathogens. Receptors and ligands of the tumor necrosis factor (TNF) family play important roles in controlling CD8 T cell activation and survival during a viral immune response. The role of specific TNF receptor (TNFR) family member in antiviral immunity depends on the stage of the immune response and can vary with the virus type and its virulence. The trimeric

nature of this cytokine receptor cooperates with the tumor necrosis factor receptor type1-associated DEATH domain protein (i.e., TNFRSF1A-associated via death domain [TRADD]).

By contrast, the TNFR-associated factor (TRAF) modulates primarily the inflammatory response. The receptor-interacting protein (RIP) kinases are a group of threonine/serine protein kinases with relatively conserved kinase domain that mediate broad-based regulation of CD8+ cells activation events. CD120, which constitute two distinct variants of the TNFR superfamily – respectively, CD120a: tumor necrosis factor receptor 1 (TNFR1, aka TNFR superfamily member 1A), and CD120b: tumor necrosis factor receptor 2 (TNFR2, aka TNFR superfamily member 1B) binds to the pro-inflammatory cytokine TNF- $\alpha$ . CD137, the tumor necrosis factor receptor superfamily member 9 (TNFRSF9), also referred to as moiety 4-1BB induced by lymphocyte activation, is expressed by activated CD8+ T cells preferentially, and thus signify engagement of anti-viral immune surveillance. The co-stimulatory immune checkpoint molecule, CD27 is also required for generating and sustaining CD8-mediated viral immunity. In the same vein, the first apoptosis signal (FAS) receptor, sometimes known as apoptosis antigen 1 (APO-1 or APT) is CD95, an apoptotic death membrane receptor that engages the cytoplasmic pathway, rather than the mitochondrial pathway of programmed cell death. CD358 is the tumor necrosis factor receptor super family member 21 (TNFRSF21), the death receptor-6, and closely interacts with TRADD. Its cousin molecule, TNFRSF25, death receptor-3, facilitates the antigen-dependent response by acting as critical driver of CD8-mediated viral immunity: in brief, co-stimulation of TNFRSF25 with a vaccine antigen sharply enhances vaccine-stimulated viral immunity [2, 32].

Following activation of TLRs by the viral components, TBK1 associates with TRAF3 and TANK through diverse scaffolding molecules of the cytoskeleton (e.g., FADD, TRADD, MAVS, SINTBAD) to phosphorylate IRF3 and IRF7, and the multi-functional ATP-dependent RNA helicase, DDX3X, to eventually homodimerize and nuclear translocate the IRFs, for regulating the transcriptional activation of pro-inflammatory and antiviral IFN's genes. Case in point, TBK1 mediates the phosphorylation of certain viral proteins (e.g., Borna disease virus P protein, a multi-helical protein that regulates viral RNA synthesis) and the transport of nucleotide oligomerization domain (NOD)-like receptors (NLRs)). NLR's are critical intracellular sensors component for the recognition of pathogen-associated molecular patterns, and trigger, for instance, the production of IL1- $\beta$ , itself involved in dsRNA stimulation. It is largely for that reason that NLR's play such a central role in DNA vaccines. Small RNAs (sRNAs) also modulate anti-viral immunity, probably by a very similar set of mechanisms [2, 33].

The initial anti-viral immune response by innate immunity produces a rapid rise in pro-inflammatory cytokines (e.g., IL6, IL1- $\beta$ , TNF- $\alpha$ ), which trigger a relatively short-lived initial burst of fever and inflammation and cellular immune migration factors (e.g., IL-8). IL10 levels increase, and so the levels of soluble intracellular adhesion molecule (sICAM)-1, and soluble vascular cell adhesion molecule (sVCAM)-1. This pattern reflects early excessive, and ultimately detrimental endothelial activation in virally infected patients, with consequential increased

plasminogen activator inhibitor 1 (PAI-1). A slower process of cellular pathology ensues, which includes myeloid and endothelial cell infection and cytopathology, followed by a sharp rise in fever indicating systemic inflammatory responses. As the damage to the infected cells progresses, loss in vascular integrity leads to increased permeability of blood vessels with transudates increasingly rich in micronutrients, red blood cells. The organism is ultimately overwhelmed by a combination of inflammatory factors and virus-induced cell damage, which together can lead to death from liver and kidney failure complicated by septic shock. This follows type I IFNs, which include IFN- $\alpha$ , IFN- $\beta$ , IFN- $\kappa$ , IFN- $\delta$ , IFN- $\varepsilon$ , IFN- $\tau$ , IFN- $\omega$ , and IFN- $\zeta$ , impairment [2, 34].

The human immune deficiency virus (HIV), and other related retroviruses can escape systemic immune surveillance by infecting the central nervous system (CNS) [2, 35–40]. It is questionable as to what extent the CNS is an immune-privileged organ protected by the blood-brain barrier (BBB). Recent evidence suggests converging pathways through which viral infection, and its associated immune surveillance processes, may alter the integrity of the blood-brain barrier, and lead to inflammation, swelling of the brain parenchyma and associated neurological syndromes. Here, we expand upon the recent “gateway theory”, by which viral infection and other immune activation states may disrupt the specialized tight junctions of the BBB endothelium making it permeable to immune cells and factors. The model we outline here builds upon the proposition that this process may be initiated by cytokines of the IL-17 family, and recognizing the intimate balance between TH17 and TH9 cytokine profiles systemically. We argue that immune surveillance events, in response to viruses such as the Human Immunodeficiency Virus (HIV), cause a TH17/TH9 induced gateway through blood brain barrier, and thus lead to characteristic neuroimmune pathology. It is possible and even probable that the novel TH17/TH9 induced gateway, which we describe here, opens because of any state of immune activation and sustained chronic inflammation, whether associated with viral infection or any other cause of peripheral or central neuroinflammation. This view could lead to new, timely and critical patient-centered therapies for patients with neuroimmune pathologies due to infection by HIV or other viruses [2, 39, 40].

## 2 Non-immune Modulators of Cellular Immunity

### 2.1 Allostasis

The immune system is subject to finely intertwined physiological regulation. Psychoneuroendocrine factors modulate immune responses, and soluble factors produced by APCs, T cells and other immune cells during response to an antigen alter psychological states, responses by the central and peripheral nervous systems, and production, release and function of hormones and endocrine peptides.

Disruptions in circadian patterns, and sleep deprivation are detrimental to immune function. Complex feedback loops involving ILs, IFNs and other soluble immune factors play an important role in regulating certain phases of sleep, namely non-rapid eye movement (REM) sleep. That is to say, the concerted set of finely modulated events that signify immune surveillance to viral infection response alter the sleep cycle, including an increase in slow-wave sleep relative to REM sleep, and sleep deregulation interferes with immunity, IL balance (e.g., TH1/TH2 cytokines balance), microenvironmental balance and consequentially the regulation of T cell maturation (e.g., TH17/TH9 subpopulation, Tregs), which determine the number, proportion, distribution and functionality of targeted cytotoxic T lymphocytes (CTLs) again specific viral pathogens. Malnutrition, under-nutrition, dietary over-load (i.e., over-nutrition), and diet in general, which has important correlates to tissue healing, repair and regeneration, also seriously impact the processes and events that determine immune surveillance efficiency. The physiological significance of the latter is most evident in the intimate relationship of the immune system with bone metabolism, which is now recognized as the field of osteoimmunology in and of itself. Taken together with the elements outlined above, it is now clear that immune surveillance in general, and immune surveillance to viral infections specifically, is modulated by a complex set of intertwined physiological factors, which is referred to as osteo-psychoneuroendocrine-immunology [41–48].

From embryogenesis, to tissue repair and regeneration, and aging, cellular immune surveillance interacts intimately with bone metabolism, the central and the peripheral nervous systems, the endocrine system and all other recognized bodily processes that contribute to homeostasis, the state of physiological balance. In brief, immune-related processes participate in, and control and regulate most, if not all, events throughout the lifespan that signify the return of the organism from a state of imbalance, or heterostasis, to harmoniously balanced physiological responses, known as homeostasis. The regulated progression from heterostasis to homeostasis is recognized as allostasis, the ability of the organism to maintain stability through change. Allostasis refers, in brief, to a constellation of biological events whose concerted action is driven by the organism's need to regain stability following a challenge. The allostasis load is the overall expenditure that the totality of these events cost to the organism. In type I allostasis, the organism successfully reestablishes physiological balance; in type II allostasis, the challenge is greater than the permissible overall cost expenditure, and homeostasis is not re-established [45, 47].

Specifically, with respect to viral immune surveillance, one can conceive of a viral infection that challenges immune homeostasis. The pathophysiological state that follows viral infection, the manifestation of the symptoms of a viral disease, is a state of virally induced immune heterostasis. The concerted events cellular and humoral events, including those outlined broadly in the section above, which are modulated by a large variety of osteo-psychoneuro-endocrine regulatory inputs, correspond to the process of immune allostasis. When immunity resolves the pathology of the viral infection, then the process in question was an immune allostasis type I. On the contrary, when the concerted immune events do not successfully

resolve the pathobiology of the viral infection, such as in the case of Human Immunodeficiency (HIV) virus infection, then the process is an immune allostasis type II [43–48].

## 2.2 *Salient Models of Non-immune Immuno-Modulators*

The field of non-immune immune-modulators is vast and complex. It is beyond the scope of this chapter to review every dimension of neuro-immune and endocrin-immune cross-talk. Suffice to say that we and others have demonstrated the role of cranial nerves, of the sympathetic and the para-sympathetic nervous systems, of a large variety of neuropeptides, as well as hypothalamic, anterior and posterior pituitary, thyroid, gonadal, adrenocortical and a plethora of other hormones involved in sleep, fat catabolism, bone metabolism and others, in regulating innate and antigen-dependent, soluble and cell-mediated immune surveillance processes at the level of physiological regulation, cell activation, proliferation and migration, as well as molecular and epigenetic pathways. The relationships among these diverse systems of psychobiologic modulation of immunity and viral immune surveillance in general and in patients with HIV/AIDS has been extensively investigated, and continues to be actively explored [41–60].

## 3 Toward a Biostatistical Algorithm

### 3.1 *Multiple Regression*

In statistical modeling, simple or multiple linear or nonlinear regression comprise a set of statistical processes for estimating the relationships among variables. Specifically, regression serves to predict an outcome  $Y$ , based on quantifiable predictor variables [61, 62]. In brief, regression analysis contributes to the understanding of how the typical value of the dependent variable (or ‘criterion variable’) changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Stated somewhat differently, regression is a biostatistical tool to estimate the conditional expectation of the dependent variable,  $Y$ , given quantifiable values of independent variables. The function of the independent variables is the regression function to be estimated, which when resolved quantifies  $Y$ , the outcome variable.

It is beyond the scope of this chapter to discuss the parametric assumptions, which must be satisfied to entertain a multiple regression model. It is also beyond the purpose of this chapter to examine in great details each of the inferential model options proffered by a multiple regression analysis. These topics are covered in detail elsewhere [61–64].

All major statistical software packages perform regression analysis and inference. Specialized regression softwares are also available for use in fields such Bayesian inference.

### **3.2 Cellular Immune Surveillance Predictive Model**

The statistical process of regression was first developed by Sir Francis Galton (1822–1911) to describe certain biological phenomena, such as predicting that the heights of descendants of tall ancestors tend to regress down towards a normal average. This inherent behavior of all biological processes to regress toward the mean is often described as well by the concept of central tendency. In time, Galton's original conceptualization of regression as describing biological phenomena was expanded into related sciences, including psychobiology and educational psychology. Pearson, one of his students, and Yule, one of Pearson's students extended the use of regression more widely to a more general statistical context. It is befitting that we propose here to return to Galton's original purpose of regression to embrace the biological phenomenon of viral immunity.

Simply stated, we propose that the state of immune homeostasis is challenged upon viral infection, at which time a state of heterostasis ensues. A myriad of immune events and non-immune regulatory events, some of which were highlighted above, come into play to induce the organism to gain a higher level of immune homeostasis characterized by immunity specifically toward the challenging infectious virus.

A cellular immune surveillance multiple regression predictive model can be constructed biostatistically, in which the outcome,  $Y$ , is simply the newly attained state of homeostasis characterized by immunity toward the infecting virus. The predictors of the model include all the heterostasis events - that is, the immune and the non-immune events that converged to produce immunity, to the same extent as the initial viral challenge, and other factors related to the state of heterostasis of the organism, including unhealthy eating habits, sleep deprivation, stress, and the like, act in concert to delay, hamper and counter  $Y$ .

Taken together, the application of this biostatistical technique to an organism's allostatic processes makes the problem a relatively easy-to-use multiple regression ( $Y = \Sigma\Pi + \Sigma N$ ). In the equation, the outcome ( $Y$ ) is the summation of positive variables that push allostasis forward ( $\Pi$ ) and negative variables ( $N$ ) that interfere or dampen allostasis. Thus, we can predict quantifiably the extent to which the outcome variable, representing the homeostatic state of health regained after a viral infection, is effected when either of the variables are changed, while holding the other constant. The prediction will be interpreted as the fine, coordinated, and time-regulated processes that occur throughout all the relevant interacting variables during immune surveillance.

## 4 Artificial Intelligence, “Immune-Tweening”: The Future of Immunity

### 4.1 Artificial Intelligence

Intelligence comprises many functions and has, therefore, been defined based on diverse criteria by philosophers, sociologists, neuroscientists and educators across the centuries. Broadly speaking, it may be fair to subsume intelligent behavior as purposeful and grounded in one or several of the following attributes: logic, understanding, self-awareness, learning, emotional knowledge, reasoning, planning, creativity, and problem-solving. Intelligence, in brief, may be described as the ability to perceive or infer information, to retain and recall knowledge appropriately to favor socio-psycho-physiological adaption and survival within a challenging environment or socio-cognitive context.

We might infer then that an environmental or a cognitive imbalance provoked by a challenge to our socio-cultural as well as ethno-cognitive schemata induces, as it were, a state of socio-psycho hererostasis, which we know full well can and is associated with psychophysiological heterostasis [46]. In that analogy, intelligence is the conduit, the mechanism, the tool, the allostatic process by which the intelligent organism regains socio-psycho-physiological balance, adaption and survival in the form of a return to homeostasis in response to an environmental or socio-cognitive challenge.

In our current age of computer-aided information, a new science has emerged that aims at utilizing computerized machines to assist intelligent organisms in their allostatic process of socio-psycho-physiological adaptation to environmental or socio-cognitive challenges. This new and improved development and utilization of machines for computer-aided information storage and retrieval, as well as computer-aided deductive, inductive and abductive reasoning and decision-making is broadly referred to as ‘machine-intelligence’, or more commonly ‘artificial intelligence’ (AI). Colloquially, AI describes the events by which a computerized machine mimics cognitive and meta-cognitive functions that we associate with the function and the power of the human mind; viz., learning, retaining in memory, and recalling appropriately from memory for the ultimate purposeful behavior of problem-solving [65, 66].

The power of AI has increased exponentially since its inception and its formal establishment as an academic discipline in the mid 1950’s. Its scope and applications have concomitantly multiplied, from automatic airplane and ship navigating pilots, to self-driving automobiles, global positioning systems and autonomous self-adjusting routing, military simulations, interactive voice-controlled home automation systems, and a plethora of others, which together make AI applications a technology that is routine in everyday life.

In the domain of the biological and life sciences, AI now encompasses multiple domains from robotics, such as for understanding and replicating human speech, limb reconstruction and the ability to move and manipulate objects and a variety of

other structural anatomico-physiological repair, to neural networking for aiding, augmenting, or repairing. This also extends to the replacing of reasoning, knowledge representation, planning, learning, natural language processing, and even perception. Machines – that is, AI – excel in statistical methods and probabilities, computations and mathematical optimizations, and traditional symbolic representations, which are embedded in neural networks and neuroscience.

An ethical dilemma follows, of course, which philosophers of science, ethicists, sociologists and psychologists actively debate. For example, if AI can replicate and surpass the human mind, then what is, really, the nature of the mind, reason, religious faith, myth, fiction and philosophy, which humans have produced for the past thousands of years? Will AI, intelligent machines soon replace humans in the workforce, with its consequential tragic societal and economic disruptions? Might AI become a danger to its own very inventors? Might it become too intelligent, as it were, and surpass human intelligence? Might AI threaten the very existence of humanity as it continues to progress unabated?

It is beyond the scope of this chapter to discuss these issues. Nonetheless, these problems are real, timely and critical. They are serious questions that pertain to every domain of society touched by AI, including AI applications in the health sciences.

Bioinformatics is that branch of AI that has direct implications in healthcare. From its role in experimental molecular biology [67–69] to analysis of gene and protein expression and regulation, to imaging studies, interpretation of the biological pathways and networks relevant to systems biology, text mining for comparative effectiveness research, to, in the context of this chapter on immunity, providing a catalogue, storage and retrieving system of B cell epitopes and immunity in general [70, 71], bioinformatics and AI in general is timely and critical domain of modern biology, the life and the health sciences.

## 4.2 “*Immune-Tweening*”: The Future of Immunity

As we consider the future of immunity in the age of AI, and its sub-discipline of bioinformatics, our understanding of the immunological sciences, and its applications and implications to the health sciences in general, and to modern healthcare in particular are a matter of urgency. They must be grounded in the development and implementation of the very computer advancements that AI and bioinformatics proffer, which enable efficient access to, use and management of, various types of information. New and improved algorithms and biostatistical analyses will have to be developed to go beyond locating a gene within a sequence, predicting a cluster of differentiation (CD) and its function during T cell activation, proliferation, migration and maturation, and clustering protein sequences into families of IL’s, IFN’s and other soluble immune factors.

The future of immunity lies in novel breakthroughs in bioinformatics that will enable deeper understanding of the fundamental biology of immune and non-immune

regulation of immune surveillance. This requires an integrative systems biology perspective that integrates and intertwines the response of cells and soluble factors of the innate and the antigen-dependent immune system, with regulatory processes directed and controlled by psychoneuroendocrine pathways and bone metabolism, in what would truly be well-coordinated and regulated psycho-neuroendocrine-osteo-immunity, as we originally proposed [72] and discussed specifically in the context of patients with HIV/AIDS [45].

Future advances in bioinformatics as it pertains to immunity will consider the fact that, the immune response to a viral pathogen, for instance, engenders an allostatic process in the organism, as we noted above, whose elements - both immune and non-immune, both cellular and soluble – are increasingly better understood. When immunity to the invading pathogen is attained, and new state of homeostasis is regained, which, in the context of bioinformatics, we could call Y.

To be clear, our knowledge and understanding of the entire spectrum of the psycho-neuroendocrine-osteo-immune events and processes that determine and ensure the success outcome of this allostatic process to Y are still incomplete. Much more immunological, psychoneuroimmunological and osteoimmunological basic, pre-clinical and clinical research is needed. However, AI also informs the process of ‘tweening’: the process by which an algorithm can be instructed to generate the intermediate frames between two images or two elements of knowledge – that is, from a biostatistical viewpoint, generate an estimate of the missing data.

We propose that the future of immunity lies in our cogent utilization of AI, and specifically bioinformatics to generate psycho-neuroendocrine-osteo-immune algorithms, completed by immune-tweening, to predict the successful outcome of the allostatic process of immune response to a viral pathogen. This can be envisaged to apply to a variety of viral antigens, such as HIV, or related retroviruses including Zika and Ebola, or new and aggressive variants of orthomyxoviruses, such as the various genera of influenza virus (isavirus, thogotovirus and quaranjavirus, and influenza virus genus A, B, C, and D, with genus A being the most dangerous to humans for causing influenza). There are several subtypes of influenza A viruses, which all can infect humans with broad spectrum of severity. They are classified, based on the viral surface proteins hemagglutinin (HA or H) and neuraminidase (NA or N). Sixteen H subtypes and nine N subtypes of influenza A virus have been identified to this date, but active virology research uncovers more H and N subtypes continuously. Type A influenza subtypes, determined by their H & N signatures are the most virulent human pathogens and cause most severe, often fulminant and lethal disease. Suffice to remember, the Spanish flu and the Swine flu (H1N1), the Asian flu (H2N2), the pandemic flu threat (H5N1) and the like in the last century alone.

Well-informed bioinformatics, advanced systems immunology databases (e.g., Genbank, UniProt, InterPro, Pfam, KEGG, BioCyc, GenoCAD, Bcipep), comparative orthology analysis (i.e., the study of the physiological and functional correspondence among pertinent genes, and interactomic structures), statistical techniques such as multiple regression as outlined above, and carefully derived psycho-neuroendocrine-osteo-immune algorithms (e.g., k-means and hierarchical

clustering; network analysis; high-throughput and high-fidelity quantification of fractal dimension analysis and sub-cellular localization, and related bioinformatics), and appropriate immune-tweening will, in the near future, permit us to describe and characterize the success of the allostatic process of immune response to these viral pathogens, and any immunogens, to a state of immunophysiological recovery, Y. As we proposed above, Y might be predicted roughly as the sum of factors and events that are positive ( $\Pi$ ) and push allostasis forward (i.e., the orderly process of immune activation and maturation), and those events that are negative ( $N$ ) factors and events and interfere with attaining Y. Simplistically, Y is the sum of all interacting the  $\Pi$ 's and  $N$ 's during the psycho-neuroendocrine-osteo-immune immune process ( $Y = \Sigma\Pi + \Sigma N$ ).

The question that bioinformatics will address in the next decades then becomes, can we define and characterize the constituents of  $\Sigma\Pi$  and of  $\Sigma N$ , and effectively use immune-tweening for those constituents that still escape our knowledge? Can we devise AI capsules with a bioinformatics-informed timed release of the appropriate IL's IFN's and other soluble immune and non-immune regulatory factors? Can we engineer immune and non-immune cells that can be injected in the patients at exact moments pre-determined by a patient-tailored psycho-neuroendocrine-osteo-immune immune algorithm? Can bioinformatics inform the development of artificial  $\Pi$ 's and  $N$ 's, that may push the organism's response more securely through the allostatic process that will lead to Y, the homeostatic state of health regained following an infection with HIV, any other retrovirus, any of the influenza subtypes, even those we have not yet characterized, or any other infectious agent?

These are questions for the future – the near future. This is, we propose, the future of immunity and immunophysiological surveillance.

## References

1. Chiappelli F. Immunophysiological role and clinical implications of non-immunoglobulin soluble products of immune effector cells. *Adv Neuroimmunol.* 1991;1:234–40.
2. Barkhodarian A, Thames AD, Du AM, Jan AL, Nahcivan M, Nguyen MT, Sama N, Chiappelli F. Viral immune surveillance: toward a TH17/TH9 gate to the central nervous system. *Bioinformation.* 2015;11:47–54.
3. Murphy K, Weaver C. *Janeway Immunobiology*. 9th ed. New York, NY: Garland Science/Taylor & Francis Group; 2017.
4. Chiappelli F, Manfrini E, Gwirtsman H, Garcia C, Pham L, Lee P, Frost P. Steroid receptor-mediated modulation of CD4+CD62L+ cell homing. Implications for drug abusers. *Ann NY Acad Sci.* 1994;746:421–5.
5. Chiappelli F, Kung MA. Immune surveillance of the oral cavity and lymphocyte migration: relevance for alcohol abusers. *Lymphology.* 1995;28(4):196–207.
6. Angeli V, Randolph GJ. Inflammation, lymphatic function, and dendritic cell migration. *Lymphat Res Biol.* 2006;4(4):217–28.
7. Liao S, von der Weid PY. Lymphatic system: an active pathway for immune protection. *Semin Cell Dev Biol.* 2015;38:83–9.
8. Randolph GJ, Ivanov S, Zinselmeyer BH, Scallan JP. The lymphatic system: integral roles in immunity. *Annu Rev Immunol.* 2017;35:31–52.

9. Ryter A. Relationship between ultrastructure and specific functions of macrophages. *Comp Immunol Microbiol Infect Dis.* 1985;8(2):119–33.
10. Langermans JA, Hazenbos WL, van Furth R. Antimicrobial functions of mononuclear phagocytes. *J Immunol Methods.* 1994;174(1–2):185–94.
11. Withers DR. Innate lymphoid cell regulation of adaptive immunity. *Immunology.* 2016;149:123–30.
12. Burnet FM. Cellular immunology: self and notself. Cambridge: Cambridge University Press; 1969.
13. Bretscher P, Cohn M. A theory of self-nonself discrimination. *Science.* 1970;169(3950):1042–9.
14. Chiappelli F, Gormley GJ, Gwirstman HE, Lowy MT, Nguyen LD, Nguyen L, Esmail I, Strober M, Weiner H. Effects of intravenous and oral dexamethasone on selected lymphocyte subpopulations in normal subjects. *Psychoneuroendocrinology.* 1992;17(2–3):145–52.
15. Chiappelli F, Kung M, Lee P, Pham L, Manfrini E, Villanueva P. Alcohol modulation of human normal T-cell activation, maturation, and migration. *Alcohol Clin Exp Res.* 1995;19(3):539–44.
16. Abbas AK, Murphy KM, Sher A. Functional diversity of helper T lymphocytes. *Nature.* 1996;383(6603):787–93.
17. Grewal IS, Flavell RA. CD40 and CD154 in cell-mediated immunity. *Annu Rev Immunol.* 1998;16(1):111–35.
18. Langman RE, Cohn M. A minimal model for the self-nonself discrimination: a return to the basics. *Semin Immunol.* 2000;12(3):189–95.
19. Guermonprez P, Valladeau J, Zitvogel L, Théry C, Amigorena S. Antigen presentation and T cell stimulation by dendritic cells. *Annu Rev Immunol.* 2002;20(1):621–67.
20. Holtmeier W, Kabelitz D. gammadelta T cells link innate and adaptive immune responses. *Chem Immunol Allergy.* 2005;86:151–83.
21. Girardi M. Immunosurveillance and immunoregulation by gammadelta T cells. *J Invest Dermatol.* 2006;126(1):25–31.
22. Andersen MH, Schrama D, Thor Straten P, Becker JC. Cytotoxic T cells. *J Invest Dermatol.* 2006;126(1):32–41.
23. Copeland KF, Heeney JL. T helper cell activation and human retroviral pathogenesis. *Microbiol Rev.* 1996;60(4):722–42.
24. Kawai T, Akira S. Innate immune recognition of viral infection. *Nat Immunol.* 2006;7(2):131–7.
25. Restifo NP, Gattinoni L. Lineage relationship of effector and memory T cells. *Curr Opin Immunol.* 2013;25(5):556–63.
26. Sall FB, Germini D, Kovina AP, Ribrag V, Wiels J, Toure AO, Iarovaia OV, Lipinski M, Vassetzky Y. Effect of environmental factors on nuclear organization and transformation of human B lymphocytes. *Biochemistry (Mosc).* 2018;83(4):402–10.
27. Kurosaki T, Kometani K, Ise W. Memory B cells. *Nat Rev Immunol.* 2015;15(3):149–59.
28. Seliger B, Ritz U, Ferrone S. Molecular mechanisms of HLA class I antigen abnormalities following viral infection and transformation. *Int J Cancer.* 2006;118(1):129–38.
29. Walker B, McMichael A. The T-cell response to HIV. *Cold Spring Harb Perspect Med.* 2012;2(11):a007054.
30. Braciale TJ, Hahn YS. Immunity to viruses. *Immunol Rev.* 2013;255(1):10.1111–12109.
31. Wortzman ME, Clouthier DL, McPherson AJ, Lin GH, Watts TH. The contextual role of TNFR family members in CD8(+) T-cell control of viral infections. *Immunol Rev.* 2013;255(1):125–48.
32. Harris JF, Micheva-Viteva S, Li N, Hong-Geller E. Small RNA-mediated regulation of host-pathogen interactions. *Virulence.* 2013;4(8):785–95.
33. Tacchetti C, Favre A, Moresco L, Meszaros P, Luzzi P, Truini M, Rizzo F, Grossi CE, Ciccone E. HIV is trapped and masked in the cytoplasm of lymph node follicular dendritic cells. *Am J Pathol.* 1997;150(2):533–42.
34. Price RW, Brew B, Saitis J, Rosenblum M, Scheck AC, Cleary P. The brain in AIDS: central nervous system HIV-1 infection and AIDS dementia complex. *Science.* 1988;239(4840):586–92.
35. Chiappelli F, Kung MA, Villanueva P. Neuropsychiatry of drugs of abuse and AIDS. *J Neuroimmunol.* 1996;69:48–9.

36. Minagar A, Commins D, Alexander JS, Hoque R, Chiappelli F, Singer EJ, Nikbin B, Shapshak P. NeuroAIDS: characteristics and diagnosis of the neurological complications of AIDS. *Mol Diagn Ther*. 2008;12(1):25–43.
37. Shapshak P, Chiappelli F, Commins D, Singer E, Levine AJ, Somboonwit C, Minagar A, Pellionisz AJ. Molecular epigenetics, chromatin, and NeuroAIDS/HIV: translational implications. *Bioinformation*. 2008;3(1):53–7.
38. Chiappelli F, Shapshak P, Commins D, Singer E, Minagar A, Oluwadara O, Prolo P, Pellionisz AJ. Molecular epigenetics, chromatin, and NeuroAIDS/HIV: immunopathological implications. *Bioinformation*. 2008;3(1):47–52.
39. Chiappelli F. Psychoneuroimmunology of immune reconstitution inflammatory syndrome (IRIS): the new frontier in translational biomedicine. *Transl Biomed*. 2015;6:11–4.
40. Chiappelli F, Bakhordarian A, Thames AD, Du AM, Jan AL, Nahcivan M, Nguyen MT, Sama N, Manfrini E, Piva F, Rocha RM, Maida CA. Ebola: translational science considerations. *J Transl Med*. 2015;13:11.
41. Chiappelli F, Franceschi C, Ottaviani E, Solomon GF, Taylor AN. Neuroendocrine modulation of the immune system. In: Greger R, Koepchen HP, Mommaerts W, Winhorst U, editors. *Human physiology: from cellular mechanisms to integration*. New York: Springer; 1996. p. 1707–29, Section L Chapter 86.
42. Chiappelli F, Abanomy A, Hodgson D, Mazey KA, Messadi DV, Mito RS, Nishimura I, Spigleman I. Clinical, experimental and translational psychoneuroimmunology research models in oral biology and medicine. In: Ader R, et al., editors. *Psychoneuroimmunology*, vol. III: Academic Press; 2001. p. 645–70, Chapter 64.
43. Prolo P, Chiappelli F, Fiorucci A, Dovio A, Sartori ML, Angeli A. Psychoneuroimmunology: new avenues of research for the 21st century. *Ann NY Acad Sci*. 2002;966:400–8.
44. Chiappelli F, Prolo P, Fiala M, Cajulis O, Iribarren J, Panerai A, Neagos N, Younai F, Bernard G. Allostasis in HIV infection and AIDS. In: Minagar PA, Shapshak P, editors. *Neuro-AIDS*: Nova Science Publisher, Inc.; 2006. p. 121–65, Chapter VI.
45. Barkhordarian A, Ajaj R, Ramchandani MH, Demerjian G, Cayabyab R, Danaie S, Ghodousi N, Iyer N, Mahanian N, Phi L, Giroux A, Manfrini E, Neagos N, Siddiqui M, Cajulis OS, Brant X, Shapshak P, Chiappelli F. Osteoimmunopathology in HIV/AIDS: a translational evidence-based perspective. *Pathol Res Int*. 2011, Article ID 359242 epub 21 May 2011.
46. Chiappelli F, Kutschman MM. Current and Future Directions in Psychobiology. In: Chiappelli F, editor. *Advances in Psychobiology*. Hauppauge, NY: NovaScience Publisher, Inc.; 2018, Chapter 1.
47. Chiappelli F, Cajulis OS. Psychobiological views on “stress-related oral ulcers”. *Quintessence Int*. 2004;35:223–7.
48. Turner-Cobb JM. Psychological and stress hormone correlates in early life: a key to HPA-axis dysregulation and normalization. *Stress*. 2005;8(1):47–57.
49. Chiappelli F, Trignani S. Neuroendocrine-immune interactions: implications for clinical research. *Adv Biosci*. 1993;90:185–98.
50. Chiappelli F, Manfrini E, Franceschi C, Cossarizza A, Black K. Steroid regulation of cytokines: relevance for TH1→TH2 shift? *Ann NY Acad Sci*. 1994;746:204–16.
51. Chiappelli F, Liu NQ. Non-mammalian models of neuroendocrine-immune modulation: relevance for research in oral biology and medicine. *Int J Oral Biol*. 1999;24:47–61.
52. Elenkov II, Wilder RL, Chrousos GP, Vizi ES. The sympathetic nerve—an integrative interface between two supersystems: the brain and the immune system. *Pharmacol Rev*. 2000;52(4):595–638.
53. Romeo HE, Tio DL, Rahman SU, Chiappelli F, Taylor AN. The glossopharyngeal nerve as a novel pathway in immune-to-brain communication: relevance to neuroimmune surveillance of the oral cavity. *J Neuroimmunol*. 2001;115:91–100.
54. Yun AJ, Lee PY, Bazar KA. Modulation of host immunity by HIV may be partly achieved through usurping host autonomic functions. *Med Hypotheses*. 2004;63:362–6.
55. Chiappelli F, Alwan J, Prolo P, Christensen R, Fiala M, Cajulis OS, Bernard G. Neuroimmunity in stress-related oral ulcerations: a fractal analysis. *Front Biosci*. 2005;10:3034–41.

56. Angeli A, Dovio A, Sartori ML, Masera RG, Ceoloni B, Prolo P, Racca S, Chiappelli F. Interactions between glucocorticoids and cytokines in the bone microenvironment. *Ann NY Acad Sci.* 2002;966:97–107.
57. Kenney MJ, Ganta CK. Autonomic nervous system and immune system interactions. *Compr Physiol.* 2014;4:1177–200.
58. Chiappelli F, Bakhtdarian A, Bach Q, Demerjian GG. Translational psychoneuroimmunology in oral biology & medicine. *For Immunopathol Dis Therap.* 2016;6:119–32.
59. Khakshooy A, Chiappelli F. Hypothalamus-pituitary-adrenal - cell-mediated immunity regulation in the immune restoration inflammatory syndrome. *Bioinformation.* 2016;12:28–30.
60. Diazzi C, Brigante G, Ferrannini G, Ansaldi A, Zirilli L, De Santis MC, Zona S, Guaraldi G, Rochira V. Pituitary growth hormone (GH) secretion is partially rescued in HIV-infected patients with GH deficiency (GHD) compared to hypopituitary patients. *Endocrine.* 2017;55:885–98.
61. Chiappelli F. Fundamentals of evidence-based health care and translational science. Heidelberg: Springer; 2014.
62. Khakshooy A, Chiappelli F, editors. Practical biostatistics in translational healthcare. New York, NY: Springer; 2018.
63. Draper NR, Smith H. Applied regression analysis. 3rd ed. Hoboken, NJ: John Wiley; 1998.
64. Bingham NH, Fry JM. Regression: linear models in statistics. Heidelberg: Springer; 2010.
65. Donald BR. Algorithms in structural molecular biology. Cambridge, MA: The MIT Press; 2011.
66. Russell SJ, Norvig P. Artificial intelligence: a modern approach. Englewood Cliffs, NJ: Prentice Hall; 2003.
67. Hogeweg P, Searls DB. The roots of bioinformatics in theoretical biology. *PLoS Comput Biol.* 2011;7(3):e1002021.
68. Sim AYL, Minary P, Levitt M. Modeling nucleic acids. *Curr Opin Struct Biol.* 2012;22(3):273–8.
69. Simonyan V, Goecks J, Mazumder R. Biocompute objects—a step towards evaluation and validation of biomedical scientific computations. *PDA J Pharm Sci Technol.* 2017;7(2):136–46.
70. Saha S, Bhasin M, Raghava GP. Bcipep: a database of B-cell epitopes. *BMC Genomics.* 2005;6:79.
71. Lund O, Nielsen M, Lundegaard C, Kesmir C, Brunak S. Immunological bioinformatics. Cambridge, MA: The MIT Press; 2005.
72. Chiappelli F. Osteoimmunopathology: evidence-based perspectives from molecular biology to systems biology. New York: Springer; 2011.

# Emerging Technologies for Antiviral Drug Discovery



**Badireddi Subathra Lakshmi, Mohan Latha Abillasha,  
and Pandjassarame Kanguane**

**Abstract** Drug discovery for viral diseases is a continuing and expanding undertaking for improved public health. Development of effective and specific antiviral drugs is still a huge challenge. Increased viral outbreaks during the last several decades necessitate novel methods in combating them. The application of classical drug discovery approaches, supported by emerging technologies such as high performance computing (HPC) aided molecular geometric optimization based screening along with deep learning (hierarchical learning) using machine learning techniques (model build with known sample data) like artificial neural networks (ANN) is warranted towards personalized and community medicine in combating viral outbreaks to advance public health. Molecular information on the mechanisms of viral pathogenesis gathered from the scientific literature and stored in specialized relational databases play a critical role in gene discovery (linking host and viral genes to disease), target validation (often protein targets for disease control), and molecular screening of small molecule inhibitors to protein targets. An integrated approach relating data from clinical virology, genomics, proteomics, gene expression analysis, sequence profiling, structure-function analysis, molecular binding colorimetric (dye based) assays and computer-aided small molecule screening is critical in viral combat as well as public health. This chapter outlines the available resources (data and meta-data (derived data) in databases) and technologies (tools and algorithms) to refine current trends in order to accelerate the drug discovery process towards combating viruses.

**Keywords** Drug discovery · New chemical entity · Docking tools · Databases · Geometric optimization · High performance computing (HPC) · Artificial neural network (ANN) · Deep learning · Viral diseases · Viral outbreaks · Viral combat

---

B. S. Lakshmi (✉) · M. L. Abillasha  
Department of Biotechnology, Anna University, Chennai, India  
e-mail: [lakshmib@annauniv.edu](mailto:lakshmib@annauniv.edu)

P. Kanguane  
Biomedical Informatics (P) Ltd, Puducherry, India

### Core Message

This chapter outlines the application of the drug discovery pipeline enabled with emerging technologies to accelerate the identification and development of new chemical compounds as drug-like molecules. These include small molecule screening enabled with deep learning (hierarchical learning) using a machine learning technique like Artificial Neural Network (ANN) supported by high performance computing (HPC) tools against defined viral protein targets to fight viral diseases. This is highly appropriate in the context of personalized and community medicine in combating viral outbreaks.

## 1 Introduction

During the last few decades, the application of biotechnology has dramatically accelerated biomedicine for improved public health. Major advances in technologies such as high throughput molecular screening, computer-aided geometric optimization based molecular screening, genomics and proteomics techniques have revolutionized the drug discovery processes in the treatment and control of emerging viral diseases. However, there is still a need for drugs for many of the viral diseases. Emerging technologies such as computer-aided geometric optimization based molecular screening enabled with deep learning (hierarchical learning) using ANN is useful in accelerating the viral drug discovery process.

## 2 Viral Databases and Dataset Analysis

Virus related data available in the form of databases for meta-data (derived data) analysis provide valuable information in the understanding of disease mechanism caused by viruses towards target definition in drug discovery. Table 1 lists the recent viral databases and datasets for disease analysis. ViPR [1], VIRALZONE [2] and KEGG databases [3] provide information on viruses causing different viral diseases. Classification of viruses using sequence data is provided by the VIRUS\_DB database [4]. Insight into disease mechanisms using viral data analysis of specific viruses such as Hepatitis C virus [5–7], Hepatitis B virus [8–10], Dengue virus [11, 12], Human Immuno-Deficiency virus [13], Hepatitis A virus [14], Influenza [15, 16] and respiratory syncytial virus [17] are currently available. Other specific databases for the understanding of viral pathogenesis and mechanism of infection are also available [18–22]. Information gathered from data made available in such databases is critical for understanding pathogenesis and help in target definition following gene discovery.

**Table 1** Viral databases and datasets

Virus database/dataset	URL	References
ViPR	<a href="https://www.viprbc.org/brc/home.sp?decorator=vipr">https://www.viprbc.org/brc/home.sp?decorator=vipr</a>	[1]
VIRALZONE	<a href="https://viralzone.expasy.org/">https://viralzone.expasy.org/</a>	[2]
KEGG disease DB	<a href="https://www.genome.jp/kegg/disease/">https://www.genome.jp/kegg/disease/</a>	[3]
VIRUS_DB	<a href="http://yaulab.math.tsinghua.edu.cn/VirusDB/">http://yaulab.math.tsinghua.edu.cn/VirusDB/</a>	[4]
Taiwanese national claims database (HCV)	<a href="https://nhird.nhri.org.tw/en/">https://nhird.nhri.org.tw/en/</a>	[5]
Research UK Clinical Database and Biobank (HCV)	<a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5837619/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5837619/</a>	[6]
HCV among syringe exchange clients:	<a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5404946/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5404946/</a>	[7]
HBV National Database of Japan.	<a href="https://www.ncbi.nlm.nih.gov/pubmed/29770539/">https://www.ncbi.nlm.nih.gov/pubmed/29770539/</a>	[8]
VAERS	<a href="https://vaers.hhs.gov/">https://vaers.hhs.gov/</a>	[9]
HBV	<a href="https://www.ncbi.nlm.nih.gov/pubmed/28008620">https://www.ncbi.nlm.nih.gov/pubmed/28008620</a>	[10]
DenvInt	<a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5648114/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5648114/</a>	[11]
Denv Ab DB	<a href="http://denvabdb.bhsai.org/denvabdb/">http://denvabdb.bhsai.org/denvabdb/</a>	[12]
French Hospital Database on HIV (ANRS-C4).	<a href="https://www.ncbi.nlm.nih.gov/pubmed/29635465">https://www.ncbi.nlm.nih.gov/pubmed/29635465</a>	[13]
International HAVNet hepatitis A virus database	<a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6144472/">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6144472/</a>	[14]
VirusMap	<a href="https://www.ncbi.nlm.nih.gov/pubmed/28529079">https://www.ncbi.nlm.nih.gov/pubmed/28529079</a>	[15]
Influenza Research Database:	<a href="https://www.fludb.org/brc/home.sp?decorator=influenza">https://www.fludb.org/brc/home.sp?decorator=influenza</a>	[16]
Healthcare resource in USA	<a href="https://www.ncbi.nlm.nih.gov/pubmed/29678177">https://www.ncbi.nlm.nih.gov/pubmed/29678177</a>	[17]
Viruses STRING:	<a href="https://jensenlab.org/training/stringapp/">https://jensenlab.org/training/stringapp/</a>	[18]
A Reference Viral Database (RVDB)	<a href="https://hive.biochemistry.gwu.edu/rvdb">https://hive.biochemistry.gwu.edu/rvdb</a>	[19]
TBEVhostDB	<a href="http://icg.nsc.ru/TBEVHostDB/">http://icg.nsc.ru/TBEVHostDB/</a>	[20]
ICTV	<a href="https://talk.ictvonline.org/">https://talk.ictvonline.org/</a>	[21]
MRPrimerV	<a href="http://mrprimerv.com/">http://mrprimerv.com/</a>	[22]

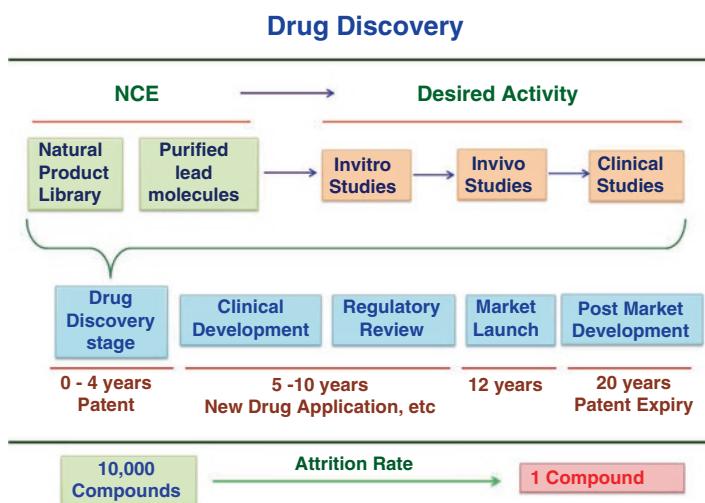
### 3 Viral Diseases

Anti-viral drugs are not available for all viruses. Therefore, it is important to advance drug discovery and development for viral diseases. In addition, improvements in current anti-viral drugs are needed to improve effectiveness and specificity of anti-viral drugs. Antiviral drugs with improved efficacy are needed for viruses such as Borna disease virus [23], Epstein Barr virus [24–27], Ebola virus [28–31], Heartland virus [32], Hepatitis B virus [33], Hepatitis C virus [34–39], Human Immunodeficiency virus [40–42], Herpes virus [43–45], Human Papilloma Virus

[46], Influenza Virus [47], respiratory syncytial virus [48, 49] and Zika virus [50, 51]. Available data provide valuable information for the definition of protein targets for several viral diseases. Nonetheless, such information is currently either limited or unknown for many viruses. Thus, database supported comprehensive data with enriched information on the molecular mechanisms of viral pathogenesis are needed for many viruses causing human diseases.

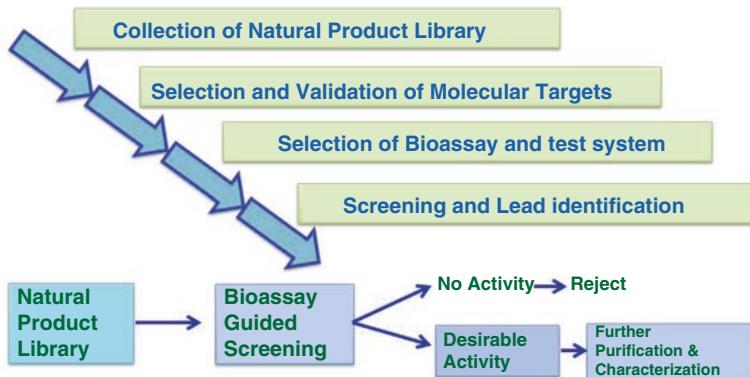
## 4 Drug Discovery Process

The classical drug discovery process of isolating novel compounds from natural sources for a given protein target is illustrated in Fig. 1. This process involves screening (high throughput or computer-aided), gene discovery (linking genes to disease) and target validation (identifying proteins for disease control) and identification of a new chemical entity (NCE) as inhibitors to defined targets. Screening new compounds from natural sources is an important component in drug discovery. The steps involved in the extraction of natural compounds for processing in the drug discovery pipeline are shown in Fig. 2. The laboratory scale identification and analysis of candidate compounds, followed by ranking with importance, is demonstrated in Fig. 3. We constantly use these drug discovery techniques for the identification of novel prime candidate compounds for the potential treatment of type II diabetes [52–67], breast cancer [68–77] and inflammation [78–80]. Thus, although the drug discovery paradigm from natural products such as plants has been optimized for a few diseases, the drug discovery process for viruses requires greater development and refinement.



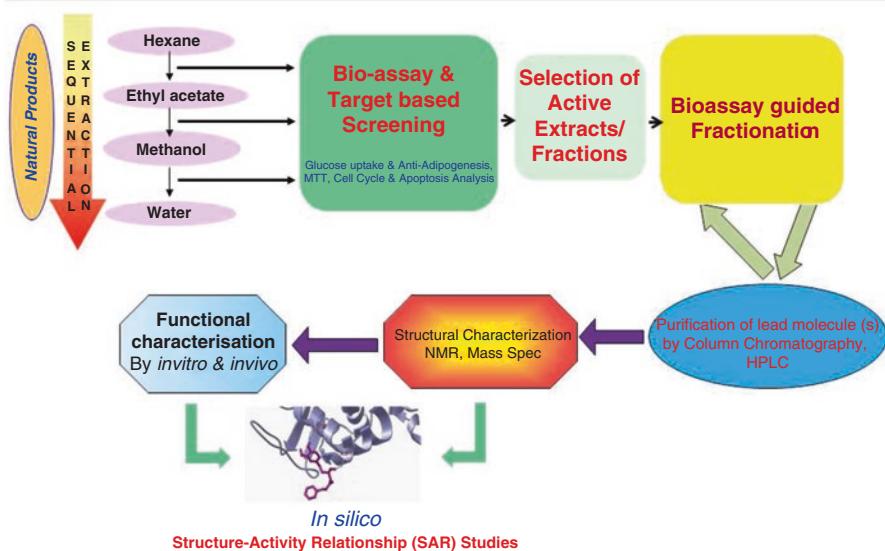
**Fig. 1** The classical drug discovery process with different stages is illustrated. The drug discovery stage consists of NCE development and target assay for activity analysis. *NCE* new chemical entity

### Stages in Natural Products to Lead Molecules



**Fig. 2** Processing of natural products is outlined. The stages involved in the conversion of natural products to lead molecules are shown. This includes the development of natural product library, bioassay optimization and screening of new compounds

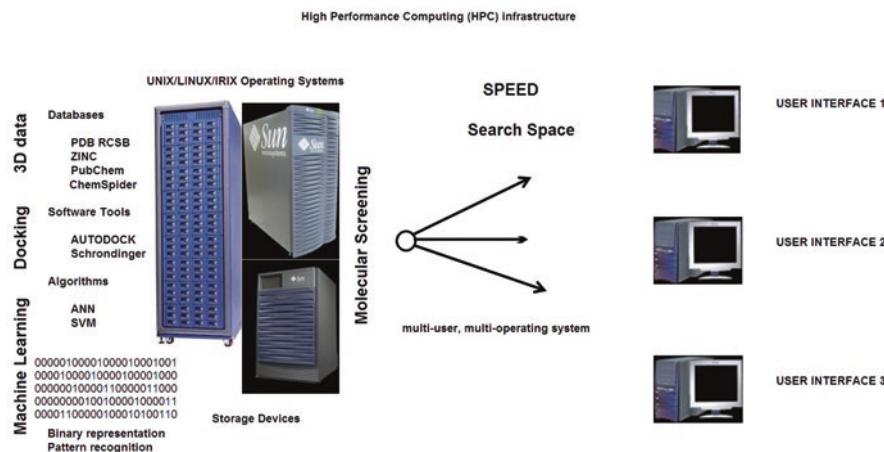
### STRATEGY



**Fig. 3** Screening of novel compounds from plant materials for a given target is demonstrated. The steps involved in screening are illustrated

## 5 Emerging Technologies for Drug Discovery

The development and application of high performance computing (HPC) in the drug discovery process has been a crucial step in recent years to accelerate discovery (Fig. 4). This involves the use of high end (high speed) sophisticated computers installed with geometric optimization based molecular screening tools (e.g. (Schrodinger™ [81] and AUTODOCK [82]) against several millions of small molecule compounds (with known three dimensional structure co-ordinates) stored in ligand databases such as ZINC™ [83], Pubchem [84], ChemsSpider [85]. Currently, there are more than 35 million compounds in the ZINC™ database, around 60 million compounds in Pubchem, and around 50 million compounds in ChemsSpider. The computer aided screening process helps in accelerating the drug discovery process by significantly reducing the search time in a considerable manner. The application of emerging techniques such as artificial neural network (ANN) for screening molecular entities against defined protein targets remains highly helpful especially during disease outbreaks where combat care is needed. The application of deep learning (hierarchical learning) using ANN (machine learning techniques using non-linear optimization algorithms) for drug discovery has gained momentum in recent years [86–103]. Machine learning methods have been applied in pharmaceutical research with the use of Bayesian methods coupled with deep learning algorithm for successful drug discovery [86–89]. These algorithms help in bioactivity prediction, *de novo* molecular design, synthesis and biological image analysis [90]. Deep learning (hierarchical learning that uses trained dataset for recognition) is also



**Fig. 4** High Performance Computing (HPC) infrastructure equipped with databases and software tools is shown. PDB Protein Data Bank, RCSB Research Collaboratory for Structural Bioinformatics, ANN Artificial neural network, SVM Support vector machine, 3D data three dimensional structural data; UNIX/LINUX/IRIS are operating systems. ZINC, PubChem and ChemSpider are small molecule LIGAND databases with known three dimensional structure data

widely used in drug delivery, biologically compatible material development, and regenerative medicine [91]. Thus, machine Learning (ML) techniques like ANN, support vector machine (SVM) are often used in novel drug discovery [92].

ANN is also used for understanding the quantitative structure-activity relationship (QSAR) framework, virtual screening of compounds, receptor modeling, formulation development, pharmaco-kinetics studies and for mathematical modeling [93–98]. These methods are also applied in the understanding of several neglected diseases like Chagas disease, sleeping sickness and leishmaniasis [99]. SVM helps to predict the drug-likeness and agro-chemical-likeness for large collection of compounds by out-performing with other known methods [100]. The characterization of absorption, distribution, metabolism, excretion (ADME), and toxicity of the potential drug candidates is optimized for drug discovery using these techniques [101, 102]. Further, personalized medicine is also possible by linking the genotypes of patients to drug sensitivity [103]. Thus, the importance of ANN and SVM in drug discovery will be critical in coming years with improved accuracy, specificity, and sensitivity providing high positive predictive value and low negative predictive value. Deep learning provides improvements in the analysis of large chemical data sets for fishing an NCE in the drug discovery process. We anticipate that the use of ANN in drug discovery for molecular screening and ADME evaluation will become common in the near future.

## 6 Viral Drug Development and Challenges

Drug discovery for viral diseases has progressed for several decades. This is especially significant in the context of virus outbreaks where the environment has been perturbed by natural as well as man-made harm. Several approaches to combat viral diseases are currently being evaluated and applied. They include, for example, the broad development of viral anti-chemokines [104] and anti-viral protease inhibitors [105]. The use of virtual screening for several viral targets [106] is gaining momentum in recent years to accelerate the discovery process. Definition of specific viral targets for the development of anti-viral drugs has been explored in many cases. These include for example, the development of antiviral drugs for HCV targets [107], viral encephalitis proteins [108], HIV-1 proteins [109] and Influenza M2 ion channel protein [110]. Nonetheless, our understanding of viral pathogenesis for the definition of appropriate drug targets towards small molecule screening is under continuous improvement. Moreover, the use of drug discovery pipelines, supported by HPC enabled with deep learning using ANN is expected to accelerate the development of anti-viral drugs in combating viral diseases and unanticipated outbreaks.

## 7 Conclusions

Molecular understanding of viruses and the diseases they cause is necessary to define protein drug targets for drug discovery. The classical drug discovery process is highly limited in combating viral outbreaks. Therefore, the importance of emerging technologies supported by HPC maintaining big data in molecular databases, which are capable of storing millions of three dimensional structural co-ordinates of small molecules along with deep learning for chemical features using ANN has become vital in recent years. Thus, we outlined the drug discovery process, from natural sources using emerging tools, including screening models, which are enabled by deep learning using ANN with HPC support.

**Acknowledgement** We wish to express our sincere appreciation to all members of Tissue Culture & Drug Discovery Lab, Centre for Food Technology, Department of Biotechnology, Anna University, Chennai and Biomedical Informatics (P) Ltd. for many discussions on the subject of this chapter. We thank Dr. Paul Shapshak, Dr. Meena Kishore Sakarkar and Dr. Peter Natesan Pushparaj for their critical comments, suggestions and useful edits on the content of this chapter, which helped to make it contextual.

## References

1. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, et al. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* 2011;40(D1):D593–8.
2. Hulo C, De Castro E, Masson P, Bougueret L, Bairoch A, Xenarios I, et al. ViralZone: a knowledge resource to understand virus diversity. *Nucleic Acids Res.* 2010;39(suppl\_1):D576–82.
3. Kanehisa M, Furumichi M, Tanabe M, Sato Y, Morishima K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* 2016;45(D1):D353–61.
4. Dong R, Zheng H, Tian K, Yau SC, Mao W, Yu W, et al. Virus database and online inquiry system based on natural vectors. *Evol Bioinforma.* 2017;13:1176934317746667.
5. Tung CH, Lai NS, Li CY, Tsai SJ, Chen YC, Chen YC. Risk of rheumatoid arthritis in patients with hepatitis C virus infection receiving interferon-based therapy: a retrospective cohort study using the Taiwanese national claims database. *BMJ Open.* 2018;8(7):e021747.
6. McLauchlan J, Innes H, Dillon JF, Foster G, Holtham E, McDonald S, et al. HCV Research UK Steering Committee. Cohort profile: the Hepatitis C Virus (HCV) Research UK clinical database and biobank. *Int J Epidemiol.* 2017;46(5):1391–h.
7. Hochstatter KR, Hull SJ, Stockman LJ, Stephens LK, Olson-Streed HK, Ehlenbach WJ, et al. Using database linkages to monitor the continuum of care for hepatitis C virus among syringe exchange clients: experience from a pilot intervention. *Int J Drug Policy.* 2017;42:22–5.
8. Fujita M, Sugiyama M, Sato Y, Nagashima K, Takahashi S, Mizokami M, Hata A. Hepatitis B virus reactivation in patients with rheumatoid arthritis: analysis of the National Database of Japan. *J Viral Hepat.* 2018;25(11):1312–20.
9. Mouchet J, Bégaud B. Central demyelinating diseases after vaccination against hepatitis B virus: a disproportionality analysis within the VAERS database. *Drug Saf.* 2018;41(8):767–74.
10. Jin YJ LJW. Therapeutic priorities for solitary large hepatocellular carcinoma in a hepatitis B virus endemic area; an analysis of a nationwide cancer registry database. *J Surg Oncol.* 2017;115(4):407–16.

11. Dey L, Mukhopadhyay A. DenvInt: a database of protein–protein interactions between dengue virus and its hosts. *PLoS Negl Trop Dis.* 2017;11(10):e0005879.
12. Chaudhury S, Gromowski GD, Ripoll DR, Khavruskii IV, Desai V, Wallqvist A. Dengue virus antibody database: systematically linking serotype-specificity with epitope mapping in dengue virus. *PLoS Negl Trop Dis.* 2017;11(2):e0005395.
13. Melliez H, Mary-Krause M, Bocket L, Guiguet M, Abgrall S, De Truchis P, Katlama C, Martin-Blondel G, Henn A, Revest M, Robineau O. Risk of progressive multifocal leukoencephalopathy in the combination antiretroviral therapy era in the French hospital database on human immunodeficiency virus (ANRS-C4). *Clin Infect Dis.* 2018;67(2):275–82.
14. Kroneman A, de Sousa R, Verhoef L, Koopmans MP, Vennema H. Usability of the international HAVNet hepatitis A virus database for geographical annotation, backtracing and outbreak detection. *Eur Secur.* 2018;23(37) <https://doi.org/10.2807/1560-7917.ES.2018.23.37.1700802>.
15. Xie Y, Luo X, He Z, Zheng Y, Zuo Z, Zhao Q, et al. VirusMap: a visualization database for the influenza A virus. *J Genet Genomics.* 2017;44(5):281–4.
16. Zhang Y, Avermann BD, Anderson TK, Burke DF, Dauphin G, Gu Z, He S, Kumar S, Larsen CN, Lee AJ, Li X. Influenza Research Database: an integrated bioinformatics resource for influenza virus research. *Nucleic Acids Res.* 2017;45(D1):D466–74.
17. Amand C, Tong S, Kieffer A, Kyaw MH. Healthcare resource use and economic burden attributable to respiratory syncytial virus in the United States: a claims database analysis. *BMC Health Serv Res.* 2018;18(1):294.
18. Cook H, Doncheva N, Szklarczyk D, von Mering C, Jensen L. *Viruses.* STRING: a virus-host protein-protein interaction database. *Viruses.* 2018;10(10):519.
19. Goodacre N, Aljanahi A, Nandakumar S, Mikailov M, Khan AS. A Reference Viral Database (RVDB) to enhance bioinformatics analysis of high-throughput sequencing for novel virus detection. *mSphere.* 2018;3(2):e00069–18.
20. Ignatjeva EV, Igoshin AV, Yudin NS. A database of human genes and a gene network involved in response to tick-borne encephalitis virus infection. *BMC Evol Biol.* 2017;17(2):259.
21. Lefkowitz EJ, Dempsey DM, Hendrickson RC, Orton RJ, Siddell SG, Smith DB. Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.* 2018;46(D1):D708–17.
22. Kim H, Kang N, An K, Kim D, Koo J, Kim MS. MRPrimerV: a database of PCR primers for RNA virus detection. *Nucleic Acids Res.* 2016;45(D1):D475–81.
23. Azami M, Jalilian FA, Khorshidi A, Mohammadi Y, Tardeh Z. The association between Borna Disease Virus and schizophrenia: a systematic review and meta-analysis. *Asian J Psychiatr.* 2018;34:67–73.
24. Ru Y, Chen J, Wu D. Epstein-Barr virus post-transplant lymphoproliferative disease (PTLD) after hematopoietic stem cell transplantation. *Eur J Haematol.* 2018;101(3):283–90.
25. Bolland CM, Cohen JI. How I treat T-cell chronic active Epstein-Barr virus disease. *Blood.* 2018;131(26):2899–905.
26. Arjunaraja S, Angelus P, Su HC, Snow AL. Impaired control of Epstein–Barr virus infection in B-cell expansion with NF-κB and T-cell Anergy disease. *Front Immunol.* 2018;9:198.
27. Hoeger B, Serwas NK, Boztug K. Human NF-κB1 Haploinsufficiency and Epstein–Barr virus-induced disease—molecular mechanisms and consequences. *Front Immunol.* 2018;8:1978.
28. Trehan I, De Silva SC. Management of Ebola virus disease in children. *Infect Dis Clin N Am.* 2018;32(1):201–14.
29. Méren A, Bigaillon C, Delaune D. Ebola virus disease: biological and diagnostic evolution from 2014 to 2017. *Med Mal Infect.* 2018;48(2):83–94.
30. Fischer WA, Vetter P, Bausch DG, Burgess T, Davey RT, Fowler R, et al. Ebola virus disease: an update on post-exposure prophylaxis. *Lancet Infect Dis.* 2018;18(6):e183–92.
31. Richards GA, Baker T, Amin P, Council of the World Federation of Societies of Intensive and Critical Care Medicine. Ebola virus disease: report from the taskforce on tropical diseases by the World Federation of Societies of Intensive and Critical Care Medicine. *J Crit Care.* 2018;43:352–5.

32. Brault AC, Savage HM, Duggal NK, Eisen RJ, Staples JE. Heartland virus epidemiology, vector association, and disease potential. *Viruses*. 2018;10(9):E498.
33. Choi YM, Lee SY, Kim BJ. Naturally occurring hepatitis B virus reverse transcriptase mutations related to potential antiviral drug resistance and liver disease progression. *World J Gastroenterol*. 2018;24(16):1708.
34. Moorman AC, Rupp LB, Gordon SC, Zhong Y, Xing J, Lu M, et al. Long-term liver disease, treatment, and mortality outcomes among 17,000 persons diagnosed with chronic hepatitis C virus infection: current chronic hepatitis cohort study status and review of findings. *Infect Dis Clin N Am*. 2018;32(2):253–68.
35. Ridruejo E, Mendizabal M, Silva MO. Rationale for treating hepatitis C virus infection in patients with mild to moderate chronic kidney disease. *Hemodial Int*. 2018;22(Suppl1):S97–S103.
36. Al-Rabadi L, Box T, Singhania G, Al-Marji C, Agarwal A, Hall I, Gordon CE, Tran H. Rationale for treatment of hepatitis C virus infection in end-stage renal disease patients who are not kidney transplant candidates. *Hemodial Int*. 2018;22:S45–52.
37. Ortiz GA, Trivedi HD, Nader C. Pharmacokinetics and drug interactions of medications used to treat hepatitis C virus infection in the setting of chronic kidney disease and kidney transplantation. *Hemodial Int*. 2018;22:S22–35.
38. Matsuura K, Tanaka Y. Host genetic variations associated with disease progression in chronic hepatitis C virus infection. *Hepatol Res*. 2018;48(2):127–33.
39. Wijarnpreecha K, Chedsachai S, Jaruvongvanich V, Ungprasert P. Hepatitis C virus infection and risk of Parkinson's disease: a systematic review and meta-analysis. *Eur J Gastroenterol Hepatol*. 2018;30(1):9–13.
40. Jiménez-Sousa MA, Martínez I, Medrano LM, Fernández-Rodríguez A, Resino S. Vitamin D in Human immunodeficiency virus infection: influence on immunity and disease. *Front Immunol*. 2018;9:458.
41. Pinto DSM, da Silva MJLV. Cardiovascular disease in the setting of human immunodeficiency virus infection. *Curr Cardiol Rev*. 2018;14(1):25–41.
42. Tsabedze N, Vachiat A, Zachariah D, Manga P. A new face of cardiac emergencies: human immunodeficiency virus-related cardiac disease. *Cardiol Clin*. 2018;36(1):161–70.
43. Harris SA, Harris EA. Molecular mechanisms for herpes simplex virus type 1 pathogenesis in Alzheimer's disease. *Front Aging Neurosci*. 2018;10:48.
44. Hogestyn JM, Mock DJ, Mayer-Proschel M. Contributions of neurotropic human herpesviruses herpes simplex virus 1 and human herpesvirus 6 to neurodegenerative disease pathology. *Neural Regen Res*. 2018;13(2):211.
45. Farooq AV, Paley GL, Lubniewski AJ, Gonzales JA, Margolis TP. Unilateral posterior interstitial keratitis as a clinical presentation of herpes simplex virus disease. *Cornea*. 2018;37(3):375–8.
46. Bacik LC, Chung C. Human papillomavirus-associated cutaneous disease burden in human immunodeficiency virus (HIV)-positive patients: the role of human papillomavirus vaccination and a review of the literature. *Int J Dermatol*. 2018;57(6):627–34.
47. Dayakar S, Pillai HR, Thulasi VP, Jayalekshmi D, Nair RR. Comparative study of molecular approaches for the detection of influenza virus from patient samples using real-time PCR: prospective disease burden study in Kerala (India) from 2010 to 2016. *Curr Infect Dis Rep*. 2018;20(8):24.
48. Ivey KS, Edwards KM, Talbot HK. Respiratory syncytial virus and associations with cardiovascular disease in adults. *J Am Coll Cardiol*. 2018;71(14):1574–83.
49. Karron RA, Zar HJ. Determining the outcomes of interventions to prevent respiratory syncytial virus disease in children: what to measure? *Lancet Respir Med*. 2018;6(1):65–74.
50. Alcendor DJ. Zika virus infection and implications for kidney disease. *J Mol Med (Berl)*. 2018;96(11):1145–51.
51. Shehu NY, Shwe D, Onyedibe KI, Pam VC, Abok I, Isa SE, Egah DZ. Pathogenesis, diagnostic challenges and treatment of zika virus disease in resource-limited settings. *Niger Postgrad Med J*. 2018;25(2):67–72.

52. Muthusamy VS, Anand S, Sangeetha KN, Sujatha S, Arun B, Lakshmi BS. Tannins present in Cichoriumintybus enhance glucose uptake and inhibit adipogenesis in 3T3-L1 adipocytes through PTP1B inhibition. *Chem Biol Interact.* 2008;174(1):69–78.
53. Lakshmi BS, Sujatha S, Anand S, Sangeetha KN, Narayanan RB, Katiyar C, et al. Cinnamic acid, from the bark of *Cinnamomum cassia*, regulates glucose transport via activation of GLUT4 on L6 myotubes in a phosphatidylinositol 3-kinase-independent manner. *J Diabetes.* 2009;1(2):99–106.
54. Shilpa K, Sangeetha KN, Muthusamy VS, Sujatha S, Lakshmi BS. Probing key targets in insulin signaling and adipogenesis using a methanolic extract of *Costusciptus* and its bioactive molecule, methyl tetracosanoate. *Biotechnol Lett.* 2009;31(12):1837.
55. Sangeetha KN, Sujatha S, Muthusamy VS, Anand S, Nithya N, Velmurugan D, et al. 3 $\beta$ -taraxerol of *Mangiferaindica*, a PI3K dependent dual activator of glucose transport and glycogen synthesis in 3T3-L1 adipocytes. *Biochim Biophys Acta.* 2010;1800(3):359–66.
56. Muthusamy VS, Saravanababu C, Ramanathan M, Raja RB, Sudhagar S, Anand S, et al. Inhibition of protein tyrosine phosphatase 1B and regulation of insulin signalling markers by caffeoyl derivatives of chicory (*Cichoriumintybus*) salad leaves. *Br J Nutr.* 2010;104(6):813–23.
57. Anand S, Muthusamy VS, Sujatha S, Sangeetha KN, Raja RB, Sudhagar S, et al. Aloe emodin glycosides stimulates glucose transport and glycogen storage through PI3K dependent mechanism in L6 myotubes and inhibits adipocyte differentiation in 3T3L1 adipocytes. *FEBS Lett.* 2010;584(14):3170–8.
58. Sathyia S, Sudhagar S, Priya MV, Raja RB, Muthusamy VS, Devaraj SN, et al. 3 $\beta$ -Hydroxylup-20 (29)-ene-27, 28-dioic acid dimethyl ester, a novel natural product from *Plumbagozeylanica* inhibits the proliferation and migration of MDA-MB-231 cells. *Chem Biol Interact.* 2010;188(3):412–20.
59. Baskaran SK, Goswami N, Selvaraj S, Muthusamy VS, Lakshmi BS. Molecular dynamics approach to probe the allosteric inhibition of PTP1B by chlorogenic and cichoric acid. *J Chem Inf Model.* 2012;52(8):2004–12.
60. Sangeetha KN, Shilpa K, Kumari PJ, Lakshmi BS. Reversal of dexamethasone induced insulin resistance in 3T3L1 adipocytes by 3 $\beta$ -taraxerol of *Mangiferaindica*. *Phytomedicine.* 2013;20(3–4):213–20.
61. Shilpa K, Dinesh T, Lakshmi BS. An in vitro model to probe the regulation of adipocyte differentiation under hyperglycemia. *Diabetes Metab J.* 2013;37(3):176–80.
62. Shilpa K, Dinesh T, Lakshmi BS. Response: an in vitro model to probe the regulation of adipocyte differentiation under hyperglycemia (Diabetes Metab J 2013;37:176–80). *Diabetes Metab J.* 2013;37(4):298–9.
63. Jayashree B, Bibin YS, Prabhu D, Shanthirani CS, Gokulakrishnan K, Lakshmi BS, et al. Increased circulatory levels of lipopolysaccharide (LPS) and zonulin signify novel biomarkers of proinflammation in patients with type 2 diabetes. *Mol Cell Biochem.* 2014;388(1–2):203–10.
64. Posa JK, Selvaraj S, Sangeetha KN, Baskaran SK, Lakshmi BS. p53 mediates impaired insulin signaling in 3T3-L1 adipocytes during hyperinsulinemia. *Cell Biol Int.* 2014;38(7):818–24.
65. Thiagarajan G, Muthukumaran P, Sarath Kumar B, Muthusamy VS, Lakshmi BS. Selective inhibition of PTP 1B by vitalboside a from *Syzygiumcumini* enhances insulin sensitivity and attenuates lipid accumulation via partial agonism to PPAR  $\gamma$ : in vitro and in Silico investigation. *Chem Biol Drug Des.* 2016;88(2):302–12.
66. Sangeetha KN, Sujatha S, Muthusamy VS, Anand S, Shilpa K. Current trends in small molecule discovery targeting key cellular signaling events towards the combined management of diabetes and obesity. *Bioinformation.* 2017;13(12):394.
67. Muthukumaran P, Thiagarajan G, Babu RA, Lakshmi BS. Raffinose from *Costusspeciosus* attenuates lipid synthesis through modulation of PPARs/SREBP1c and improves insulin sensitivity through PI3K/AKT. *Chem Biol Interact.* 2018;284:80–9.
68. Senthil V, Ramadevi S, Venkatakrishnan V, Giridharan P, Lakshmi BS, Vishwakarma RA, et al. Withanolide induces apoptosis in HL-60 leukemia cells via mitochondria mediated cytochrome c release and caspase activation. *Chem Biol Interact.* 2007;167(1):19–30.

69. Sudhagar S, Sathy S, Anuradha R, Gokulapriya G, Geetharani Y, Lakshmi BS. Inhibition of epidermal growth factor receptor by ferulic acid and 4-vinylguaiacol in human breast cancer cells. *Biotechnol Lett*. 2018;40(2):257–62.
70. Sudhagar S, Sathy S, Gokulapriya G, Lakshmi BS. AKT-p53 axis protect cancer cells from autophagic cell death during nutrition deprivation. *Biochem Biophys Res Commun*. 2016;471(4):396–401.
71. Kennedy RK, Naik PR, Veena V, Lakshmi BS, Lakshmi P, Krishna R, Sakthivel N. 5-methyl phenazine-1-carboxylic acid: a novel bioactive metabolite by a rhizosphere soil bacterium that exhibits potent antimicrobial and anticancer activities. *Chem Biol Interact*. 2015;231:71–82.
72. Sathy S, Sudhagar S, Lakshmi BS. Estrogen suppresses breast cancer proliferation through GPER/p38 MAPK axis during hypoxia. *Mol Cell Endocrinol*. 2015;417:200–10.
73. Sangeetha KN, Lakshmi BS, Devaraj SN. Dexamethasone promotes hypertrophy of H9C2 cardiomyocytes through calcineurin B pathway, independent of NFAT activation. *Mol Cell Biochem*. 2016;411(1–2):241–52.
74. Sathy S, Sudhagar S, Sarathkumar B, Lakshmi BS. EGFR inhibition by pentacyclic triterpenes exhibit cell cycle and growth arrest in breast cancer cells. *Life Sci*. 2014;95(1):53–62.
75. Ramadevi Mani S, Lakshmi BS. G1 arrest and caspase-mediated apoptosis in HL-60 cells by dichloromethane extract of Centrosemapubescens. *Am J Chin Med*. 2010;38(06):1143–59.
76. Sudhagar S, Sathy S, Pandian K, Lakshmi BS. Targeting and sensing cancer cells with ZnOnanoprobes in vitro. *Biotechnol Lett*. 2011;33(9):1891–6.
77. Sudhagar S, Sathy S, Lakshmi BS. Rapid non-genomic signalling by 17 $\beta$ -oestradiol through c-Src involves mTOR-dependent expression of HIF-1 $\alpha$  in breast cancer cells. *Br J Cancer*. 2011;105(7):953.
78. Gayathri B, Manjula N, Vinaykumar KS, Lakshmi BS, Balakrishnan A. Pure compound from Boswelliaserrata extract exhibits anti-inflammatory property in human PBMCs and mouse macrophages through inhibition of TNF $\alpha$ , IL-1 $\beta$ , NO and MAP kinases. *Int Immunopharmacol*. 2007;7(4):473–82.
79. Subramanian K, Selvakumar C, Vinaykumar KS, Goswami N, Meenakshisundaram S, Balakrishnan A, Lakshmi BS. Tackling multiple antibiotic resistance in enteropathogenic Escherichia coli (EPEC) clinical isolates: a diarylheptanoid from Alpiniaofficinarum shows promising antibacterial and immunomodulatory activity against EPEC and its lipopolysaccharide-induced inflammation. *Int J Antimicrob Agents*. 2009;33(3):244–50.
80. Balakrishnan G, Janakarajan L, Balakrishnan A, Lakshmi BS. Molecular basis of the anti-inflammatory property exhibited by cyclo-penta nophenanthenol isolated from Lippianodiflora. *Immunol Investig*. 2010;39(7):713–39.
81. <https://www.schrodinger.com/>.
82. <http://autodock.scripps.edu/resources/adt>.
83. <http://zinc.docking.org/>.
84. <http://pubchem.ncbi.nlm.nih.gov>.
85. <http://chemspider.com>.
86. Gaweñ E, Hiss JA, Brown JB, Schneider G. Advancing drug discovery via GPU-based deep learning. *Expert Opin Drug Discov*. 2018;13(7):579–82.
87. Jing Y, Bian Y, Hu Z, Wang L, Xie XQ. Deep learning for drug design: An artificial intelligence paradigm for drug discovery in the big data era. *AAPS J*. 2018;20(3):58.
88. Gaweñ E, Hiss JA, Schneider G. Deep learning in drug discovery. *Mol Inf*. 2016;35(1):3–14.
89. Korotcov A, Tkachenko V, Russo DP, Ekins S. Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets. *Mol Pharm*. 2017;14(12):4462–75.
90. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. *Drug Discov Today*. 2018;23(6):1241–50.
91. Baskin II, Winkler D, Tetko IV. A renaissance of neural networks in drug discovery. *Expert Opin Drug Discovery*. 2016;11(8):785–95.
92. Lima AN, Philot EA, Trossini GH, Scott LP, Maltarollo VG, Honorio KM. Use of machine learning approaches for novel drug discovery. *Expert Opin Drug Discov*. 2016;11(3):225–39.

93. Dobchev D, Karelson M. Have artificial neural networks met expectations in drug discovery as implemented in QSAR framework? *Expert Opin Drug Discovery*. 2016;11(7):627–39.
94. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today*. 2015;20(3):318–31.
95. Patel J. Science of the science, drug discovery and artificial neural networks. *Curr Drug Discov Technol*. 2013;10(1):2–7.
96. Reutlinger M, Schneider G. Nonlinear dimensionality reduction and mapping of compound libraries for drug discovery. *J Mol Graph Model*. 2012;34:108–17.
97. Hecht D, Fogel GB. A novel in silico approach to drug discovery via computational intelligence. *J Chem Inf Model*. 2009;49(4):1105–21.
98. Pozzan A. Molecular descriptors and methods for ligand based virtual high throughput screening in drug discovery. *Curr Pharm Des*. 2006;12(17):2099–110.
99. Scotti L, Ishiki H, Mendonca Junior FJ, da Silva MS, Scotti MT. Artificial neural network methods applied to drug discovery for neglected diseases. *Comb Chem High Throughput Screen*. 2015;18(8):819–29.
100. Zernov VV, Balakin KV, Ivaschenko AA, Savchuk NP, Pletnev IV. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J Chem Inf Comput Sci*. 2003;43(6):2048–56.
101. Gombar VK, Silver IS, Zhao Z. Role of ADME characteristics in drug discovery and their in silico evaluation: in silico screening of chemicals for their metabolic stability. *Curr Top Med Chem*. 2003;3(11):1205–25.
102. Liu R, Sun H, So SS. Development of quantitative structure– property relationship models for early ADME evaluation in drug discovery. 2. Blood-brain barrier penetration. *J Chem Inf Comput Sci*. 2001;41(6):1623–32.
103. Shi LM, Fan Y, Lee JK, Waltham M, Andrews DT, Scherf U, Paull KD, Weinstein JN. Mining and visualizing large anticancer drug discovery databases. *J Chem Inf Comput Sci*. 2000;40(2):367–79.
104. Murphy PM. Viral anti-chemokines: from pathogenesis to drug discovery. *J Clin Invest*. 2000;105(11):1515–7.
105. Hsu JT, Wang HC, Chen GW, Shih SR. Antiviral drug discovery targeting to viral proteases. *Curr Pharm Des*. 2006;12(11):1301–14.
106. Murgueitio MS, Bermudez M, Mortier J, Wolber G. In silico virtual screening approaches for anti-viral drug discovery. *Drug Discov Today Technol*. 2012;9(3):e219–25.
107. Schinazi RF, Bassit L, Gavagnano C. HCV drug discovery aimed at viral eradication. *J Viral Hepat*. 2010;17(2):77–90.
108. Dacheux L. Editorial [hot topic: the challenge of viral encephalitis: from etiological diagnosis to efficient antiviral drug discovery (Guest Editor: Laurent Dacheux)]. *Infect Disord Drug Targets*. 2011;11(3):205.
109. De Clercq E. Human viral diseases: what is next for antiviral drug discovery? *Curr Opin Virol*. 2012;2(5):572–9.
110. Hari Narayana Moorthy NS, Poongavanam V, Pratheepa V. Viral M2 ion channel protein: a promising target for anti-influenza drug discovery. *Mini Rev Med Chem*. 2014;14(10):819–30.

# Wavelet-based Multifractal Spectrum Estimation in Hepatitis Virus Classification Models by Using Artificial Neural Network Approach



**Yeliz Karaca**

**Abstract** Fractal and multifractal geometries have been applied extensively in various medical signals which exhibit fractal characteristics. Application of such geometries rests on the estimation of fractal features. Within this framework, various methods have been proposed for the estimation of the multifractal spectral or fractal dimension of a particular signal. Wavelet transform modulus maxima (WTMM) is one of the methods employed for the detection of fractal dimension of a signal. It was developed for the characterization of signal singularities. Hepatitis, inflammation of the liver, may prove to be a serious disease with serious potential risks. This study proposes an alternative method for the classification of hepatitis virus as per die/live with the use of two aspects, namely multifractal analysis and Artificial Neural Network (ANN). As the first aspect, for the multifractal analysis, Wavelet Transform Modulus Maxima (WTMM) (Multifractal Spectrum estimation) was used with the following stages: (a) WTMM was applied to the hepatitis dataset (self-similar and significant attributes were identified) and wtmm\_hepatitis dataset was generated. (b) Continuous Wavelet Transform was applied on the hepatitis dataset (hepatitis\_dataset) and wtmm\_hepatitis dataset. The second aspect is related to the application of Feed Forward Back Propagation (FFBP) algorithm which is an ANN application with the following steps: (i) FFBP algorithm was applied to both hepatitis dataset (hepatitis\_dataset) and (wtmm\_hepatitis dataset) to identify the classification as per die/live (ii) The attributes proven to be the most effective were determined based on the results (sensitivity, specificity and accuracy rate). The highest level of accuracy has been obtained from the wtmm\_hepatitis dataset. The main contribution of this study is that it has proven to provide an alternative in multifractal spectrum estimation by determining the self-similar and significant attributes through WTMM for the first time in the literature. The proposed method in the study aims at bringing a new frontier in the related fields by placing emphasis on the

---

Y. Karaca (✉)

University of Massachusetts Medical School, Worcester, MA, USA

e-mail: [yeliz.karaca@ieee.org](mailto:yeliz.karaca@ieee.org); [yeliz.karaca@umassmemorial.org](mailto:yeliz.karaca@umassmemorial.org)

significance of significant attributes' characterization to obtain optimal accuracy rates for the solution of problems.

**Keywords** Hepatitis virus · Diagnosis · Multifractal analysis · Wavelet transform modulus maxima · Multifractal spectrum · Singularity · Wavelet continuous transform · ANN · Lipschitz exponent

### Key Phrases

Health decision-making, deterioration of liver due to hepatitis, Wavelet-based multifractal analysis for detecting attributes, classification accuracy performance by Artificial Neural Networks for hepatitis cases, Lipschitz Exponent ( $\alpha$ ) measuring with WTMM, detection of singularity with Wavelet.

## 1 Introduction

Artificial Neural Networks (ANN) is a type of artificial intelligence applied in various areas, ranging from medicine, engineering, economics, image processing, and other relevant areas. ANN is made up of neurons which work in a similar way to the brain, and utilizes processing of the brain as a basis to develop algorithms, which can be employed to model complex patterns, address classification, and prediction problems. These networks are applied extensively to solve medical problems in particular. Correspondingly, fractal and multifractal methods are also widely used in data analysis problems in general and particularly in the medical area to diagnose, classify the diseases and for other similar purposes. Accordingly, Wavelet Transform Modulus Maxima (WTMM) approach is useful while analysing multifractal data (data that have a multiple fractal dimension) [1, 2]. A mathematical model for infectious disease is a significant theoretical method in epidemiology. It has been used to simulate the hepatitis B prevalence and assess different immunization strategies. Besides this, mathematical modelling of infectious diseases' spread continues to contribute through important insights regarding the behaviours and control of the diseases. It has recently become an important tool in understanding the dynamics of diseases apart from decision making processes for timely intervention.

Among many serious health issues, hepatitis is one of the most serious medical problems [3]. Known to be one of the deadliest diseases, hepatitis is reported to cause 1.5 million deaths worldwide. Hepatitis is characterized by the inflammation and destruction of liver cells [4]. Five hepatotropic viruses, which are hepatitis A [5], B [6, 7], C [8], D [9] and E virus, are addressed [10]. Other than medical reasons, psychological factors also play a role in hepatitis, which are categorized as non-treatment-related psychiatric symptoms, including but not limited to irritability, depression, delirium, and even mania [11].

Medical diagnosis of this medical problem is difficult, and many factors should be taken into account regarding its diagnosis [4]. Early detection increases the

chances of healing. In this regard, diagnosis and classification of the disease come to the foreground. The use of technology, particularly artificial intelligence (AI), offers many benefits including reduction of cost and time, increasing the quality of life through early diagnosis, thus enhancing clinical experience, as well as reduction of misdiagnosis [12].

As one of the types of artificial intelligence, ANN is applied to address various diseases. In recent years, studies concerning with the use of ANN in the medical area have been performed. Regarding the classification of hepatitis using ANN, several relevant examples from the literature can be provided. One study [3] examines the factors that describe increasing risk of hepatitis-C virus. The authors used a three-stage procedure which yielded a classification accuracy rate of 89%. In another study related to hepatitis virus, [13] introduced an automatic diagnosis system based on Neural Network, which dealt with feature extraction and classification. The ANN diagnosis system classification accuracy of the study regarding hepatitis virus was obtained at around 99.1% and 100% for training data and testing data, respectively. The study of [12] provides an artificial neural network based approach for the diagnosis of hepatitis virus. Certain factors that were likely to influence the performance of patients were outlined and these were used as input variables for the ANN model. The test data evaluation of the study revealed that the ANN model managed to predict the diagnosis accurately with a rate of 93%. Another study [14] dealt with the issues concerning Extreme Learning Machine (ELM) which is an extension of feed forward Neural Network model. They proposed the hybrid algorithms, namely Weighted ELM and Weighted ELM with Ant Colony Optimization with continuous domain algorithm (ACOR). Comparative analysis was conducted over the hepatitis dataset, and the experimental results showed that Weighted ELM with ACOR was superior to the other proposed methods. Finally, the study of [4] aimed to propose an accurate method for the diagnosis of hepatitis disease by making use of the benefits regarding the ensemble learning. As the methods, they used Non-linear Iterative Partial Least Squares to carry out the reduction of data dimensionality, Self-Organizing Map technique to cluster task and ensembles of Neuro-Fuzzy Inference System to predict hepatitis. Decision trees were also used for the selection of the most significant features in experimental dataset. The analyses of the study showed that their analyses were superior to ANFIS, K-Nearest Neighbours, Support Vector Machine and Neural Network.

The diagnosis of the diseases plays a very important role in the treatment process and survival rates of the patients afflicted with the diseases. Correspondingly, the applications developed and performed regarding the detection, classification and prediction of the diseases are of great importance. When recent studies are inquired, it is seen that Multiple Sclerosis (MS), neurological disorder of the auto-immune system, on which quite a lot of studies exist. A related study [15] proposed a classification for the subgroups of MS disease based on the multifractal method through the use of the Self-Organizing Map (SOM) algorithm. Consequently, a cluster analysis was obtained by identifying pixels from affected regions based on MR images by multifractal methods, and diagnosing the subgroups of MS through artificial neural networks. Another study [16] had the aim of identifying the self-similar and

homogenous pixels by the application of the Diffusion Limited Aggregation (DLA) onto the MR images of the patients. Through the study, it has been demonstrated that through the ANN algorithms' application, the most significant pixels can be identified within the relevant dataset. The authors proposed using Wavelet feature descriptor and combined it with an Artificial Neural network for classification of MS subgroups with the application of wavelet transform. Feed forward neural network and feed forward back propagation network were also used. The results obtained showed that the feed forward back propagation neural network yielded a better classification result than feed forward. As for other groups of illnesses, where classification is of vital importance, one study [17] is related to the early detection for lung cancer utilizing Wavelet Feature Descriptor and Feed Forward Back Propagation Neural Networks Classifier. The results reveal that the feed forward back propagation neural network yields more accurate classification results compared to those obtained from feed forward.

In line with the parallel developments and technological advancements, a growing interest has been seen in Wavelet Analysis [18, 19] and Wavelet Coefficient Method [20, 21] fractal and multifractal analysis to address the needs arising in the medical area [22] related to signals, image processing and other relevant methods and fields. There are different approaches with regard to multifractal behaviour. Accordingly, WTMM approaches have different uses and fields of application in recent years. To cite some studies from recent years, [23]'s study presents the measurements of Lipschitz exponent utilizing wavelet transform with a new area that is based on objective function. The results of the study experiments show the method is more robust and also more precise. The study by [24] aimed at trying to understand the key features of more complex model of Lipschitz regularity, and develop robust but simple methods for estimation. The study by [2] reflects another aspect in terms of approach with respect to multifractal behaviour. Their study is concerned with World Stock Indexes. The study made use of two aspects to fix difference. Wavelet Transform Modulus Maxima approach, including two basic aspects: Wavelet aspect and Multifractal formalism. The use of the approaches in indexes indicates the interdisciplinary flexibility of the methods. The study of [25] is concerned with Modulus maxima derived from the continuous wavelet transform. The authors developed an R-Wave detector and tested it by using patient signals as a result of which they attained high sensitivity rates at 99.7% and a Positive Predictive Value at 99.68%.

The approach in this study aims to be broader and more comprehensive since it includes two different approaches that have been used when compared with other studies done with hepatitis data. The hepatitis dataset (hepatitis\_dataset) analysed in this study includes 150 patients in two different classes which are live (in 120 cases with 80%) and die (in 30 cases with 20%). Classification with regard hepatitis has not been the concern or the scope of this current study. The dataset of the study includes 19 attributes (13 binary and 6 attributes with 6–8 discrete values). As for the conducting of this study, it encompasses two stages, as elaborated below: The first approach is related to multifractal analysis, and it is provided using the Wavelet Transform Modulus Maxima approach that encompasses two main aspects: Wavelet

aspect (Continuous Wavelet Transform), as well as Multifractal Analysis estimation (Multifractal Spectrum estimation). (a) Multifractal Analysis (Multifractal spectrum estimation) is applied on the hepatitis\_dataset; and significant self-similar attributes were identified. Thus, the dataset named “wtmm\_hepatitis dataset” was generated. (b)Wavelet analysis (Continuous Wavelet Transform) was applied on hepatitis\_dataset and wtmm\_hepatitis dataset. The second aspect of the study is concerned with Artificial Neural Network (ANN), the application of Feed Forward Back Propagation (FFBP), which includes the following steps: (i) FFBP was applied on the hepatitis\_dataset and (wtmm\_hepatitis dataset) to identify the classification as per survival. (ii) FFBP algorithm was employed for the classification of die/live situation and the related classification allowed to do the calculation of overall accuracies (about sensitivity, specificity, accuracy rates); hence, comparative analyses were done. (iii) The attributes proven to be the most effective for the classification as per die/live were determined.

Among the studies that have been mentioned above, the study provides a different outlook in that the significant, self-similar and efficient attributes as obtained by WTMM method have been handled for the classification of die/live for hepatitis through the application of ANN. In this regard, this study is one of its kind and it aims at bridging a gap in the literature by the use of these methods and the algorithm.

The paper is organized in the following way: Section 2 is on Materials and Methods. Section 3 is concerned with Experimental Results and Discussion, which is provided under three subsections: 3.1 on the WTMM analyses for hepatitis dataset with Lipschitz Exponent (LE) measuring, 3.2 on the application of 1D Continuous Wavelet Transform on the hepatitis datasets and 3.3 dealing with application of FFBP Algorithm. Finally, Section 4 provides the conclusion of the study.

## 2 Materials and Methods

### 2.1 Patient Details

In the study related to hepatitis virus, the hepatitis dataset (hepatitis\_dataset) was used and the data were obtained from the UC Irvine Machine Learning Repository [26].

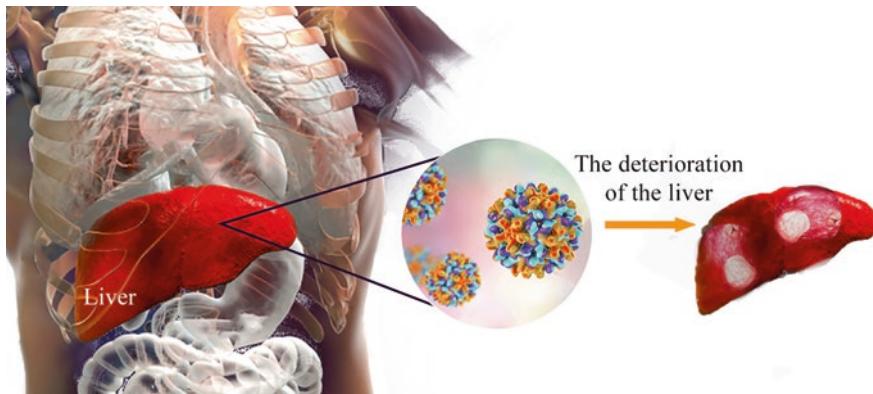
Details are presented in Table 1. The hepatitis dataset consists of 19 attributes which are influential in affecting the course of the disease. Including 150records of patients who are in two different classes (e.g.: die in 30 cases with 20% and live in 120 cases with 80%. 19 attributes (13 binary and six attributes with 6–8 discrete values) are included in the dataset.

The proposed method in this study is evaluated on a real-world dataset the data of which have been obtained from the UC Irvine Machine Learning Repository [26].

Figure 1 as may be seen above presents the progression of hepatitis and the damage in the liver caused by hepatitis virus.

**Table 1** Hepatitis dataset

Distribution of hepatitis as per survival	Attributes	Data size
Die (30) Live (120)	Age, Gender Steroid, Antivirals, Fatigue, Malaise, Anorexia, Liver Big, Liver Firm, Spleen Palpable, Spiders, Ascites, Varices Bilirubin, Alk Phosphate, Sgot, Albumin, Protome, Histology	150 × 19

**Fig. 1** Deterioration caused by hepatitis

As can be seen from Fig. 1, hepatitis stems from the substance named bilirubin, a yellow substance that the body forms while renewing the red blood cells. It is also one of the byproducts of red blood cells. Normally, every day 1% of the red blood cells get retired, thus they are replaced by the new ones. The old ones are processed in the liver and eliminated during which most of the bilirubin is discarded from the body by excretion. If there is a breakdown of blood cells which exceeds the amount that the liver can cope with, then bilirubin (yellow pigments) start to accumulate in the body. When this amount becomes significant, hepatitis starts to emerge and the deterioration of liver is seen.

## 2.2 Methods

This study has provided contributions in terms of two approaches. Accordingly, the aim of this paper is to determine the significant, self-similar and distinguishing attributes along with the accurate classification of hepatitis as per survival (die/live) based on the dataset. The steps in this study from two perspectives depend on the following:

The first one is multifractal analysis, and it is provided by utilizing the Wavelet Transform Modulus Maxima approach that encompasses two main aspects: Wavelet aspect (Direct Continuous Wavelet Transform), as well as Multifractal Analysis estimation (Multifractal Spectrum estimation). The steps for the application of this approach are provided in detail below:

- (a) Multifractal Analysis (Multifractal spectrum estimation) is applied on the hepatitis dataset (hepatitis\_dataset) and significant self-similar attributes were identified. Thus, the dataset named wtmm\_hepatitis dataset was generated.
- (b) Wavelet analysis (Continuous Wavelet Transform) was applied on hepatitis\_data set and wtmm\_hepatitis dataset. In this stage, the datasets were compared visually based on the obtaining of the analyses.

The second stage of the study with regard to Feed Forward Back Propagation (FFBP), which is one of the ANN algorithms, is comprised of the following aspects. The steps for the application of this approach are provided in detail below:

1. FFBP algorithm, was applied on the hepatitis\_dataset and wtmm\_hepatitis dataset for the purpose of identifying the classification as per survival.
2. FFBP algorithm was used for the classification of die/live aspect and the relevant classification enabled the calculation of overall accuracies based on sensitivity, specificity, accuracy rates; thus, comparative analyses were performed.
3. The attributes which have proven to be most effective for the classification as per die /live were determined based on the results concerning sensitivity, specificity as well as accuracy rates.

Computations and figures were done on Matlab in this study.

## 2.2.1 Multifractal Analysis

Multifractals is regarded as an extension of fractals. A multifractal object is more multifaceted since it is always invariant by translation even though the dilatation factor required be able to distinguish the detail from the whole object depending on the detail being observed. Regarding fractal dimension (referred to as FD or D) estimation, many methods are available for the approximation of the multifractal spectrum as wavelets [27, 28].

### 2.2.1.1 Wavelet Transform

$\psi$  is a real function that is a wavelet if its integral is zero Eq. (1) [28].

$$\int_{-\infty}^{\infty} \psi(x) dx = 0 \quad (1)$$

The continuous wavelet transform of a function  $f \in L^2(R)$  related to the wavelet  $\psi$  is defined in Eq. (2).

$$Wf(u,s) = \int_{-\infty}^{\infty} f(t) \frac{1}{\sqrt{s}} \psi^* \left( \frac{t-u}{s} \right) dt \quad (2)$$

The complex conjugate of  $\psi$  is denoted as  $\psi^*$  [28]. A wavelet  $\psi(x)$  is stated to have  $n$  vanishing moments provide that for all positive integers  $k < n$ , it fulfils the condition stated in Eq. (3).

$$\int_{-\infty}^{\infty} x^k \psi(x) dx = 0 \quad (3)$$

One of the widespread wavelets in practice is the  $n$ -th derivative of the Gaussian function, which is denoted in Eq. (4).

$$\psi_n(x) = -\frac{d^n}{dx^n} e^{-\frac{x^2}{2}} \quad (4)$$

The number of vanishing moments is significant due to offering an upper bound measurement for singularity characterization while conducting a wavelet singularity analysis [28–33].

### *Detection of Singularity with Wavelet*

Lipschitz Exponent (LE) is a measurement of singularity strength. It was shown that LE can be calculated by WTMM as put forth Mallat et al. [27, 29]. The wavelet transform of a function  $f(u)$  is the local maxima of wavelet transform modulus  $Wf(u)$ , which can be explained in the following terms [23, 27, 29, 34]:

- A local maximum is the point  $(u_0, s_0)$  in that  $\frac{(\partial Wf(u, s_0))}{\partial u}$  has a zero-crossing at  $u = u_0$ , once  $u$  changes.
- A modulus maximum is the point  $(u_0, s_0)$  in which  $|Wf(u, s_0)| < |Wf(u_0, s_0)|$  when  $u$  belongs to the left or the right of the neighbourhood of  $u_0$ ; in addition,
- $|Wf(u, s_0)| \leq |Wf(u_0, s_0)|$  when  $u$  belongs to the opposite neighbourhood of  $u_0$ .
- A maxima line calls any of the connected curves within the scale space  $(u, s)$ ; and along this, all points are modulus maxima.

In this study, it is supposed the  $\psi$  has a compact support, and is  $n$  times differentiable in a continuous manner; and is the smoothing function's  $n$ th derivatives.

### *Lipschitz Exponent ( $\alpha$ ) Measuring with WTMM*

The Lipschitz exponent can be calculated through the use of the following Algorithm 1 [23, 29] in accordance with the theorem (which may be resorted to in references 23–34).

**Algorithm 1** Measurement of the Lipschitz exponent.

---

Step 1. Compute the straight line  $l(\log_2(s))$  which connects  $(\log_2(s_{small}))$  and  $\log_2 | Wf(u, s_{small}) |$  and  $(\log_2(s)_{max}, \log_2 | Wf(u, s_{max}) |)$ .

If  $l(\log_2(s)) \geq \log_2 | Wf(u, s) |$  return the intercept  $\log_2(A)$  and slope  $\alpha$  of  $l(\log_2(s))$ , proceed with step 6, or else proceed with step 2.

Step 2. Let  $s = s_{max}$  and  $f(A, \alpha) = C$  in which  $C$  is a constant which is large enough.

Step 3. Calculate tangent  $l(\log_2(s))$  at  $\log_2(s), \log_2 | Wf(u, s) |$ .

If  $l(\log_2(s)) \geq \log_2 | Wf(u, s) |$  proceed with Step 4. Or not, proceed with Step 6.

Step 4. Calculate record of the result  $f$  and intercept  $\log_2(A)$  and slope  $\alpha$  of  $l(\log_2(s))$ .

If  $f < f(A, \alpha), f(A, \alpha) = f$  and  $LE = \alpha$ .

If  $s = s_{min}$ , go along with Step 6, or else, go along with Step 5.

Step 5.  $s = s - \Delta \log_2(s)$ , proceed with Step 3.

Step 6. Output  $LE = \alpha$

---

Since the a priori knowledge of  $\alpha$  is used, the Algorithm 1 used in this study looks for the optimal result along  $\log_2 | Wf(u, s) |$  curve merely, and the problem of initialization of  $\alpha$  and  $A$  is possible to be prevented [23, 27, 34].

#### 2.2.2 Numerical Experiments

Some results regarding the hepatitis dataset are provided in Table 2. For each of the variable in the hepatitis dataset, Wavelet Transform Modulus Maxima (WTMM) methods have been applied. It is observed that for the irregular data, like the ones handled here, which belong to correlate local irregularity at specific attributes in hepatitis dataset with hepatitis classification as per die/live. Within this framework, there is a wide range of sampled functions of 150 length and an estimate of local regularity for each 19 hepatitis attribute has been obtained (see Table 1).

In this study, the following conclusions can be inferred from numerical experiments; numerical experiments were obtained by Wavelet Transform Modulus Maxima methods and FFBP algorithm which is one of the ANN algorithms.

The steps of multifractal method applied on the hepatitis\_dataset can be summarized as follows:

In step 1, to capture the pointwise LE, the maxima was focused on while analysing regular with Wavelet Transform Modulus Maxima (WTMM) was applied on the hepatitis\_dataset and significant self-similar attributes were identified. Subsequently, the dataset named wtmm\_hepatitis dataset was formed. In addition, Wavelet aspect (Continuous Wavelet Transform) was applied on both of the datasets.

**Table 2** Outline of the proposed strategy

Input: hepatitis_dataset = [150, 19]
Put the hepatitis as per die/live as a target data into a matrix $O[2 \times 1]$
Stage 1. hepatitis_dataset = [150 × 19]
Stage 2. Wavelet Transform Modulus Maxima method as 1-D attributes Applied Gaussian wavelet as 1-D applied to Wavelet hepatitis_dataset Obtained wtmm_hepatitis dataset = [150, 15]
FFBP algorithm were applied on the hepatitis_dataset = [150 × 19] and wtmm_hepatitis dataset = [150, 15] and classification of hepatitis was made as per Hepatitis (die/live)
Stage 3. FFBP Classification 5 × 5 cross validation Divide input data (Hepatitis_dataset = [150 × 19]), (wtmm_hepatitis dataset = [150, 15]) randomly for 70% train, %15 test, 15% validation and target data $O[2 \times 1]$ into five different folds.
<b>for</b> $j = 1 : 5$ Make use of the $j - th$ fold for test, and the remaining two folds are merged as the training set. Record the results of the classification over $j - th$ fold. <b>end</b>
<b>Output:</b> Make a summary of the results of classification over each fold for hepatitis was made as per Hepatitis (die/live) Report overall accuracy

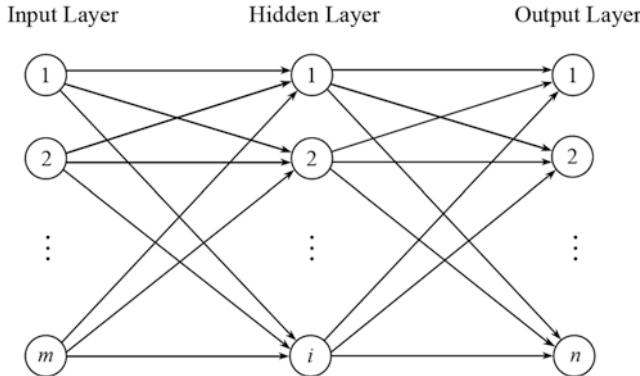
In the second approach of the study, Feedforward Back Propagation algorithm was applied on the hepatitis dataset (hepatitis\_dataset) and (wtmm\_hepatitis dataset) so as to identify the classification as per survival (die/live). This classification allowed for the calculation of overall accuracies as per sensitivity, specificity, accuracy rates; thus, comparative analyses were performed to detect the most effective attributes for the classification as per survival regarding hepatitis.

### 2.2.3 Artificial Neural Network Algorithm

Artificial neural networks (ANNs) are computing systems which have been inspired based on the biological neural networks which comprise animal brains. The procedure that is utilized to perform the learning process in a neural network is named as the training algorithm. Feed Forward Backpropagation Algorithm is one of the ANN algorithms, which has been employed in this study [35].

#### 2.2.3.1 Feed Forward Back Propagation Algorithm

A feedforward neural network is an artificial neural network in which the connections between the units do not form a cycle [35, 36]. Accordingly, it is dissimilar from the recurrent neural networks. The feedforward neural network is known to be the first and most simple kind of artificial neural network that has been developed.



**Fig. 2** The General Network Structure for FFBP Algorithm

In such networks, the information flows in only one direction, forward in the following direction: through the hidden nodes (if any) from the input nodes towards the output nodes. There are neither loops nor cycles in the network. A single-layer perceptron network is the simplest type of neural network and this includes a single layer of output nodes. It is known that the inputs are fed directly to the outputs through a series of weights. That is because it is regarded as the simplest sort of feed-forward network. The sum of the weights' inputs as well as products are computed in each node. In addition, if the value exceeds a certain threshold (which typically happens to be 0), then the neuron fires and assumes the activated value (which generally happens to be 1). If this condition is not fulfilled, it assumes the deactivated value (which has typically -1 value). Such kind of neurons concerning activation function are named as linear threshold units or artificial neurons as well (see Fig. 2).

The algorithm's network architecture is defined accordingly, and also the weights are included. Once the input examples that have  $m$ -dimension are entered,  $x_i = [x_1, x_2, \dots, x_m]^T$  can be seen. Correspondingly, the examples of output that are desired with  $n$ -dimension is specified by  $d_k = [d_1, d_2, \dots, d_n]^T$  (see Fig. 2).  $x_i$  values, the neurons' output values in  $i$ th layer ( $n$ ), the total input that is supposed to come to a neuron in  $j$  layer is administered in line with Eq. (5) [38] (see (Fig. 2)) [35, 36].

$$net_j = \sum_{i=1}^m w_{ji} \cdot x_i \quad (\text{from } i.\text{node to } j.\text{node}) \quad (5)$$

The  $j$  neuron output in the hidden layer (transfer function output) is calculated as provided in Eq. (6) [35].

$$y_j = f_j(net_j) \quad j = 1, 2, \dots, J \quad (6)$$

The total input that is going to come to  $k$  neuron in the output layer is calculated according to Eq. (7).

$$net_k = \sum_{j=1}^J w_{kj} \cdot y_j \quad (7)$$

The calculation of the non-linear output of a  $k$  neuron in the output layer is done as shown in Eq. (8).

$$o_k = f_k (net_k), \quad k = 1, 2, \dots, n \quad (8)$$

The output obtained from the network is compared to the actual output and  $e_k$  error is computed Eq. (9) [35].

$$e_k = (d_k - o_k) \quad (9)$$

$d_k$  denotes the target of any “ $k$ ” neuron in the output layer and  $o_k$  denotes the outputs obtained from the network. The weights which were obtained from the output layer are updated. The calculation of the total square error is made according to Eq. (10) for each of the examples [35, 36, 37],

$$E = \frac{1}{2} \sum_k (d_k - o_k)^2 \quad (10)$$

In this section of the study, FFBP algorithm was applied to the hepatitis\_dataset and wtmm\_hepatitis dataset for the hepatitis classification (die/live).

### 3 Experimental Results and Discussion

There are three main parts in this study concerning 3.1 is concerned with the WTMM analyses for the hepatitis dataset with LE measuring; 3.2 is on the application of 1D Continuous Wavelet Transform on the hepatitis datasets; and 3.3 deals with application of the FFBP Algorithm.

For the analyses concerning the experiments of this study, Matlab was utilised [38].

Two different approaches have been made use of in the study. The first approach is multifractal analysis, and it is presented using the Wavelet Transform Modulus Maxima approach that has two basic aspects: Multifractal Analysis estimation (Multifractal Spectrum estimation) as well as Wavelet aspect (Continuous Wavelet Transform). The second stage of the study concerning Artificial Neural Network (ANN).

The steps for the application of these two stages are presented with their details in Table 2.

The steps for application as presented in Table 3 are provided in the further Sections 3.1, 3.2 and 3.3.

**Table 3** Network Properties of the Training for FFBP Algorithm

ANN network properties	ANN algorithms network properties value
Input data	hepatitis_dataset ( $150 \times 19$ ), wtmm_hepatitis dataset ( $150 \times 15$ )
Training function	Scaled Conjugate Gradient
Adaption learning function	Learning Gradient Descent
Transfer function	Tansig
Performance function	Mean Squared Error
Hidden layer numbers	12
Transfer function	Sigmoid
Epoch	1000 iteration

### 3.1 The WTMM Analyses for Hepatitis Dataset with LE Measuring

In this study, the significant and self-similar attributes in the hepatitis dataset were identified. The sampled functions have a length of 150 and local regularity for each of the 19 hepatitis attributes is to be identified (see Table 1). To capture the point-wise LE, the maxima sufficient was focused on when analysing a strongest singularity.

As can be seen from Fig. 3, the WTMM first makes the calculation of the continuous wavelet transform by using the second derivative of a Gaussian wavelet. The wavelet that fulfils this criterion is the Mexican hat wavelet. Subsequently, WTMM makes the determination of the modulus maxima for each of the scale. The WTMM is to be used for the hepatitis dataset in this study in which 150 samples exist to make the determination of maxima in an accurate way. The presence of strong local singularities is marked by negative LE and for normal fluctuation positive LE is valid.

The modulus maximum definition at point  $x_0$  [23, 28, 29, 34],

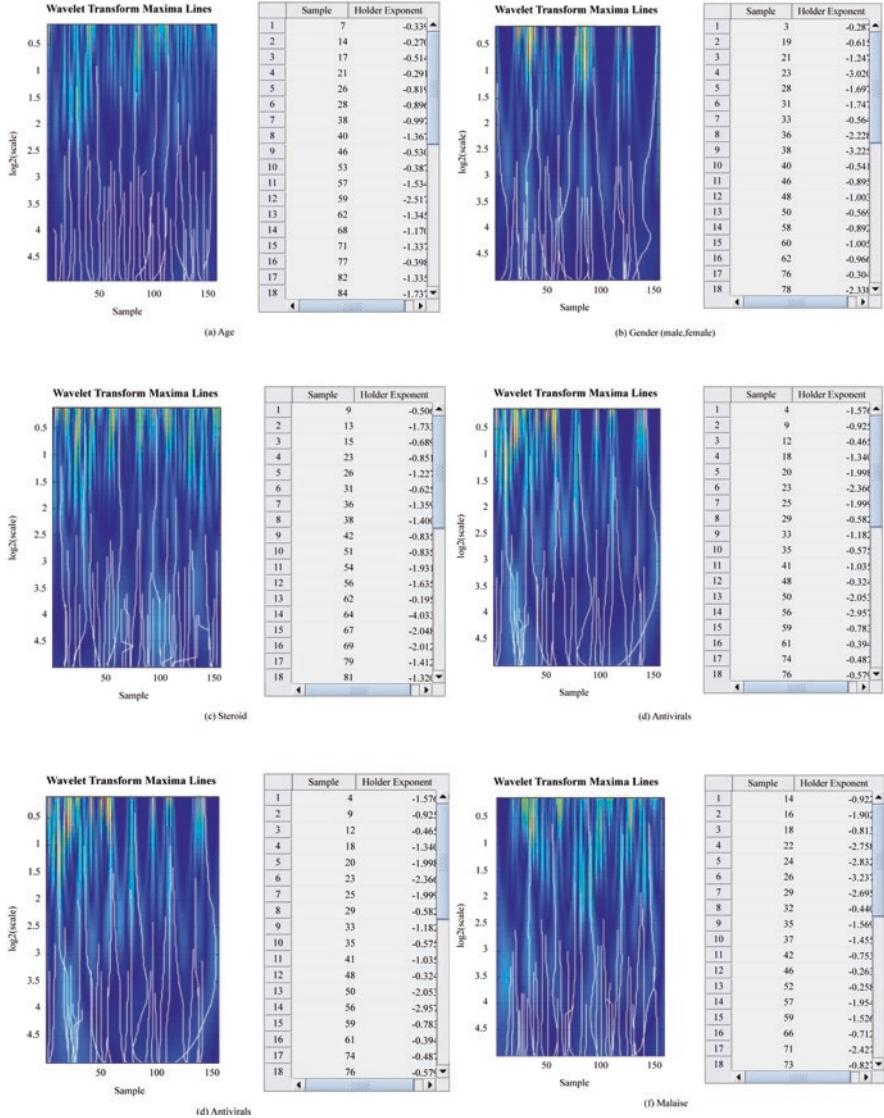
$$|Wf(u, s_0)| < |Wf(u_0, s_0)|$$

in which  $x$  is in the left or right neighbourhood of  $x_0$ . In a case when  $x$  is in opposite neighbourhood of  $x_0$ , the relevant definition is as in [23, 28, 29]

$$|Wf(u, s_0)| \leq |Wf(u_0, s_0)|$$

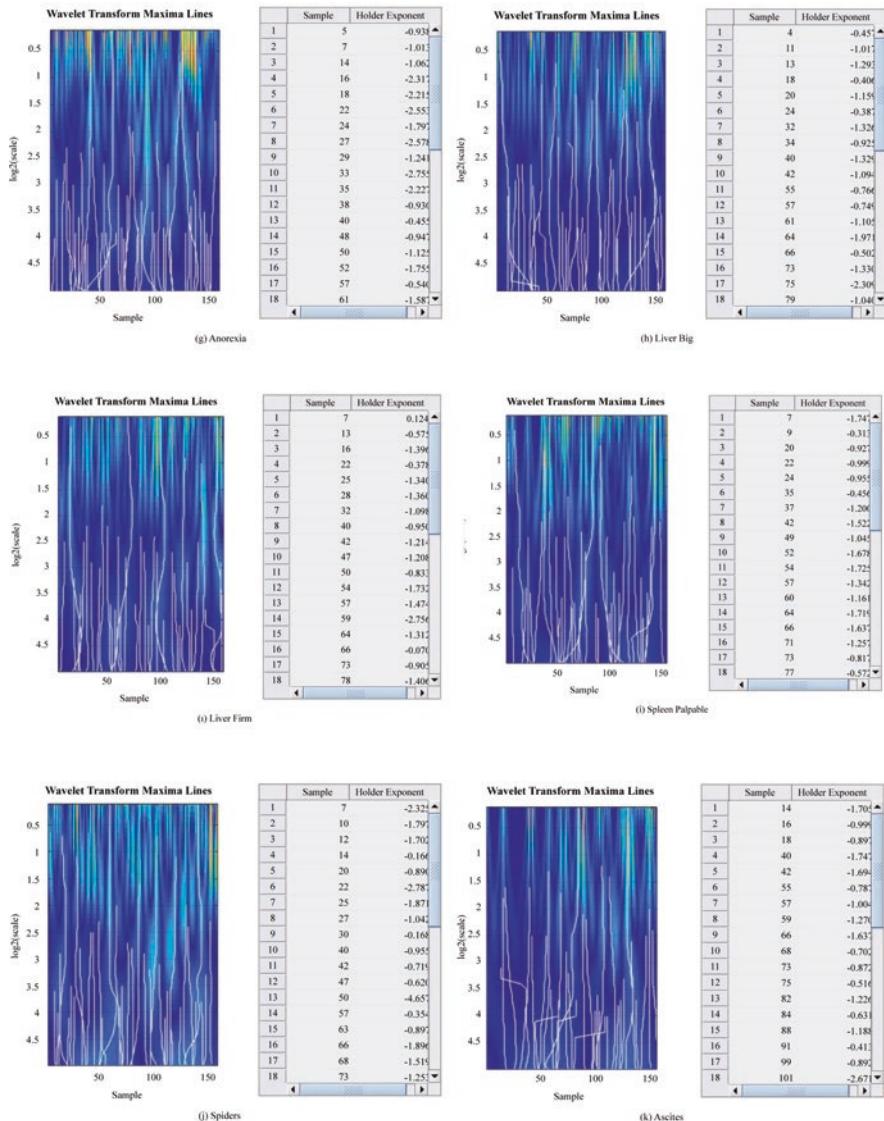
The WTMM to find additional maxima recurs for values in that particular scale. The WTMM goes on through finer scales, meanwhile checking if the maxima align between scales or not. Should a maximum converge towards the finest scale, it will be a true maximum, indicating a singularity at the relevant point.

Accordingly, wtmm\_hepatitis dataset was obtained the LE and plot the modulus maxima in Fig. 3. The LE at samples 1 and 150 are very close to the values which

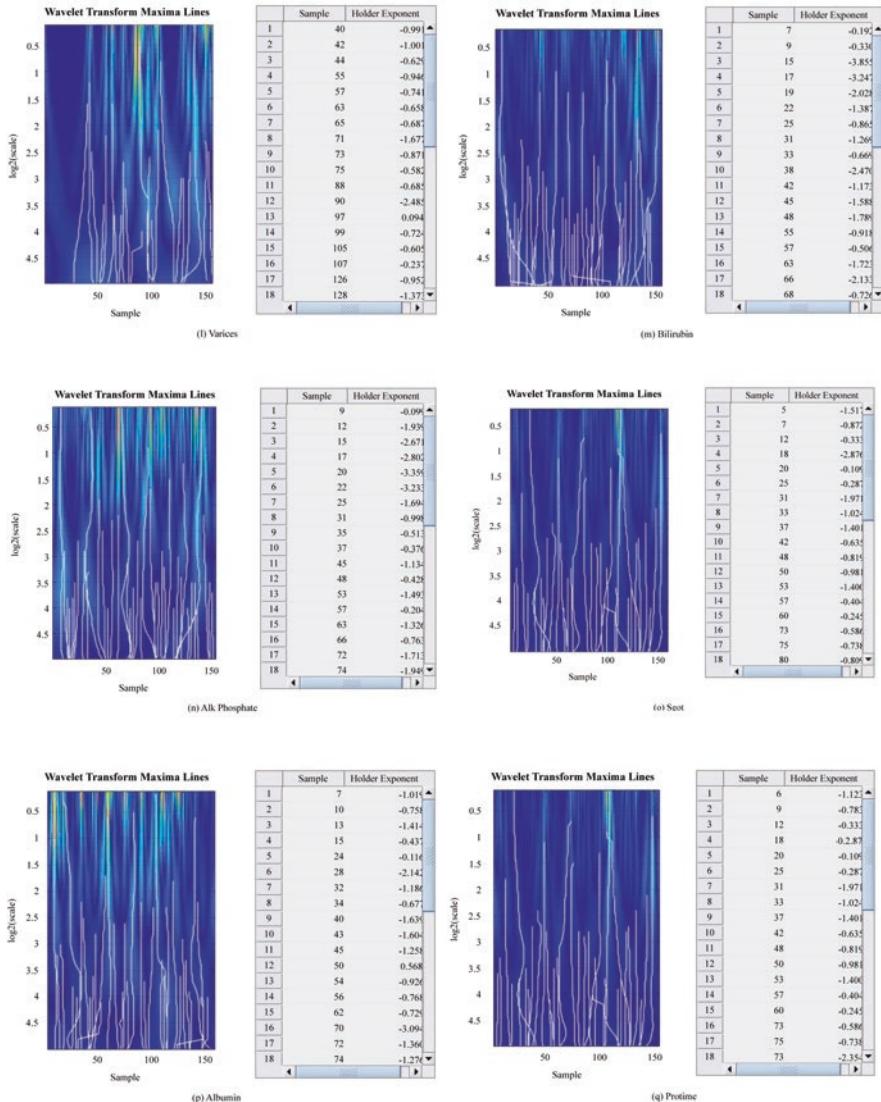


**Fig. 3** (a) log-log characteristics of Modulus Maxima for the attributes in the hepatitis\_dataset. (b) log-log characteristics of Modulus Maxima for the attributes in the hepatitis\_dataset. (c) log-log characteristics of Modulus Maxima for the attributes in the hepatitis\_dataset. (d) log-log characteristics of Modulus Maxima for the attributes in the hepatitis\_dataset

are stated in the strongest regularity data. For each of the 19 attributes (concerning 150 patients with hepatitis), WTTM has been applied, details of their application results are depicted in Fig. 3.

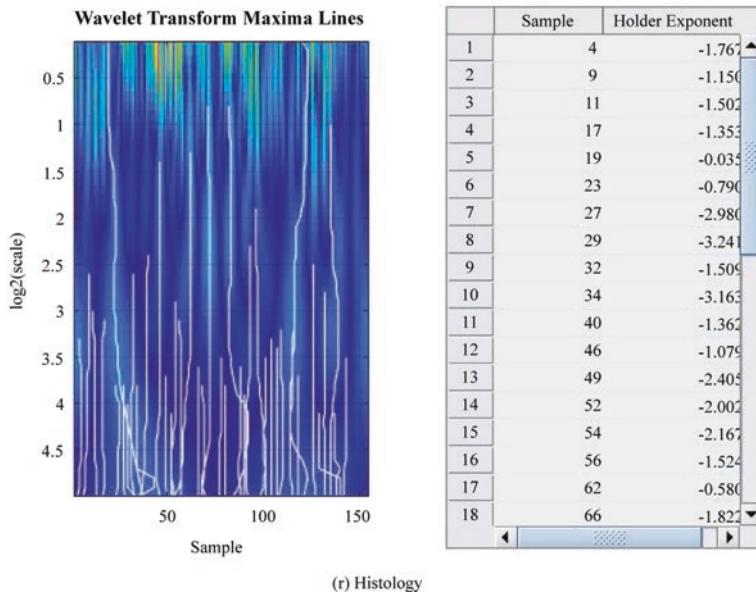
**Fig. 3** (continued)

As a result of the analyses conducted, among 19 attributes, 15 of them have been found to be significant. These 15 attributes are: Age, Gender, Steroid, Antivirals, Malaise, Anorexia, Liver Big, Liver Firm, Spleen Palpable, Spiders, Ascites, Varices, Alk Phosphate, Albumin, and Histology.

**Fig. 3** (continued)

### 3.2 Application of 1D Continuous Wavelet Transform on the Hepatitis Datasets

1D Continuous Wavelet transform of function was applied on the hepatitis\_dataset ( $150 \times 19$ ) and wtmm\_hepatitis dataset ( $150 \times 15$ ) and the results of these applications are presented Fig. 4.



(r) Histology

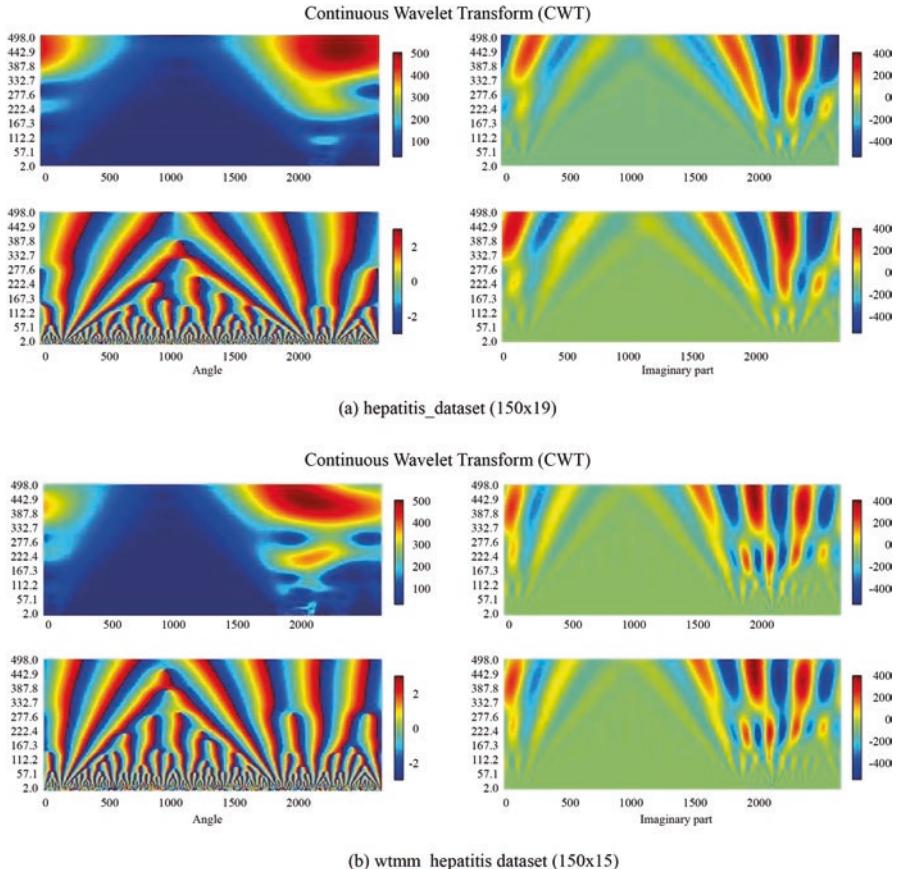
**Fig. 3** (continued)

Continuous Wavelet transform of function  $f(x)$  presented in Fig. 4a, the discontinuity is seen in a clear way from the fact that  $|Wf(s, x)|$  remains constant approximately over a large range of scales, in the abscissa neighbourhood, within the range of 100–2200. A negative Lipschitz exponent corresponds to sharp irregularities in which the wavelet transform modulus goes up in value at fine scales. Wavelet transform maxima increase in a proportional way to  $s-1$ , over a large range of scales in the equivalent neighbourhood.

Continuous Wavelet transform of function  $f(x)$  shown in Fig. 4b, the discontinuity is seen in a clear way from the fact that  $|Wf(s, x)|$  remains constant approximately over a large range of scales, in the abscissa neighbourhood within the range 200–1900. A negative Lipschitz exponent corresponds to sharp irregularities in which the wavelet transform modulus increases at fine scales. The wavelet transform maxima increase in a proportional way to  $s-1$ , over a large range of scales in the equivalent neighbourhood.

### 3.3 Application of FFBP Algorithm

In this section of the study, FFBP algorithm, which is one of the important ANN algorithms, has been applied on hepatitis\_dataset ( $150 \times 19$ ) and wtmn\_hepatitis dataset ( $150 \times 15$ ) for the classification of hepatitis in terms of die/live aspect. The common parameters that yield the overall accuracy results (sensitivity, specificity, accuracy rate) in the application are presented in Table 3.

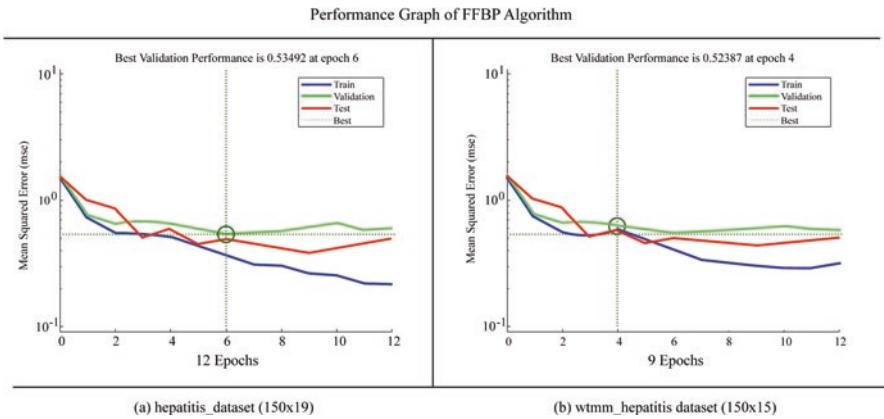


**Fig. 4** 1D Continuous Wavelet Transform of function (a) hepatitis\_dataset ( $150 \times 19$ ) (b) wtmm\_hepatitis dataset ( $150 \times 15$ )

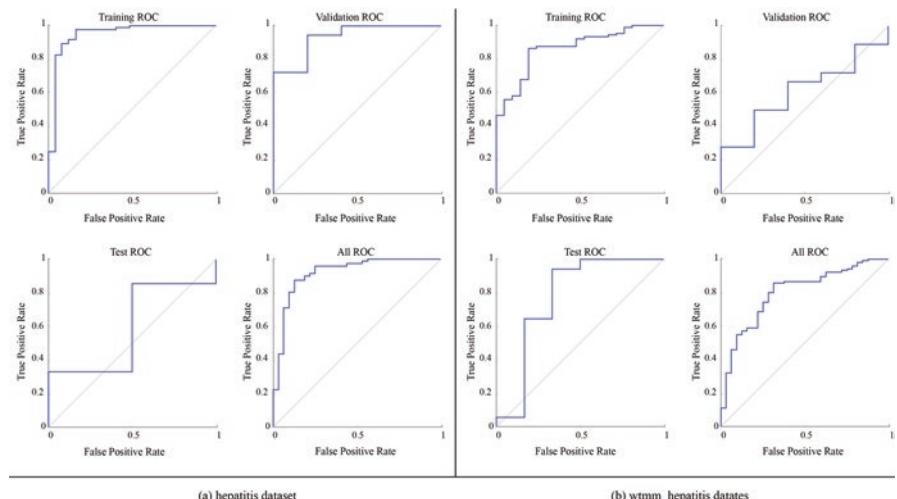
The performance graph that is derived from the classification of Hepatitis (die/live aspect) for hepatitis\_dataset ( $150 \times 19$ ) and wtmm\_hepatitis dataset ( $150 \times 15$ ) by FFBP algorithm is provided in Fig. 5.

Figure 5 provides the results of Mean Squared Error as obtained from the FFBP algorithm application both on the hepatitis\_dataset ( $150 \times 19$ ) and on the wtmm\_hepatitis dataset ( $150 \times 15$ ).

The best validation performance obtained from the FFBP algorithm training procedure for the hepatitis\_dataset ( $150 \times 19$ ) is 0.53492 (see Fig. 5a) and for wtmm\_hepatitis dataset ( $150 \times 15$ ), the result is 0.52387 (see Fig. 5b). Based on the best validation performance result (see Fig. 5), the accuracy rate for wtmm\_hepatitis dataset ( $150 \times 15$ ) has proven to be higher compared to that of hepatitis\_dataset ( $150 \times 19$ ) in terms of classification.



**Fig. 5** Performance graph of FFBP algorithm for (a) hepatitis\_dataset ( $150 \times 19$ ) (b) wtmm\_hepatitis dataset ( $150 \times 15$ )



**Fig. 6** Results of the Training ROC, Validation ROC analysis Test ROC analysis and all ROC analysis for (a) the hepatitis\_dataset (b) wtmm\_hepatitis dataset

The results of the Training ROC, Validation ROC analysis Test ROC analysis and All ROC analysis as obtained from the classification of Hepatitis as per (die/live) based on the application of FFBP algorithm on hepatitis\_dataset ( $150 \times 19$ ) and wtmm\_hepatitis dataset ( $150 \times 15$ ) are presented in Fig. 6.

### 3.4 The FFBP Algorithm Classification Results

The overall accuracy results (sensitivity, specificity, accuracy rate) regarding the application of the FFBP algorithm on the hepatitis\_dataset and wtmm\_hepatitis dataset for the classification of hepatitis virus as per (die/live) application are presented in Table 4.

For the classification of the hepatitis dataset, the ANN algorithm (FFBP) was applied on the relevant data sets, which are the hepatitis\_dataset and wtmm\_hepatitis dataset. Moreover, the confusion matrices of the most accurate rates obtained (see Table 4) are provided in Fig. 7.

Figure 7 shows the highest accuracy rate with FFBP application in the confusion matrix for the hepatitis\_dataset and for the wtmm\_hepatitis dataset.

Overall, the aim of the study is to make the classification of hepatitis as per die/live. A higher level of accuracy regarding classification of the disease will contribute a great deal to the life quality of the patients since early and correct diagnosis will be highly significant. For this, the determination of the most significant attributes in the hepatitis dataset (see Table 1) makes a difference. For the accurate classification of hepatitis, in the study the more significant attributes in the dataset have been determined in line with two approaches: the first one is multifractal analysis which includes the following steps:

**Table 4** Overall accuracy results of FFBP application on the hepatitis datasets

Hepatitis datasets		Sensitivity (%)	Specificity (%)	Accuracy rate (%)
hepatitis_dataset		98.2	99.9	79
wtmm_ hepatitis dataset		99.1	85	<b>82</b>



**Fig. 7** The confusion matrices for (a) hepatitis\_dataset (b) wtmm\_hepatitis dataset

- (a) The multifractal Analysis (Multifractal spectrum estimation) was used and applied on the hepatitis\_dataset ( $150 \times 19$ ) so that significant and self-similar attributes could be identified. In this way, the dataset named wtmm\_hepatitis dataset ( $150 \times 15$ ) was generated, the results of which can be seen in Sect. 3.1.
- (b) Wavelet analysis (Continuous Wavelet Transform) was applied on the hepatitis\_data set and also wtmm\_hepatitis dataset. In this stage of applying on both datasets, the datasets were compared visually based on the obtaining of the analyses, which can be seen in detail in Sect. 3.2.

The steps of all these procedures mentioned above were applied on the hepatitis data set ( $150 \times 19$ ) (see Table 1), and from the applications of steps (a) and (b), the significant attributes have been obtained. The significant and self-similar attributes that belong to wtmm\_hepatitis dataset ( $150 \times 15$ ) are: Age, Gender, Steroid, Antivirals, Malaise, Anorexia, Liver Big, Liver Firm, Spleen Palpable, Spiders, Ascites, Varices, Alk Phosphate, Albumin, and Histology.

In order to evaluate the self-similar and significant attributes' role in the accuracy for the classification of the hepatitis data, a second stage was performed, and this second stage of the study is related to the Feed Forward Back Propagation (FFBP) algorithm. This includes the following steps:

Firstly, the FFBP algorithm was applied both on the hepatitis\_dataset ( $150 \times 19$ ) and wtmm\_hepatitis dataset ( $150 \times 15$ ) for the purpose of identifying the classification as per survival.

Secondly, the FFBP algorithm was used for the classification of die/live aspect. This classification ensured the calculation of overall accuracies based on sensitivity, specificity, accuracy rates; hence, comparative analyses were performed. Hepatitis\_dataset ( $150 \times 19$ ) yielded an accuracy rate of 79% and wtmm\_hepatitis dataset ( $150 \times 15$ ) yielded a rate of 82%. This obviously has shown that the wtmm\_hepatitis dataset which comprised of self-similar and significant attributes is better at classification in terms of accuracy compared to the larger dataset, namely hepatitis\_dataset.

## 4 Conclusion

Fractal and multifractal geometries are well-known mathematical concepts widely applied in various fields in science to attain accurate solutions. This subject matter is of interest for many researchers in the community. The main contribution of this study is that it has provided an alternative direction with the use of two approaches regarding hepatitis virus. The use of the Wavelet Transform Modulus Maxima method and application of ANN for the classification of die/live for hepatitis provide these two approaches used together. In this respect, the use of these two approaches has been done for the first time in the literature, when compared with the other works (3, 4, 5, 13, 14). Thus, this study provides contribution to the field. The dataset at stake is comprised a total of 150 hepatitis patients (whose attributes were analysed), and the analyses have been conducted in a comparative manner. This can

be regarded as another contribution of the study. There are two datasets in this study: hepatitis\_dataset (150x19) and wtmm\_hepatitis dataset (150x15). The latter dataset is comprised of self-similar and significant attributes as a result of the WTMM application. Moreover, ANN was applied on both hepatitis\_dataset and wtmm\_hepatitis dataset, thus, classification of die/live has been done for hepatitis through accuracy rates. The results show that the accuracy rate for wtmm\_hepatitis dataset (150x15) has proven to be higher than that of hepatitis\_dataset (150x19) in terms of classification. Consequently, a higher level of accuracy concerning classification of the disease will contribute a great deal to the life quality of the patients as accurate diagnosis in due time will be highly significant. The results of this study pinpoint the importance of selecting self-similar and significant attributes for the classification of a disease in medical field which might result in deaths. The problem examined in this paper may be solved via alternative methodologies in the future. As a result of the models proposed in this study, it has attempted to provide a new direction along with its methodology to deal with vital situations efficiently in the related fields .

**Acknowledgement** The author is genuinely grateful to Professor Carlo Cattani for his academic support and guidance.

## References

1. Venkatakrishnan P, Sangeetha S. Singularity detection in human EEG signal using wavelet leaders. *Biomed Signal Process Control.* 2014;13:282–94. <https://doi.org/10.1016/j.bspc.2014.06.002>.
2. Puckovs A, Matvejevs A. Wavelet transform modulus maxima approach for world stock index multifractal analysis. *University. Inf Technol Manage Sci.* 2013;15(1):76–86. <https://doi.org/10.2478/v10313-012-0016-5>.
3. Yasin H, Jilani TA, Danish M. Hepatitis-C classification using data mining techniques. *Int J Comput Appl.* 2011;24(3):1–6. ISSN 0975-8887.
4. Nilashi M, Ahmadi H, Shahmoradi L, Ibrahim O, Akbari E. A predictive method for hepatitis disease diagnosis using ensembles of neuro-fuzzy technique. *J Infect Public Health.* 2019;12(1):13–20. <https://doi.org/10.1016/j.jiph.2018.09.009>.
5. Almuneef MA, Memish ZA, Balkhy HH, Qahtani M, Alotaibi B, Hajeer A, et al. Epidemiologic shift in the prevalence of Hepatitis A virus in Saudi Arabia: a case for routine Hepatitis A vaccination. *Vaccine.* 2006;24(27):5599–603. <https://doi.org/10.1016/j.vaccine.2006.04.038>.
6. Al-Thaqafy MS, Balkhy HH, Memish Z, Makhdum YM, Ibrahim A, Al-Amri A, et al. Hepatitis B virus among Saudi National guard personnel: seroprevalence and risk of exposure. *J Infect Public Health.* 2013;6(4):237–45. <https://doi.org/10.1016/j.jiph.2012.12.006>.
7. Al-Thaqafy MS, Balkhy HH, Memish Z, Makhdum YM, Ibrahim A, Al-Amri A, Al-Thaqafi A. Improvement of the low knowledge, attitude and practice of hepatitis B virus infection among Saudi National Guard personnel after educational intervention. *BMC Res Notes.* 2012;5(1):597. <https://doi.org/10.1186/1756-0500-5-597>.
8. Shepard CW, Finelli L, Alter MJ. Global epidemiology of hepatitis C virus infection. *Lancet Infect Dis.* 2005;5(9):558–67. [https://doi.org/10.1016/S1473-3099\(05\)70216-4](https://doi.org/10.1016/S1473-3099(05)70216-4).
9. Wu JC, Chen TZ, Huang YS, Yen FS, Ting LT, Sheng WY, et al. Natural history of hepatitis D viral superinfection: significance of viremia detected by polymerasechain reaction. *Gastroenterology.* 1995;108(3):796–802.

10. Haagsma EB, van den Berg AP, Porte RJ, Benne CA, Vennema H, Reimerink JH, et al. Chronic hepatitis E virus infection in liver transplant recipients. *Liver Transp.* 2008;14(4):547–53. <https://doi.org/10.1002/lt.21480>.
11. Mistler LA, Brunette MF, Marsh BJ, Vidaver RM, Luckoor R, Rosenberg SD. Hepatitis C treatment for people with severe mental illness. *Psychosomatics.* 2006;47(2):93–107. <https://doi.org/10.1176/appi.psy.47.2.93>.
12. Metwally NF, AbuSharekh EK, Abu-Naser SS. Diagnosis of hepatitis virus using artificial neural network. *Int J Acad Pedagogical Res.* 2018;2(1):1–7. ISSN: 2000-004X.
13. Jilani TA, Yasin H, Yasin MM. PCA-ANN for classification of Hepatitis-C patients. *Int J Comput Appl.* 2011;14(7):1–6. ISSN: 0975-8887.
14. Priya S, Manavalan R. Optimum parameters selection using ACO R algorithm to improve the classification performance of weighted extreme learning machine for hepatitis disease dataset. In: 2018 International Conference on Inventive Research in Computing Applications (ICIRCA). IEEE; 2018. p. 986–91. <https://doi.org/10.1109/ICIRCA.2018.8597232>.
15. Karaca Y, Cattani C. Clustering multiple sclerosis subgroups with multifractal methods and self-organizing map algorithm. *Fractals.* 2017;25(4):1740001. <https://doi.org/10.1142/S0218348X17400011>.
16. Karaca Y, Cattani C, Karabudak R. ANN classification of MS subgroups with diffusion limited aggregation. In: Gervasi O, et al., editors. International Conference on Computational Science and Its Applications. ICCSA 2018. Lecture notes in computer science, vol. 10961. Cham: Springer; 2018. p. 121–36. [https://doi.org/10.1007/978-3-319-95165-2\\_9](https://doi.org/10.1007/978-3-319-95165-2_9).
17. Arulmurugan R, Anandakumar H. Early detection of lung cancer using wavelet feature descriptor and feed forward back propagation neural networks classifier. In: Hemanth D, Smys S, editors. Lecture notes in computational vision and biomechanics, vol. 28. Cham: Springer; 2018. p. 103–10. [https://doi.org/10.1007/978-3-319-71767-8\\_9](https://doi.org/10.1007/978-3-319-71767-8_9).
18. Karaca Y, Aslan Z, Siddiqi AH. 1D Wavelet and partial correlation application for MS subgroup diagnostic classification. Classification. In: Manchanda P, Lozi R, Siddiqi A, editors. Industrial mathematics and complex systems. Industrial and applied mathematics. Singapore: Springer; 2017. p. 171–86. [doi.org/10.1007/978-981-10-3758-0\\_11](https://doi.org/10.1007/978-981-10-3758-0_11).
19. Parey A, Singh A. Gearbox fault diagnosis using acoustic signals, continuous wavelet transform and adaptive neuro-fuzzy inference system. *Appl Acoust.* 2019;147:133–40. <https://doi.org/10.1016/j.apacoust.2018.10.013>.
20. Karaca Y, Aslan Z, Cattani C, Galletta D, Zhang Y. Rank determination of mental functions by 1D wavelets and partial correlation. *J Med Syst.* 2017;41(1):1–10. <https://doi.org/10.1007/s10916-016-0606-2>.
21. Karaca Y, Sertbaş A, Bayrak Ş. Classification of erythematous – squamous skin diseases through SVM kernels and identification of features with 1-D continuous wavelet coefficient. In: Gervasi O, et al., editors. International Conference on Computational Science and Its Applications. ICCSA 2018. Lecture notes in computer science, vol. 10961. Cham.: Springer; 2018. p. 107–20. [https://doi.org/10.1007/978-3-319-95165-2\\_8](https://doi.org/10.1007/978-3-319-95165-2_8).
22. Cattani C. Fractals and hidden symmetries in DNA. *Math Probl Eng.* 2010;2010:31. 507056. <https://doi.org/10.1155/2010/507056>.
23. Venkatakrishnan P, Sangeetha S, Sundar M. Measurement of Lipschitz exponent (LE) using wavelet transform modulus maxima (WTMM). *Int J Sci Eng Res.* 2012;3:6. ISSN 2229-5518.
24. Izadi H, Innanen K, Lamoureux MP. Continuous wavelet transforms and Lipschitz exponents as a means for analysing seismic data. *CREWES Res Rep.* 2011;23:1–8.
25. Legarreta IR, Addison PS, Grubb N, Clegg GR, Robertson CE, Fox KAA, Watson JN. R-wave detection using continuous wavelet modulus maxima. *Comput Cardiol.* 2003;1(30):565–8. <https://doi.org/10.1109/CIC.2003.1291218>.
26. Blake CL, Merz CJ. UCI repository of machine learning databases. 1996. Available from: <https://archive.ics.uci.edu/ml/index.php>. Accessed 2 Jan 2019.
27. Mallat S. A wavelet tour of data processing. USA: Elsevier, Academic Press; 1999.
28. Peng ZK, Chu FL, Peter WT. Singularity analysis of the vibration signals by means of wavelet modulus maximal method. *Mech Syst Signal Process.* 2007;21(2):780–94. <https://doi.org/10.1016/j.ymssp.2005.12.005>.

29. Mallat S, Hwang WL. Singularity detection and processing with wavelets. *IEEE Trans Inf Theory*. 1992;38(2):617–43. <https://doi.org/10.1109/18.119727>.
30. Vrscay ER. A generalized class of fractal-wavelet transforms for image representation and compression. *Can J Electr Comput Eng*. 1998;23(1–2):69–83. <https://doi.org/10.1109/CJECE.1998.7102047>.
31. Tu GJ, Karstoft H. Logarithmic dyadic wavelet transform with its applications in edge detection and reconstruction. *Appl Soft Comput*. 2015;26:193–201. <https://doi.org/10.1016/j.asoc.2014.09.044>.
32. Yuan YT, Li BF, Ma H, Lin J. Ring-projection-wavelet-fractal signatures: a novel approach to feature extraction. *IEEE Trans Circuits Syst II: Analog Digital Signal Process*. 1998;45(8):1130–4. <https://doi.org/10.1109/82.718824>.
33. Jaffard S, Lashermes B, Abry P. Wavelet leaders in multifractal analysis. In: Qian T, Vai MI, Xu Y, editors. *Wavelet analysis and applications. Applied and numerical harmonic analysis*. Birkhäuser Basel; 2006. p. 201–46. [https://doi.org/10.1007/978-3-7643-7778-6\\_17](https://doi.org/10.1007/978-3-7643-7778-6_17).
34. Bujanovic T, Abdel-Qader I. On wavelet transform general Modulus maxima metric for singularity classification in mammograms. *Open J Med Imaging*. 2013;3(1):17. <https://doi.org/10.4236/ojmi.2013.31004>.
35. Karaca Y, Cattani C. Computational methods for data analysis. De Gruyter; 2018. ISBN: 978-3-11-049636-9.
36. Hagan MT, Menhaj MB. Training feedforward networks with the Marquardt algorithm. *IEEE Trans Neural Netw*. 1994;5(6):989–93. <https://doi.org/10.1109/72.329697>.
37. Saeedi E, Hossain MS, Kong Y. Feed-forward back-propagation neural networks in side-channel information characterisation. *J Circuits Syst Comput*. 2019;28(1):1950003. <https://doi.org/10.1142/S0218126619500038>.
38. The MathWorks. MATLAB (R2018b). Natick: The MathWorks, Inc.; 2018.

# Computational Coarse Protein Modeling of HIV-1 Sequences Using Evolutionary Search Algorithm



Sandhya Parasnath Dubey and Seetharaman Balaji

**Abstract** There are extensive research works on HIV-1 genome and its encoded proteins. The genome comprises of nine genes that code for at least 15 proteins. Although these sequences are available to the public, the structure information is incomplete. The structure information is vital to understand the pathogenesis as well as for preventive measures. There are experimental efforts such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy to solve the structures of HIV-1 proteins. However, there are some limitations with these methods, for instance, membrane associated proteins are difficult to crystallize and NMR has size limitation. Moreover, these methods are very expensive and time consuming. Hence, computational methods can be of use. This chapter deals with a computational protein structure prediction (PSP) based on the primary structure. One of the popular approaches for modeling coarse protein structure is Dill's HP-model. This work presents a revised HP model for HIV-1 proteins. These proteins were modeled over 2D square lattice with optimal conformation using evolutionary programming. The modeled conformations were also evaluated against the experimental structure.

**Keywords** Human immune virus-1 · HIV-1 · NP-complete problem · Computational protein structure prediction · 2D square lattice · Protein HP-model · Protein folding · Coarse protein modeling · *Ab initio/de novo* protein modeling · Evolutionary search algorithm · Evolutionary programming

---

S. P. Dubey · S. Balaji (✉)

Department of Computer Applications, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India

Department of Biotechnology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India

### Key Concepts

There is a huge information gap between the number of protein sequences and structures in the public databases. This raises the concern in understanding the structure of proteins as well as their function or pathogenicity. There are experimental efforts to solve the structures of HIV-1 proteins. However, there are some limitations with these methods that are very expensive and time consuming. This chapter demonstrates protein structure prediction of HIV-1 sequences using evolutionary search algorithm to understand the coarse structure and folding of HIV-1 proteins. These efforts can accelerate high-throughput structure prediction with a reasonable accuracy.

## 1 Introduction

Human immune virus 1 (HIV-1) has been a major public health threat since its revelation in the United States in 1981 [1]. As per the global health observatory data of WHO there are about 35 million people died due to HIV infections and more than 70 million people are infected with HIV virus. Globally, 36.9 million people were living with HIV at the end of 2017 [2]. There are extensive research on HIV-1 genome and its encoded proteins right from its discovery [3]. The genome comprises of nine genes that code for 15 proteins [4]. The sequences are publicly made available [5] so as to contribute to the ongoing studies on understanding the pathogenesis and prevention of HIV-1. There are experimental efforts such as cryo-electron microscopy, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy to solve the structures of HIV-1 proteins [6]. However, there are some limitations with these methods, for instance, membrane associated proteins are difficult to crystallize, very low signal-to-noise ratio in cryo-electron microscopy and size limitations of NMR [7]. Moreover, these methods are very expensive and time consuming.

There are ~0.2% of solved structures of the known protein sequences are available in protein data bank (PDB) [8, 9]. With these limitations it is difficult to determine the protein structures (experimentally) of all deposited sequences. As more and more sensitive next-generation sequencing techniques are available, the protein sequence-structure gap started widening [10]. Hence, computational methods can be of use in predicting experimentally unsolved protein folds [11] to bridge the information gap between sequence and structure data. Computationally, there are three major protein structure prediction methods viz., *ab initio/de novo* protein modeling, homology-based modeling, and fold recognition/threading [12]. The *ab initio* technique is solely based on the amino acids constituents of proteins, whereas the other two methods requires templates (homologous crystal structures) for modeling [13]. The *ab initio* method attempts to predict the near native structure of proteins purely based on the physicochemical properties of amino acids (AAs).

One of the popular approach for modeling coarse protein structure is Dill's HP-model [14]. The method uses a potential scoring function based on AAs properties such as hydrophobicity (H) and polarity (P), abbreviated as 'HP'-model. This

method has been modified by many researchers for better prediction accuracy [15]. In this chapter, we present an application of revised HP model [16] to predict HIV-1 proteins through an evolutionary programming. Evolutionary programming is global optimization algorithm and inspired by the theory of evolution. Details on Evolutionary programming and its application for PSP problem are available in various research articles [17–20]. There are about 15 proteins available in PDB repository for HIV-1. In this chapter, eight of them such as virion infectivity factor, protease, matrix protein, negative regulatory factor protein, intergrase, virus protein U, capsid and envelope proteins. All these proteins were modeled over a two-dimensional (2D) square lattice using evolutionary search algorithm to obtain optimal conformation. These conformations were compared with the experimentally solved structures available in PDB.

## 2 Materials and Methods

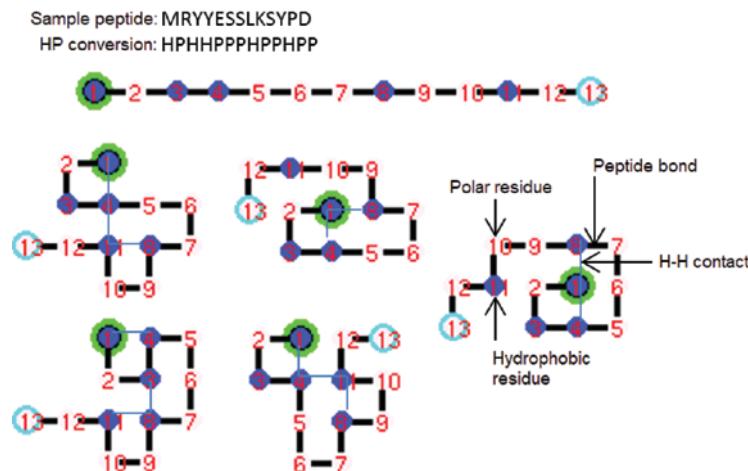
Hydrophobic-polar (HP) model is one of the most explored model for the *de novo* PSP [14] problem. The HP model is an abstract representation of a coarse PSP. In the HP model, amino acids are classified into two groups hydrophobic (H)  $H \in \{I, L, V, F, W, M, C, A, Y, G\}$  and polar (P),  $P \in \{P, S, T, R, H, K, N, Q, D, E\}$  [21]. The PSP problem is addressed in three stages (i) modeling of structure, (ii) potential scoring function, and (iii) development of search algorithm.

### 2.1 Modeling of Proteins

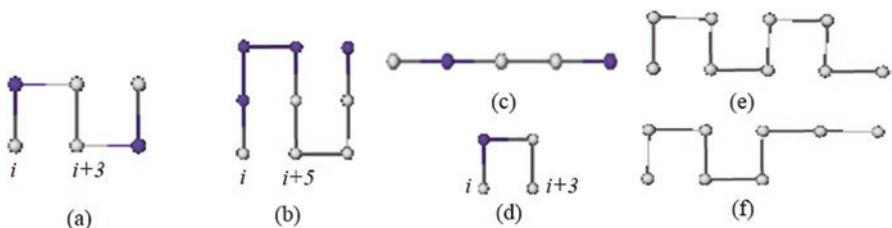
Originally HP-model was proposed to fold sequence  $\text{Seq } S \in \{\text{HP}\}^l$ ;  $l$  is length of the sequence (i.e. number of amino acid present in the protein sequence) over a 2D-square lattice. Folding is constrained such that each lattice point is occupied by single residue and the consecutive residue of HP sequence placed at the unit distance (i.e. they are one of the nearest neighbor), such that the conformation is a non-cyclic graph (as depicted in Fig. 1).

In HP-model, protein structure is represented using a two or three-dimensional lattice. Bond length and angle are limited to the properties of the lattice. As in the case of 2D-square lattice, the bond length is fixed to a unit value for every amino acid and the bond angle is restricted to one out of four values i.e., 0, 90, 180 and 270 because a square repeats itself at 90 degree intervals.

In the proposed modeling, the folding pattern of HP string is based on amino acid propensity values that favors a three-state secondary structure prediction [22]. The secondary structure constraints were incorporated in the algorithm that restricts the string folding either into ‘S’-like pattern or a ‘straight chain’. The former represents the  $\alpha$ -helix and the latter a  $\beta$ -sheet. The possibility of  $\alpha$ -helix is predicted based on the interactions (Fig. 2) between  $i$  and  $i + 3$  residues (Fig. 2a) or between  $i$  and  $i + 5$  residues (Fig. 2b).



**Fig. 1** Folding in HP model: A sample peptide and its corresponding HP sequence with possible conformations. Hydrophobic amino acids are represented as blue nodes whereas the polar ones are white. The covalent bond (peptide bond) is depicted by a thick black line and the H-H contacts by a thin blue line. Number represents the residue position in the sequence



**Fig. 2** Basic secondary structure notion on 2D square lattice (a, b) helix with hydrophobic interaction between  $i, i + 3$  and  $i, i + 5$  residues, respectively (c) sheet (d) turn (e) helix (f) turn

On the contrary, ‘sheet’ does not take part in such interactions. Hence, it appears as a ‘straight chain’ (Fig. 2c). In the case of turns, experimentally ‘ $\beta$ -turn’ class is predominant due to interaction between  $i$  and  $i + 3$  residues, the same is used in the proposed approach. It appears like the alphabet ‘C’ or ‘U’ (Fig. 2d). There are cases where two consecutive turns in opposite directions can appear as ‘S’, this cannot be mistaken as  $\alpha$ -helix. To overcome this conflicting situation the ‘S’-like pattern can be considered as a ‘ $\alpha$ -helix’, only if the consecutive residues are perpendicular as shown in Fig. 2e, whereas if the move proceeds in the same direction then it is considered as ‘turn’ (Fig. 2f).

## 2.2 Scoring Function

Lattice model is grounded on the concept of thermodynamic free energy, which explains that the native state of proteins need a minimum free energy [23]. Thus, the folded protein conformations are evaluated using the free energy value. In the lattice

$$\begin{array}{c}
 \textbf{a} & \textbf{b} & \textbf{c} \\
 E = \begin{bmatrix} * & H & P \\ H & -1 & 0 \\ P & 0 & 0 \end{bmatrix} & E' = \begin{bmatrix} * & H & P \\ H & -2 & +1 \\ P & +1 & +1 \end{bmatrix} & E^r = \begin{bmatrix} * & H & h & P & N \\ H & -4 & -2 & 0 & 0 \\ h & -2 & -4 & 0 & 0 \\ P & 0 & 0 & 1 & -1 \\ N & 0 & (c) & -1 & 1 \end{bmatrix}
 \end{array}$$

**Fig. 3** Energy matrix (a) HP, (b) Shifted HP model, and (c) HhPN model

model free energy is quantified using interaction among amino acids. Two amino acids interact if they are topological neighbor and are not connected by the peptide bond, the existence of peptide bond is restricted between consecutive amino acids of protein sequences. There are four types of interaction presents in the lattice model namely H-H, P-P, H-P and P-H. In the HP model, the value for the H-H interaction is  $-1$  whereas other three, P-P, H-P and P-H have value of  $0$  involving one attractive interaction (H-H) and three neutral interaction (P-P, H-P, and P-H). Energy matrix for the HP model is shown in Fig. 3a.

Although there are various energy matrix such as MJ-model [24], BMF-model [25], HOP-model [26], and shifted HP model [27] to describe different types of interactions, HP model and its extensions made it more predominant in the research community over these elaborated models. Shifted HP-model [27] is extension of HP model involving both attraction and repulsive interaction. Unlike HP-model, protein conformations in shifted HP model are not maximally compact. Shifted HP-model derived from the concept of existence of potential binding site, uses for investigating the ligand binding problem [28] which enhance the biological relevance of this modeling method. The shifted energy matrix is given in Fig. 3b.

HP and shifted energy model were only based on the hydrophobic and polar properties of amino acids. To make the energy function more flexible with other amino acid properties, we have considered the revised energy function to assess the folded conformation. The proposed energy function derived by considering the additional properties of amino acid such as carbon content, disulfide bond, hydrophobicity scale, codon frequency and BLOSUM62 similarity matrix [29]. The derived energy function is termed as HhPN energy function with following classification of amino acids, highly hydrophobic  $H \in \{I, L, V, F, W, M, C\}$ , partial hydrophobic  $h \in \{A, Y, G, P, S\}$ , positive charged amino acids  $P \in \{T, R, H, K, N, Q\}$ , and negative charged amino acids  $N \in \{D, E\}$ . The energy matrix associated with the HhPN model is depicted in Fig. 3c.

### 2.3 Search Algorithm

Although lattice model has reduced the complexity of the original PSP problem [30], it exhibits an inordinately huge search space [31]. For instance, the number of SAW conformation for sequence of length  $l$  is directly proportional to  $\mu^l$  [32], where  $\mu$  is coordination number of lattice. It has been proved that the number of possible SAW conformations for small sequence of length 30 is more than the element of the universe

[33]. Hence, it has been proven that the HP model based PSP problem is ‘NP-complete problem’ [34] implying that the impracticality of polynomial time algorithm. To deal with the ‘NP-complete’ issue of the proposed model, we are using evolutionary programming to obtain the optimal conformation. Furthermore, the energy landscape of protein structures have many crest and trough which results into a local optimal solution [35]. To overcome the trap in the local optima, we have used a hybrid search algorithm. This algorithm is comprised of Hill climbing and evolutionary search algorithm [36]. The details on this algorithm and its implementation method are available in our previous work [36]. Pseudocode for used evolutionary algorithm is given in Algorithm 1.

## 2.4 Mathematical Formulation

Mathematically, PSP is modeled as minimization problem where the objective is to bring most of the H residue at the unit distance with other H residues. Such H-H pair contributes to the stability of the structure by reducing the free energy [14], i.e., if two H residues are at unit distance and are not consecutive in the input string then they contribute for free energy (designated by a value of  $-1$ ).

**Algorithm 1** The Pseudocode for Evolutionary programming

---

Input: Population\_size, Sequence\_length // problem size is defined as Sequence\_length

Output: Conf<sub>best</sub> // individuals are protein conformation

1. Population $\leftarrow$ InitializePopulation(Population\_size, Sequence\_length)
2. EvaluatePopulation(Population)
3.  $Conf_{best} \leftarrow GetBestSolution(Population)$
4. **While**(not termination condition())
  5. Children $\leftarrow$ Null
  6. **For** ( $Parent_i \in Population$ )
    7.  $Child_i \leftarrow Mutate(Parent_i)$
    8. Children  $\leftarrow$  Children +  $Child_i$
  9. **End**
  10. EvaluatePopulation(Children)
  11.  $Conf_{best} \leftarrow GetBestSolution(Children)$
  12. Union  $\leftarrow Population + Children$
  13. **For** ( $Conf_i \in Union$ )
    14. **For** (1 to Population\_size)
      15.  $Conf_j \leftarrow Selection(Union)$
      16. **If** ( $E(Conf_i) < E(Conf_j)$  // E is fitness function)
        17.  $Conf_{i\_win} \leftarrow Conf_{i\_win} + 1$
      18. **End**
    19. **End**
    20. **End**
    21. **End**
    22. Population  $\leftarrow SelectBestByWin(Union, Population\_size)$
    23. Return ( $Conf_{best}$ )

---

Formally, the PSP problem is defined as a triplet  $(S, f, \Omega)$ , where ' $S$ ' is the search space which holds collection of different possible conformations, ' $f$ ' is the objective function (in terms of free energy, the E-value), and  $\Omega$  is the set of imperatives that must be satisfied to get an optimal solution. The objective is to discover an ideal arrangement, which is the arrangement  $S_x$  with the smallest objective value under the condition that all imperatives are satisfied. Triplet  $(S, f, \Omega)$ , is defined as ' $S$ ' a set of lattice conformations for given HP sequences, ' $f$ ' objective function to be minimized defined as Eq. 1

$$\text{Minimize } f = \sum_{i,j:i+1 < j} e_{ij} \quad (1)$$

$$\text{Where, } e_{ij} = \begin{cases} -1, & \text{if } H - H \text{ contact} \\ 0, & \text{Otherwise} \end{cases}$$

$\Omega$ : Set of following constraints that need to be satisfied when modeling the PSP on lattice:

- Self-avoiding walk (SAW): each amino acid must occupy only one lattice point, which no other amino acid can share,
- Adjacent amino acids of primary structure must be at the unit distance

For HhPN revised energy function, PSP is defined as follows:

$$\text{Minimize } E \sum_{i,j:i+1 < j} c_{ij} e_{ij} + \sum_{i,j:j=i \pm 2 \text{ and } N \leq 3} c'_{ij} e_{ij} \quad (2)$$

$$\text{Where, } c_{ij} = \begin{cases} 1, & \text{if residue } i \text{ and } j \text{ are neighbour on lattice} \\ 0, & \text{otherwise} \end{cases}$$

$$c'_{ij} = \begin{cases} 0.7, & \text{if residue } i \text{ and } j \text{ are diagonally neighbour} \\ 0, & \text{otherwise} \end{cases}$$

$$\text{and } e_{ij} = \begin{cases} -4, & \text{if } S_i = S_j = H \text{ or } S_i = S_j = h \\ -2, & \text{if } S_i = H \text{ and } S_j = h \text{ or vice-versa} \\ -1, & \text{if } S_i = P \text{ and } S_j = N \text{ or vice-versa} \\ 1, & \text{if } S_i = S_j = P \text{ or } S_i = S_j = N \\ 0, & \text{otherwise} \end{cases}$$

Unlike to original HP model, revised energy matrix has also considered the diagonal interactions which solve the parity problem associated with the square lattice [37]. Diagonal interaction is depicted in parameter  $c'_{ij}$  whereas along square side interaction using  $c_{ij}$ .

### 3 Result and Discussion

The enzymatic, regulatory and accessory proteins of HIV-1 involved in integration, maturation and assembly were modeled based on the proposed algorithm. The accuracy of the predicted conformations was compared with the available structures from PDB. The percentage Q3 accuracy is listed for the test set (Table 1). The statistical parameters for each structure are listed in Table 2. The F1 score obtained is ranging from 0.74 to 0.96 with an average  $r^2$  of 0.81 between experimental and predicted structures. The experimentally determined structures were visualized and displayed using PyMOL (v.2) software [38]. The proposed program achieves an average Q3 score of 76.8%. For each of the Q3 parameter i.e., for helix (H), sheet (S) and coil (C), the average score is 81.7%, 80.5% and 67.2%, respectively. The accuracy of the  $\alpha$ -helix and  $\beta$ -sheet mappings were in similar range. The percentage accuracy was reduced for proteins with shorter  $\beta$ -sheet (with one or two residues), as observed in the sequences (PDB IDs) 1HIW, 1HPV, 1AVV, and 5HGP (Table 1). However, in the experimentally solved structures, shorter sheets do exist but it is difficult to predict using the proposed method. On a contrary to this 3H47, 2EZ0 and 3DCG have showed consistent performance for all the three states (Table 2).

- (a) **Virus protein U (PDB 2N28):** It is a small protein of 81 amino acids encoded by HIV-1. In the experimental structure, Zhang et al. [39] have reported that, 44 amino acids acquire the helical pattern resulting into a four  $\alpha$ -helix and 14 amino acids take part in turn, for rest of the residues no secondary structure is assigned. However, while visualizing a conformation, residues without defined secondary structure are represented as loop. The modeled conformation has correctly identified and folded the 38 out of 44 amino acids as  $\alpha$ -helix (Fig. 4a) with the  $r^2$  value of 0.99. The Q3 score for helix is 86.4% and for coil it is 69.2%. However, the Q3 score is not assigned for sheet, and the turn is difficult to model in the computational method due to single amino acid. Moreover, the Q3 accuracy of the model conformation is 77.8% and the statistical accuracy is 80.2%.
- (b) **Protease (PDB 1HPV):** It is classified under “all  $\beta$ -topology” [40]. It is made up of nine strands involving 53 amino acids, reported by Kim et al. [41]. In the modeled conformation 40 amino acids formed  $\beta$ -sheet with Q3 value of 75.5%, in agreement with the experimentally solved structure. Four amino acids participate in the  $\alpha$ -helix formation and 21 for loop, whereas the remaining 21 residues were not assigned with any structural information. The modeled conformation had similar helical region, however, the size of the helix is longer due to constrained folding, where it requires minimum of six amino acids to fold into  $\alpha$ -helix (Fig. 4b). The Q3 score for 1HPV were 100%, 75.5%, and 60% for helix, sheet, and coil, respectively. The Q3 score is also supported by statistical parameters.
- (c) **Virion infectivity factor (3DCG):** It is an essential protein for viral replication, the protein is composed of 118 amino acids and it is solved using X-ray diffraction method. Stanley et al. [42] reported that the structure (3DCG) is

**Table 1** Q3-Score for HIV-1 protein structure prediction

PDB ID	Length (AA)	Residue distribution			Modeled structure			Q3 Accuracy parameter (%)		
		Helix	Sheet	Coil	Helix	Sheet	Coil	Helix	Sheet	Coil
2N28	81	9–27, 35–39, 41–48, 59–70	AB	6, 8, 28, 30–31, 50–51, 54–58, 71–72, 75–76, 78	8–13, 16–27, 31–39, 44–52, 60–71	AB	6,8,28, 30, 50–51, 54, 58, 72, 75	86,4	AB	69,2
IHPV	99	87–90	2–4, 10–15, 18–24, 31–33, 43–48, 53–66, 69–77, 84–85, 96–98	5–7, 16–17, 26–27, 29–30, 35, 40, 50–51, 67–68, 78–83, 91–94	85–90	9–15, 18–25, 29–33, 45–56, 67–81, 96–99	5–7, 16–17, 26–27, 35, 40, 67–68, 82–83, 85, 91–94	100	75,5	60
3DCG	118	24–35, 39–41	2–9, 10, 12–19, 23, 42–46, 49–50, 55–56, 67–68, 70–71, 73–78, 79–80, 84–85, 89–90	47–48, 53–54, 57–61, 64–66, 69	23–32, 37–42, 84–89	1–8, 11–16, 43–57 59–65, 69–80,	Since sheet and turn are adjacent to each other with each having two residues, here for modelling only sheet considered.	80	84	N/A
IHW	133	10–16, 31–43, 48–52, 54–67, 70–71, 95–96 N/A	19–20, 27–28, 45–46, 21–26, 44, 53, 68–69, 90, 108–109	17–18, 81–17, 30–47, 51–62, 66–80, 81–86, 93–104	24–32, 35–40, 47–52	2–9, 12–21, 59–63, 74–78	47–48, 53–54, 57–58, 64–66, 69	73,3	77,8	76,9

(continued)

**Table 1** (continued)

PDB ID	Length (AA)	Residue distribution			Modeled structure			Q3 Accuracy parameter (%)		
		Helix	Sheet	Coil	Helix	Sheet	Coil	Helix	Sheet	Coil
1AVV	151	81–93, 104–117, 187–189, 194–198, 200–202	76–77, 101–102, 119, 126, 133, 135, 143–145, 181–185	96–99, 122, 131–132, 136–140, 190–191	80–94, 106–115, 151–156, 188–195, 198–209	7–14, 118–125, 128–133, 140–146,	1–6, 15–79, 95–105, 116–117, 126–127, 134–139, 147–150, 157–179, 186–187, 196–197	78.9	100	57.1
5KRS	153	94–107, 118–133, 150–168, 172–185, 195–206	60–68, 71–78, 83–88, 112–114, 135–137	58–59, 69–70, 79–82, 89–93, 108, 116–117, 140–141, 169–171	79–104, 118–133, 145–150, 163–182, 199–206	60–75, 108–114, 140–144, 183–194	58–59, 108, 116–117, 140–141	72.6	68.6	66.8
3H47	231	17–30, 36–43, 49–57, 62–83, 101–104, 111–117, 126–145, 150–152, 161–174, 189–192, 196– 205, 211–217	2–4, 10–12	31–35, 44–46, 61, 94, 105–109, 120–122, 148, 157–158, 194, 207–208	18–26, 36–43, 49–60, 61–87, 101–108, 11–118, 133–141, 162–170, 186–191, 199–208, 211–220, 221–231	N/A	31–35, 44–46, 58, 94, 105–109, 120–122, 148, 157–158, 194, 207–208	77	N/A	62.5
2EZO	123	4–54, 81–121	AB	56, 58–59, 62–65, 75, 79–80	3–14, 15–23, 25–36, 37–45, 46–54, 76–81, 86–103, 104–188	AB	56, 58–59, 62–65, 75	90.2	N/A	80

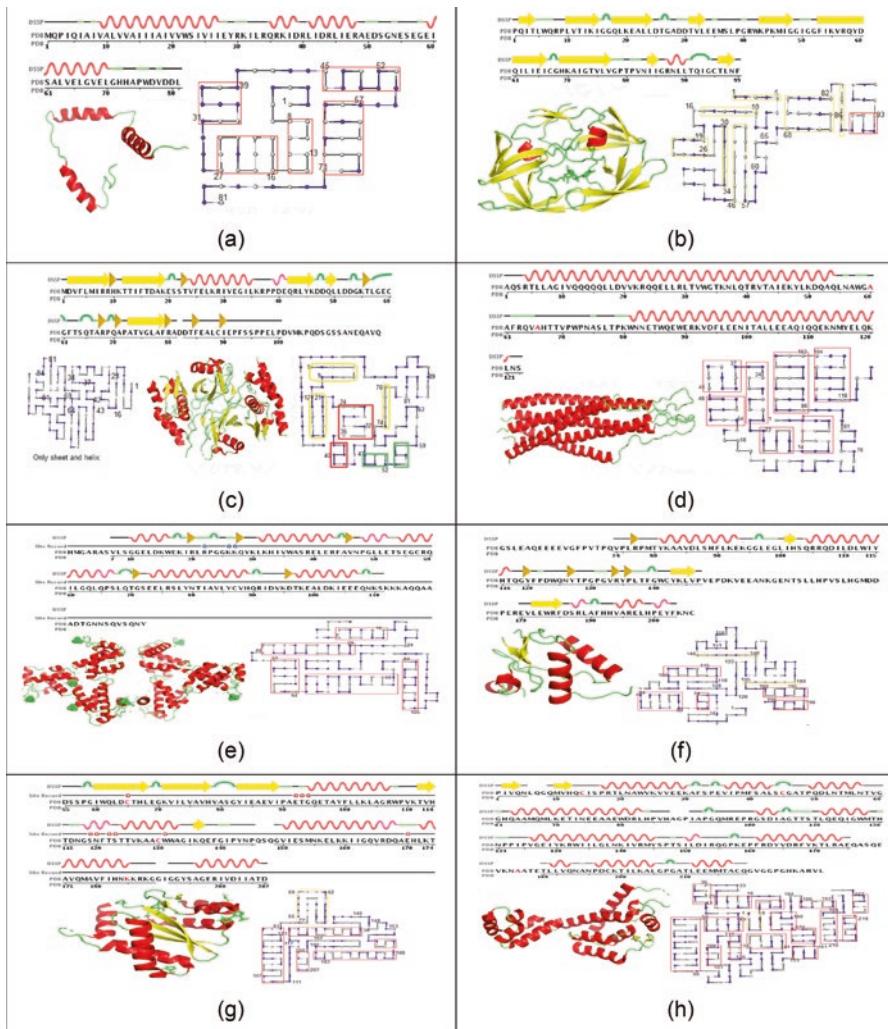
**Table 2** Statistical parameters for structural evaluation of tested proteins

PBD ID	L	E	TP	FP	FN	TN	Sn	Sp	PPV	NPV	Ac	F1
2N28	81	58	38	10	6	27	86.4	72.9	79.2	81.8	80.2	82.6
1HPV	99	50	43	13	12	31	78.2	70.5	76.8	72.1	74.8	77.5
3DCG	118	36	42	15	8	53	84.0	77.9	73.7	86.9	80.5	78.5
1HIW	123	29	55	20	12	46	82.1	69.7	73.3	79.3	75.9	77.5
1AVV	133	38	43	11	6	91	87.8	89.2	79.6	93.8	88.7	83.5
5KRS	151	60	70	19	31	33	69.3	63.5	78.7	51.5	67.3	73.7
3H47	153	78	94	19	26	92	78.3	82.9	83.2	77.9	80.5	80.7
5HGP	231	65	93	37	29	72	76.2	66.1	71.5	71.3	71.4	73.8
2EZO	118	39	88	6	1	28	98.9	82.4	93.6	96.5	94.3	96.2

*L* Length (number of amino acids), *E* Energy value, *TP* True Positive, *FP* False Positive, *FN* False Negative, *TN* True Negative, *Sn* Sensitivity, *Sp* Specificity, *PPV* Positive Predictive Value, *NPV* Negative Predictive Value, *Ac* Accuracy, *F1* F1-score

32%  $\beta$ -sheet and 12%  $\alpha$ -helix, whereas the structural information for the rest 56% residues is absent. However, 32% of  $\beta$ -sheet is comprised of both  $\beta$ -strand and bridge. For the sake of convenience we treated them alike because in most cases either bridge is a part of  $\beta$ -strand or unable to recognize by our program, especially if it is preceded or followed by any other structural pattern ( $\alpha$ -helix or turn). Hence, while modeling two conformations were assigned as optimum solution, one with all the three secondary states and another with only  $\alpha$ -helix and  $\beta$ -sheet. The Q3 score of all the three state predictions were 73.3%, 77.8%, and 76.9% for  $\alpha$ -helix,  $\beta$ -sheet, and coil, respectively. This showed a closer mapping with the experimental structure and had sheets and turns adjacent to each other. The statistical accuracy for the predicted conformation is 80.5% with  $r^2$  value of 0.69.

- (d) **Matrix protein (1HIW):** It is a trimeric protein with 133 amino acid residues, solved using X-ray diffraction method by Hill et al. [43]. However, the structural information is limited to 53%, out of which 50% of residues contributes for  $\alpha$ -helix and only 3% for  $\beta$ -sheets, loop forming amino acid reported in the predicted structure are 11%. This protein presents a good opportunity for computational modeling of the rest of the amino acids (i.e., nearly 40% of the primary structure is unsolved). Although, it is a major challenge for 1HIW, the computational model fails to predict the  $\beta$ -sheet forming residues correctly. This is due to five individual amino acids, each have the propensity to form  $\beta$ -sheets. Hence, in the case of 1HIW, the proposed approach correctly predicted the helical pattern with Q3 score of 82%. Similar results were obtained based on statistical parameters.
- (e) **Nef protein (1AVV):** The structural details of 1AVV protein obtained using X-ray diffraction method [44]. The deposited structure has 25% helical region involving 38 amino acids and 10% sheet composed of 8 strands involving 16 amino acids. Whereas, the rest 65% of structural details is absent. Computationally modeled structure has shown a good agreement with the



**Fig. 4** Structural conformation for tested proteins (a) 2N28, (b) 1HPV, (c) 3DCG, (d) 2EZO, (e) 1HIW, (f) 1AVV, (g) 5KRS, (h) 3H47

experimentally obtained structure with Q3 score of 78.9%, 100%, and 57.1% for  $\alpha$ -helix,  $\beta$ -sheet and turn, respectively. The statistical behavior of folded conformation indicate a good correlation between all the parameters with an average value of 80%. The statistical accuracy of 88.7% has been attained for the modeled conformation.

- (f) **Integrase (5KRS):** It is an essential factor for virus replication, and an important multifunctional therapeutic target. It is composed of 153 amino acids and it is classified as hydrolase transferase/inhibitor protein. Its native structure is

obtained using the X-ray diffraction method [45]. The major part of this protein is helical with 47% of amino acid contributes to  $\alpha$ -helix formation. About 18% form  $\beta$ -sheets that is in turn made up of five strands and it is composed of 9, 8, 6, 3, and 3 amino acids residues in each strand. The last two strands are difficult to predict by our method due to the constraints employed (a minimum of four amino acids to predict  $\beta$ -strands). On mapping the secondary structure of the modeled conformation over experimental structure, it has given a Q3 score of 72.6% and 68.6% for  $\alpha$ -helix and  $\beta$ -sheet, respectively. The lower sheet percentage is due to short stretch of  $\beta$ -strands and this is also reflected in the statistical accuracy of 67.3%.

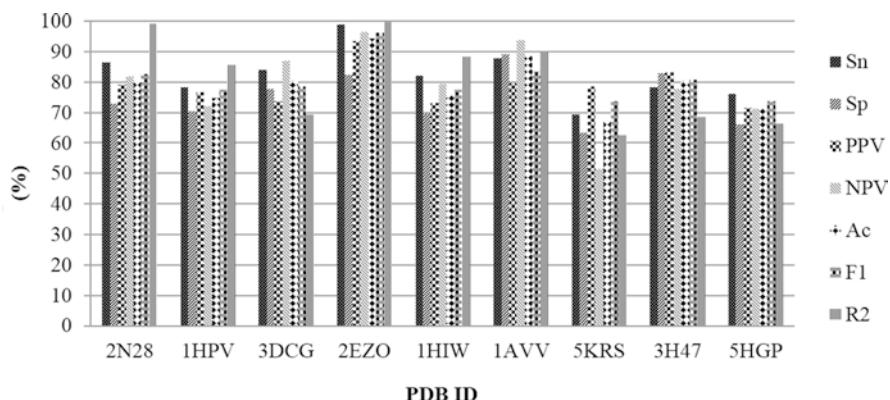
- (g) **Capsid protein (3H47):** The structural details for the capsid protein was solved by Pornillos et al., (2009) using X-ray diffraction method [46]. The majority of the protein is helical (involving 122 amino acids), and few turns, whereas for the rest of the sequence, the structure is not assigned. The terminal six amino acids contribute for  $\beta$ -sheets. The modeled conformation has shown a good agreement with the experimental structure with a Q3 score of 77% for  $\alpha$ -helix, whereas the method could not identify sheets involving only 3 amino acids. The modeled conformation has shown an accuracy of 80.5%. There is another structure for the same protein (HIV-1 CA) is available in PDB (5HGP) deposited by Jacques and James [47]. There were differences in the constituting amino acids. Hence, this was also modeled similar to 3H47, the majority of the structure is comprised of  $\alpha$ -helix with 53% of amino acids resulting into 12 helices. The proposed method could not identify short stretch of  $\beta$ -sheets. On comparing the modeled conformation with experimental structure the Q3 score of 77% and 62.5% has been obtained for  $\alpha$ -helix and coil, respectively. Similar results have been found on statistical evaluation with an accuracy of 71.4%.
- (h) **The envelope protein (PDB: 2EZ0):** The transmembrane region (representing gp41) of simian immunodeficiency viruses (SIV) is modeled. As it falls under the same lentivirus lineage [48], this may provide insights into the viral entry inhibition and also enables structure-based drug design approaches. The helical core of SIV gp41 is similar to the helical core of HIV-1 gp41. The ectodomain of SIV gp41 provides insights into the binding site of gp120 and mechanism of cell fusion. This is one of the largest protein structures determined by NMR [49]. The protein structure showcases 74% of  $\alpha$ -helix and 10% for the loop, whereas the rest 16% is unsolved. Computationally the predicted structure in comparison with the NMR structure, has got a Q3 value of 90.2% for  $\alpha$ -helix and 80% for random coil. Similarly, on statistical comparison, the modeled conformation has shown sensitivity of 98%, which indicate that the computational method is efficient and correctly predict the secondary structure. The specificity is 82.4%, the PPV is 93.6%, and the NPV is 96.5%. These values demonstrate the efficiency of the computational method. Furthermore, the accuracy of the conformation obtained is 94.3%.

### 3.1 Statistical Analysis

The molecular interactions that stabilize the structures of the predicted model were compared with the experimentally determined ones and the correctness of the interactions were given in terms of ‘sensitivity’. Higher the sensitivity value lessens the chance of missing interaction. With the tested HIV-1 sequences, the least sensitivity obtained is 69.3%, and the highest is being 98.9% (Table 2). On an average, the proposed method has consistently given the sensitivity of 82% (Fig. 5). The specificity of the approach is also verified i.e., the case where the prediction approach should not represent any false interaction. The approach, on an average ruled out 75% of the wrong interactions.

The occurrence of interactions in the predicted model conformation is in accordance with the experimental structure and it is given by positive predictive value (PPV). This parameter also indicates the significance of conformation by assessing the number of true positives in the model conformation. Higher the value better the model conformation, whereas, negative predictive value (NPV) is the percentage of residues arranged on the lattice such that they do not contribute to the interaction as in experimentally determined structure. The higher the value of NPV, the better is the model conformation. The proposed approach has given an average of 79% for both PPV and NPV. The overall accuracy (F1 score) of proposed method is about 80%.

Among the eight protein sequences tested, the envelop protein (2EZO) outperformed all other sequences based on the parameters studied. It has also given similar results on both quantification parameters (Q3 and statistical). However, the longest sequences considered in this study is of length 231 (3H47 and 5HGP). These sequences also performed well with an average accuracy of 75.9%. However, the difference between the accuracy of these two sequences is in order of 10. This difference is due to the constituting amino acids, and the limit of constraint folding implemented in our approach.



**Fig. 5** Statistical parameters for tested proteins

### 3.2 The Correlation Between Predicted and Actual Structures

The correlations between the predicted and actual structures for all the tested HIV-1 protein sequences is given by  $r^2$  value. This signifies how well the model conformation fit with the experimentally determined structure (especially the AA contacts), a linear regression line is drawn and verified the amino acid distributions around the line (Fig. 6).

The higher the  $r^2$  value better the model conformation, it depict the case where the model conformation is able to predict correct structure. The  $r^2$  ranges from 0 to 1 (i.e. 0–100%). The approach predicted correctly for proteins having “all  $\alpha$ -topology” as in the case of Vpu (2N28), Nef (1AVV) and the envelope protein (2EZO), the  $r^2$  is above 0.9. Similarly, the matrix protein (1HIW) is also helix dominated protein with 50% helical region and 3% sheet ( $r^2 = 0.87$ ). However, the sheets were not depicted in the model conformation as they are made up of single residues each. The total number of amino acids contributing helical pattern is 44, whereas in the modeled conformation it was only 38 amino acids. However, this presents one of the best modeling cases. The protease (1HPV) is a  $\beta$ -sheet dominated protein with 53 amino acids, and it is predicted to contain nine strands ( $r^2 = 0.83$ ). Due to computational constraints, the model conformation fail to identify four short strands. The virion infectivity factor (3DCG) has one  $\alpha$ -helix and four  $\beta$ -sheets. Besides, it has 9 more strands contributed by single AA residue (not represented in the modeled conformation,  $r^2 = 0.69$ ). The remaining tested sequences have all three structural states (Fig. 6).

## 4 Conclusion

In this chapter, the applicability of the computational coarse structure prediction for HIV-1 protein sequences using evolutionary search algorithm is demonstrated to obtain an optimal protein conformation. The tested sequences were evaluated with the experimentally solved structure using the test parameters Q3-score as well as statistical analysis. The comparison of the predicted structure with the experimentally reported ones indicated that the coarse level computational predictions can be applied to HIV-1 protein structure prediction with a reasonable accuracy of 80%. The prediction accuracy could further be improved, if the computational method was able to predict the shorter  $\beta$ -strand (with one or two residues). These short segments do exist in the experimentally solved structures but it was difficult to predict using the proposed method. This scenario may be tackled by implementing artificial intelligence or machine learning methods by introducing more training sets. In this study, we have showed the usefulness of the evolutionary search approach for the coarse structure prediction to bridge the gap between the sequence and structure information of proteins. The improvement in these efforts can accelerate high-throughput structure-based drug design approaches.

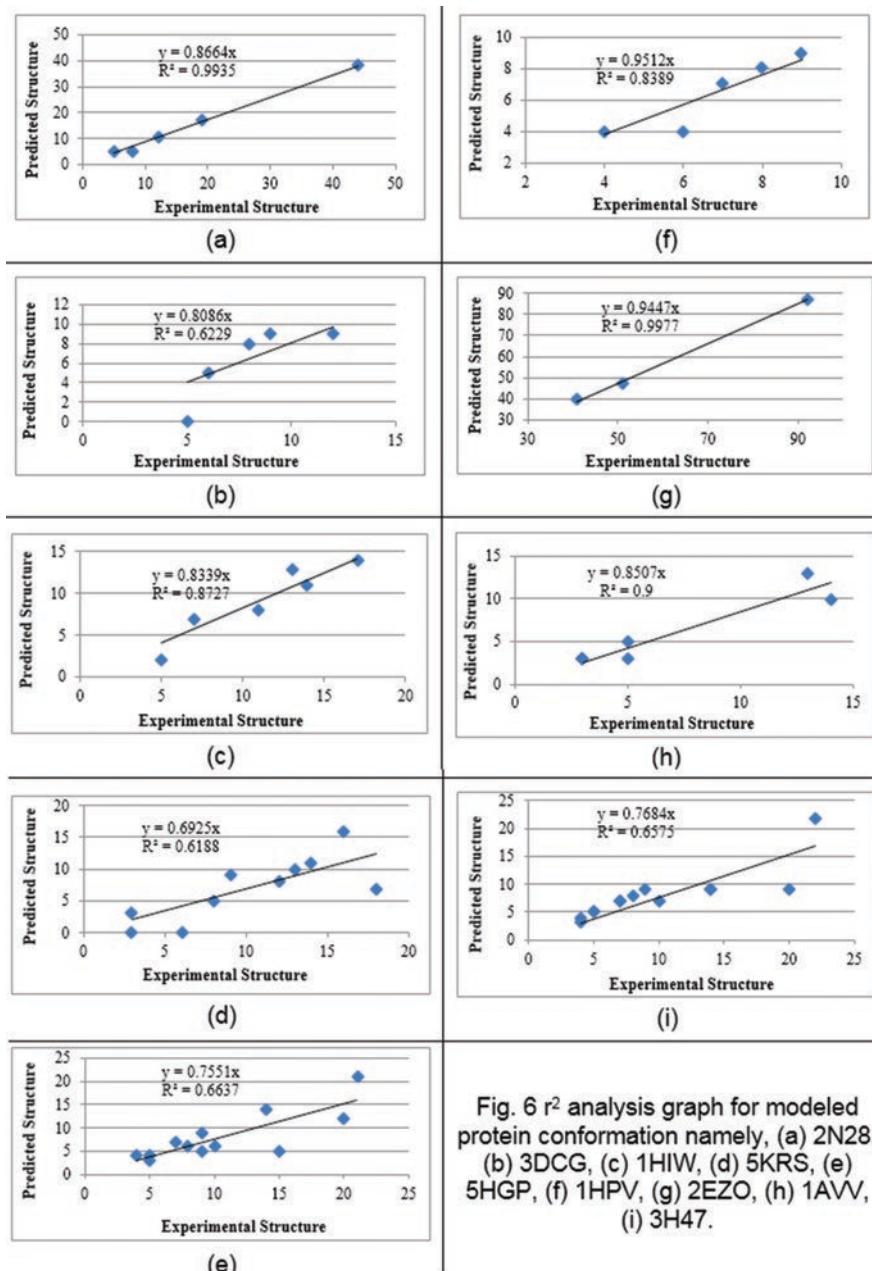


Fig. 6  $r^2$  analysis graph for modeled protein conformation namely, (a) 2N28, (b) 3DCG, (c) 1HIW, (d) 5KRS, (e) 5HGP, (f) 1HPV, (g) 2EZO, (h) 1AVV, (i) 3H47.

**Fig. 6** The correlation between predicted and experimental structures, (a) 2N28, (b) 3DCG, (c) 1HIW, (d) 5KRS, (e) 5HGP, (f) 1HPV, (g) 2EZO, (h) 1AVV, (i) 3H47

**Acknowledgements** The corresponding author acknowledges the grant (No. VGST/GRD-533/2016-17/241) received from Karnataka Science and Technology Promotion Society (KSTePS), India for supporting the ‘Centre for Interactive Biomolecular 3D-literacy (C-in-3D)’ under the VGST scheme – Centres of Innovative Science, Engineering and Education (CISEE) for the year 2016-17.

## References

1. Chu M, Zhang W, Zhang X, Jiang W, Huan X, Meng X, Zhu B, Yang Y, Tao Y, Tian T, Lu Y. HIV-1 CRF01\_AE strain is associated with faster HIV/AIDS progression in Jiangsu Province, China. *Sci Rep.* 2017;7(1):1570.
2. Global Health Observatory (GHO) data, available at: <http://www.who.int/gho/hiv/en/>. Last accessed 30 Oct 2018.
3. Li G, De Clercq E. HIV genome-wide protein associations: of 30 years of research. *Microbiol Mol Biol Rev.* 2016;80(3):679–731.
4. Frankel AD, Young JA. HIV-1: fifteen proteins and an RNA. *Annu Rev Biochem.* 1998;67:1–25.
5. HIV database <http://www.hiv.lanl.gov>. Last accessed 30 Oct 2018.
6. Hoque MT. Genetic algorithm for ab initio protein structure prediction based on low resolution models. Ph.D. dissertation, Monash University, Faculty of Information Technology. Gippsland School of Information Technology; 2008.
7. Kaptein R, Boelens R, Scheek RM, Van Gunsteren WF. Protein structures from NMR. *Biochemistry.* 1988;27(15):5389–95.
8. Protein data bank, [http://www.rcsb.org/pdb/home/home.do/](http://www.rcsb.org/pdb/home/home.do). Last accessed 23 Mar 2018.
9. Kc DB. Recent advances in sequence-based protein structure prediction. *Brief Bioinform.* 2016;18(6):1021–32.
10. Schwede T. Protein modeling: what happened to the “protein structure gap”? *Structure.* 2013;21(9):1531–40.
11. Dubey SP, Kini NG, Balaji S, Kumar MS. A comparative study on single and multiple point crossovers in a genetic algorithm for coarse protein modeling. *Crit Rev Biomed Eng.* 2018;46(2):163–71.
12. Denise C. Structural GENOMICS exploring the 3D protein landscape. *Biomed Comput Rev, Simbios.* 2010;10:11–8.
13. Lee J, Freddolino PL, Zhang Y. Ab initio protein structure prediction. In: *Protein structure to function with bioinformatics*. Springer, Dordrecht; 2017. p. 3–35.
14. Lau KF, Dill KA. A lattice statistical mechanics model of the conformational and sequence spaces of proteins. *Macromolecules.* 1989;22(10):3986–97.
15. Hoque T, Chetty M, Sattar A. Extended HP model for protein structure prediction. *J Comput Biol.* 2009;16(1):85–103.
16. Dubey SP, Kini NG, Balaji S, Kumar MS. Protein structure prediction on 2D square HP lattice with revised fitness function. In: *Advances in computing, communications and informatics (ICACCI)*, IEEE, Udupi; 2017. p. 1732–36.
17. Unger R, Moult J. Genetic algorithms for protein folding simulations. *J Mol Biol.* 1993;231(1):75–81.
18. Rashid MA, Iqbal S, Khatib F, Hoque MT, Sattar A. Guided macro-mutation in a graded energy based genetic algorithm for protein structure prediction. *Comput Biol Chem.* 2016;61:162–77.
19. Li B, Li Y, Gong L. Protein secondary structure optimization using an improved artificial bee colony algorithm based on AB off-lattice model. *Eng Appl Artif Intell.* 2014;27:70–9.
20. Dubey SP, Kini NG, Balaji S, Kumar MS. A review of protein structure prediction using lattice model. *Crit Rev Biomed Eng.* 2018;46(2):147–62.
21. Dotu I, Cebrian M, Van Hentenryck P, Clote P. On lattice protein structure prediction revisited. *IEEE/ACM Trans Comput Biol Bioinform.* 2011;8(6):1620–32.

22. Chou PY, Fasman GD. Prediction of protein conformation. *Biochemistry*. 1974;13(2):222–45.
23. Anfinsen CB. Principles that govern the folding of protein chains. *Science*. 1973;181(4096):223–30.
24. Kapsokalivas L, Gan X, Albrecht A, Steinhöfel K. Two local search methods for protein folding simulation in the HP and the MJ lattice models. In: *Bioinformatics research and development*. Berlin, Heidelberg: Springer; 2008. p. 167–79.
25. Berrera M, Molinari H, Fogolari F. Amino acid empirical contact energy definitions for fold recognition in the space of contact maps. *BMC Bioinformatics*. 2003;4(1):8.
26. Shi G, Wüst T, Li YW, Landau DP. Protein folding of the HOP model: a parallel Wang—Landau study. *J Phys Conf Ser*. 2015;640(1):012017. IOP Publishing.
27. Blackburne BP, Hirst JD. Evolution of functional model proteins. *J Chem Phys*. 2001;115(4):1935–42.
28. Cutello V, Nicosia G, Pavone M, Timmis J. An immune algorithm for protein structure prediction on lattice models. *IEEE Trans Evol Comput*. 2007;11(1):101–17.
29. Pearson WR. Selecting the right similarity-scoring matrix. *Curr Protoc Bioinformatics*. 2013;43(3.5):1–9.
30. Hoque MT, Chetty M, Sattar A. Protein folding prediction in 3D FCC HP lattice model using genetic algorithm. In: *Evolutionary Computation, 2007. CEC 2007. IEEE Congress on 2007 Sep 25*. IEEE. p. 4138–45.
31. Koliński A, Gront D, Kmiecik S, Kurcinski M, Latek D. Modeling protein structure, dynamics and thermodynamics with reduced representation of conformational space. In: *NIC Workshop 2006: From computational biophysics to system biology; 2006*, vol. 34.
32. Hoque MT, Chetty M, Sattar A. Genetic algorithm in ab initio protein structure prediction using low resolution model: a review. In: *Biomedical data and applications*. Berlin, Heidelberg: Springer; 2009. p. 317–42.
33. Dill KA, Bromberg S, Yue K, Chan HS, Ftebig KM, Yee DP, Thomas PD. Principles of protein folding—a perspective from simple exact models. *Protein Sci*. 1995;4(4):561–602.
34. Berger B, Leighton T. Protein folding in the hydrophobic-hydrophilic (HP) model is NP-complete. *J Comput Biol*. 1998;5(1):27–40.
35. Bryngelson JD, Onuchic JN, Socci ND, Wolynes PG. Funnels, pathways, and the energy landscape of protein folding: a synthesis. *Proteins*. 1995;21(3):167–95.
36. Dubey SP, Balaji S, Kini NG, Sathish Kumar M. A novel framework for ab initio coarse protein structure prediction. *Adv Bioinforma*. 2018;2018:7607384.
37. PyMOL by Schrodinger available at: <https://pymol.org/2/>. Last accessed 30 Oct 2018.
38. Seeliger D, de Groot BL. Ligand docking and binding site analysis with PyMOL and Autodock/Vina. *J Comput Aided Mol Des*. 2010;24(5):417–22.
39. Zhang H, Lin EC, Das BB, Tian Y, Opella SJ. Structural determination of virus protein U from HIV-1 by NMR in membrane environments. *Biochim Biophys Acta*. 2015;1848(11):3007–18.
40. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol*. 1995;247(4):536–40.
41. Kim EE, Baker CT, Dwyer MD, Murcko MA, Rao BG, Tung RD, Navia MA. Crystal structure of HIV-1 protease in complex with VX-478, a potent and orally bioavailable inhibitor of the enzyme. *J Am Chem Soc*. 1995;117(3):1181–2.
42. Stanley BJ, Ehrlich ES, Short L, Yu Y, Xiao Z, Yu XF, Xiong Y. Structural insight into the human immunodeficiency virus Vif SOCS box and its role in human E3 ubiquitin ligase assembly. *J Virol*. 2008;82(17):8656–63.
43. Hill CP, Worthy lake D, Bancroft DP, Christensen AM, Sundquist WI. Crystal structures of the trimeric human immunodeficiency virus type 1 matrix protein: implications for membrane association and assembly. *Proc Natl Acad Sci U S A*. 1996;93(7):3099–104.
44. Franken P, Arold S, Padilla A, Hoh E, Strub MP, Boyer M, Jullien M, Dumas C, Bodeus M, Benarous R. HIV-1 Nef protein: purification, crystallizations, and preliminary X-ray diffraction studies. *Protein Sci*. 1997;6(12):2681–3.

45. Patel D, Antwi J, Koneru PC, Serrao E, Forli S, Kessl JJ, Feng L, Deng N, Levy RM, Fuchs JR, Olson AJ, Engelman AN, Bauman JD, Kvaratskhelia M, Arnold E. A new class of allosteric HIV-1 integrase inhibitors identified by crystallographic fragment screening of the catalytic core domain. *J Biol Chem.* 2016;291:23569–77.
46. Pornillos O, Ganser-Pornillos BK, Kelly BN, Hua Y, Whitby FG, Stout CD, Sundquist WI, Hill CP, Yeager M. X-ray structures of the hexameric building block of the HIV capsid. *Cell.* 2009;137(7):1282–92.
47. Jacques DA, McEwan WA, Hilditch L, Price AJ, Towers GJ, James LC. HIV-1 uses dynamic capsid pores to import nucleotides and fuel encapsidated DNA synthesis. *Nature.* 2016;536(7616):349.
48. Joy JB, Liang RH, Nguyen T, Mccloskey RM, Poon AFY. Origin and evolution of HIV. In: Shapshak P, Sinnott JT, Somboonwit C, Kuhn JH, editors. *Global virology I. Identifying and investigating viral diseases.* New York: Springer; 2015.
49. Caffrey M, Cai M, Kaufman J, Stahl SJ, Wingfield PT, Covell DG, Gronenborn AM, Clore GM. Three-dimensional solution structure of the 44 kDa ectodomain of SIV gp41. *EMBO J.* 1998;17(16):4572–84.

# Drug Development for Hepatitis C Virus Infection: Machine Learning Applications



Sajitha Lulu Sudhakaran, Deepa Madathil, Mohanapriya Arumugam, and Vino Sundararajan

**Abstract** Hepatitis C virus (HCV) infection is one of the leading causes of mortality and morbidity, and is widely reported for its association with the development of liver cirrhosis, hepatocellular cancer, and liver failure. Most of the reported cases of *hepatitis C* end up with a chronic form of the infection, existing as a large threat for public health and can be prevented by evading or eradicating the virus through effective drug development. Conventional medicines that are both safe and easily affordable, have not yet been developed for the treatment of chronic HCV infection. Apart from only identifying novel drugs, it is equally important to explore their effectiveness by ascertaining drug target accuracy, which is a crucial part of any drug development program. Moreover, it is highly critical to understand the activity and molecular basis of drug resistance of various drugs, as they may retain activity against a broad spectrum of drug resistant viral variants. Drug discovery and design are highly complex, time consuming, and expensive endeavors. Therefore, it is crucial to incorporate new technologies for this process. Modern drug design strategies include ligand-based (LBDD) and structure-based drug design (SBDD) methods to develop new drug candidates. Machine Learning (ML) approaches are extensively applied in drug design processes for HCV and most common applications include classifying drug targets into druggable and non-druggable, prioritizing drug targets, discovering novel inhibitors, predicting diseases by using risk factors as classifiers, *in silico* ADMET prediction, etc. However, a few studies using Machine Learning

---

S. L. Sudhakaran (✉) · M. Arumugam

Department of Biotechnology, School of BioSciences and Technology,

Vellore Institute of Technology, Vellore, Tamilnadu, India

e-mail: [ssajithalulu@vit.ac.in](mailto:ssajithalulu@vit.ac.in)

D. Madathil

Department of Sensor and Biomedical Technology, School of Electronics Engineering,

Vellore Institute of Technology, Vellore, Tamilnadu, India

V. Sundararajan

Department of Biosciences, School of BioSciences and Technology,

Vellore Institute of Technology, Vellore, Tamilnadu, India

approaches have been reported for prediction of biological activity from multivariate models, prediction of binding site secondary structural modes of docking, and virtual screening.

The most common ML techniques applied in HCV drug discovery, comprise techniques such as random forest, SVM, Decision tree, Genetic algorithms, K-Nearest Neighbor's, Naive Bayesian classifiers, Particle swarm optimization, as well as multilinear regression models. These tools are widely used in drug discovery studies as they are readily accessible, both as open source and commercial distributions, statistically consistent, computationally efficient, and relatively straight-forward to implement and interpret. Moreover, data-mining software enables users to implement these algorithms through graphical user interfaces and can also be written and executed using packages such as R, Matlab, and Octave. Datamining and Machine Learning approaches hence seem as promising aid for Drug Development studies on HCV infection.

**Keywords** HCV · Drug resistance · Machine learning methods · SVM · Decision tree · Genetic algorithms · K-nearest neighbors · Naive Bayesian classifiers · Particle swarm optimization · And multilinear regression models

### Key Concepts

- Machine learning methods in drug discovery
- Prioritization of druggable targets by Genetic Algorithm and SVM
- Prediction of advanced liver fibrosis by Alternating Decision Tree (ADT)
- Discovery of novel HCV inhibitors by Random Forest method
- Random Forest based Virtual screening for identifying novel inhibitors
- ADMET property predictions by MLR and PLS, Secondary structure predictions by stochastic supervised machine learning algorithms
- Machine learning algorithms implemented in HCV research: – SVM, Decision tree, Genetic algorithms, K-Nearest Neighbors, Naive Bayesian classifiers, Particle swarm optimization, and Multilinear regression methodology.

## 1 Introduction

### Machine Learning (ML) Approaches in Drug Designing for Hepatitis C Virus (HCV)

The drug discovery process includes application of multiple methodologies for identifying and designing new bioactive molecules, which could be developed into novel drugs. A significant protocol in drug designing is the accurate prediction of biological activity of chemical compounds from their molecular descriptors called as quantitative structure-activity relationships or QSAR studies. It has been clearly reported that Molecular descriptors help in understanding different aspects

of chemical and biological interactions in drug discovery and design. The quality of a QSAR model depends on the accuracy of input data, selection of descriptors, and statistical tools, and most importantly validation of the developed model. The validation process can be performed using strategies including internal validation (or cross-validation, external validation (test compound set, not used in the model training), and data randomization.

## ***1.1 ML in Classifying Drug Targets as Druggable and Non-druggable***

Reaction to a molecule is often unpredictable in complex biological systems. The interference of chemical substances like drug is often unfavorable. Rapid advancement in pharmaceutical biotechnology over the years and the enhanced understanding of biology has led to more effective drug design. While the approved drugs increased over the past decade, they have been outpaced by the more rapid increased cost of drug development [1].

Druggability is the characteristic feature of a molecule that interacts with a biological target and generates a desirable pharmacological response [2]. Widely reported druggable targets are proteins while nucleic acids are slowly being studied [2]. According to Gashaw et al., an ideal drug target should possess the following properties: promising assay for high throughput screening, ability to modify a disease, low impact on the regulation of physiological conditions or other diseases, differential expression across the body for specific targeting, and existence of a biomarker to identify the efficacy of its binding [3].

*In vitro* evaluation of all proteins or nucleic acid fragments for their druggability is an overwhelming task. Insufficient knowledge of the pathogenesis of disease at the molecular level, further worsens the situation. It is therefore unfeasible for *in vitro* evaluation of all drug biomarkers before being able to first prioritize them. Consequently, computational models that can predict drug targets with high sensitivity and specificity on a genome-wide scale would be highly encouraged. With the advancements in technology, we now have access to a plethora of data, including protein-protein interaction (PPI), gene regulatory networks, metabolic regulatory networks, and protein and gene expression profiles. Although consolidating these diverse data sets is still challenging, progress has been made in the past few years. Combining these data with machine learning aids building predictive models. Such analyses have the potential of identifying biologically relevant patterns that confer druggability to potential drug targets [4].

### **1.1.1 Machine Learning Algorithms**

ML algorithms such as SVMs, decision trees, ensemble of classifiers, logistic regression, radial basis function, and Bayesian networks are commonly used in the classification of druggable and non-druggable compounds [5]. DrugBank, which

employs SVMs makes druggability predictions by considering learning features such as connectivity degree, cluster coefficient, distance based measures and topological coefficient [6]. Therapeutic Target Database (TTD) exploit SVM-recursive feature elimination method for feature selection and SVM-RBF methods for prediction. The learning features used by TTD include row chromosomal copy number, mutation occurrence and closeness centrality [7]. PubChem, ChemBL and BindingDB use rank based method and learning features include combination of kernel, correlation diffusion and differential gene expression [8].

### 1.1.2 ML in Prediction of Advanced Liver Fibrosis (HCV Patients)

Classification and predictive learning machine method, Alternating Decision Tree (ADT) was used in the prediction of advanced liver fibrosis in chronic HCV patients. CART [9] and C4.5, which are conservative boosting decision tree algorithms generates intricate decision tree structures which are difficult to interpret. Predictions made by ADTree are easy to interpret. An alternating decision tree is comprised of decision nodes and prediction nodes. Decision nodes identify a collection of attributes. The branches between the nodes indicates the possible values that these attributes can have in the observed samples. Prediction nodes have a numeric score. Prediction nodes exist as both root and leaves in ADT [10].

## 1.2 *ML in Discovering Novel Inhibitors*

Discovery of novel NS5B-polymerase inhibitors was done by combining random forest, multiple e- pharmacophore modeling and docking [11]. Random Forest Modelling requires descriptors, which were calculated with Dragon 6.0. The “random-Forest” package in R aids the construction of random forest models, which is essential for the classification of HCV NS5B polymerase inhibitors and non-inhibitors. RF method produces large number of decision trees and the ensemble learning method is employed for the classification of samples. Data which are not used for building tree is termed as Out Of Bag (OOB) data. OOB data that gives an internal validation of RF that are used to elucidate the prediction accuracy of the RF model.

### 1.2.1 E-Pharmacophore Generation and Validation

E-pharmacophore models were built by considering co-crystal structures of HCV NS5B polymerase (PDB ID: 3HHK, 3SKA, 2BRK, 4DRU, 2GIR and 3PHE). Discovery of higher affinity ligands were achieved by building pharmacophore models of co-crystal ligands, exhibiting affinity values ranging from 2.4 nM to 140 nM for all three regions of NS5B polymerase. The pharmacophore sites were

produced using the six chemical features: hydrogen-bond acceptor (A), hydrogen-bond donor (D), hydrophobic (H), negative ionizable (N), positive ionizable (P), and aromatic ring (R).

### **1.2.2 Molecular Docking**

Molecular docking provides significant insights to the binding mode and interaction of ligands with target protein. The co-crystal structures (PDB ID: 3HHK, 3SKA, 2BRK, 4DRU, 2GIR and 3PHE) employed for pharmacophore generation was used to generate the energy grid.

### **1.2.3 Virtual Screening**

RF-based virtual screening (RB-VS), the e-pharmacophore-based virtual screening (PB-VS) and the docking-based virtual screening (DB-VS) methods were employed for the identification of novel NS5B polymerase inhibitors. The RB-VS stage, aids the screening of a chemical library, including 441,574 compounds from the InterBioScreen database. The compounds that passed through the RB-VS stage were processed by a second filtering of PB-VS. In the PB-VS stage, screening is based on the matching hypothesis. Molecules were required to match each site in the hypothesis. Further, DB-VS stage, is associated with docking methods to screen the compounds.

## **1.3 *ML in Insilco ADMET Prediction***

Pharmaco-Kinetics (PK) studies include elucidation of safety and tolerance of drugable compounds [12]. This is achieved by considering Absorption, Distribution, Metabolism and Toxicity Clinical evaluation of ADME include detection of bioavailability, half-life and Plasma Protein Binding [13] But ML based approaches are successful in identifying ADME properties of xenobiotics. ML techniques such as Partial Least Squares, Multiple Linear Regression and Decision Trees plays significant role in identifying the relationships between structural attributes of chemical compounds with their PK properties [14].

### **1.3.1 Multiple Linear Regression**

Multiple Linear Regression method derives the relationship between independent variables and dependent variable [15]. This is achieved by fitting a linear equation to the experimental data. Even though this method is simple, robust and easy to

comprehend, only small data sets could be analyzed by this method. MLR is not applicable to data sets that have orthogonal descriptors. Hence the Partial Least Square method is employed to model datasets with orthogonal descriptors [16].

### ***1.4 ML in Prediction of Secondary Structure and Binding Site***

The knowledge about protein structures aids in understanding its functions in great depth. Stability and biological activity of a protein are highly related to protein folding characteristics [17]. Defined prediction of protein secondary structure helps in understanding protein folding. Circular Dichroism, which is an analytical technique that is employed in the identification of secondary structural elements (helix, sheet, and turns) in proteins [18]. Machine learning algorithms also helps in the prediction of secondary structural features in protein. An ensemble technique, based on two stochastic supervised machine learning algorithms such as Maximum Entropy Markov Model (MEMM) and Conditional Random Field (CRF) was built for protein secondary structure prediction [19].

### ***1.5 ML in Docking and Virtual Screening***

Virtual Screening is a technique, which allows searching small molecule libraries to identify the best binders for a target protein. Machine learning approaches such as Neural Networks, Support Vector Machines (SVM) and Random Forest (RF) are used in virtual screening applications [20]. Accuracy of docking was enhanced by introducing machine-learning based scoring functions.

## **2 Machine Learning Techniques in Hepatitis C Virus (HCV) Drug Discovery**

### ***2.1 Random Forest Method***

Random forest is a machine learning algorithm proposed by Leo Breiman in 2000's. This method aids for building a predictor ensemble with a set of decision trees that grow in randomly selected subspaces of data. The conception of random forest ideology by Breiman was influenced by work done by Amit and Geman [21] on geometric feature selection, the random subspace method of Ho [22] and the random split selection approach proposed by Dietterich [23]. According to Breiman's approach, each tree in the collection is formed by first selecting at random, at each node, a small group of input coordinates (also called features or variables hereafter)

to split on and, secondly, by calculating the best split based on these features in the training set. The tree is grown using CART methodology to maximum size, without pruning. This subspace randomization scheme is blended with bagging [24] to resample, with replacement, the training data set each time a new individual tree is grown. Random Forest methodology is fast and easy to implement, produce highly significant predictions, ability to handle very large number of input variables and offer resistance to overfitting [25]. Random Forest methodology was used to generate predictive models for disease progression in chronic HCV patients [26]. The RF model generated provide insights to intensity of clinical monitoring required and provide prognostic information to patients. Predictor variables used in this study include, demographics, viral characteristics, clinical characteristics (including relevant comorbidities), laboratory test results and histology. The genotyping of HCV was also achieved by implementing RF method [27].

## 2.2 *Support Vector Machine (SVM) Method*

Classification of data plays significant role in protein structure predictions, microarray gene expression, Ligand based studies such as QSAR. Classification methods were based on traditional statistics. But, this method is applicable only for sample size tending to infinity. Classification of finite samples has been achieved by implementing machine learning methodologies. SVM is a powerful machine learning methodology proposed in early 90s by Vapnik. SVM is a supervised learning method, which could be employed for classification and regression. A Concurrent minimization of the empirical classification error and maximization the geometric margin is considered as an exceptional property of SVM. Hence, SVM is termed as Maximum Margin Classifiers. The algorithm of SVM is based on the Structural risk Minimization (SRM) [28]. SVM map input vector to a higher dimensional space where a maximal separating hyperplane is constructed. Two parallel hyperplanes are constructed on each side of the hyperplane that separate the data. The separating hyperplane is the hyperplane that maximize the distance between the two parallel hyperplanes. An assumption is made that the larger the margin or distance between these parallel hyperplanes the better the generalization error of the classifier will be [29].

SVM is applied in the most difficult classification schemes such as text categorization, hand-written digit recognition, image classification and object detection, tone recognition, microarray gene expression data analysis and data classification. The parameters in SVM algorithm include choice of kernel functions, the standard deviation of the Gaussian kernel, relative weights associated with slack variables to account for the non-uniform distribution of labeled data, and the number of training examples [30].

Classification of HCV genotypes were done by implementing SVM methodology. SVM algorithm exploits a set of positively and negatively labeled training vectors to develop a classifier that could be used to classify new unlabeled test samples. SVM models were generated for nucleotide positions of HCV regions such as 5'

**Table 1** Machine learning algorithms implemented for HCV research

Machine learning method	HCV research
Random forest	Generate predictive models by incorporating longitudinal data Create models to predict hepatitis virus immunoassay outcome HCV genotyping using statistical classification approach Identification of HCV NS5B polymerase inhibitors Network analysis of HBV & HCV induced hepatocellular carcinoma
Support vector machine (SVM)	Classification of HCV NS5B polymerase inhibitors Assay features of HCV's amino acids using SVM algorithm Binding site prediction in HCV protein complexes using SVM Classification models of HCV NS3 protease inhibitors based on SVM Hepatitis virus immunoassay outcome predictions using SVM Diagnosis of liver disease caused by HCV infection by SVM algorithm Classification of HCV genotypes by SVM algorithm Prediction of interferon efficiency in HCV treatment by SVM algorithm
Decision tree	Prediction of responsiveness of HCV patients in drug treatment. Prediction of advanced liver fibrosis in HCV patients Comparison of treatment strategies of HCV treatment Prediction of HCV response to treatment strategies
Genetic algorithm	QSAR study of HCV NS5B polymerase inhibitors Hybrid genetic algorithm for making treatment decisions in HCV
k-nearest neighbours	Molecular field analysis on human HCV NS5B polymerase inhibitors: 2,5-disubstituted imidazole[4,5-c]pyridines 3D-QSAR study on benzimidazole derivatives acting as HCV inhibitors
Naïve Bayesian classifier	Prediction of liver disease using Naïve Bayesian algorithm Hepatitis disease diagnosis using Naïve Bayesian classifiers
Particle swarm optimization	Classification system for HCV diagnosis

NCR, CORE, E1 and NS5B to predict most common HCV subtypes and genotypes. Models generated for features encoding NS5B and E1 regions tend to possess more predictive power than the other regions.

Table 1: Depicts machine learning algorithms implemented for HCV research.

### 2.3 Decision Tree

A decision tree is a relevant classification method in data mining classification. Classification, description, and generalization of data has been done by employing Decision trees. Numerous disciplines such as signal processing, pattern recognition,

decision theory and statistics utilize the application of Decision trees [31]. The algorithm relies on the construction of a decision tree from a training set TS, which is a set of cases, or tuples in the database terminology. Each case identifies values for a collection of characteristics and for a class. Each characteristic may have either discrete or continuous values. Moreover, the special value unknown is allowed to denote unspecified values. The class may have only discrete values [31]. Decision Tree was employed to predict responsiveness of HCV patients to drug treatment. Decision tree model was generated based on Single Nucleotide Polymorphisms (SNPs) calculated in a genome wide association study. Test subjects used for generating decision tree learning include 142 Japanese patients with HCV genotype 1. Decision tree thus obtained is able to predict with high probability whether or not a new patient will be helped by the recommended treatment [32].

## 2.4 *Genetic Algorithm*

Genetic algorithms belong to stochastic search algorithms, which works on a population of possible solutions. Genetic Algorithms are based on mechanism of population genetics and selection. New solutions are obtained by considering genetic operator functions such as mutation and cross over. Hence, from a population of probable solutions, better solutions are made [33]. Genetic algorithms help in identifying optimal solution for a given research problem and the methodology is easy to implement and comprehend.

The combination of Genetic Algorithm and Multiple Linear Regression was used to identify physico-chemical features of chemical compounds inhibiting HCV NS5B polymerase. In this study, 72 derivatives of indole-5-carboxamide were considered. Feature selection was accomplished by Genetic Algorithm and regression analysis was done by MLR [34]. Genetic Algorithm finds potential applications in QSAR and molecular docking studies.

## 2.5 *K-Nearest Neighbors*

KNN classification method was proposed by Fix and Hodges [35]. The classification scheme employed by KNN algorithm is based on closest training examples in the feature space. KNN is classified as instance-based learning, or lazy learning where the function is only approximated locally and all computation is delayed until classification [36]. The KNN is the essential and modest classification technique employed when there is insufficient knowledge about data distribution. This rule is retained the entire training set during learning and assigns to each query a class represented by the majority label of its k-nearest neighbors in the training set. The Nearest Neighbor rule (NN) is the simplest form of KNN when K = 1. In this method each sample is classified based on the correspondingly to its surrounding samples. Hence, classification of unknown sample, could be predicted by studying the

classification of its nearest neighbor samples. Given an unknown sample and a training set, all the distances between the unknown sample and all the samples in the training set can be computed. The distance with the smallest value corresponds to the sample in the training set closest to the unknown sample. The performance of KNN is analyzed by considering choice of k and the distance metric applied [37].

The prediction of in-hospital mortality in HCV patients undergone primary liver cancer surgery was done by neural network based approach. The test subjects for this study include patients who undertook liver surgery during the period from 1998 to 2009. Neural network based model generated from this test subjects were evaluated by considering the area under the receiver operating characteristic (AUROC) curves and Hosmer-Lemeshow (H-L) statistics. The accuracy rate was calculated and compared using paired T-tests. Significance of input parameters in the system model and variables were done by conducting global sensitivity analysis [38].

## 2.6 *Naïve Bayesian Classifier*

A Naïve Bayesian classifier is an efficient supervised learning method which relies on Bayes' theorem. This approach is devoted to derive a probabilistic classification model from previously evaluated characteristic features [39]. One of the most prominent application of Naïve Bayesian Classifier is applied for antispam mail filtering, which trains the filter to automatically separate spam mail and legitimate messages in a binary manner. Vijayarani et.al, utilized naïve Bayesian classifiers as well as SVM to classify major liver diseases such as liver cancer, cirrhosis and hepatitis based on observed symptoms. The classification results confirmed the accuracy of Naïve Bayesian approach compared to SVM algorithm. This approach was also utilized to identify risk factors involved in the onset and progression of HCV disease [40].

## 2.7 *Particle Swarm Optimization*

Particle Swarm Optimization (PSO) was first explained by James Kennedy and Russel Eberhart in 1995 [41]. PSO is a swarm-intelligence-based method. This method is termed to be approximate and non-deterministic. The ideology of PSO is derived from observation of swarming habits of birds or fish and evolutionary computation. Optimization techniques employed in Particle Swarm method identifies parameters which provide maximum or minimum value of a target function. PSO algorithm is produces multiple solutions at a time. Solutions thus generated are evaluated by an objective function to determine the fitness. Each solution is represented by a particle in the search space. The particles swarm in search space to find optimal solution. Each solution in search space is identified by its position, velocity and individual best position. Modified PSO was utilized for the classification of HCV infected patients. This classification scheme employs the usage of convergence factor, inertia weight, position and velocity of particle.

## 2.8 *Multilinear Regression (MLR) Models*

MLR method explores correlation between predictor and response variables. Response variables address the effect of predictor variables in target function [42]. MLR uses two or more predictor variables to obtain outcome. MLR is simple and easy to interpret. But MLR method is vulnerable to descriptors which are correlated to one another, making it incapable of deciding which correlated sets may be more significant to the model. New methodologies of MLR include Best Multiple Linear Regression (BMLR), Heuristic Method (HM), Genetic Algorithm based Multiple Linear Regression (GA-MLR), Stepwise MLR, Factor Analysis MLR. BMLR is instrumental for variable selection and QSAR/QSPR modelling.

QSAR studies were reported for HCV NS3/4A protease inhibitors by MLR and SVM based methodologies. The bioactivity of NS3/4A protease inhibitors were reflected from the correlation coefficient. Correlation coefficient defined by MLR for training and test set was 0.87 and 0.85 respectively. Another QSAR study was reported to identify physicochemical properties of HCV NS5B polymerase by employing combination of Genetic Algorithm and MLR (GA-MLR) [43].

## 3 Conclusions

Machine Learning algorithms play vital role in significant aspects of drug discovery such as classification, regression, virtual screening, secondary structure prediction and ADMET prediction. Machine Learning algorithms are implemented in HCV research to discover novel therapeutic agents which could circumvent drug resistance acquired by virus. Machine Learning algorithms such as Random Forest, Support Vector Machine (SVM), Genetic algorithm, Decision Tree, k-Nearest neighbors, Particle Swarm Optimization and Naive Bayesian Classifier are reported for HCV research. Machine Learning approaches were utilized for HCV genotyping, classification of HCV inhibitors, immunoassay outcome predictions, binding site predictions and prediction of interferon efficiency in the treatment of HCV infected patients. Machine learning algorithms produce acceptable output values by performing leverage statistical analysis. Hence, Machine Learning algorithms could be implemented to resolve complex and dynamic problems in Biology.

## References

1. Schuhmacher A, Gassmann O, Hinder M. Changing R&D models in research-based pharmaceutical companies. *J Transl Med.* 2016;14:105.
2. Imming P, Sinning C, Meyer A. Drugs, their targets and the nature and number of drug targets. *Nat Rev Drug Discov.* 2006;5:821–34.
3. Gashaw I, Ellinghaus P, Sommer A, Asadullah K. What makes a good drug target? *Drug Discov Today.* 2011;16:1037–43.

4. Costa PR, Acencio ML, Lemke N. A machine learning approach for genome-wide prediction of morbid and druggable human genes based on systems-level data. *BMC Genomics*. 2010;11:S9.
5. Chen H, Engkvist O, Wang Y, Olivecrona M, Blaschke T. The rise of deep learning in drug discovery. *Drug Discov Today*. 2018;23:1241–50.
6. Zhu M, Gao L, Li X, Liu Z, Xu C, Yan Y, et al. The analysis of the drug–targets based on the topological properties in the human protein–protein interaction network. *J Drug Target*. 2009;17:524–32.
7. Jeon J, Nim S, Teyra J, Datti A, Wrana JL, Sidhu SS, et al. A systematic approach to identify novel cancer drug targets using machine learning, inhibitor design and high-throughput screening. *Genome Med*. 2014;6:57.
8. Bolton EE, Wang Y, Thiessen PA, Bryant SH. PubChem: integrated platform of small molecules and biological activities. *Annu Rep Comput Chem*. 2008;4:217–41.
9. Barros RC, Basgalupp MP, de Carvalho ACPLF, Freitas AA. Automatic design of decision-tree algorithms with evolutionary algorithms. *Evol Comput*. 2013;21:659–84.
10. Hashem S, Esmat G, Elakel W, Habashy S, Abdel Raouf S, Darweesh S, et al. Accurate prediction of advanced liver fibrosis using the decision tree learning algorithm in chronic hepatitis C Egyptian patients. *Gastroenterol Res Pract*. 2016;2016:1–7.
11. Wei Y, Li J, Qing J, Huang M, Wu M, Gao F, et al. Discovery of novel hepatitis C virus NS5B polymerase inhibitors by combining random forest, multiple e-pharmacophore modeling and docking. *PLoS One*. 2016;11:e0148181.
12. Barton HA, Pastoor TP, Baetcke K, Chambers JE, Diliberto J, Doerrer NG, et al. The acquisition and application of absorption, distribution, metabolism, and excretion (ADME) data in agricultural chemical safety assessments. *Crit Rev Toxicol*. 2006;36:9–35.
13. Vrbanac J, Sauter R. ADME in Drug Discovery, A Comprehensive Guide to Toxicology in Nonclinical Drug Development (2nd Ed)2017;39–67
14. Matarollo VG, Gertrudes JC, Oliveira PR, Honorio KM. Applying machine learning techniques for ADME-Tox prediction: a review. *Expert Opin Drug Metab Toxicol*. 2015;11:259–71.
15. Alexopoulos EC. Introduction to multivariate regression analysis. *Hippokratia*. 2010;14:23–8.
16. Cramer RD. Partial least squares (PLS): its strengths and limitations. *Perspect Drug Discovery Des*. 1993;1:269–78.
17. Yon JM. Protein folding: a perspective for biology, medicine and biotechnology, *Braz J Med Biol Res*, April 2001;34(4):419–435.
18. Greenfield NJ. Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc*. 2006;1:2876–90.
19. Muggleton S, King RD, Sternberg MJE. Protein secondary structure prediction using logic. *Protein Eng*. 1992;7:647–57.
20. Lavecchia A. Machine-learning approaches in drug discovery: methods and applications. *Drug Discov Today*. 2015;20:318–31.
21. Amit Y, Geman D. Shape quantization and recognition with randomized. *Trees*. 1997;9:1545–88.
22. Ho TK. The random subspace method for constructing decision forests. *IEEE Trans Pattern Anal Mach Intell*. 1998;20:832–44.
23. Dietterich TG. An experimental comparison of three methods for constructing ensembles of decision trees: bagging, boosting, and randomization. *Mach Learn*. 2000;40:139–57.
24. Buja A, Stuetzle W. Observations on bagging. *Stat Sin*. 2006;16:323.
25. Ziegler A, König IR. Mining data with random forests: current options for real-world applications. *Wiley Interdiscip Rev Data Min Knowl Discov*. 2014;4:55–63.
26. Zhang Y, Lok ASF, Higgins PDR, Konerman MA, Waljee AK, Zhu J. Improvement of predictive models of risk of disease progression in chronic hepatitis C by incorporating longitudinal data. *Hepatology*. 2015;61:1832–41.
27. Ping Qiu, Xiao-Yan Cai, Wei Ding, Qing Zhang, Ellie D Norris, and Jonathan R Greene, HCV genotyping using statistical classification approach, *J Biomed Sci*. 2009; 16(1): 62.
28. Srivastava DK, Lekha B. Data classification using support vector machine. *J Theor Appl Inf Technol*. 2005;12:1–7.

29. Understanding Support Vector Machine algorithm from examples (along with code). Available at <https://www.analyticsvidhya.com/blog/2017/09/understaing-support-vector-machine-example-code/>.
30. Chapter 2: SVM (Support Vector Machine)—Theory—Machine learning 101—Medium. Available at <https://medium.com/machine-learning-101/chapter-2-svm-support-vector-machine-theory-f0812effc72>.
31. Kareem IA, Duaimi MG. Improved accuracy for decision tree algorithm based on unsupervised discretization. *Int J Comput Sci Mob Comput.* 2014;36:176–83.
32. Kawamura Y, Takasaki S, Mizokami M. Using decision tree learning to predict the responsiveness of hepatitis C patients to drug treatment. *FEBS Open Bio.* 2012;2:98–102.
33. Shapiro J. Genetic algorithms in machine learning. Berlin, Heidelberg: Springer; 2001. p. 146–68.
34. Rafiee H, Khanzadeh M, Mozaffari S, Bostanifar MH, Avval ZM, Aalizadeh R, et al. QSAR study of HCV NS5B polymerase inhibitors using the genetic algorithm-multiple linear regression (GA-MLR). *EXCLI J.* 2016;15:38–53.
35. Fix E, Hodges JL. Discriminatory analysis. Nonparametric discrimination: consistency properties. *Int Stat Rev/Rev Int Stat.* 1989;57:238.
36. Mitchell TM. Instance-based Learning, Machine Learning. McGraw-Hill publishers, ISBN: 0070428077 (March 1, 1997).
37. Chomboon K, Chujai P, Teerarassamme P, Kerdprasop K, Kerdprasop N. An empirical study of distance metrics for k-nearest neighbor algorithm. In: The proceedings of the 2nd international conference on industrial application engineering 2015; 2015, p. 280–285.
38. Shi H-Y, Lee K-T, Lee H-H, Ho W-H, Sun D-P, Wang J-J, et al. Comparison of artificial neural network and logistic regression models for predicting in-hospital mortality after primary liver cancer surgery. *PLoS One.* 2012;7:e35781.
39. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, et al. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J Chem Inf Model.* 2006;46:462–70.
40. Vijayarani S, Dhayanand S. Liver disease prediction using SVM and Naïve Bayes algorithms. *Int J Sci Eng Technol Res.* 2015;4:816–20.
41. Kennedy J, Eberhart R. Particle Swarm Optimization, Computational Intelligence PC Tools, 1996 by Academic Press Professional (APP).
42. Salleh FHM, Zainudin S, Arif SM. Multiple Linear Regression for Reconstruction of Gene Regulatory Networks in Solving Cascade Error Problems, Advances in Bioinformatics, 2017, 1–15.
43. Qin Z, Wang M, Yan A. QSAR studies of the bioactivity of hepatitis C virus (HCV) NS3/4A protease inhibitors by multiple linear regression (MLR) and support vector machine (SVM). *Bioorg Med Chem Lett.* 2017;27:2931–8.

# Modern Developments in Short Peptide Viral Vaccine Design



Christina Nilofer, Mohanapriya Arumugam, and Pandjassarame Kanguane

**Abstract** Vaccine design and development against viral diseases is multifaceted. Classical vaccines (live-attenuated vaccines, inactivated vaccines, subunit, recombinant, polysaccharide, conjugate vaccines, and toxoid vaccines) are often less effective for many viral diseases. Hence, short peptide (10–20 amino acid residues) vaccine components exploiting T-cell mediated immunity have been recognized as alternative solutions. This involves the specific binding of short antigen peptides to allele (gene variant) specific host human leukocyte antigens (HLA). Allele-specific HLA typing among different ethnic groups has gained momentum in recent years through advances in sequencing (Next Generation Sequencing (NGS)), High Performance Computing (HPC), machine learning techniques such as Artificial Neural Networks (ANN) and Support Vector Machine (SVM) techniques. More than 20,000 HLA alleles have been typed, defined, named, and made available at the IMGT®/HLA database (Immuno Polymorphism Database – ImMunoGeneTics Database/Human Leukocyte Antigen) for public access. Identification of short peptide antigens capable of binding specifically to the human host HLA alleles is now possible using computer-aided HLA-peptide binding prediction methods. This is achieved using three dimensional HLA structure based molecular modeling, and known HLA-peptide binding data enabled machine learning techniques like ANN and SVM. The former provides broad coverage across HLA alleles and the later offers high accuracy with high specificity for limited HLA alleles. Thus, the

---

C. Nilofer

Peptide Vaccine Design, Biomedical Informatics (P) Ltd,  
Iulan Sandy Annex, Puducherry, India

Biotechnology, School of Bio Sciences and Technology, VIT University,  
Katpadi, Vellore, Tamil Nadu, India

M. Arumugam

Biotechnology, School of Bio Sciences and Technology, VIT University,  
Katpadi, Vellore, Tamil Nadu, India

P. Kanguane (✉)

Peptide Vaccine Design, Biomedical Informatics (P) Ltd,  
Iulan Sandy Annex, Puducherry, India  
e-mail: [kanguane@bioinformation.net](mailto:kanguane@bioinformation.net)

combined use of structural features, molecular modeling, machine learning techniques, and other applied mathematical models including Quantitative matrices (QM), Bayesian Networks (BN), and Hidden Markov Models (HMM) help in the effective design of short peptide vaccine components and immune therapeutics for the prevention and control of diseases caused by viruses. Hence, we outline recent advances in HLA-peptide binding prediction for short peptide vaccine design.

**Keywords** Epitope · T-cell receptor · NGS · HPC · SVM · HMM · ANN · Quantitative matrix · HLA typing · Sequence · Nomenclature · ANN · Vaccine · Short vaccine peptide · Vaccine design · Alleles · Polymorphism · Molecular modeling · Structure

## Abbreviations

ANN	Artificial Neural Network
EA	Evolutionary Algorithm
HLA	Human Leukocyte Antigens
HMM	Hidden Markov Model
HPC	High Performance Computing
IMGT	the international ImMunoGeneTics information systemdatabase
NGS	Next Generation Sequencing
QM	Quantitative Matrices
QSAR	Quantitative Structure Activity Relationship
SVM	Support Vector Machine

### Core Message

Short peptide (10–20 amino acids long) viral vaccine design, exploiting T-cell mediated immunity (immune response by the activation of T cells) of the human host is now possible through the identification of epitopes (antigenic determinants capable of stimulating an immune response) using HLA-peptide binding prediction methods. This is precisely feasible by the combined use of known three-dimensional HLA-peptide structures, peptide binding pocket features, molecular modeling methods and machine learning techniques including Artificial Neural Network (ANN) and Support Vector Machine (SVM). This is also possible using other mathematical models such as Quantitative Matrices (QM), Bayesian networks (BN), and Hidden Markov Models (HMM). Thus, the accurate design of short peptide vaccine constituents and immune therapeutics for the prevention and control of viral diseases is promising using computer-aided mathematical models. We outline current development and progress in HLA-peptide binding prediction for epitope design towards short peptide vaccine development in this chapter.

## 1 Introduction

Short antigen peptides (10–20 amino acid residues long) capable of binding host HLA molecules in an allele (gene variant) specific manner through the exploitation of T-cell immunity (utilizing cytotoxic (CD8<sup>+</sup>) and helper (CD4<sup>+</sup>) T-cells) is currently possible [1–4]. This is an alternate solution where live-attenuated vaccines, inactivated vaccines, subunit, recombinant, polysaccharide, conjugate vaccines and toxoid vaccines are not effective for many viral diseases. The design of short peptide epitopes is accelerated through HLA-peptide binding prediction using computer-aided structure and applied mathematical models. It should be noted that HLA molecules are highly polymorphic (sequence dissimilarity) and the peptide binding groove significantly varies among alleles for specific binding [5–11].

## 2 HLA Genes, Typing, and Alleles

Human HLA genes are located on the short arm (p21) of chromosome 6 and are highly polymorphic. They are classified into three classes, HLA classes I, II, and III. HLA classes I & II are further sub-grouped into three and six genes, respectively. They are HLA-A, HLA-B, HLA-C of HLA class I and HLA-DPA1, HLA-DPB1, HLA-DQA1, HLA-DQB1, HLA-DRA, HLA-DRB of HLA class II [12]. The HLA class I processes and presents intracellular peptides and HLA class II presents extracellular peptides to the immune system. HLA class III genes trigger inflammation and other immune responses [13].

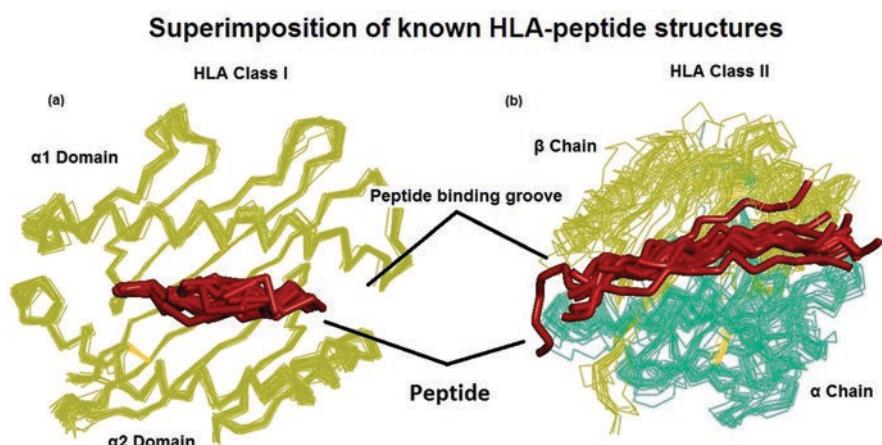
Techniques used to perform HLA typing includeserological cytotoxicity (conventional method) [14], flow cytometry, Polymerase Chain Reaction (PCR), and Next Generation Sequencing (NGS) [15]. Among them, PCR and NGS techniques are widely used for HLA typing of individual blood samples. The immune system uses HLA protein markers to differentiate between self- and non-self proteins. Thus, HLA typing is important for transplants between donor and recipient. Identical HLA markers between the donor and the recipient are critical for transplants, to circumvent graft versus host disease (GvHD). Additionally, minor histocompatibility (miHA) antigen peptides among identical HLA donor-recipient pairs are important in the context of bone marrow transplants (BMT) and GvHD [16].

Sequences of known HLA alleles are available at IPD-IMGT®/HLA (Immuno Polymorphism Database – ImMunoGeneTics Database/Human Leukocyte Antigen) maintained by EMBL-EBI [17]. There are approximately 20,088 HLA alleles (as of December 2018), among which, 14,800 are HLA class I and 5288 are HLA class II alleles. The extensive diversity of HLA alleles allows the immune system to fight a wide range of viral diseases through T-cell immunity (utilizing cytotoxic (CD8<sup>+</sup>) and helper (CD4<sup>+</sup>) T-cells).

### 3 HLA-Peptide Binding

Short antigen peptides bind HLA molecules with high specificity utilizing T cell mediated immune response using T cytotoxic and T helper cells. The specific binding of antigen peptides with allele specific HLA molecules is critical for T-cell mediated immune response. Therefore, it is important to understand the molecular principles of HLA-peptide binding. Our understanding of HLA-peptide binding has improved considerably using known HLA-peptide structures and their analysis [18]. The binding of peptides to class I and class II HLA alleles is illustrated in Fig. 1 generated using known HLA-peptide structure complexes given in Table 1 [19]. The physical and chemical features of HLA-peptide binding are studied using known structures accumulated in databases [20].

Peptide binding patterns are different in class I and class II HLA alleles. The peptides have a limited conformation in the class I groove, while they have an extended conformation in the class II groove. However, the binding interaction between the HLA molecule and peptide is predominantly van der Waals (vdW), specifically supported by hydrogen bonds and electrostatics calculated using the software PPCHECK [21, 22] as shown in Fig. 2. It is also known that the interactions are limited to the groove between HLA and the bound peptide. These interactions are mainly between sidechain – sidechain of HLA and peptide [23].



**Fig. 1** Superimposition of known HLA-peptide structures is shown for 64 HLA-peptide complexes. This is generated using the software tool Discovery Studio Visualizer™ version 16.1. The data used for superimposition is given in Table 1

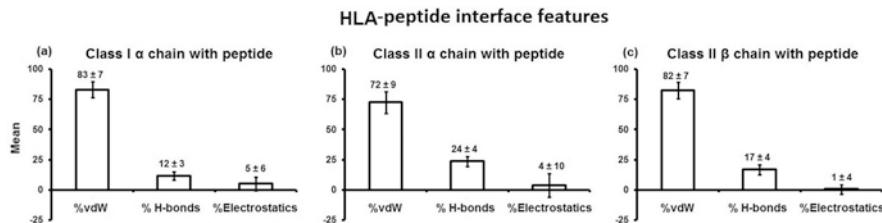
**Table 1** Known HLA-peptide structural complexes available at the RCSB – Protein Databank (PDB) [19]

PDB ID	Class	HLA allele	Peptide sequence
1W72	Class I	A*0101	EADPTGHSY
3BO8	Class I	A*0101	EADPTGHSY
1BD2	Class I	A*0201	LLFGYPVYV
1DUZ	Class I	A*0201	LLFGYPVYV
1AKJ	Class I	A*0201	ILKEPVHGV
1HHG	Class I	A*0201	TLTSCNTSV
1HHI	Class I	A*0201	GILGFVFTL
1HHJ	Class I	A*0201	ILKEPVHGV
1HHK	Class I	A*0201	LLFGYPVYV
1I4F	Class I	A*0201	GVYDGREHTV
3UTS	Class I	A*0201	ALWGPDPAAA
3UTT	Class I	A*0201	ALWGPDPAAA
1I7R	Class I	A*0201	FAPGFFPYL
1I7T	Class I	A*0201	ALWGVFPVL
1I7U	Class I	A*0201	ALWGFVPVL
1IM3	Class I	A*0201	LLFGYPVYV
1P7Q	Class I	A*0201	ILKEPVHGV
1QRN	Class I	A*0201	LLFGYAVYV
1QSE	Class I	A*0201	LLFGYPRYV
1QSF	Class I	A*0201	LLFGYPVAV
3FQN	Class I	A*0201	YLDSGIHSGA
3FQR	Class I	A*0201	YLDSGIHSGA
3FQT	Class I	A*0201	GLLGSPVRA
3FQU	Class I	A*0201	GLLGSPVRA
3FQW	Class I	A*0201	RVASPTSGV
3FQX	Class I	A*0201	RVASPTSGV
1JHT	Class I	A*0201	ALGIGILTV
1B0G	Class I	A*0201	ALWGFPPVL
1I1F	Class I	A*0201	FLKEPVHGV
1I1Y	Class I	A*0201	YLKEPVHGV
1AO7	Class I	A*0202	LLFGYPVYV
1X7Q	Class I	A*1101	KTFPPTEPK
2HN7	Class I	A*1101	AIMPARFYPK
3BVN	Class I	B*1402	RRRWRLTV
1HSA	Class I	B*2705	ARAAAAAAA
1JGE	Class I	B*2705	GRFAAAIAK
3BP4	Class I	B*2705	IRAAPPPLF
1JGD	Class I	B*2709	RRLLRGHNQY
1K5N	Class I	B*2709	GRFAAAIAK
1OF2	Class I	B*2709	RRKWRRWHL
3BP7	Class I	B*2709	IRAAPPPLF

(continued)

**Table 1** (continued)

PDB ID	Class	HLA allele	Peptide sequence
1ZSD	Class I	B*3501	EPLPQGQLTAY
3LN4	Class I	B*4103	AEMYGSVTEHPSPSPL
3LN5	Class I	B*4104	HEEAVSVDRVL
3DX6	Class I	B*4402	EENLLDFVRF
1SYS	Class I	B*4403	EEPTVIKKY
3DX7	Class I	B*4403	EENLLDFVRF
3DX8	Class I	B*4405	EENLLDFVRF
3DXA	Class I	B*4405	EENLLDFVRF
1.00E+27	Class I	B*5101	LPPVVAKEI
1A1M	Class I	B*5301	TPYDINQML
1A1O	Class I	B*5301	KPIVQYDNF
1A9B	Class I	B*5301	LPPLDITPY
1A9E	Class I	B*5301	LPPLDITPY
3VH8	Class I	B*57:01	LSSPVTKSF
3VRI	Class I	B*57:01	RVAQLEQVYI
3VRJ	Class I	B*57:01	LTTKLTNTNI
2RFX	Class I	B*5701	LSSPVTKSF
3UPR	Class I	B*5701	HSITYLLPV
1IM9	Class I	CW*0401	QYDDAVYKL
1EFX	Class I	CW3	GAVDPLLAL
1QQD	Class I	CW4	QYDDAVYKL
3CDG	Class I	E	VMAPRTLFL
2D31	Class I	G	RIIPRHLQL
2DYP	Class I	G	RIIPRHLQL
3KYN	Class I	G	KGPPAALTL
3KYO	Class I	G	KLPAQFYIL
2NNA	Class II	DQ8	QQYPSGEGSFQPSQENPQ
1JK8	Class II	DQ8	LVEALYLVCGERGG
4GG6	Class II	DQA1	QQYPSGEGSFQPSQENPQ
1S9V	Class II	DQA1*0501	LQPFPQPELPY
1UVQ	Class II	DQB1*0602	EGRDSMNPSTKVSWAA VGGGGSLVPRGSGGGG
1KLG	Class II	DR1	GELIGILNAAKVPAD
1KLU	Class II	DR1	GELIGTLNAAKVPAD
1ZGL	Class II	DR2	VHFFKNIVTPRTPGG
1A6A	Class II	DR3	PVSKMRMATPLLMQA
2SEB	Class II	DR4	AYMRADAAAGGA
2Q6W	Class II	DRA	AWRSDEALPLGS
1SJH	Class II	DRA*0101	PEVIPMFSALSEG
1T5W	Class II	DRA	AAYSDQATPLLLSPR
2IAN	Class II	DRA1	GELIGTLNAAKVPAD
1H15	Class II	DRA1*0101	GGVYHFVKKHVHES
2FSE	Class II	DRB1*0101	AGFKGEQGPKGEPG



**Fig. 2** The characteristic chemical features at the HLA-peptide interface are shown for 64 HLA-peptide interfaces. The interface is van der Waals (vdW) dominated with limited H-bonds and electrostatics. MHC-peptide interfaces were analyzed using an interface analysis tool named PPCCheck [19, 20]. PPCCheck identifies non-covalent interactions based on the distance between atoms of the two interacting HLA and peptide. Thus, interface size, van der Waals (vdW), electrostatics, H-bonds and total energies were calculated for each of 64 HLA-peptide interfaces (Table 1)

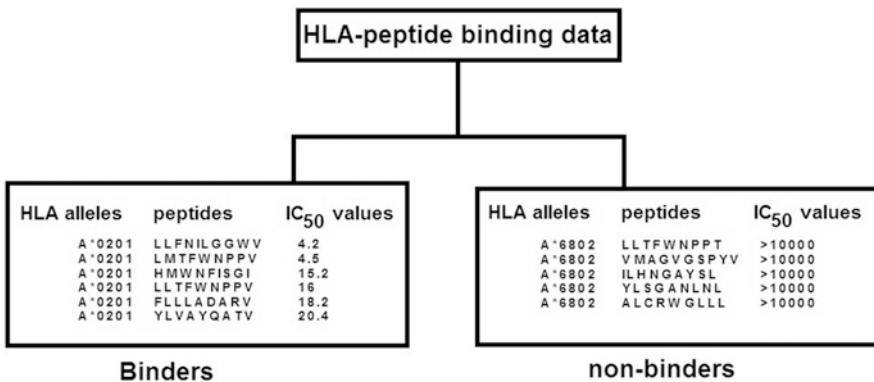
## 4 HLA-Peptide Binding Data

Allele specific HLA-peptide binding data are generated in immunological studies using competitive binding assays, measured using inhibitory concentration 50 ( $IC_{50}$ ) values. Thousands of HLA specific peptide binding data with known  $IC_{50}$  values are available in the literature controlled by several publishers including Elsevier, Springer, and John Wiley Inc. Those HLA-peptide binding data that are available in the literature are manually collected, organized, and stored in several databases such as MHCPEP [24], MHCBN [25] and SYFPEITHI [26]. These databases contains data on HLA-peptide binding and non-binding grouped based on  $IC_{50}$  values (Fig. 3). These data are useful in developing data driven models for HLA-peptide binding prediction.

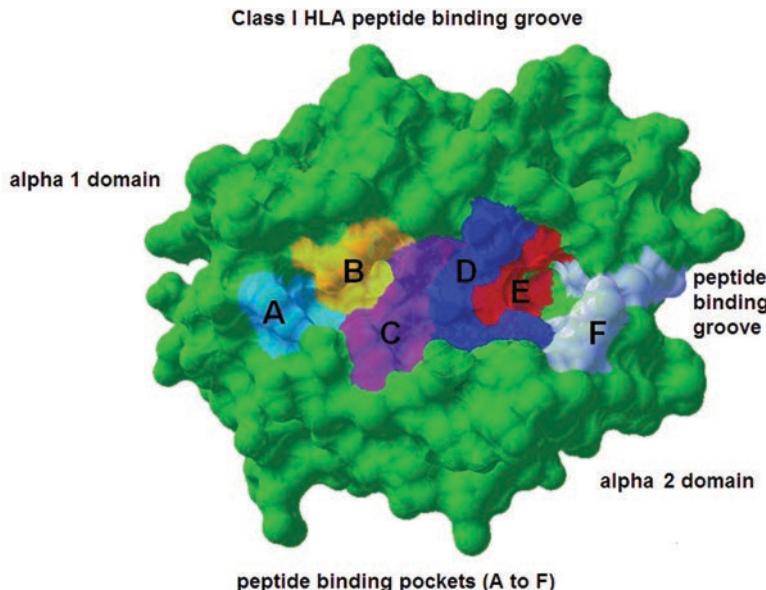
## 5 HLA Peptide Binding Groove

HLA molecules are polymorphic among alleles, and the peptide binding grooves have different amino acid compositions in the A to F pockets (Fig. 4). The pockets are distributed in several combinations across different alleles creating vast pocket diversity. However, there are identical pockets distributed across alleles reducing the functional diversity among known alleles [27]. Moreover, the grouping of peptide binding grooves using their electrostatics characteristics is contextual [28]. Thus, the peptide binding property of the groove among HLA alleles overlaps leading to functional similarity among them.

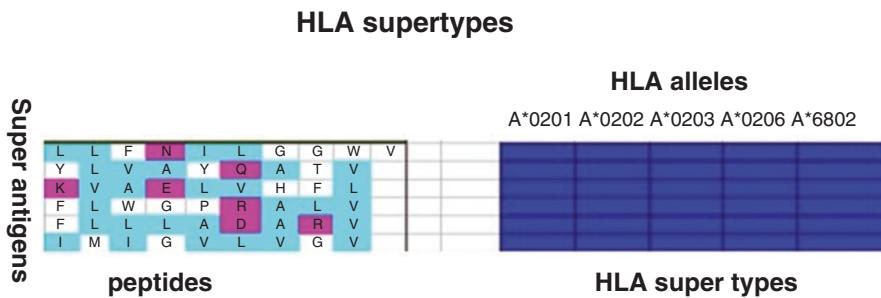
## HLA-peptide data



**Fig. 3** HLA-peptide binding data with known IC<sub>50</sub> (Inhibitory concentration 50 defined as the amount of peptide required to inhibit 50% of standard peptide) values are grouped into binders and non-binders. Peptides with low IC<sub>50</sub> values are grouped as binders and those peptides with high IC<sub>50</sub> are non-binders



**Fig. 4** Peptide binding pockets marked A to F in class I HLA A\*0201 groove is shown. The peptide binds to the groove by anchoring in these pockets. The pockets in the groove have varying shapes, sizes and chemical properties for different HLA alleles caused by polymorphism. This image was generated using Discovery Studio™



**Fig. 5** An illustration of HLA supertypes is shown. Multiple HLA alleles capable of binding a single peptide antigen exhibit HLA super types with overlapping peptide binding function. Antigen short peptides capable of binding multiple HLA alleles are super antigens. Dark blue = peptide binders; pink shades = dissimilar aminoacid; light blue = similar aminoacid

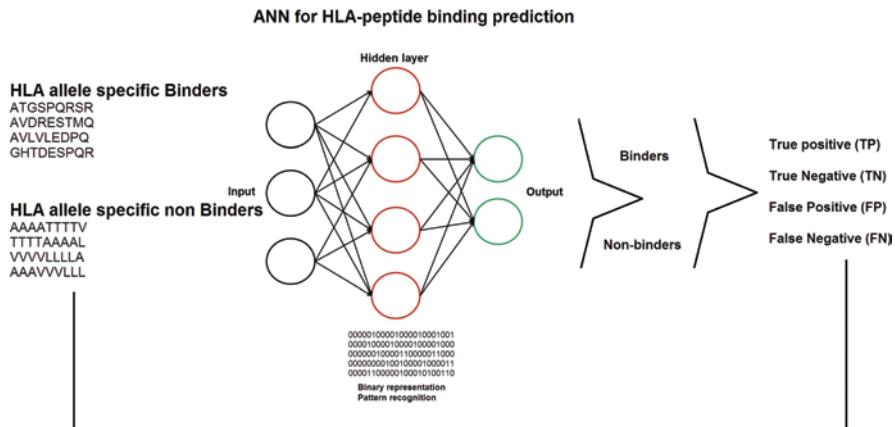
## 6 HLA Super-Types

HLA alleles are extremely polymorphic among ethnic groups, and consequently, the peptide binding specificity varies. However, the majority of alleles is disbursed within a few HLA supertypes, where different members of a supertype family bind similar peptides, and also exhibit distinct repertoires (Fig. 5). A framework for grouping alleles into supertypes from only sequence information is available [29, 30]. The structural basis for HLA supertypes is also demonstrated [31]. The grouping of HLA alleles into different categories of supertypes is important in the understanding of antigenic peptide selection, and discrimination during T-cell mediated immune response. This has utility in epitope design for the development of HLA based vaccines and immune therapeutics.

## 7 Machine Learning Algorithms for HLA-Peptide Binding Prediction

Machine learning techniques such as Artificial Neural Network (ANN) and Support Vector Machine (SVM) provide the ability to routinely learn and improve from experience (known data) without being explicitly programmed. It uses computer programs to access data for learning by intelligence (cleverness).

Artificial Neural Network (ANN) is a well-trained communicating network made of nodes (artificial neurons) and edges (connecting links). Signals pass from one node to another when the strength (weight) of the signal crosses a particular threshold value. Each signal navigates through multiple middle (the hidden) layers to perform different functions between the input and the output layers as shown in Fig. 6. ANN is a framework of learning algorithms to process complex input data, to obtain the desired output. ANN is a data-driven model, trained to recognize



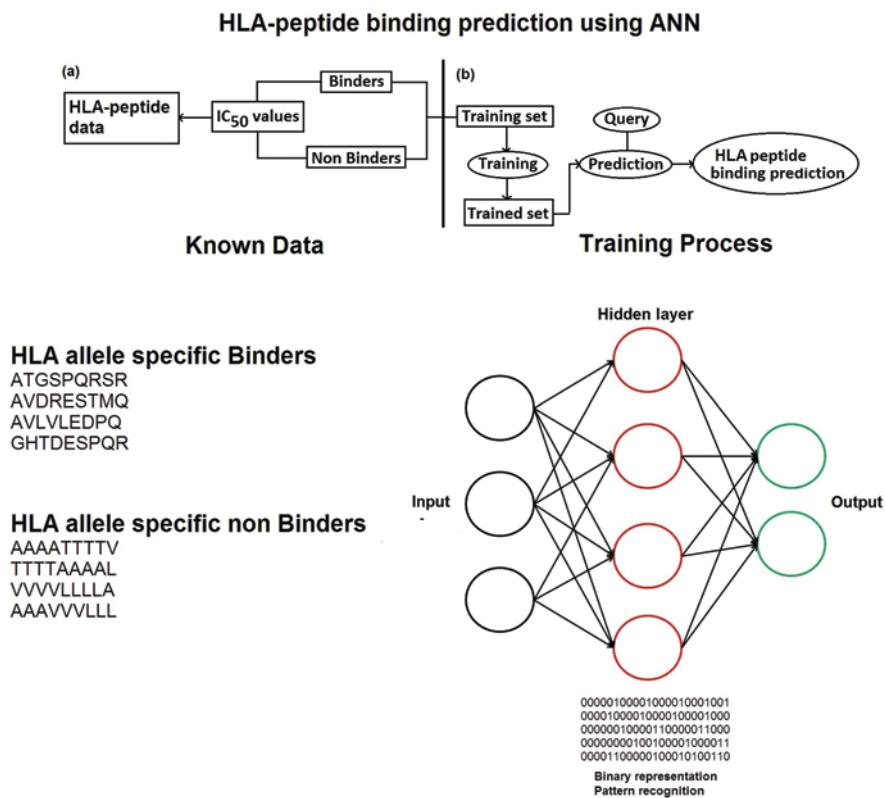
**Fig. 6** An Artificial Neural Network (ANN) is an interrelated group of nodes. Each node is connected by an arrow from the output of one to the input of another. The hidden layer is the trained (taught) layer created using known data having specific patterns representing a natural phenomenon. This is illustrated using known HLA-peptide binders and non-binders. The hidden layer is the trained layer with known patterns specific for HLA binders and non-binders

patterns and relationships based on various complex inputs [32]. The application of ANN in HLA-peptide binding prediction is illustrated in Fig. 7. It uses known groups of HLA specific peptide binders and non-binders to train the ANN model for prediction through pattern recognition.

Support Vector Machine (SVM) also known as Support Vector Networks is a supervised machine learning algorithm mainly used for classification and regression analysis of data [33, 34]. SVM consists of two kinds of classifiers namely linear classifier and non-linear classifiers (Fig. 8). The linear classifier is used in two-dimensional space separating the input data into different groups and non-linear classifiers are used for multi-dimensional space. Prediction of HLA-peptide binding is achieved by classifying query peptides into binders and non-binders using SVM models.

## 8 HLA-Peptide Binding Prediction

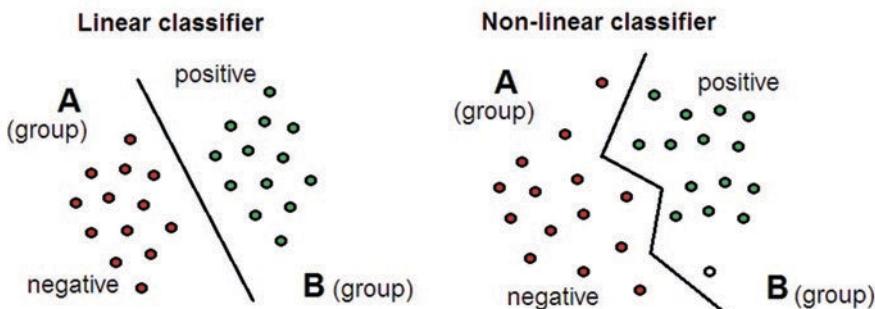
HLA-peptide binding prediction is an important step in short peptide epitope design for vaccine development. Identification of short peptide antigens capable of binding specifically to human HLA alleles is now possible using computer-aided HLA-peptide binding prediction methods [35–37]. This is achieved using three dimensional HLA structure based molecular modeling and HLA-peptide binding data enabled machine learning techniques like ANN, SVM, BN, HMM, QM and PDF (Fig. 9).



**Fig. 7** HLA peptide binding prediction using ANN is shown. (a) The binding and non-binding peptides are identified using HLA-peptide binding assay using IC<sub>50</sub> values. (b) The information on binding and non-binding peptides is used to train ANN to identify the binding ability of each query peptide. Therefore, use of ANN in HLA peptide binding prediction minimize cost and accelerate the screening process to assist laboratory experimentation

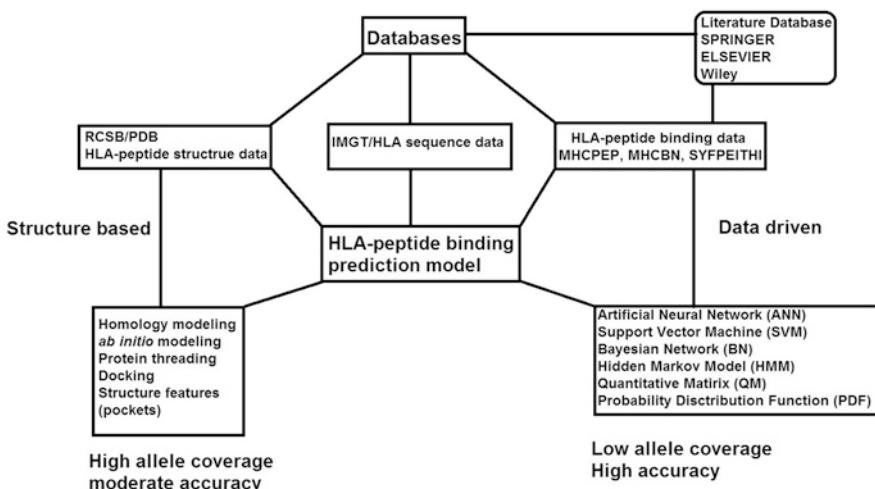
Table 2 lists some of the HLA-peptide binding prediction tools and techniques [38–57]. HLA-peptide binding prediction is possible through three dimensional structure based molecular modeling and data driven statistical models such as ANN, SVM, BN, HMM, QM and PDF. The structure-based models use known structures with molecular modeling, side chain packing, threading, docking, hot spots prediction and structure features [5–11]. This approach provides broad coverage across all known HLA alleles with adequate prediction accuracy using large computer resources. It uses available three dimensional HLA-peptide structures to build HLA-peptide models for evaluating HLA-peptide binding interaction in epitope design [7].

## Support Vector Machine (SVM)



**Fig. 8** Support Vector Machine (SVM) is a data classification model which groups data into categories (group A (negative) and group B (positive)). It consists of two kinds of classifier namely linear classifier and non-linear classifier. Linear classifier clusters test data based on features and non-linear classifier clusters test data based on trained (previously taught) data with known information

## HLA-peptide binding prediction models



**Fig. 9** Development of HLA-peptide binding prediction models is shown using a block diagram. This is possible using structure based and data driven prediction models illustrated in this diagram. The structure based model uses known HLA-peptide structures with techniques such as homology modeling, threading, docking and quantitative matrices generated by structural pockets. This method offers HLA-peptide binding prediction for all the 20,000 alleles. The data driven models uses binding data generated for each HLA allele for training using ANN, SVM, BN, HMM, QM and PDF. This method offers prediction for only a handful of alleles where HLA-peptide binding data is known from prior experimentation

**Table 2** HLA-peptide binding prediction and other related servers

Server	Model type with description	HLA class	Reference
CTLPred	SVM and ANN	Class I	[36]
ProPred1	Promiscuous HLA class-I binding peptide	Class I	[37]
MAPPP	Antigenic peptide processing prediction	Class I	[38]
nHLAPred	Artificial Neural Network (ANN)	Class I	[39]
BIMAS	Matrix	Class I	[40]
LPPEP	Linear programming for HLA A2	Class I	[41]
SVMHC	Support Vector Machine (SVM)	Class I	[42]
NetMHC	Artificial Neural Network (ANN)	Class I	[43]
MHCPred	Quantitative matrices (QM)	Class I	[44]
MMPRED	Quantitative matrices (QM)	Class I	[45]
PREDEP	Position specific scoring matrices (PSSM)	Class I	[46]
T-epitope	Virtual pockets in structures	class I	[47]
SYFPEITHI	Motifs matrices (MM)	Class I/II	[48]
RANKPEP	Position specific scoring matrices (PSSM)	Class I/II	[49]
MHCbench	Prediction comparison	Class I/II	[50]
ProPred	Quantitative matrices (QM)	Class II	[51]
Epipredict	Quantitative matrices (QM)	Class II	[52]
ProPred2	Promiscuous HLA binding peptide	Class II	[53]
HLADR4Pred	SVM and ANN	Class II	[54]
MHC2Pred	Support Vector Machine (SVM)	Class II	[55]
NetChop	Artificial Neural Network (ANN)	Cleavage sites	[56]
PAProC	Proteasome cleavage prediction	Proteasome cleavage	[57]

The data-driven methods use known HLA-peptide binding data with  $IC_{50}$  values generated from immunological assays for developing models using ANN [58–65], QSAR [66], SVM [67], probability distribution function (PDF) [68], Bayesian Network (BN) [69], Hidden Markov Model (HMM) [70] and other hybrid models combining QM, ANN and SVM are also available [71, 72]. This approach provides limited HLA coverage for a number of class I and class II alleles with known binding data from immunological assays. However, the accuracy, specificity, and sensitivity provided by these models are high. Thus, the combined use of three dimensional structures [5–11], the structure features at the peptide binding groove, molecular modeling and machine learning techniques [58–72] like Artificial Neural Network (ANN) and Support Vector Machine (SVM) as well as applying other applied mathematical models like Hidden Markov Models (HMM), Bayesian networks (BN) and Quantitative Matrices (QM) is highly promising in the design and development of short peptide vaccine candidates for further consideration in the context of viral diseases.

## 9 Conclusions

The accurate design and development of short peptide viral vaccine candidates is now feasible using HLA-peptide binding prediction models. This is possible using molecular modeling, peptide threading, structural features and machine learning algorithms (ANN, SVM, BN, HMM, QM and PDF) for improved accuracy with high sensitivity and specificity.

**Acknowledgement** We wish to express our sincere appreciation to all members of Biomedical Informatics (P) Ltd. and Department of Biotechnology, Anna University, Chennai for discussion on the subject of this chapter for Global Virology III. We thank Paul Shapshak, PhD, for his critical comments, suggestions and useful edits of the content of this chapter.

## References

1. Kangueane P, Viswaoorani K, Nilofer C, Manimegalai S, Sivagamy M, Kangueane U, Sowmya G, Sakharkar MK. In: Shapshak P, Levine A, Foley B, Somboonwit C, editors. Chapter 35: short oligo-peptide T-cell epitopes in HIV-1/AIDS vaccine development: current status, design, promises and challenges, book chapter in global virology II – HIV and NeuroAIDS. 2017 1st ed. New York: Springer; 2018. 978-1-4939-7288-3 (ISBN).
2. Kangueane P, Sowmya G, Anupriya S, Dangeli S, Mathura VS, Sakharkar MK. Short peptide vaccine design: promises and challenges in book titled “Global virology I – identifying and investigating viral diseases”, vol. 1. New York: Springer; 2015. p. 1–14. ISBN 978-1-4939-2410-3.
3. Kangueane P, Kayathri R, Sakharkar MK, Flower DR, Sadler K, et al. Designing HIV gp120 peptide vaccines: rhetoric or reality for NeuroAIDS, Chapter 9. In: Goodkin K, Shapshak P, Verma A, editors. The spectrum of Neuro-AIDS disorders: pathophysiology, diagnosis, and treatment. ISBN: 9781555813697; ISBN10: 1555813690. Washington, DC: ASM Press; 2008. p. 105–19.
4. Shapshak P, Kangueane P, Fujimura RK, Commins D, Chiappelli F, Singer E, Levine AJ, Minagar A, Novembre FJ, Somboonwit C, Nath A, Sinnott JT. Editorial NeuroAIDS review. AIDS. 2011;25:123–41.
5. Sowmya G, Vaishnai A, Kangueane P. Structure modeling based computer aided T-cell epitope design. Bio-Algorithms and Med-Systems. 2008;4:5–13.
6. Kangueane P, Sakharkar M. HLA-peptide binding prediction using structural and modeling principles. Methods Mol Biol. 2007b;409:293–9.
7. Kangueane P, Sakharkar M, Lim K, Hao H, Lin K, Chee R, Kolatkar P. Knowledge-based grouping of modeled HLA peptide complexes. Hum Immunol. 2000;61:460–6.
8. Zhao B, Mathura V, Rajaseger G, Moothhal S, Sakharkar M, Kangueane P. A novel MHCp binding prediction model. Hum Immunol. 2003b;64:1123–43.
9. Mohanapriya A, Lulu S, Kayathri R, Kangueane P. Class II HLA-peptide binding prediction using structural principles. Hum Immunol. 2009;70:159–69.
10. Kangueane P, Sakharkar M. T-Epitope Designer: a HLA-peptide binding prediction server. Bioinformation. 2005b;1:21–4.
11. Zhao B, Sakharkar KR, Lim CS, Kangueane P, Sakharkar MK. MHC-peptide binding prediction for epitope based vaccine design. Int J Integr Biol. 2007;1(2):127–40.
12. <https://ghr.nlm.nih.gov/primer/genefamily/hla>.

13. Blackwell JM, Jamieson SE, Burgner D. HLA and infectious diseases. *Clin Microbiol Rev.* 2009;22:370–85.
14. Sheldon S, Poulton K. HLA typing and its influence on organ transplantation. *Methods Mol Biol.* 2006;333:157–74.
15. Berger A. HLA typing. *BMJ.* 2001;322:218.
16. Ren E, Kangueane P, Kolatkar P, Lin M, Tseng L, Hansen J. Molecular modeling of the minor histocompatibility antigen HA-1 peptides binding to HLA-A alleles. *Tissue Antigens.* 2000;55:24–30.
17. <https://www.ebi.ac.uk/ipd/imgt/hla/allele.html>.
18. Kangueane P, Sakharkar M, Kolatkar P, Ren E. Towards the MHC-peptide combinatorics. *Hum Immunol.* 2001;62:539–56.
19. <https://www.rcsb.org/>.
20. Govindarajan K, Kangueane P, Tan T, Ranganathan S. MPID: MHC-Peptide Interaction Database for sequence-structure-function information on peptides binding to MHC molecules. *Bioinformatics.* 2003;19:309–10.
21. Sukhwani A, Sowdhamini R. Oligomerization status and evolutionary conservation of interfaces of protein structural domain superfamilies. *Mol BioSyst.* 2013;9:1652–61.
22. Sukhwani A, Sowdhamini R. PPCheck: a webserver for the quantitative analysis of protein-protein interfaces and prediction of residue hotspot. *Bioinform Biol Insights.* 2015;9:141–51.
23. Adrian P, Rajaseger G, Mathura V, Sakharkar M, Kangueane P. Types of inter-atomic interactions at the MHC-peptide interface: identifying commonality from accumulated data. *BMC Struct Biol.* 2002;2:2.
24. Brusic V, Rudy G, Harrison LC. MHCPEP: a database of MHC-binding peptides. *Nucleic Acids Res.* 1994;22(17):3663–5. PubMed PMID: 7937075; PubMed Central PMCID: PMC308338.
25. Bhasin M, Singh H, Raghava GP. MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics.* 2003;19(5):665–6.
26. Rammensee HG, Friede T, Stevanovic S. MHC ligands and peptide motifs: 1st listing. *Immunogenetics.* 1995;41:178–228.
27. Zhao B, Png A, Ren E, Kolatkar P, Mathura V, Sakharkar M, Kangueane P. Compression of functional space in HLA-A sequence diversity. *Hum Immunol.* 2003a;64:718–28.
28. Kangueane P, Sakharkar M. Grouping of class-I HLA alleles using electrostatic distribution maps of the peptide binding grooves. *Methods Mol Biol.* 2007a;409:175–81.
29. Mohanapriya A, Nandagond S, Shapshak P, Kangueane U, Kangueane P. A HLA-DRB super-type chart with potential overlapping peptide binding function. *Bioinformation.* 2010;4:300–9.
30. Kangueane P, Sakharkar M, Rajaseger G, Bolisetty S, Sivasekari B, Zhao B, Ravichandran M, Shapshak P, Subbiah S. A framework to sub-type HLA supertypes. *Front Biosci.* 2005a;10:879–86.
31. Kangueane P, Sakharkar M. Structural basis for HLA-A2 supertypes. *Methods Mol Biol.* 2007c;409:155–62.
32. Agatonovic KS, Beresford R. Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research. *J Pharm Biomed Anal.* 2000;22:717–27.
33. Cortes C, Vapnik V. Support-vector networks. *Mach Learn.* 1995;20:273–97.
34. <http://www.statsoft.com/textbook/support-vector-machines>.
35. Adams HP, Kozioz JA. Prediction of binding to MHC class I molecules. *J Immunol Methods.* 1995;18:181.
36. Gulukota K, Sidney J, Sette A, DeLisi C. Two complementary methods for predicting peptides binding major histocompatibility complex molecules. *J Mol Biol.* 1997;26:1258.
37. Trolle T, Metushi IG, Greenbaum JA, Kim Y, Sidney J, Lund O, et al. Automated benchmarking of peptide-MHC class I binding predictions. *Bioinformatics.* 2015;31:2174–81.
38. Bhasin M, Raghava GP. Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine.* 2004;22:3195–204.

39. Singh H, Raghava GP. ProPred1: prediction of promiscuous MHC Class-I binding sites. *Bioinformatics*. 2003;19:1009–14.
40. Hakenberg J, Nussbaum AK, Schild H, Rammensee HG, Kuttler C, Holzhütter HG, et al. MAPPP: MHC class I antigenic peptide processing prediction. *Appl Bioinforma*. 2003;2:155–8.
41. <http://crdd.osdd.net/raghava/nhlapred/index.html>.
42. <https://www-bimas.cit.nih.gov/>.
43. <https://zlab.bu.edu/zhiping/lpppep.html>.
44. <http://www-bs.informatik.uni-tuebingen.de/Services/SVMHC/>.
45. <http://www.cbs.dtu.dk/services/NetMHC/>.
46. Guan P, Doytchinova IA, Zygori C, Flower DR. MHCpred: a server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res.* 2003;31:3621–4.
47. Bhasin M, Raghava GP. Prediction of promiscuous and high-affinity mutated MHC binders. *Hybrid Hybridomics*. 2003;22:229–34.
48. <http://margalit.huji.ac.il/Teppred/mhc-bind/index.html>.
49. <http://www.bioinformation.net/ted/>.
50. <http://www.syfpeithi.de/>.
51. Reche PA, Glutting JP, Reinherz EL. Prediction of MHC class I binding peptides using profile motifs. *Hum Immunol.* 2002;63:701–9.
52. <http://crdd.osdd.net/raghava/mhcbench/reference.html>.
53. <http://crdd.osdd.net/raghava/hldr4pred/>.
54. <http://crdd.osdd.net/raghava/mhc2pred/info.html>.
55. <http://crdd.osdd.net/raghava/mhc2pred/help.html>.
56. Keşmir C, Nussbaum AK, Schild H, Detours V, Brunak S. Prediction of proteasome cleavage motifs by neural networks. *Protein Eng.* 2002;15:287–96.
57. Nussbaum AK, Kuttler C, Hadeler KP, Rammensee HG, Schild H. PAPrC: a prediction algorithm for proteasomal cleavages available on the WWW. *Immunogenetics*. 2001;53:87–94.
58. Brusic V, Zelezniakow J. Artificial neural network applications in immunology. Proceedings of the 1999 International Joint Conference on Neural Networks IJCNN'99. 1999;2034. ISBN:0-7803-5532-6
59. Malik M, Sauer D, Brunmark AP, Yuan L, Vitiello A, Jackson MR, et al. Application of an artificial neural network to predict specific class I MHC binding peptide sequences. *Nat Biotechnol.* 1998;16:753.
60. Buus S, Lauemøller SL, Worning P, Kesmir C, Frimurer T, Corbet S, et al. Sensitive quantitative predictions of peptide-MHC binding by a ‘Query by Committee’ artificial neural network approach. *Tissue Antigens*. 2003;62:378–84.
61. Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V. Neural models for predicting viral vaccine targets. *J Bioinforma Comput Biol.* 2005;3:1207–25.
62. Brusic V, Rudy G, Honeyman G, Hammer J, Harrison L. Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network. *Bioinformatics*. 1998;14:121–30.
63. Choi SE, Park CW, Sohn YH, Ko SY, Oh HB, Kim GH, et al. Artificial neural network weights of residues for the serological specificities of HLA. *Int J Immunogenet.* 2011;38:269–75.
64. Bellgard MI, Tay GK, Hiew HL, Witt CS, Ketheesan N, Christiansen FT, et al. MHC haplotype analysis by artificial neural networks. *Hum Immunol.* 1998;59:56–62.
65. Honeyman MC, Brusic V, Stone NL, Harrison LC. Neural network-based prediction of candidate T-cell epitopes. *Nat Biotechnol.* 1998;16:966–9.
66. Brusic V, Bucci K, Schönbach C, Petrovsky N, Zelezniakow J, Kazura JW. Efficient discovery of immune response targets by cyclical refinement of QSAR models of peptide binding. *J Mol Graph Model.* 2001;19:405–11.
67. Zhang GL, Bozic I, Kwok CK, August JT, Brusic V. Prediction of supertype-specific HLA class I binding peptides using support vector machines. *J Immunol Methods*. 2007;320:143–54.
68. Soam SS, Khan F, Bhasker B, Mishra BN. Prediction of MHC class I binding peptides using probability distribution functions. *Bioinformation*. 2009;3:403–8.

69. Astakhov V, Cherkasov A. Prediction of HLA-A2 binding peptides using Bayesian network. *Bioinformation*. 2005;1:58–63.
70. Noguchi H, Kato R, Hanai T, Matsubara Y, Honda H, Brusic V, et al. Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. *J Biosci Bioeng*. 2002;94(3):264–70.
71. Singh SP, Mishra BN. Prediction of MHC binding peptide using Gibbs motif sampler, weight matrix and artificial neural network. *Bioinformation*. 2008;3:150–5.
72. Basin M, Raghava GP. A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J Biosci*. 2007;32:31–42.

# Artificial Life and Therapeutic Vaccines Against Cancers that Originate in Viruses



María Elena Escobar-Ospina and Jonatan Gómez

**Abstract** The construction of artificial life processes that seek to contribute to the development of therapeutic vaccines to treat human cancers, which have their origin in infectious processes caused by viruses, requires research on three fronts. On the one hand, to know the life cycle of the virus under study, as well as to recognize the mechanisms and strategies that it can implement to attack its host and proceed to infect it. On the other hand, to acknowledge the components, mechanisms and strategies that the immune system develops to identify the presence of a stranger and prepare to repel it, in response to the kind of attack that poses. And finally, to design the strategies, that exogenously, allow activating the host's immune system so that it prepares answers with objectives aimed at counteracting the injuries caused by the virus that attacks it. This chapter describes in general, the methods that through a process of artificial life, allow to simulate the interactions that arise between human immune system, pathogen (viruses as etiological agents of cancer), and therapeutic vaccines to treat lesions that originates in the activity of this type of pathogen.

**Keywords** Artificial life · Artificial immune system · Cancer · Virus · Simulation model · Therapeutic vaccine · Toll-like receptors · Cytokines

## 1 Introduction

The world today has several methods of prevention and treatment for the control of diseases that originated in infections caused by viruses. Nonetheless, several types of cancer associated with viruses continue to be a global health concern, with particular emphasis on developing countries.

---

M. E. Escobar-Ospina (✉) · J. Gómez

Department of Engineering – Systems and Computing, Universidad Nacional,  
Bogotá, Colombia

e-mail: [meescobaro@unal.edu.co](mailto:meescobaro@unal.edu.co); [jgomezpe@unal.edu.co](mailto:jgomezpe@unal.edu.co)

Based on the most recent estimate of incidence, mortality and cancer prevalence in the world presented by Globocan project in 2012 [1], IARC (International Agency for Research on Cancer) reports the proportions of cancer cases attributable to infections worldwide, including all infectious agents, between both sexes, which originated in Helicobacter Pylori with a participation of 35.7%, Human Papillomavirus (HPV) with 29.4%, Hepatitis B Virus (HBV) with 19.1%, Hepatitis C Virus (HCV) with 7.7%, and in the other infectious agents with 8.1% [2].

In the same way, the southeast Asia is reported as the region with the most important proportion (55.1%) of cancer cases attributable to infections caused by HPV; the western pacific region with the most significant proportion of cancer cases originating in infections generated by Helicobacter Pylori (45.6%) and by HBV (31.2%); the eastern Mediterranean region with the most considerable proportion of cancer cases attributable to infections caused by HCV (23.7%); and the highest rate for other infectious agents is reported in the African Region (26.7%) [2].

The elevated incidence of cancers associated with the viruses reveals the importance of exploring alternative ways to confront the challenge that this relationship implies. Therefore, it is essential to provide tools that allow implementing novel mechanisms, which are based on the study of viral integration patterns from the relationship between the host and pathogen. Tools that support us to advance in the understanding of their mechanisms and that allow us to produce better controls in related diseases.

From our experience in the development of artificial life models inspired by complex biological systems, we present in this chapter an overview of the methods we used to understand and simulate the interactions that arise between the immune system, viral pathogen and therapeutic vaccines. Regarding the pathogen, we focus on DNA viruses that are integrated into the host and that are considered etiological agents of cancer. In therapeutic vaccines, we concentrate on those that seek to treat the lesions that originate the activity of DNA viruses.

In the second section of this chapter and following a structure from the general to the particular, we initially present key biological backgrounds that allow us to link complex systems, biological systems and the immune system. This offers us the basis to introduce below some key concepts of artificial life, which as an interdisciplinary field of research, provides us the possibility of designing modern intelligent tools that can contribute to *in-silico* experimentation processes. Subsequently, we present the linked complexity to viral pathogens and their association (supported by evidence-based medicine) with the development of certain types of cancers. Therefore, it is necessary to comprehend how viruses are classified, recognize their life cycle, follow the way in which they are integrated into the human host, and to identify the mechanisms that allow them to induce infectious lesions and generate optimal environments to promote the development of certain types of cancers. In addition, we show some immunotherapeutic methods used to counteract cancers of viral origin, especially those based on the design of therapeutic vaccines.

Derived from this biological background and with a view to the computational component, the third section describes the sequence on which we base the construction of an artificial life model that allows us to integrate these three subsystems (immune system, viruses, and vaccines), in order to produce software tool that aims to support the research and design of therapeutic vaccines against cancers that originate in DNA viruses that are integrated into the human host. With the aim to acquire a more proper understanding of the biological cycle associated with these three subsystems and synthetically reproducing interactions similar to those of the true world we explain the steps we follow in the construction of a functional artificial life prototype, the way to experiment as well as the process to obtain and analyze the results that it produces.

## 2 Background

Particularly in humans, biological systems by definition are also considered as complex system [3]. Among several systems that integrate the human body, the immune is one of the biological systems that make it up. Therefore, the biological immune system is also considered a complex system [4].

Although there is an established relationship between the immune, biological and complex systems, it is equally necessary to define the characteristics that identify them and the limits that bias them particularly when trying to observe each one of them as an integral part of an exclusive domain.

### 2.1 *Complex Systems*

Merelli and colleagues (2015) define complex systems as a considerable number of finite entities that interact with each other, which by themselves represent equally complex systems that have autonomous strategies and behaviors [5]. Under a network approach, Sayama (2015) defines complex systems as a massive quantity of microscopic components that are interacting with each other in non-trivial ways, where the essential information resides in the relationship between the parties and not necessarily in the parts by themselves [6].

According to Mitleton-Kelly (2003) the theories of complexity provide a conceptual structure, a manner of thinking, and a way of seeing the world. He states that complex behavior arises from the interrelation, interaction and inter-connectivity, between the elements of a system and connecting the system and its environment. He also explains each of the principles of complexity that are based on the generic characteristics of all complex systems, highlighting the following: connectivity and inter-dependence, co-evolution, dissipative structures, exploration of the space of possibilities, feedback, self-organization, emergence and cre-

ation of a new order, chaos and complexity, and self-similarity [7]. Other authors present complex systems as systems that are made up of interacting parts that possess the capacity to generate another scheme of collective behaviors through self-organization. This means that they can be generated by way of the spontaneous formation of temporal structures, spatial structures or functional structures. In addition, they are presented as adaptive systems as they evolve and may contain self-directed feedback loops. Complex systems are much more than the sum of their parts. These systems are characterized by being extremely sensitive to initial conditions as well as to emergent behaviors that are not easily predictable. The fact that the collective behavior of any system cannot be inferred simply from the understanding of the conduct of its components has already been recognized. This has led to define concepts and modeling tools described in terms of complexity [4, 8].

Two key concepts characterize complex systems: emergency and self-organization. The first concept is associated with non-trivial relationships of the properties of the system at microscopic and macroscopic scales. The second concept is associated with the spontaneous formation of structures and behaviors over time [6]. Complex systems exhibit emergency patterns that are unpredictable from the inspection of particular elements. The emergency is described as unexpected, unpredictable, or surprising. This means that modeled systems exhibit behaviors that are not explicitly constructed in the model. The unpredictability is due to the non-linear effects that result from the interactions of entities that produce simple behaviors [4]. Emerging properties represent patterns and regularities that arise through interactions between smaller or simpler entities in a system that by itself does not exhibit such proposals. The combination of structure and emergency is seen as self-organization [3]. Self-organization is a dynamic process through which a system spontaneously forms non-trivial macroscopic structures and behaviors over time. Complex systems can arise and evolve through self-organization. In this way, they are neither completely established nor totally random, which allows them to develop emergent behaviors at macroscopic scales [6].

Complex biological systems are also presented as adaptive systems because they exhibit characteristics of adaptation to a dynamic environment [3]. Complex adaptive systems (CAS) are defined as a collection of particular components or agents, with the freedom to act in ways that are not completely predictable. Their interactions are intertwined in such a way that through these actions an agent can affect the conditions in which other agents behave [9]. Some of the most general properties that stand out in CAS are: having multi-agent structures, self-organization, co-evolution, emergence, and adaptation [10, 11].

The dynamics of a complex system is affected by feedback processes that can be positive or negative. Regularly, positive feedback increases the activity of the system, and negative feedback does exactly the opposite. Cooperation in complex systems involves signal flows that allow activating the dynamics of the system while moving through it. The models built in networks have tolerated the implementation

of exceptionally useful schemes for the study of complex systems, since they allow the visualization of cooperative characteristics. Their interactions can be conducted by information flows through of the network or by way of non-directed flows, cases in which agents interacts across asymmetric relationships [9, 11].

Complex systems possess unique properties that allow them to differentiate themselves from other systems. They tend to be generic when they are understood through functional analysis. Its structures span several scales and its components are interdependent and interact in non-linear ways [3].The analysis of complex systems considers two approaches: one is based on differential equations, and the other focuses on the interactions between components. Under the first approach a global and abstract description is obtained. With the second approach, a more comprehensive view can be obtained through the use of agent-based simulations models [5].

Sayama (2015) considers the investigation of complex systems can be developed in four themes: pattern formation, evolution and adaptation, networks and collective behavior. The formation of patterns arises from a process of self-organization that involves space and time. Evolution and adaptation are related to general actions that contribute to explain biological mechanisms and promote non-biological processes, which have learning dynamics and creative abilities. Networks and collective behavior emerge as fields of research that rely on the capacity that is currently available to obtain elevated traffic and high-resolution data. This allows the scientist to analyze the structure of networks at multiple scales and also to develop dynamic models that can explain collective behavior [6].

## 2.2 *Complex Biological Systems*

Biological systems can be defined from several points of view. Alcocer-Cuarón and colleagues (2014) describe biological systems with a focus based on bioengineering and framed in hierarchical, structural and functional concepts. They describe biological systems as a set of differentiated and self-organized components that interact in pairs among themselves, through an environment in networks, isolated from other grouping when establishing limits. Its relationship with other systems is described as a closed-circuit in stable state. From the hierarchical point of view, a biological system is formed by sub-systems, depending on the physical scale of the specific system. At the same time, these subsystems are integrated by subsystems of second order, which in turn are made up of subsystems of the third order. From the structural point of view, a biological system is characterized by its limits (that is, the barrier that isolates, separates and preserves it), and by the elements and means of both internal and external origin. In addition, it is emphasized that their identity and autonomy depend on the presence of such limits. From a functional point of view, biological systems can be classified into one of the following categories: reception, integration and response. The medium

plays a fundamental role in the identification and separation of elements through interconnected networks [12].

Both biological systems and their processes are considered complex. These systems are rich in information, and the related processes do not seek to a concept of precision or optimum, but rather to functional and survival events [13]. They consist of enormous amounts of components. Its primary challenge is to perform any analysis, since the processes that govern them are not linear. This allows them to have a broad repertoire of distinct behaviors with which an organism can respond to a disturbance. During development other difficulties appear. Their emergence does not obey to a particular cause but may be the consequence of a combination of slight alterations in several of its components [14]. Such is the case that occurs in diseases that originate in viral type infections. The immune responses that evidence the presence of the disease can take a long time to manifest in relation to the moment of the stimulus. This may involve knowing the current state of the biological system and its historical background. Recovery from a severe infection depends largely on the preconditions of the organism. The changes that occur may be the collective result of early infectious processes and responses of the body to its detection.

In biology, it has been argued that systems have always been present. Diverse disciplines, including immunology, suggest that mathematical tools are important when going from identifying the components to trying to understand their collective functioning [15]. The integration of data and the construction of models represent part of the essential activities that strengthen the advances obtained in research, both biological and technological. The scientific community suggests that the fundamental way in which biology is constructed can change the manner to take advantage of the enormous amount of biological data that are available. It is pointed out that the academic use of this information is fundamental to promote the objective of improving human health and expand the understanding of the mechanisms underlying life. Some health priorities point to the integration of heterogeneous biological data sets, the unification of experiments and computation, and the development of methods for systems analysis in biology. All these high priority objectives for the integration of biological data can be gathered through the creation of biological systems models [4].

The quantification of the complexity of a biological organism can be grouped into several categories: structural, functional, sequence, and network. In a measure of structural complexity, one typically tries counting the number of parts and their connections. The challenges arise when defining a scalar that classifies all possible structures, which is not necessarily predictive of the complex function, although it is predicted to be true in most cases. Ideally the measure of biological could be functional, that is, reflect how the organism works in a complex world. Because function is understood as a product of natural selection, it is frequently implicitly assumed that a measure of functional complexity could result in evolution. Regularly, this type of measurement is based on information theory. Regarding sequence complexity, all forms of life on earth contain a genetic code that is responsible for the generation of its appearance and function. Because of the limitations of using, the

full-length sequence, the focus on sequence complexity measurement has been based on mathematical estimations, or more specifically, on the complexity of symbolic chains that have derived in some other theories such as: Kolgomorov, compositional, or information. The difficulty in assessing the complexity of biological networks arises because they typically contain thousands of nodes and several thousand borders, and often seem to be unstructured. The best way to evaluate the complexity of a network is to measure the complexity of the set of rules used to construct it. In biology, this set of rules is encoded in the genome. Therefore, a first estimate of the order of complexity of a network could be given by the complexity of the genome that produces it [4].

The management of time and space equally represent a challenge, given that biological systems can operate simultaneously at several scales. The temporal scale is governed by physical and chemical processes, which occur in fractions of milliseconds and shorter times. The biochemical procedures can run on time scales from seconds to minutes. Therefore, the effort of modeling biological processes against the time scale may be quite simplified. As for the spatial scale, all biological processes have a molecular component, and their measurement system can be of the order of Ångström or nanometers. At the cellular level, it can be treated with a scale of micrometers to millimeters, and in some cases of exception in centimeters. From the variability in spatial measurements, the representations of biological systems often focus on one or two scales of space and time. Such an event can become another challenge during the construction of this type of models [14]. That is why certain characteristics and properties can solely be known by the probability distribution level, and therefore it seems natural to incorporate stochasticity into models of such systems [4].

The resulting models of biological systems are presented in two varieties. Some focus on specific systems, which include all the functional and numerical details. Others try to help understand the fundamental and generic characteristics of the organization of biological systems. The opportunity presented for biological systems results from the confluence of three scientific frontiers. The first frontier faces the course of the rapid and vast accumulation of exhaustive biological information at the physiological, cellular, molecular and sub-molecular levels. Part of this information includes: genome quantification, expression patterns, protein identification, characterization of molecular networks interactions, and global evaluation of the system under consideration. The second frontier results from innovation in other fields of science that have allowed the emergence of techniques that help to test, detect and measure. Such is the case of bioengineering and robotics. These fields of research have allowed thousands of biomarkers to be measured in the bloodstream, and this has contributed to the purposes sought by medicine, medicaments and biotechnology. The third frontier is found in the co-evolution of mathematics, physics and computational techniques. The advance in algorithms has allowed the simulation and optimization of extremely complex biological flows. This achievement allows us to obtain an approximation to the dynamics of complex non-linear systems, which constitute

a primordial source for prediction. Accomplishing these goals suggest promising intervention in medications and vaccines [14].

Within the context of biological systems, we now concentrate directly on the immune system of human beings. From the descriptions of Alcocer and colleagues (2014) previously presented (item 1.2.2), we consider use of analogies in order to evince the manner in which the immune system is framed in those concepts when it is also conceived as a complex system. Alcocer-Cuarón and colleagues (2014) point out that biological systems are organized in hierarchical networks that interact [12]. In the case of human beings, the immune system (IS) resembles a subsystem of the first order. The immune organs and tissues represent part of a subsystem of second-order (such as: thymus, bone marrow and lymphatic tissue.). The different cell populations are assimilated to a third order subsystem. The behaviors of protein populations, transcription factors and cytokines, resemble networks that interact within the immune system. From the functional point of view and according to the classification of Alcocer-Cuarón and colleagues (2014), the IS classifies as a biological system of the first order. Nevertheless, some of its internal processes can be classified as a reception and response subsystems, which are considered third-order subsystems.

The elements that constitute a biological system have associated fundamental functional processes to support and develop the system, these are: transfer, replication and integration. The transfer processes establish the interaction between the system and its environment. The processes of replication determine and control de growth, development, regeneration, reproduction and immunity of system. The integration processes link all the elements of the subsystem [12]. Following the analogy with SI in our case, the transfer elements correspond to the population of Toll-like receptors, effector and adapters molecules. The replication components are correlated with the transcription factors. The elements of integration are equivalent to the signaling pathways.

Regarding the functions themselves, Alcocer and colleagues (2014) indicate that all biological systems are flows whether matter, energy or information [12]. Following our exercise with the SI, these flows may be related to downstream and upstream signaling pathways, depending on the interactions that arise between their distinct components from the host-pathogen relationship.

The functions of all biological systems interact in pairs [12]. In the human SI there are several examples that coincide with this statement. An example under this context is found in the process of somatic hypermutation when changes of base pairs are introduced in the V(D)J area within the recombination procedure of the variable region of the immunoglobulin genes.

In biological systems, interactions at low-levels emerge as objects that express their properties at a more superior level [3]. All interactions between biological systems occur through closed circuits within the medium, which makes it possible to regulate and develop functions that enable them to complete these cycles [12]. In the SI processes are developed that perfectly apply to this description. Such is the case of the reactions that take place in germinal centre where affinity selection

occurs through somatic hypermutation (SHM). In addition, include the processes of clonal expansion and diversification of the B-cell receptor, by means of class switch recombination (CSR).

The medium in biological systems helps the development of organic processes through the exchange of matter, energy and information, using interactions with the outside and with other systems [12]. This description ideally applies to the interactions that arise between a therapeutic vaccine and the response obtained from the IS, depending on the medium of administration. In this way the immune response may vary if the vaccine is applied to the host intramuscularly, orally, nasally or intravenously.

### ***2.3 Biological Immune Systems***

According to Harvard Medical School (2015) the science of immunology accompanies the study of the development, anatomical functions and malfunctioning of the immune system, which result relevant in the understanding of human disease [16]. Typically, immunology is the study of the immune system. This field includes research around the development of T-cells and B-cells, humoral and cell-mediated responses, involvement of cytokines and chemokines, co-stimulation, hematopoiesis and presentation of antigens, among other aspects [17].

In accordance with “American Cancer Society” (ACS), the immune system as such is a collection of organs, cells and special substances that help protect the host from infections and other diseases. Immune cells and substances travel through the body to protect it from germs that cause disease and also help combat cancer [18]. The immune system is defined by Parkin and colleagues (2001) as an interactive network of lymphoid organs, cells, humoral factors and cytokines [19]. Although its fundamental function is to induce host protection, not all of its immune responses are protective. In many cases, the characteristics of the pathogen that poses a challenge to the host can determine the type of immune response [20]. Therefore, is essential to comprehend the concepts and rules that govern the immune system, especially when it comes to developing preventive and therapeutic therapies that focus on the dynamic of their defense mechanisms.

The human beings immune system is composed of an enormous diversity of populations of cells, molecules, interactions and processes. All these components generate communication with other systems (such as: nervous, endocrine, respiratory and circulatory), which turn out to be as complex as the immune system itself. These systems work in coordination in order to preserve the host in a stable state. However, one of the fundamental roles of the immune system is to defend the host from infectious agents, among which are: viruses, fungi and parasites.

The immune system is considered an emerging paradigm that aims at a systematic and quantitative understanding, and in this order two approaches have been proposed. On one hand, we attempt establishing the molecular and cellular components as well as their interactions, based on collections of unbiased data where

high-performance measurements are used that are supported by biomedical technologies. On other hand, we try comprehend the principles of its operation by formulating hypothesis based on data, an approach that has led to the construction of mathematical models that helps explain and predict the dynamics of the system. The first approach is known as “data-based modeling.” The second approach is known as “hypothesis-based modeling”. Arazi and colleagues (2013) consider that the interactions between these two approaches may be fundamental in the future development of biology and medicine [21].

There are two types of immunity, innate and adaptive. Innate immunity represents a first defense mechanism and has a prominent role in the initiation and regulation of adaptive immunity. Within the cells of the adaptive immune system are the lymphocytes, which correspond to red and white blood cells, more specifically groups of T-cells and B-cells. These cellular populations play fundamental roles in recognition processes and specific immune responses.

### 2.3.1 Innate Immunity

In innate immunity the immune response of an organism does not depend upon previous sensitization to an antigen from an infection or vaccine. The innate response is characterized by rapid reaction to pathogens and is mediated by several cell types including: basophils, stem cells, neutrophils, macrophages, polymorphic leukocytes, natural killer, keratinocytes, and antigen presenting cells (APCs) [17, 22]. APCs possess certain receptors that allow them to detect extracellular molecules such as heat shock proteins, DNA, RNA and other structures which are released by damaged cells or by the pathogen. Once the innate immune effectors are activated, certain cytokines are released that can produce some effects on the abnormal cells [23].

Innate immunity is neither antigen-specific nor does it generate immunological memory, but it prepares the ideal environment for an appropriate adaptive immune response to be induced to confront the threat of a pathogen. The basic result of innate immune activation includes production of soluble pro-inflammatory mediators, recruitment of immune cells at sites of infection, triggering of the complement cascade and initiation of the adaptive immune response [17]. The innate immune system plays a vital role in the processes of initiation and regulation of immune responses. Specialized cells of innate immunity have evolved, since they recognize and bind to molecular patterns that are only found in microorganisms. However, the innate immune system is by no means a whole solution for the protection of the body [24].

The innate immune system depends on the production and specific release of soluble mediators like cytokines, chemokines and interferons, which participate in the activation and recruitment of immune cells [25]. Innate immunity involves the coordinated action of receptor families, known as pattern recognition receptors (PRRs) or microbial sensors. They respond to a broad range of microorganisms

through the detection of specific conserved molecules or patterns. The innate immune response is activated by specialized groups of receptors found in macrophages, stem cells, dendritic cells, natural killer cells and polymorphonuclear leukocytes. The receptors bound to their ligands induce the activation of different signaling pathways involving effector molecules, needed to send danger signals or eradicate the pathogen [26]. Keratinocytes (KCs) form the majority of cells in the epidermis of the skin, the first line of defense against percutaneous pathogens [27]. The KCs are perceived as true innate immune cells and not only as a passive protective barrier [28]. The regulatory processes of activation and inactivation of this cellular population are coordinated by growth factors and cytokines produced by KCs and other types of cells around them [29].

### 2.3.2 Adaptive Immunity

Hoffman (2011) presents adaptive immunity as the ability of the immune system to learn and exhibit memory, both concepts in terms of establishing an immune state in relation to an antigen. It also considers the possibility that this immunity results specifically tolerant to that antigen [30]. Adaptive immunity is also defined as the invocation of cellular immune responses, in contrast to innate immunity that is distinguished by being a non-specific immunity [17].

The adaptive immune response incorporates more specific additional mechanisms, which allows better confronting a particular threat and then controlling the infection to try to eliminate it. Adaptive (or acquired) immunity is directed against specific invaders and produces adaptive immune cells that are modified by exposure to such invaders. The adaptive immune system primarily involves lymphocytes, characterized by humoral and cell-mediated immunity, capable of identifying a broad repertoire of antigens. B-cells mediate a humoral response and the creation of antibodies that can bind specifically with foreign antigens, including bacteria and microbial toxins. Cell-mediated immunity represents a more complex system that involves the production of cytotoxic T-lymphocytes, macrophages and activated NK cells, which are capable of releasing cytokines in response to antigens mediated by T-cells. Any capable substance of generating as such a response from the lymphocytes is named an antigen or immunogen. The antigens are not the same invading microorganisms. These are substances like toxins or enzymes in microorganisms that the immune system considers foreign. The antigens are small fragments of pathogen. Adaptive immune responses are normally directed against the antigen that causes them and these are known as antigen-specific responses. The T-cells recognize a pathogen only after the antigens have been processed and presented, typically in combination with a self-receptor, called the major histocompatibility complex (MHC). Cell-mediated immunity is essential to fight intracellular organisms, and also to perform tumor surveillance, mediating rejection of transplants and combating infections caused by fungi and viruses [17, 24].

The successful recognition of antigens triggers both immune effector and memory responses. An effector immune response includes both populations T-cells, CD8+ and CD4+. The population of CD8+ T-cells removes the cells with foreign antigens. The population of CD4+ T-cells differentiates several classes of effector cells, including those that can activate macrophages, cytotoxic T-cells and B-cells. The B-cells effector response involves plasma cells that secrete antibodies capable of neutralizing or eliminating a foreign agent. The memory response occurs when T-cells and B-cells are activated by exposure to a foreign antigen. The activation of these cells results in the proliferation and preservation of the antigen-specific receptor, such that successive exposure to the foreign agent triggers a robust immune response [26].

Adaptive immunity involves interactions between variable regions [30]. During the adaptive immune response, lymphocyte populations undergo a characteristic process of three phases that include: clonal expansion, which develops through a series of cell divisions; contraction or cessation of expansion; and finally elimination, in which the majority of the lymphocytes die by apoptosis [31].

### 2.3.3 Differentiation, Proliferation and Cellular Apoptosis

Some specialized cells of immune system evolve to recognize and bind to molecular patterns found in microorganisms, and others are modified to expose such invaders to the control of immunity. These cells are exposed to several interactions processes, which allow them to differentiate, proliferate and execute programmed death cycles. These cellular processes can trigger various responses that help the recognition and destruction of specific substances, which form part of the dynamics required by the immune system to activate its defense mechanisms.

Tarlinton (2012) has suggested that choosing the fate of cells can be a stochastic, directed, inherited process or combination of these, depending on the circumstances [32]. This process begins with stem cells present in the bone marrow, which depending on the stimuli they receive can activate the ability they have to differentiate into early myeloid or lymphoid progenitors. These early progenitors then pass to their corresponding common progenitor. The common lymphoid progenitor gives rise to the lineages of T-cells, B-cells and NK cells. The common myeloid progenitor gives rise to populations of monocytes and dendritic cells [33, 34].

Monocytes represent a conserved population of leukocytes that are present in all vertebrates. These are defined by their location, phenotype and morphology, as well as by genetic characteristics and microRNA expression signatures. In humans, monocytes correspond to 10% of the nucleated cells in the blood. It has been suggested that monocytes act as a temporary reservoir precursor for tissue-resident mononuclear phagocytes. Monocytes and their descendants emerge as a third cell system, with high dynamism and plasticity. Through physiological processes, monocytes contribute to the derivation of cells like DCs and macrophages, although it is also believed that monocytes can have functions as short-lived effector cells within tissues [35].

Differentiation and functional specialization of myeloid population may be regulated to ensure that these cells perform their own function. The acquisition of a particular cellular identity during cell differentiation from myeloid precursors (monocytes, macrophages, dendritic cells) is coordinated by several regulatory elements. These elements form gene expression programs that include transcription factors, epigenetic regulators and post-transcriptional mechanisms. Regarding differentiation, the function of innate immune cells is coordinated by the PU.1 transcription factor, which leads to chromatin remodeling and labeling of specific DNA sites [36]. In humans, a monocytes subdivision has been proposed between two important subgroups, phenotypic and functionally distinct. The first subgroup, defined by CD14low CD16+, seems to be dedicated to the surveillance of endothelial integrity and these cells effectively act as luminal blood macrophages. The second subgroup, defined as CD14+, includes classical monocytes functions at sites of injury and replenishment of compartments of DCs and peripheral macrophages [35]. Monocytes circulate in the blood, bone marrow and spleen, and do not proliferate in a stable state. This population represents immune effector cells equipped with pattern recognition receptors (PRRs) and chemokine receptors, which during infection mediate the migration process from the blood to the tissues. Monocytes produce inflammatory cytokines and capture toxic cells and molecules. During inflammation, monocytes can be differentiated between macrophages and DCs. The processes of differentiation and migration are apparently determined by the inflammatory medium and by the patterns recognition receptors that are associated with the pathogens [37].

Monocytes and macrophages are white blood cells, also known as “accessory-cells” or “adherent-cells”. In contrast to B-cells and T-cells, they do not possess an antigen-specific receptor however perform a fundamental role in the regulation of the immune system. Functionally both types of cells are similar, and physically they can be differentiated by their size and distribution of tissue. Macrophages are longer than monocytes and are found in lymphoid organs, while monocytes are located mainly in the blood. Monocytes, macrophages and DCs are non-specific accessory cells that express MHC class II and interact with T-helper cells during the immune response [30]. It has been proposed that blood monocytes, most macrophage subgroups and various classes of DCs, emerge *in-vivo* from ancestors derived from hematopoietic cells with potential myeloid restricted differentiation. In the bone marrow, common myeloid progenitors (CMPs), granulocyte-macrophages precursors (GMPs) and macrophage DC progenitors (MDPs) are located. The MDPs progenitors correspond to a group of cells that proliferate in the bone marrow, share phenotypic characteristics with myeloid precursor populations and differentiate into many subgroups of macrophages and DCs. Within the bone marrow, MDPs differentiate between monocytes and common dendritic cell precursors (CDPs). Monocytes have been considered as an intermediate development between bone-marrow precursors and tissue macrophages, although these two cellular populations of the limits do not derive from monocytes in steady-state. Monocytes exit the blood and may enter tissues under condition of inflammation, where they can produce subgroups of macrophages and inflammatory DCs, with the ability to process

and present antigens to T-cells. Nevertheless, monocytes do not give rise to common DCs (cDCs) or plasmacytoid DCs (pDCs). It is the proliferation of CDPs that induce the differentiation between pDCs and cDCs precursors [37].

The development of B-cells is initiated in the fetal liver and in the bone marrow in adults; however, functional maturation occurs later in the secondary lymphoid tissues. Activation of B-cells occurs through the involvement of multiple cell surface receptors, including the B-cell receptor complex (BC), CD19 and CD45 [17]. The process of differentiation of B-cells allows showing two different schemes, one called instructional and the other intrinsic. In the first scheme, B-cells respond to signal from the environment, which are provided by the antigen and T-cells through asymmetrically distributed factors and receptors. Consequently, this signaling process allows them to regulate the mechanisms of daughter cell differentiation. In the second scheme, the B-cells implement the selection of the destination based on regulatory networks, which allows them to establish autonomy within each cell. The cells that have a relationship of siblings show a strong concordance in the selection of their destinations, both in the result and in time. Some of these cells undergo class switch recombination, and others die [32]. Memory B-cells and plasma cells are derived from the naïve B-cells. Plasma cells are classified into at least two different groups, characterized by their differences in terms of average life and physical location. Short-lived plasma cells are found in extra-follicular locations, and long-lived plasma cells are located primarily in the bone marrow [38]. The long-lived plasma cells that reside in the bone marrow produce and constitutively secrete antibodies. They differ from memory B-cells in that they may contain none or minimal levels of the B-cells receptors (BCR), and cannot be stimulated to divide or to boost any rate of antibody production. Terminally differentiated plasma cells continuously perform effector functions in the absence of antigenic stimulation, that which does not turn out to be equivalent when they are in the T-cell compartment, where the antigen becomes the principal regulator of effector function [39]. Despite the complexity linked to the B-cell lineage, it has also been pointed out that its terminal differentiation can be described as a simple probabilistic process that is governed by a gene regulatory network and modified by environmental stimuli [40].

The T-cells constitute a diverse population that develops in the bone marrow and matures in the thymus. All T-cells express a T-cell receptor (TCR) that confers them specificity and that is essential for the role that this cellular population plays in adaptive immunity [17]. Wells and colleagues (1997) show that the strength of the TCR signal directly regulates the recruitment of individual T-cells between the group that is proliferating. They also report that the degree of signaling to a response can be regulated by the co-stimulator microenvironment of the T-cell. Based on this, the authors suggest that there is a linear relationship between the TCR commitment and the frequency of the response. In addition, they show that only between 50% and 60% of the T-cells that become activated after polyclonal stimulation manage to participate in a process of division [41]. Later, a quantitative evaluation of the proliferation of T-cells *in-vivo* is carried out, showing the interrelation between cell division and other parameters of immune responses among which is the production

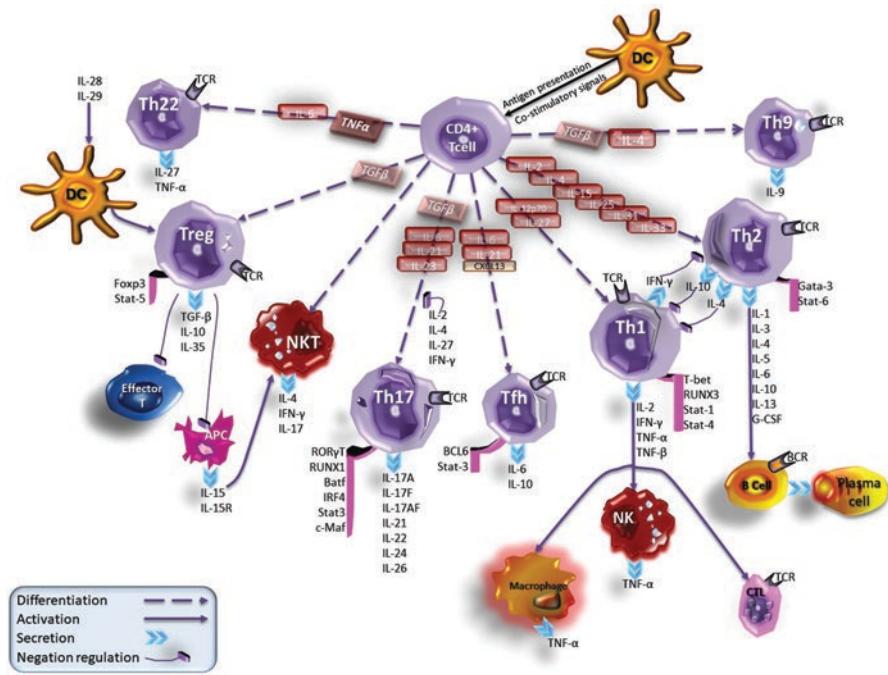
of cytokines and the availability of co-stimulation. Based on this, it is established that the link between both concepts (division and co-stimulation) form part of the regulation of the proliferative responses of this cell population [42]. Around the same time, Sebzda and colleagues (1999) work on the interactions that occur during the maturation of thymocytes that define the repertoire of T-cells [43]. Like these, many others works have allowed to characterize and distinguish the processes of development and differentiation of the various populations of T-cells.

The mammalian thymus receives stem cells from the bone marrow. About 97% of candidates T-cells die, while the remaining percentage (3%) is essential for the continuous development of the adaptive immune system [44]. Lymphoid precursors, biased or multipotent, enter the pathways of T-cell development in response to signals received from the thymic microenvironment. The Notch genes encode members of a family of receptors that mediate signaling events. Notch is a single transmembrane protein present on the cell surface. Ligands for Notch are also transmembrane proteins, and therefore, cell-to-cell contact is an essential prerequisite for triggering signaling events [45]. It is believed that Notch signaling in the thymus causes hematopoietic precursors to compromise the fate of T-cells, promote the T-cells gene expression program that prepare the TCR antigens, select the TCR-based repertoire and train it to assume functional roles like immune effectors. Nevertheless, there are yet many questions to be answered in relation to the molecular mechanisms that regulate this commitment [26].

T-lymphocytes are fundamental regulators of mammalian immune response to pathogens and tumor cells. All T-cells express a heterodimeric T-cell receptor (TCR) on their surface, which consist of alpha / beta or gamma / delta dimers. TCR confers antigen specificity on T-cells and proves to be essential for its role in the adaptive immune response and the generation of antigen-specific memory T-cells. By virtue of their function in the immune system, the T-cells are classified as: T-helper cells (effectors), and regulatory T-cells or gamma/delta T-cells [17] (see Fig. 1).

Both cell types, T and B, have receptors that recognize antigens. Antibodies on the surface of B-cells are the main, although not the only, source of antigens recognition. When they are activated, they allow initiating the processes of cell differentiation. From the plasma cells are produced and secreted antibodies in the blood and in the body fluids, in order to prevent the adverse effects of the antigen. T-cells equally have receptors, which under the context of the major histocompatibility complex (MHC), provide a means of self-recognition and effector functions that allow their interaction [46]. The interactions between B-cells and T-cells involve the upregulation of the following molecules: proteins of MHC class II; CD80 (B7.1) and CD86 (B7.2), which contact CD28 and CTL-A on T-cells; and CD40 which contacts CD40L on T-cells [47].

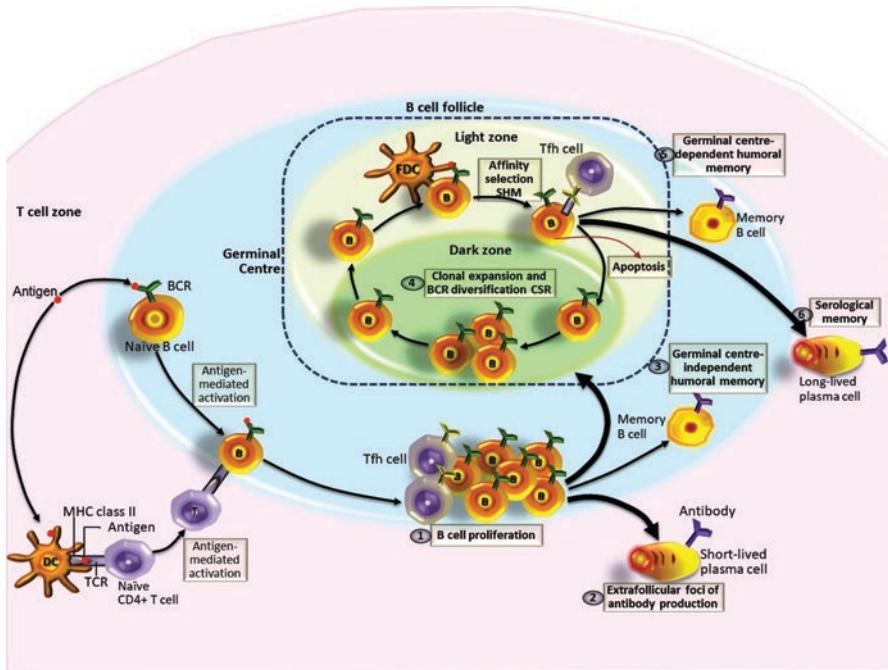
When infection caused by viruses is active, the generation of pathogen-specific IgG antibodies involves the differentiation of antigen-specific B cells. The process regularly requires the help of subgroups CD4+ T-helper cells, of which the follicular T-cells (Tfh) seem to be more dedicated to this task. The subgroup Tfh cells is characterized by the transcription factor BCL6 (*B cell lymphoma 6*), several surface



**Fig. 1** Differentiation of CD4+ T-cell types. The high plasticity profile by CD4 T-cells allows them to change from one type of cell to another. Based on their respective master regulators, cytokine profiles, chemokine responses, and interactions with other cells, these subgroups of T-cells can promote diverse types of reactions, fundamental in host defense [17, 97]

markers (CXCR5, PD-1), ICOS (*inducible T-cell costimulator*), and the cytokine IL-21, which together support the B-cell proliferation and the formation of immunoglobulins IgM, IgG and IgA. Dendritic cells (DCs) generally regulate the differentiation of T-helper cells and define the quality of the help that T-cells will provide to B-cells. The B-cells primed by the antigen, but not by DCs, induce stable co-expressions of Th1 (Tbet and IFN- $\gamma$ ) and Tfh (BCL6, IL-21, and CXCR5). In addition, the IL-6 and IL-21 cytokines derived from B-cells induce IL-21 and IFN- $\gamma$  respectively, with IL-21 being the key interleukin in antibody responses [48] (see Fig. 2).

The direct stimulation of T-helper cells, by means of activated B-cells, leads to a rapid and efficient induction and expansion of both cellular populations [49]. Naïve T-cells continuously circulate between the blood and the lymph nodes to search for antigens. The encounter between a peptide - MHC complex that is recognized by TCR, will result in the activation of T-cells prior to co-stimulation provided by mature DCs. These professional APCs collect antigens in the peripheral tissues and migrate to the lymph nodes through the veins. T-cells primed by DCs induce marker activation, cytokine secretion and cell-proliferation [50] (see Fig. 2).



**Fig. 2** Dynamics among some cell groups mediating immune responses. This figure illustrates the activation of key processes that are part of the preparation of the adaptive immune response. The activation of these processes results from the interactions that arise among certain cell populations on specific zones [318, 320]. **Zones:** T-cell zone (pink-color area); B-cell follicle (blue-color area); Germinal Centre (dotted area); light zone (yellow-color area); dark zone (green-color area). **Processes:** (1) B cell proliferation; (2) Extrafollicular antibody production; (3) Germinal centre-independent humoral memory; (4) Clonal expansion and BCR diversification; (5) Germinal centre-dependent humoral memory; (6) Serological memory

### 2.3.4 Pattern Recognition Receptors

Pattern recognition receptors (PRRs) are proteins expressed by a variety of cellular populations, especially by groups of effector cells. These groups include: macrophages, monocytes, neutrophils, DCs, NK-cells, B-cells, and specific types of T-cells. Also included are non-immune cells, such as epithelial cells and fibroblasts, responsible for detecting the presence of invaders by recognizing specific ligands that then allow activating their corresponding signaling pathways. Each of the members of these receptor families is distinguished by its ligands specificity, cellular localization and a single downstream signaling pathway.

Damage-associated molecular patterns (DAMPs) are endogenous molecules released from host cells under non-infectious conditions, such as stress, injury or cell death [51, 52]. These molecules include heat shock proteins, high-mobility group box 1 (HMGB1), uric acid crystals, heparin sulfate, messenger RNA,

surfactant protein A, and various extracellular matrix products such as fibronectin and fibrinogen [53].

The PRRs that detect viral and microbial components are known as pathogen-associated molecular pattern (PAMPs) [54]. PAMPs are exogenous molecules derived from both pathogenic and nonpathogenic microbes [52]. These are normally specific for one type of pathogenic antigen, but in general, sufficient to be present in an enormous number of particular pathogens [55]. PAMPs are recognized by Toll-like receptors (TLRs) and include molecules expressed by viruses, bacteria, fungi and protozoa [56, 57], and their molecular structures are glycoproteins, lipopolysaccharides, proteoglycans and nucleic acid motifs [58]. Viral PAMPs are composed primarily of unique nucleic acids, including double-stranded RNA (dsRNA), single-stranded RNA (ssRNA) and cytosolic DNA [54]. The innate immune system recognizes PAMPs that are expressed on infectious agent, but not on the host [59].

Both DAMPs and PAMPs activate process of immune responses [60, 61]. Each family of PRRs is differentiated by its structure, the mechanisms it uses in the process of recognizing pathogens and the initial immune responses it triggers. The activation of these responses results in the induction of pro-inflammatory cytokines, interferons type I and chemokines. Then, these products bind to membranes to recruit members of the kinase associated with the interleukin-1 receptor (IL-1R) and the family of associated factors. This process stimulates the expression of the gene through the activation of transcription factors NF- $\kappa$ B, interferon regulatory factors (IRFs) and MAP kinases, which together participate in innate and adaptive immune responses. Based on these criteria, the following five families of receptors have been identified: TLR (*Toll-like receptor*), NLR (*Nucleotide-binding oligomerization domain (NOD)-like receptor*), CLR (*C-type lectin receptor*), RLR (*Retinoic acid inducible gene-I (RIG-I)-like receptor*) [57, 58, 60, 61], and ALRs (*absent-in-melanoma (AIM)-like receptors*). The TLRs and CLRs families are located in the plasma membrane, whereas NLRs, RLRs and ALRs are intracellular receptors [62]. The TLRs detect PAMPs in the extracellular space and the endosome, and the NLRs, RLRs and ALRs function as pathogen detectors in intracellular compartments [54].

### 2.3.5 NOD-Like Receptors

NOD-like receptors (NLRs) are a class of PRRs that respond to host perturbations either from infection agents or cell stress [63]. The NLRs receptors are important molecules that participate in the recognition and initiation of immune responses to intracellular pathogens [64]. The organization of the general domain of NLRs includes: an amino-terminal effector binding region that consists of protein-protein interaction domain, such as the CARD (*caspase-recruitment*) domain, PYD (*pyrin*) domain or BIR (*baculovirus inhibitor repeat*) domain. In addition, it includes an intermediate NOD, which turns out to be necessary for nucleotide binding and self-organization; and an array of LRR (*leucine-rich repeat*) carboxyl-terminal motifs,

which are presumed to detect conserved microbial patterns and modulate NLR activity [54].

This family is made up of a group of cytoplasmic receptors that plays a fundamental role in the immune response through the recognition of PAMPs and DAMPs. In the human genome there are 22 known NLR members and these are classified into four subfamilies: NLRA, NRLB, NLRP, and NLRC. The NLRA subfamily includes a sole member, known as CIITA (*class II transactivator*). The NRLB subfamily consists of only one member, known as NAIP (*neuronal apoptosis inhibitory protein*). The NLRC subfamily consists of six members: NLRC1 (NOD1), NLRC2 (NOD2), NLRC3, NLRC4, NLRC5, and NLRX1. The NLRP subfamily consists of 14 members named from NLRP1 to NLRP14 [62]. However, there are also 34 members in the genome of mice and more than 200 members in some invertebrates. Although many of its functions are still unknown [65], it is well-known that the number of receptors and paralog genes (that is, genes of the same species that have evolved by genetic duplication), vary significantly between species [64].

This diversity among NLRs is derived from the specificity of the ligand conferred by a central NOD domain (also known as NACHT), COOH-terminal leucine-rich repeat (LRR) and a NH<sub>2</sub>-terminal protein-protein interaction domain, by means of which the activation of different biological pathways is triggered. The NACHT domain consists of seven different conserved motifs [64]. The NLRs recognize several ligands from microbial pathogens (peptidoglycan, flagellin, viral RNA, fungal hyphae, and others), host cells (ATPs, cholesterol crystals, uric acid, and others), and environmental sources (alum, asbestos, silica, alloy particles, UV radiation, skin irritants, and others). Several NLRs act as PRRs, recognizing the referred ligands and activating inflammatory responses. However, some NLRs cannot act as PRRs but instead respond to cytokines such as interferons [62].

The NLRs perform crucial roles in several biological processes, among which are: regulation of antigens presentation, detection of metabolic changes in the cell, modulation of inflammation, embryonic development, cell death, and differentiation of adaptive immune response [64]. According to functions performed by active NLRs, it is proposed to classify them into four main categories: signal transduction, transcription activation, autophagy, and inflammasome formation [62].

In signal transduction, both NLRC1 (NOD1) and NLRC2 (NOD2) active the NF- $\kappa$ B and MAPK signaling pathways, and succeed in doing so through interaction with a downstream adapter molecule known as RIP2 (*receptor interacting protein* 2). The activation of these signaling pathways plays a fundamental role in the host immune response. When NF- $\kappa$ B is activated, it can move to the nucleus and improve the transcription of pro-inflammatory cytokines; and when the MAPK signaling pathway is activated, the secretion of some pro-inflammatory cytokines begins. NOD2 can also detect viral ssRNA, which allows it to activate the production of interferon and antiviral defense. NOD1, NOD2, NLRC2, NLRC3, and NLRC4 act as negative regulators of the NF- $\kappa$ B pathway by modifying TRAF6 [62]. NLRP2 and NLRP12 are also reported as negative regulators of immune response, since they interfere with the activation of NF- $\kappa$ B. Likewise, NLRC3 is identified as a

negative regulator of T-cells activation, and also for inhibiting the activity of TRAF6 dependent on TLR4. In antiviral immunity, it has been shown that NOD1 can induce IFN type I through the activation of the ISGF3 pathway, in response to Helicobacter Pylori [64]. For its part, NOD2 is associated with the mitochondrial antiviral signaling protein (MAVS) through the induction of IFN type I in response to viral infection [54].

The regulation of the MHC class-I and class-II genes, as well as their accessory molecules, turn out to be fundamental in the adaptive immune response. Although the expression of MHC genes depends on several transcription factors such as NF- $\kappa$ B, IFNs, RFX, CREB1, ATF1 y NFY, the expression of MHC class-I request the presence of NLRC5, and the expression of MHC class-II requires the presence of NLRA (CIITA) [62]. Nevertheless, NLRA and NLRC5 do not seem to interact directly with motifs linked to DNA, but they cooperate with regulatory elements conserved MHC. NLRC5 is constitutively expressed in different tissues, but is predominantly expressed in hematopoietic cells, especially lymphocytes [64]. When NLRC5 is induced by IFN- $\gamma$  through the activation of STAT1 it acts a transactivator of the MHC class-I gene by assembling RFX, ATF1 and NFY, on the SXY module in the MHC class-I promoter [62]. In addition, NLRC5 can be upregulated modestly by IFN- $\beta$ , Poly I:C, LPS and viral infection. NLRA (CIITA) is constitutively expressed in the antigen presenting cells and is also induced by IFN- $\gamma$  by non-hematopoietic cells. When NLRA is induced by IFN- $\gamma$ , the cytokines TGF- $\beta$  and IL-10 can deregulate their expression [64]. NLRA functions as a transactivator of MHC class-II gene expression [62]. The regulation of MHC class-II genes by CIITA is complex. Its regulatory dynamic try ensuring an appropriate and specific temporal expression of the cell, probably to avoid the presentation of antigens restricted to MHC class-II, which seems to be little desired by other cells types [64].

Autophagy represents a fundamental intracellular process for cellular homeostasis and the recycling of organelles and damaged proteins, as well as for the destruction of intracellular pathogens [66]. Autophagy in innate immunity has been termed “*xenopahgy*”. This represents an autophagic pathway that focuses on bacteria and intracellular viruses, which means another mechanism in the defense of the host against cytoinvasive bacteria [64]. Autophagy can negatively regulate the activation of inflammasomes, either by removing their endogenous activators or by eliminating them directly together with their downstream cytokines [66]. NOD1 and NOD2 are involved in cytoplasmic recognition of bacterial pathogens [64]. They can induce autophagy to remove pathogens by recruiting ATG16L1 through the plasma membrane at the site of bacterial entry [62]. ATG16L1 is an essential protein for the initiation of autophagy that results in increasing activation of caspase-1, as well as in the increased production of IL-1 $\beta$  and IL-18 in macrophages after an endotoxin treatment [66]. NLRX1 located in the mitochondria, regulates autophagy induced by viruses. NOD2-mediated autophagy is important for bacterial clearance through promoting the presentation of MHC class-II antigens [62, 64]. Typically, NOD1 and NOD2 represent another layer of able host defense to coordinate an immune response by examining the cytosol due to an invasion of pathogens. NLRP4, NLRP5 and NLRP7 have been reported to play significant roles in autophagy [64]. NLRP4

is known as a negative regulator of autophagic processes, particularly when interacting with Beclin-1, a central membrane that allows coordination of the autophagic machinery [62, 64]. It has been suggested that inflammasomes and autophagy mutually regulate each other [66].

The inflammasomes are complexes of multimeric and formed proteins in a cell to orchestrate host defense mechanisms against infectious agents and physiological irregularities [65]. The inflammasomes represent a critical system for the monitoring of cytosol, a complex of macromolecular proteins that promotes the proteolytic maturation and release of pro-inflammatory cytokines (IL-1 $\beta$  and IL-18), as well as the induction of pyroptotic cell death [25]. The assembly of the inflammasome complex is initiated by NLRs receptors or absence of ALRs receptors. Both NLRs and ALRs mediate the recognition of PAMPs liberated during infections caused by viruses, bacteria, fungi, and protozoa, or DAMPs released in the course of cell damage [65]. Atypical inflammasome contains three essential elements: (i) a protein detector that belongs to a PRR family (whether it belongs to the NLR family or is absent from the ALR family); (ii) an adapter protein like ASC (*apoptosis-associated speck-like*), which contains a caspase recruitment domain; and (iii) pro-caspase-1 [66]. The inflammasome formation is triggered by sterile activators or activators associated with the pathogen. Sterile activators include DAMPs that emerge from diverse sources: those that derive from themselves, some derivates from activators associated with the pathogen, and others that are derived from stimuli caused by the environment. Activators associated with the pathogen include several PAMPs that can be derived from viruses (RNA or M2 protein), fungi and bacteria [62].

Members of NLRs represent critical components of inflammasome activation since they serve as platform for protein interactions with pro-caspase-1 [64]. Within the inflammasomes detected by PRRs, particularly members of the NLRs and ALRs families, they tend a bridge for the activation of pro-caspase-1 by means of the ASC adapter protein. For this reason, PRRs recruit ASC by means of PYD-PYD interactions and then ASC is linked to pro-caspase-1 through CARD-CARD interactions. PYD was originally discovered in the Pyrin protein [25]. At present four inflammasomes are identified, three of them are regulated by NLR members (NLRP1, NLPR3 and NLRC4), which interact with the adapter protein ASC (protein containing PYRIN-CARD) [64]. The ASC protein proves to be fundamental in the recruitment of caspase-1 whose activity induces the proteolytic processing of pro-IL-1 $\beta$  and pro-IL-18 [65]. The fourth inflammasome (AIM2) is regulated by a member of the ALRs family [64], which operates as a nucleic acid detector that responds specifically to double-stranded DNA (dsDNA) derived from either the host or the pathogen. The inflammasomes NLRP1 and NLRC4 are activated by specific PAMPs, such as *myramyl dipeptide* and *flagellin* respectively. The inflammasome NLRP3 can be activated by pathogenic microorganisms and endogenous mediator such as ROS (*reactive oxygen species*), mitochondrial DAMPs and ATP (*adenosine triphosphate*); as well as by crystalline structures, fibrillar proteins and environmental irritants. The inflammasome NLRC4 is activated after binding with specific members of the NAIP (*neuronal apoptosis inhibitory protein*) family [66]. When the six

NLRs located in the cytoplasm (NLRP1, NLRP3, NLRP7, NLRP12, NLRC4, NAIP) and two located in the nucleus (CIITA, NLRC5) detect these DAMPs or PAMPs [62]. The NLRs oligomerize and recruit ASC through interactions with the PYRIN-PYRIN domain. Consequently, pro-caspase-1 binds to ASC through CARD-CARD domains, which completes the formation of inflammasome [62, 64]. NLRP1, NLRP3 and NLRP6 activate the inflammasome through the recruitment of the ASC adapter protein by the Pyrin domain. This event triggers caspase-1, which in turn leads to the processing and release of IL-1 $\beta$  and IL-18 [64]. NLRP6 and NLRP12 have shown dual functions in certain infectious processes that depend on the context. They can activate caspase-1 but also act as negative regulators of inflammasome. However, it is unknown if there is a molecular change that privileges one role over another [65]. NLRP6 is shown as a negative regulator of inflammasome signaling and one of its functions is to prevent the clearance of both bacterial pathogens: Gram-positive and Gram-negative [67]. NLRP12 negatively regulates inflammatory responses dependent on NF- $\kappa$ B [65]. NLRC4 (without PYRIN domain) can form two types of inflammasomes. In one scenario, the recruitment of ASC by the inflammasome NLRC4 results in the production of IL-1 $\beta$  and IL-18. In another scenario, the inflammasome NLRC4 formed without the recruitment of ASC ends in pyroptosis (a pathway for programmed cell death mediated by the activation of caspase-1) [62]. NLRP10 (without LRR domain) is reported as a negative regulator of inflammasome and at the same time as an important initiator of adaptive immunity [64]. However, Eisenbarth and colleagues (2012) found evidence that NLRP10 does not work through inflammasome to modulate the activity of caspase-1 nor does it regulate other inflammasomes [63].

Three different inflammasome complexes are involved in antiviral immunity. NLRP3, RIG-I and AIM2. After cell secretion, IL-1 $\beta$  and IL-18 induce several biological and associated effects with infection, inflammation and autoimmune process. IL-1 $\beta$  participates in the generation of systemic and local responses to infection, injury and immune challenge; usually generating fever, activating lymphocytes and promoting infiltration of leukocytes in sites of injury or infection. IL-18 lacks the pyrogenic activity of IL-1 $\beta$ , and its primary function appears to represent the induction of IFN- $\gamma$  production [54]. The biological activities of IL-1 $\beta$ , IL-18 and pyroptosis are enormously beneficial for the host during an infection. However, IL-1 $\beta$  and IL-18 induced by endogenous damage signals trigger sterile inflammation, a risk factor for the development of metabolic and autoinflammatory diseases [65]. The inflammasome activation may upregulate autophagy in an attempt to protect the host from excessive inflammation [66].

### 2.3.6 AIM-Like Receptors

ALRs (*absent-in-melanoma (AIM)-like receptors*) are intracellular PRRs [62] that respond specifically to viral DNA, and are believed to be key players in antiviral immunity [64]. ALRs share a partially conserved HIN-200 (*hematopoietic interferon-inducible nuclear antigens with a 200-amino-acid-repeat*) domain at the

C-terminus that detects dsDNA. They may also contain a Pyrin domain of homotypic protein-protein interaction in the N-terminus that binds to a downstream adapter [68]. The ALRs can be directly associated with their ligand (dsDNA) through the HIN-200 domain [66]. ALRs recruit a bipartite protein known as ASC to compromise the activation of caspase-1. In macrophages or dendritic cells, ALRs that form inflammasomes induces the reorganization of cytoplasmic ASC, which is considered inflammasome assembly marker [65]. ALRs directly assemble filamentous signaling platforms, called inflammasomes, on foreign dsDNA [69]. While four ALRs members have been reported in the human genome, 13 members have been found in mice, but many of them remain uncharacterized [70]. Two principal members of the ALRs family are AIM2 (*absent in melanoma 2*) and IFI16 (*interferon gamma inducible protein 16*) [68, 326].

AIM2 is an inflammasome receptor for dsDNA, which consists of the HIN-200 and PYD domains. The first domain is positively loaded covering the dsDNA; and the second, recruit ASC to activate caspase-1. After nucleation initiated by AIM2, ASC and caspase-1 form filamentous structures, and are thought to eventually exhibit a single inflammasome speck observed in primary macrophages or dendritic cells [65]. It seems that PYD not only triggers signaling through binding to adapter proteins, but also maintains a self-inhibited AIM2 conformation [25]. AIM2 performs a fundamental role in the detection of both types of pathogens, both DNA viral and bacterial. AIM2 is oligomerized on cytosolic dsDNA and then associated with the adapter filament CARD (ASC), which promotes the polymerization of the pro-caspase-1 filament. After the final polymerization step, caspase-1 is activated by means of self-proteolysis, triggering inflammatory responses that include the maturation of cytokines (IL-1 $\beta$ ) and pyroptosis [68, 69]. Only a limited group of pathogens containing DNA activate the inflammasome AIM2 [65]. The cytosolic dsDNA of intracellular pathogens is detected by AIM2, and this in turn allows triggering the formation of inflammasome and release of IL-1 $\beta$ . AIM2 is involved in the activation and release of IL-18. Consequently, IL-18 contributes to the antiviral response by inducing IFN- $\gamma$  mediated by macrophages [71]. The dsDNA derived from the host can also be recognized by AIM2. Whether it is endogenous or microbial DNA itself, the activation of the AIM2 inflammasome triggers autophagy host adaptive responses, which are designed to limit excessive inflammation and restore cellular homeostasis [66]. AIM2 equally contributes to apoptotic cell death responses [65].

IFI16 resides predominantly in the nucleus and binds to the dsDNA. Human IFI16 orchestrates both dependent and independent functions of inflammasome, in response to viral infections [65]. IFI16 detect intracellular DNA and induces IFN- $\beta$ , regulating the activation of IRF3 and NF- $\kappa$ B. IFI16 is considered essential to restrict the replication of many viruses [68]. Like AIM2, IFI16 is also involved in the detection of viral DNA and triggers the induction of IFN- $\beta$ . In turn, IFN- $\beta$  acts in a paracrine form on IFN type I receptor, activating IRF1 and other transcription factors that express IFN- $\alpha$ . These events initiate a positive feedback loop that amplifies the response of IFNs type I and triggers a potent antiviral defense. Activation of AIM2 results in the formation of inflammasome, then processes

IL-1 $\beta$  and releases IL-18. The secretion of IFN- $\beta$ , IL-1 $\beta$  and IL-18 by keratinocytes and other skin-resident cells, then activates the innate and adaptive immune cells. IFI16 represent a negative regulator of inflammasome activation of AIM2 in keratinocytes, and this can also inhibit caspase-1 and enable AIM2 in monocytes. For its part, AIM2 can also act as a physiological inhibitor of the IFN- $\beta$  response in keratinocytes [71]. It has been observed that IFI16 initiates its inflammasomes activities in response to Kaposi sarcoma-associated herpes virus, Epstein-Barr virus and herpes simplex virus-1 [65].

Khare and colleagues (2014) found that the POP3 (*pyrin domain-only protein 3*) can maintain a balanced inflammasome response in humans. This balance is achieved by inhibiting the assembly of inflammasomes ALRs in immunogenic DNA responses. While other POPs are related to the inflammasome ASC adapter, POP3 interacts with the PYD adapter, thus preventing the recruitment of ASC [70]. The structural differences within ASC-PYD when interacting with variations determined by PYDs in the PYD-PYD interaction mechanisms, probably present difficulties for the assembly of inflammasomes and consequently suffer repercussions on the innate immune response [25].

### 2.3.7 TOLL-Like Receptors

Toll-like receptors (TLRs) are a class of transmembrane pattern recognition receptors [62] capable of identifying PAMPs and DAMPs, which form part of an alert mechanism that allows the immune system to prepare and activate its defense strategies. Until now ten of its members (TLR1 - TLR10) have been identified, both in humans and in mice, and two more (TLR11 and TLR13) only in mice [72].

TLRs represent a family whose primary responsibility within the human immune system is to recognize PAMPs expressed by some infectious agents. These receptors activate a signaling pathway known as NF- $\kappa$ B (*nuclear factor kappa-light-chain-enhancer of activated B cells*), which is responsible for regulating the expression of some cytokines through adapter molecules. The activation of this pathway is enormously significant because it allows enabling the link between the innate and the adaptive immune responses, producing inflammatory cytokines and co-stimulatory molecules. Within this environment there are also some components that activate apoptosis pathways, all of which form part of the defense mechanisms used by immune system. Additionally, TLRs are defined by some researchers as one the mechanisms that perceive the invasion of pathogenic microorganisms [73], transmembrane glycoprotein receptors that act as detectors of microbial components [52, 58], prototypes of pattern recognition receptors [74] and protective immune sentinels [73, 75], which belong to one of four classes of PRRs, which identify evolutionary conserved molecular structures [58]. These structures bind to certain compounds to trigger responses to the immune challenges they face [57], lead inflammatory pathways, coordinate effector functions of innate immunity, regulate cell proliferation, integrate repair processes and tissue regeneration [56, 60].

In addition to being recognized as receptors that are part of the innate immune system, they are also identified as immunoregulators that lead adaptive pathways. Therefore, it is currently well-known that TLRs are involved in innate and adaptive immune responses [56, 58, 60, 74–76], performing a fundamental role in the protection of the host against invading pathogens, regulating the activity of epithelial cells that act as the first line of defense in mucosal sites [60]. It has been shown that TLRs are expressed not only in immune cells but also in cancer cells [51, 60, 77, 78]. Some researchers identify a dual role in the processes of TLR signaling, as a result of the interactions between tumor-cells, immune-cells and pattern recognition molecules. Under a first role, the TLRs can induce a tumor microenvironment as a consequence of an uncontrolled signaling process which leads to the proliferation of tumor-cells, in that way evading the immune response. In a second role, under a protector effect, these receptors can induce the inhibition of tumor progression as a consequence of an antitumor response [60, 77].

In terms of their structure, TLRs are a family of transmembrane receptors that have been preserved through evolution [56]; they are structurally characterized by leucine-rich repeats and signaling domains of the Toll/IL-1 receptor (TIR) [52, 79]. They have an extracellular junction region (N-terminal) that contains multiple LRRs, to give shape to a structure similar to the horseshoe of a horse. Equally they include an intracellular region (C-terminal) that shows similarity with the intracellular domain of the IL-1R receptor, a region referred to as the TIR homology domain, which mediates the signaling events after activation of the receptors [58]. The TLRs and IL-1R typically possess a conserved region of approximately 200 amino acids in their cytoplasmic tails, which is known precisely as the Toll / IL-1R (TIR) domain. Within the TIR domain, the region of homology compresses three conserved boxes, which are crucial for signaling. The conservation of amino acid sequence between the TIR domains varies in size [73]. TIR domain adheres to a Toll-IL-1R receptor that enables cytoplasmic signaling to transmit TLR activation to various effectors within the cell. The extracellular domain (N-terminal) is composed of approximately among 16 and 28 LRRs. Each repeat contains among 20 and 30 amino acids with the conserved motif “LxxLxLxxN”. The intracellular domain (C-terminal), known as TIR, is required for the interaction and recruitment of a variety of adapter molecules to activate the signaling pathways [57]. TLRs recognize groups of structurally similar and widely distributed molecules, in contrast to the extremely selective molecular level recognition of T-cells and B-cells. The TLRs of the cell surface identify bacterial pathogens, fungi and protozoa by recognizing external molecules on these organisms, while viral infections are recognized by the presentation of nucleic acids into intracellular compartments [80]. The explanation of the crystal structure of several TLR ectodomains have provided structural signals that suggest that various PAMPs act as ligands for TLRs [81]. It has been shown that the elaborate crystal structures of several TLRs with their ligands have complex heterodimeric forms, such as occurs with TLR1-TLR2 and TLR4-MD2; and homodimeric forms such as TLR3-TLR3. Both structures are essential for the binding of a ligand and the initiation of downstream signaling pathways [57].

**TLR subtypes** The recognition of PAMPs by TLRs links the molecular signatures of the pathogen to the activation of the innate immune response [82]. TLRs play a role in the mediation of the recruitment of infected tissues and in the uptake of microorganisms by phagocytic cells. The activation of antigen-presenting cells (APCs) and the stimulation of immune responses, mediated by B-cells and T-cells, are due to TLRs [60]. On innate immune myeloid cells, the TLRs induce the secretion of pro-inflammatory cytokines to capture lymphocytes, which then mount an antigen-specific adaptive immune response to definitely try to eradicate the invaders [75]. Plasmacytoid dendritic cells (pDCs) and inflammatory monocytes have unique signaling pathways that govern antiviral responses that are probably absent in other cellular types [81].

The recognition of specific PAMPs and the identification of ligands by means of signaling pathways, trigger TLRs that lead to the expression of several genes involved in a defensive response, which in mammals initiate inflammatory processes and coordinate effector functions of tissue regeneration [56]. The concert between PRRs and TLRs secures a fundamental participation in the innate and adaptive immune responses [81]. Considering the specificities and expression of TLR ligands in the different cell types, the PAMPs recognized by these receptors and derived from bacteria, viruses, parasites and fungi, are classified into three categories: proteins, nucleic acids, and lipid-based elements [83]. The recognition of PAMPs by means of TLRs occurs in several cellular compartments, and these include: plasma membrane, endosome, lysosome, endolysosome. It is believed an appropriate cellular localization is significant for the accessibility of the ligand and the maintenance of tolerance to own molecules [81].

According to their cellular location, the TLRs are divided into two categories. Some are expressed exclusively on the cell surface and others intracellularly, the latter being transported to the intracellular vesicle by means of a transmembrane protein, located in the endoplasmic reticulum [57, 81]. Some nucleic acids recognize intracellular TLRs located within the endosomal / lysosomal compartments and endoplasmic reticulum. The intracellular TLRs include: TLR3, TLR7, TLR8 and TLR9, which specialize in the recognition of viruses through the detection of nucleic acids; and also TLR13, that is expressed only in mice. Other elements perceive extracellular TLRs that reside in the plasma membrane. The extracellular TLRs include: TLR1, TLR2, TLR4, TLR5 and TLR6, which are involved in the recognition of bacteria and fungi; and in addition, TLR10 and TLR11 expressed only in mice, all of them with the capacity to recognize PAMPs. The subcellular location of TLR4 is unique, because it is located both in the cell membrane and in the endosomal vesicle [51, 56, 57, 60, 75].

Based on the sequences of their amino acids and genomic structures, the TLRs are classified into subfamilies. The TLR2 subfamily is made up of TLR2, TLR6 and TLR1, in which TLR1 and TLR6 form heterodimers with TLR2. The TLR9 subfamily is made up of TLR7, TLR8 and TLR9, in which TLR7 and TLR8 form heterodimers with TLR9 [60]. The TLR11 family is composed of TLR11, TLR12 and

TLR13, where all its members are expressed in mice. In humans TLR11 is not functional, and TLR12 and TLR13 are completely absent [84].

With regard to the type of cells that express these receptors, the classification is as follows. TLR1 and TLR6 are ubiquitously expressed in human leukocytes. TLR2, TLR4 and TLR5, are restricted to mesenchymal cells, myelomonocytic, such as myeloid dendritic cells (mDCs), monocytes, neutrophils and basophils. Additionally, TLR2 and TLR4 are expressed in endothelial and neuronal cells and can induce pro-inflammatory response in myelocytic cells. TLR3 is expressed by means of mDCs, some groups of human B-cells, and it may be upregulated in monocytes stimulated by LPS. TLR7, TLR9 and TLR10, are expressed in plasmacytoid dendritic cells (pDCs) and in human B-cells. TLR1, TLR2 and TLR6, induce pro-inflammatory responses in myelocytic cells [51].

In contrast to the PAMPs that they recognize and the signaling they mediate, the classification is presented as follows. Lipopolysaccharides (LPS), MD2 and CD14 are linked to TLR4. Lipopeptides and lipoproteins are related to TLR2, in complex with TLR1 and TLR6. Double-stranded viral RNA (dsRNA) and its synthetic polynosinic – polycytidylic acid mimic (poly I:C) interact with TLR3. Structural epitope of bacterial flagellin is associated to TLR5. Single-stranded RNA (ssRNA) is related to TLR7 and TLR8. Single-stranded and double-stranded DNA molecules containing unmethylated CpG motifs, derived from the host or viral and bacterial pathogens, it correlates with TLR9. Ribosomal RNA in mice is associated with TLR13. The antiviral TLRs are: TLR3, TLR7, TLR8 and TLR9 [56, 57, 60, 74, 75].

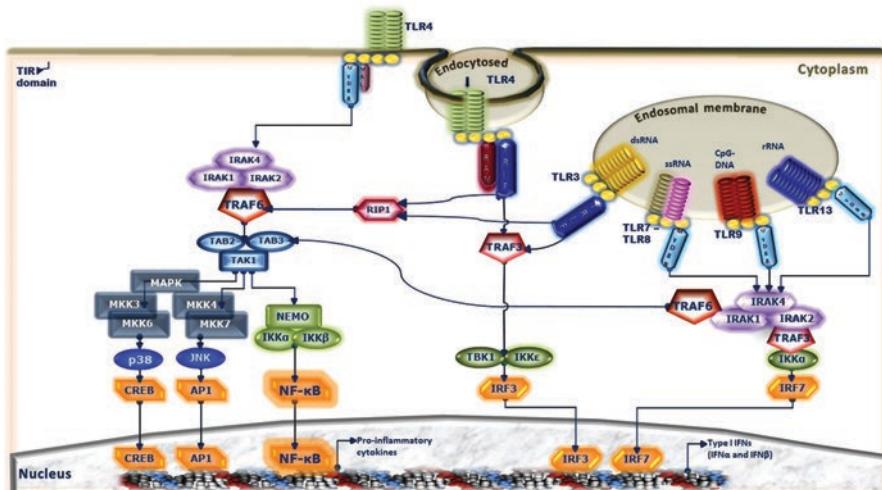
Adaptive-immunity faces the same fundamental challenge as innate-immunity, specifically when both try to differentiate self- from non-self antigens. The self-recognition through TLRs can occur and to contribute to inflammation and autoimmunity. In humans, endogenous ligands are identified by almost all TLRs, except for TLR5 and TLR10. Predominantly, endogenous ligands are molecules released from damaged tissues or apoptotic cells. In this way, diverse types of damaged tissue can produce high concentrations of endogenous TLR ligands which lead to a continuous cycle of chronic activation and defective material, via effector mechanisms activated by TLRs. These endogenous ligands can subsequently act as adjuvants or stimulate B-cell and autoreactive T-cell responses [80].

**Adaptor molecules of TLRs** Individual TLRs trigger specific biological responses. Such is the case with TLR3 and TLR4, which generate Interferon type I (IFN-I) responses and inflammatory cytokines. Cell surface receptors like TLR1-TLR2, TLR2-TLR6 and TLR5, induce primarily inflammatory cytokines. These differences are explained through the discovery of adaptors molecules that contain TIR domain, which are recruited by different TLRs to activate several signaling pathways [81]. All adapters containing TIR domains were considered soluble to cytoplasmic proteins. It was also believed they served as location motifs to facilitate delivery of the adapter to the membrane, through the interaction of a TLR with the TIR domain. However, this concept was challenged in several articles in which it

was evident that only some adapters were found in the intracellular membranes before starting the transduction process. In addition, they showed that TLR4 had the most complex signaling arrangement, capable of activating two pathways one of which illustrates the importance of cellular location for differential signaling through TLRs [74, 85].

The characterization of the signalling pathways of some TLRs and their conduction to the activation of various patterns, was given from the discovery of the independent path of MyD88 and the identification of the existence of several adaptors, which partly explain their distinct mechanisms. All these adaptors include TIR domain. TLRs and IL-1R are a family of receptors related to some innate immune response incorporating an intracellular domain called Toll-IL-1R (TIR), which selectively recognize a broad spectrum of microbial components and endogenous molecules, released by injured tissue and whose function is to recruit cytosolic adaptors that contain TIR. The TLRs bind to multiple ligands located in different types of organisms and structures, which are then activated and present adaptors to respond to the activation processes. Different adaptors are coupled to various receivers, and the adaptor used decides the signaling pathway that will be activated. The characterization of the adaptors containing TIR domain establishes essential roles in the TLR signaling pathways. According to the order of identification, the adaptors currently characterized are the following: MyD88 (*Myeloid differentiation factor 88*), TIRAP (*TIR domain-containing adaptor protein*), also known as MAL (*Myd88-adaptor-like protein*); TRIF (*TIR domain-containing adapter-inducing IFN- $\beta$* ), also known as TICAM1 (*TIR domain-containing molecule 1*); TRAM (*TRIF-related adaptor molecule*) also known as TICAM2 (*TIR domain-containing molecule 2*); SARM (*Sterile alpha and TIR motif-containing protein 1*) and BCAP (*B-cell adaptor for phosphoinositide 3-kinase adaptor protein 1*) [51, 56, 73, 85].

**TLRs signaling cascade** The key result from the interactions between the receptor and the ligand represent the coupling of the interface connecting the TIR domains and each TLR which in turn are responsible for the signaling. It is presumed that this coupling leads to a conformational change or to creating a new surface that allows the recruitment of adaptor proteins (which also contain TIR domains) to trigger the signaling pathways. The classical TLR signaling pathway leads to the synthesis of pro-inflammatory cytokines and chemokines such as IL-1 $\beta$ , IL-6, IL-18, IL-12 and TNF- $\alpha$ . A similar feature in all recognized TLRs correspond to the activation of at least three major signaling pathways: (i) MAPKs (*mitogen-activated protein kinases*), (ii) one or more IRFs (*interferon regulatory factor*), and (iii) NF- $\kappa$ B (*nuclear factor kappa-light-chain-enhancer of activated B cells*). MAPKs activate AP-1 (*activator protein 1*), a heterodimeric transcription factor comprised of Jun, Fos or ATF subunits that bind to a common DNA site. The interaction between AP-1 and NF- $\kappa$ B induces the expression of genes needed for inflammation and activation of adaptive immunity including IL-1 $\beta$ , IL-6, IL-18 and TNF (*tumor necrosis factor*). Depending on the circumstances, IRF4 or IRF7 are essential for the induction of IFN-I [56, 58, 73, 83] (see Fig. 3).



**Fig. 3** Intracellular Toll-like receptors (TLRs) signaling pathways. This figure represents the signaling pathways of TLRs that recognize viral nucleic acids, located into intracellular membranes. Next to the TIR domain (golden circles) the MyD88 (cyan), TRAM (red) and TRIM (blue) adaptors are located. From these interactions, their corresponding signaling pathways are triggered generating the activation of several of their components, among which are some kinase complexes and transcription factors. Across the signaling cascades, the pathways allowing the translocation to the nucleus of some factors that mediate the activation of certain populations of cytokines are enabled [55, 57, 73, 83]

Once the TLRs are activated by a ligand, a cascade of intracellular kinases is triggered by intermediate adapter molecules. Depending on its nature, this pathway continues with the recruitment and activation of complex IRAK, TBK, IKKs, and ubiquitin ligases (TRAF6, TRAF3 and Pellino-1). Ultimately, the coupling of the following pathways takes place: NF- $\kappa$ B, IFN-I, MAPK-p38 and MAPK-JNK [56, 75, 86].

In the TLR2 and TLR4 signaling, the MyD88 and TIRAP (MAL) adaptors lead to the activation of IRAK4 (*Interleukin-1 receptor-associated kinase 4*). ORAK4 in active state is meshing in the cascades of MAP (*mitogen-activated protein*) kinases and NF- $\kappa$ B, which leads to the induction of pro-inflammatory cytokines [83].

TLR3 initiates a TRIF-dependent pathway, which activates the expression of pro-inflammatory cytokines and IFN-I, through two independent routes. The N-terminal domain of TRIF interacts with TRAF6, while the C-terminal domain of TRIF interacts with RIP1 and activates TAK1. These two pathways can activate NF- $\kappa$ B to then induce the expression of pro-inflammatory cytokines. TLR3 triggers antiviral immune responses through the production of IFN-I and inflammatory cytokines via IRF3, suggesting that TLR3 plays an essential role in the prevention of virus [52, 81, 83].

TLR4 is the only TLR that uses four adaptors and activates the pathways dependent on MyD88 and TRIF. Initially, TLR4 recruits TIRAP in the plasma membrane and subsequently facilitates the recruitment of MyD88 to then trigger the initial activation of MAPK and NF- $\kappa$ B. TLR4 in the endosome recruits the TRAM and TRIF adaptors. This event leads to the activation of the IRF3 transcription factor and the induction of genes that produce IFNs-I. TLR4 signaling via TRIF adaptor leads to the K63-linked poly-ubiquitylation of TRAF3, subsequently promoting the IFN-I response by means of interferon regulatory factors (IRFs). TLR4 signaling via MyD88 leads to the activation of TRAF6, which modifies cIAP1 or cIAP2 with K63-linked polyubiquitin. TRAF6 in active state modifies TRAF3 with polyubiquitin linked to K48, causing its proteosomal degradation. This degradation allows the TRAF6-TAK1 complex to activate the MAPK-p38 pathway and promote the production of inflammatory cytokines. It is noteworthy that the induction of inflammatory cytokines via TLR4 signaling requires the activation of both the MyD88- and TRIF-dependent pathways. So far, the reason why one way is insufficient is a mystery [75, 81, 83, 87].

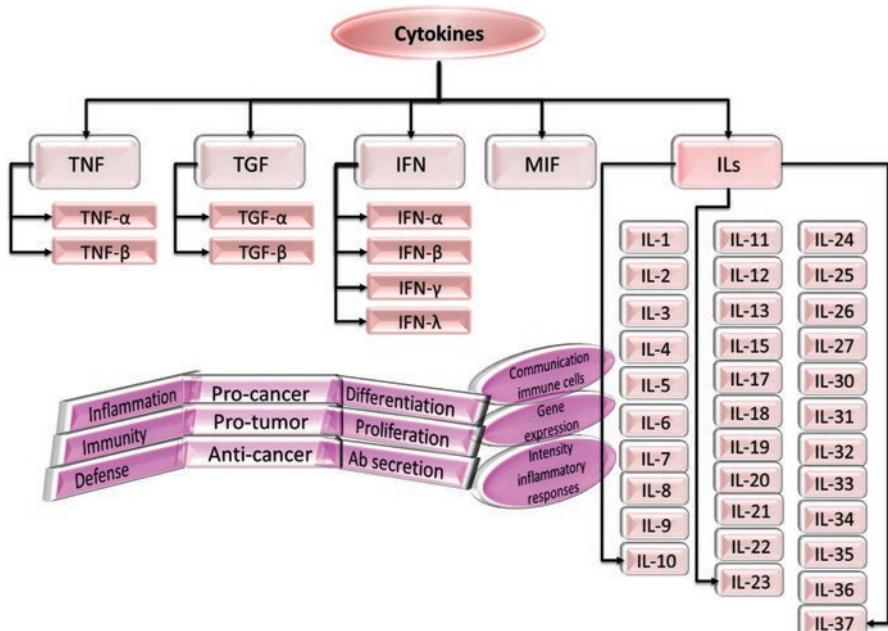
The TLR7 and TLR9 signaling pathways in pDCs have been extensively investigated for their potential to induce the production of IFN-I following viral infection. Their signaling pathways in pDCs are unique, in the sense that both require MyD88 for the induction of IFN-I. Plasmacytoid dendritic cells express IRF7 that binds to MyD88 and form a signaling complex with IRAK4, TRAF6, TRAF3, IRAK1, IRAK2 and IKK $\alpha$ . In this complex, IRF7 is phosphorylated by IRAK1 and /or IKK $\alpha$ , then dissociates from it and translocates to the nucleus. While IRAK1, IKK $\alpha$ , and TRAF3 are specifically involved in the activation of IRF7, the MyD88, IRAK4, and TRAF6 components are critical for the activation of IRF7 and NF- $\kappa$ B. Both TLR7 and TLR9 induce the secretion of inflammatory cytokines through the activation of NF- $\kappa$ B by means of MyD88. However, these TLRs can induce the expression of IFN-I through the activation of IRF7. This regulatory factor is phosphorylated by the IRAK1, IRAK4 or IKK $\alpha$  complex, and then translocated to the nucleus where it induces the transcription of IFN- $\alpha$  [52, 81]. The TLR signaling in pDCs depends on their spatiotemporal regulations related to the retention time of distinct agonists in early and late endosomes. TLR signaling in early endosomes is coupled with the production of IFN-I, whereas signaling from late endosomes is coupled with the production of pro-inflammatory cytokines and cell differentiation [88]. The TLR signaling is also deactivated by several negative controllers. These negative regulators are: IRAK-M and MyD88 short (MyD88s), which antagonize with the activation of IRAK1; FADD, which antagonizes with MyD88 or IRAKs; SHP1 and SHP2, which dephosphorylate IRAK1 and TBK1, respectively; and A20, which deubiquitylates TRAF6 and IKK [75].

The importance of the deregulation processes of TLR signaling cascades lies in the fact that they can cause severe diseases to human beings. For example, heritable deficiencies of MyD88, IRAK4, UNC93B1 o TLR3 are susceptible to recurrent bacterial or viral infections. The activation of oncogenic mutations of MyD88 can lead to lymphomas, because of its recurrent nature and its ability to activate NF- $\kappa$ B. Being able to identify patients whose tumors depend on these signals path-

ways implies that they can be benefited by therapies focused on inhibiting NF- $\kappa$ B in cancer, considering their persistent activation in many types of lymphoid cancers, as well as their relationship with solid tumors. Therapies may also be focused on IRAK4, either alone or in combination with agents targeting B-cell receptors, or NF- $\kappa$ B or Jak-STAT3 pathways. The deficiency of TLR3 in humans is associated with susceptibility to HSV-1 (*herpes simplex virus type 1*) [75, 81, 89, 90].

### 2.3.8 Cytokines, Chemokines and Growth Factors

Cytokines are chemicals made by some immune cells, which turn out to be fundamental in controlling the growth and activity of other immune cells and blood cells [18]. They are key players of the immune response and therefore maintain an outstanding participation in the understanding of pathophysiological aspects of human diseases. Therefore, the study of their networks and signaling pathways has interested the scientific community. Currently, cytokines participate in a substantial way in the research and development of new immunotherapy approaches in cancer [91] (see Fig. 4).



**Fig. 4** Families of cytokines. This figure shows the five main families of cytokines, which participate in processes of differentiation, proliferation and secretion of antibodies. Depending on their environment, they can affect the conditions of inflammation, immunity and host defense mechanisms. TNF (*tumor necrosis factor*). TGF (*transforming growth factor*). IFN (*interferons*). MIF (*macrophage migration inhibitory factor*). ILs (*interleukins*)

Cytokines, chemokines and growth factors are proteins that mediate signaling between cells and orchestrate processes of growth, proliferation and differentiation. Each of these groups is able to promote transitional states between the various stages of the cell cycle, such as the change between a resting cell and a cell that is dividing, the change from a non-secreting cell to a cell that produces antibodies or the variation of the production of antibodies from one isotype class to another. Cytokines are proteins that play a fundamental role in all immune responses both in inducing and effector phases, including cellular proliferation, intercellular communication and cell death. Chemokines are small cytokines that attract and recruit other cells in a localized area where they exert their biological effects. Chemokines are equally known as cytokines that initiate leukocyte chemotaxis. The growth factors correspond to certain types of cytokines that act on B-cells and/or T-cells. These factors provide signals that together with the signal transmitted through the specific receptor of a lymphocyte, instruct the cell to divide, differentiate or change to another state [17, 30, 46].

The cytokines signal by means of single-pass oligomers, type I transmembrane receptors, with distinct extracellular domains to bind to the ligand and intracellular domains that allow it to perform signal transduction. Cytokines receptors contain between one and three chains, one or more of which typically have limited similarity in the proximal region of the membrane (frequently referred to as box1/box2 motifs). According to the terminology, the subunit that binds to the ligand of a receptor is referred to as the alpha chain. Other subunits of signal transduction are named beta and gamma chains. All cytokines receptors are associated with one or more members of JAKs, which couple the binding of the ligand to the tyrosine phosphorylation of several signaling proteins (STATs) recruited in the receptor complex. Cytokines receptors are classified into the following five families. (i) Type I cytokines receptors: these receptors possess certain conserved motifs in their extracellular amino acid domain and lack an intrinsic tyrosine kinase protein activity. This family is divided into three subgroups, based on the ability of its members to form complexes with one of three different types of receptor signaling components (gp130, common beta chain ( $\beta$ c), and common gamma chain ( $\gamma$ c)). (ii) Type II cytokines receptors: these are multimeric receptors composed of heterologous subunits, and they are predominantly receptors for interferons. (iii) Chemokine receptors: these receptors are a seven-pass transmembrane receptor proteins belonging to a subfamily of G-protein-coupled-receptor (GPCRs). On basis of the number and spacing of the conserved cysteine residues in the N-terminal position they are subdivided in four families: CXCR (also named alpha), CCR (beta), CX3CR (delta), XCR (CR). (iv) TNF receptor family: this receptor (TNFR) is associated with procaspases through adapter proteins that can attach other inactive procaspases and trigger the caspase cascade. (v) TGF- $\beta$  receptors family: this family correspond two single-pass transmembrane receptor kinases into a kinase-active heterotetramer. These receptors are known as TGF- $\beta$  type I and II [92–95]. Cytokines induce their responses by binding to cell surface receptors with high specific affinity on target cells to initiate a series of intracellular signal transduction pathways. The receptors of several cytokines and growth factors are homologous

within their extracellular domains, whereby they can exert their effects through common pathways of transduction signals. Several of them exhibit some redundancy in their functions, share overlapping properties, as well as subunits of their cell surface receptors [96].

Proinflammatory cytokines can be induced by bacterial products by means of pattern recognition receptors. Toll-like receptors (TLRs) act at the cellular surface level and Nod-like receptors (NLRs) [97] and Aim-like receptors (ALRs) reacts at the intracellular level [25]. The activation of TLR signaling pathways leads to the release of chemokines and inflammatory cytokines [51, 56, 75, 77]. The secretion of inflammatory cytokines occupies lymphocytes to prepare an antigen-specific adaptive immune response that tries to eradicate the pathogen [75]. The activation of PRRs together with the cytokines provides essential conditions for the initiation of T-cell responses, which improves the uptake, processing and presentation of antigens. This action is achieved through the APCs via the upregulation of MHC and the expression of some molecules, including DCs, which in their maturation stage are characterized by the production of proinflammatory cytokines and the upregulation of co-stimulatory molecules. After the expression of cytokine receptors, their migration occurs. All these events together contribute to the activation of antigen-specific T-cells [51, 82, 98].

All cells that release cytokines can influence the immunoregulation process of this network. Some researchers classify inflammatory cytokines according to whether they participate in acute or chronic inflammation [96]. Other authors classify cytokines in two distinct groups. The first group corresponds to immunostimulatory cytokines, tumor suppressors or Th1-type cytokines, which induce and exhibit cell-mediated immunity. The second group corresponds to the immunoinhibitory cytokines, tumor promoters or Th2-type cytokines, which induce humoral immunity [98, 99]. Activation of B-cells and early phases of differentiation of plasma cells induce the expression of several TLRs. These receptors mediate the processes of proliferation, differentiation and anti-apoptotic effects in this type of cells that can secrete IL-6 in response to LPS, CpG DNA or synthetic ligands [100, 101].

The pattern of cytokine induction is determined by the TLR type, the nature of the ligand and the cell class that are activated. Through the induction of TLRs, some proinflammatory cytokines are released, among which are included: TNF- $\alpha$ , TNF- $\beta$ , IL-1 $\beta$  IL-6, IL-8, IL-12 e IFN-I [51, 56]. Among type I interferons, IFN- $\alpha$ 2 is the most immediate activation product [77]. In acute inflammation the following cytokines participate: IL-1, TNF- $\alpha$ , IL-6, IL-8, IL-11; and some chemokines such as: G-CSF y GM-CSF. In chronic inflammation, cytokines are classified into two groups, those that mediate humoral response and those that induce cellular responses. The first group includes the cytokines: IL-4, IL-5, IL-6, IL-7 and IL-13. The second group includes the cytokines: IL-1, IL-2, IL-3, IL-4, IL-7, IL-9, IL-10, IL-12, IFNs, TGF- $\beta$  and TNF (- $\alpha$ , - $\beta$ ) [96]. Th1-type cytokines are produced mainly by lymphocytes, APCs and NK cells. To this group belong the cytokines: IFN- $\gamma$ , TNF- $\alpha$ , IL-2 and IL-12. Th2-type cytokines are produced mainly by lymphocytes and monocytes. To this group belong the cytokines: IL-4, IL-5, IL-6 IL-8 and IL-10 [98]. Anti-inflammatory cytokines are immunoregulatory molecules that control the

response of proinflammatory cytokines. The main anti-inflammatory cytokines include receptor antagonists of the following interleukins: IL-1, IL-4, IL-6, IL-10, IL-11 e IL-13. The specific receptors for IL-1, TNF- $\alpha$  and IL-18 also function as inhibitors of proinflammatory cytokines [96].

The effects of cytokines depend on the time of release, the environment in which they can respond, the neighbors with whom they compete or have synergy, receptor density and tissue responses for each of them. These variables together with the fact that cytokines can play a dual role according to their environment, sometimes acting as anti-inflammatory and others as proinflammatory, generate a complex network of interactions. This network includes sophisticated interdependent feedback mechanisms through which it is sought to maintain control by the host [96, 102].

### 2.3.9 Transcription Factors

Transcription factors are proteins bound to specific DNA sequences. They regulate the expression of gene by controlling the transfer of DNA-RNA genetic information, either by activating or repressing gene expression. The modulation of the transcription factor in target gene lead to new molecules that appear on the cell surface, including receptors for cytokines that stimulate cell proliferation and differentiation [17].

The signaling through the T-cell receptor (TCR) allows the activation of three families of transcription factors: NFAT, AP-1 and NF- $\kappa$ B. The expression of genes by these transcription factors is unrestricted only to T-cells; on the contrary, they are found in almost all types of cells in the body [103]. The main transcription factors specifically affected by TLR signaling through MAPKs include CREB, AP-1 and NF- $\kappa$ B [104].

**Nuclear factor of activated T cells** The NFAT family consists of five members: NFAT1 (also known as NFATp or NFATc2), NFAT2 (NFATc or NFATc1), NFAT3 (NFATc4), NFAT4 (NFATx or NFATc3), y NFAT5; of all then, NFAT3 is not expressed by immune cells [103, 105]. NFAT is a transcription factor that normally resides in the cytoplasm and plays a fundamental role in the differentiation of T-cells. Its sustained signaling has been related to the promotion of Th1 cells [103]. The activation of NFAT proteins is induced by the commitment of coupled receptors to the calcium signaling pathway, such as antigen receptors that are expressed by monocytes and NK cells, or by Fc $\epsilon$  receptors that are expressed by stem cells. NFAT5 controls the expression of several cytokines, including TNF and lymphotoxin- $\beta$ , when it is induced by osmotic stress on lymphocytes. Calcium-mediated NFAT transcription factors have key roles in the regulation of many aspects of T-cell function. During T-cell activation, many cytokine genes are regulated by NFAT proteins. In turn, NFAT activation is regulated by cytokine signaling. This signaling, through the IL-2 and IL-15 receptors in human peripheral blood mononuclear cells, regulates the binding of NFAT proteins to the promoters of the

gene encoding CX3CR1 (*CX3C-Chemokine receptor 1*). IL-2 promotes the binding of NFAT2 to the CX3CR1 promoter, whereby its expression is induced; while IL-15 represses its expression, in this way inducing preferential binding with NFAT1. NFAT also focuses on other genes that control alternative functions in activated T-cells, such as cell cycle progression and cell death induced by activation. The two main signaling pathways that are induced in response to T-cell stimulation are due to the cooperation between NFAT and AP-1 proteins. The calcium signaling, responsible for the activation of NFAT proteins; and the RAS-MAPK pathway, by which the expression and activation of FOS and JUN are induced. Although AP-1 proteins may be the most important transcriptional partner for NFAT during T-cell activation, cooperation with different transcription factors could be fundamental for the regulation of other processes including: differentiation of T-helper cells, tolerance of T-cells, and development of thymocytes. The cooperation between AP-1 and NFAT proteins induces the expression of: IL-2, IFN- $\gamma$ , GM-CSF, TNF- $\alpha$ , IL-3, IL-4, IL-13, IL-15, FasL, y CD25 [105].

**Activated protein-1** The AP-1 is a wide family of protein dimeric complexes, mainly conformed by Fos (c-Fos, v-Fos, FosB, Fra1, Fra2) and Jun (c-Jun, v-Jun, JunB, JunD). Based on their potential dimerization with Fos or Jun, the following factors are also considered members of this family: ATF (ATF2, ATF3, B-ATF, JDP1, JDP2) and Maf (c-MAF, MafA, MafB, Nr1) [103]. Members of the AP-1 family are important activators of the upstream ERK signaling pathway, which are involved in proliferation, transformation, differentiation, and cell death [106].

**cAMP response element binding protein** CREB is a transcription factor known for its participation in cell proliferation, differentiation and survival. Some of its functions have been shown in immune responses, including its participation in inhibition of NF- $\kappa$ B, induction of macrophage survival; and also in proliferation, survival and regulation of B and T lymphocytes. This transcription factor differentially regulates the Th1, Th2, and Th17 responses [107]. CREB is a factor with bZIP domain, which binds to response elements of several genes involved in metabolism, transcription, immune regulation, proliferation and cell survival. Its activity is enhanced by its coactivator proteins bound to CBP and p300 [104]. The CREB family is composed of specific structural components, characterized by a transactivation domain consisting of a kinase-inducible domain (KID) and a constitutively active glutamine-rich domain (Q2), which synergizes in response to cAMP stimulation. CREB induces the transcription of immune genes that possess a CRE element, including: IL-2, IL-6, IL-10, TNF- $\alpha$ , COX-2 and MIF. This factor equally performs a specific role in the LPS/TLR4 pathway that mediates the anti-apoptotic response, dependent on NF- $\kappa$ B on macrophages, promoting their survival and improving the immune response. In addition, it plays an essential role in the production of IL-10, which in turn inhibits inflammation induced by TLR. GM-CSF (*granulocyte-macrophage colony-stimulating factor*) can induce activation of CREB, and this may be another mechanism by which CREB inhibits NF- $\kappa$ B activity, thereby decreasing pro-inflammatory responses [107].

**Nuclear factor kappa B** NF- $\kappa$ B represents a fundamental regulator of immunity [108]. This factor is maintained in the cytosol through association with its inhibitor I $\kappa$ B. In phosphorylation, I $\kappa$ B is targeted for degradation by ubiquitination, and consequently NF- $\kappa$ B becomes a free set to translocate to the nucleus where it promotes the transcription of more than 100 target genes [82]. NF- $\kappa$ B regulates the expression of fundamental genes in several biological processes, including inflammatory reactions, immune-responses, apoptosis, stress-responses, differentiation and cell proliferation [108–111]. It has been suggested its components regulate the Th1 and Th2 responses. There is consensus that all NF- $\kappa$ B members affect the proliferation of T-cells induced by TCR in some contexts. Many of the transcription factors are regulated as a result of the interaction between cytokine receptors and TCR signaling pathways. Some examples include: T-bet, Gata3, IRFs, Foxp3, and many others, which play principal roles in T-cell differentiation [103]. This factor has also been linked with functions of synaptic plasticity, memory and navigation associated with memory. However, it is pointed out that the understanding of their target genes in memory and the precision of their functions are not absolutely clear, which is why further deepening of these topics is suggested [112]. NF- $\kappa$ B also promotes the expression of several anti-apoptotic genes such as: TRAF1, TRAF2, cIAP-1 and cIAP-2 [113]. NF- $\kappa$ B is induced by various stimuli in almost all types of cells and tissues examined so far, which represents one of its most noticeable complexities. The pattern of their gene expression is usually specific to the stimulus and cell type. This suggests it may be possible to modulate NF- $\kappa$ B-dependent gene expression in one tissue, but not in others; or affect some of its target genes, but not others. Another aspect that makes it complex is based on the inactivation of the feedback loop integrated into the system, where the activation of the transcription of the genes encoding I $\kappa$ B mediated by NF- $\kappa$ B, indicates the latter is temporary. This characteristic of temporality means the regulation of feedback allows cycles of activation and inactivation, which leads to the periodic fluctuations of the nuclear activity of NF- $\kappa$ B. This property of self-regulation increases the flexibility of the NF- $\kappa$ B response allowing the type of cell or stimulus to depend on the control of one or more fluctuations, as well as the regulation of amplitude, frequency and peak width of the time-dependent response [114]. This complexity is increased by the fact that different NF- $\kappa$ B dimers exhibit differential preferences for variations in the DNA binding sequences. This causes distinct targets to be differentially induced by different NF- $\kappa$ B dimers. The binding of NF- $\kappa$ B dimers to I $\kappa$ B molecules not only prevents DNA binding, but also changes the stable state of the cytosol complex, which demands that this association be subject to dissociation and reunification processes [109].

In mammals, the family of transcription factors NF- $\kappa$ B is composed of five members: Rel-A (p65), Rel-B, c-Rel, NF- $\kappa$ B1 (p50/p105) and NF- $\kappa$ B2 (p52/p100). Its members operate as homodimers or heterodimers, which allows them to control gene expression differentially through signals generated by bacterial products, cytokines, viral expression, growth factors and stress stimuli. Based on their general structures and processing mode, the members of this family are also presented as the “NF- $\kappa$ B proteins” (p50/p105, p52/p100) and “Rel proteins” (Rel-A, Rel-B, c-Rel).

These five members share a domain of Rel homology (RHD), which is essential for dimerization, as well as for the union of related DNA elements [73, 82, 108, 109, 115]. The p50 and p52 proteins are produced by proteolytic removal of C-terminal sequences from long precursor proteins (p105 and p100, respectively), providing specificity in DNA binding to form heterodimers with other NF- $\kappa$ B subunits, which are recognized of its contribution to gene expression. Operating as homodimers (p50/p50 or p52/p52), they can positively or negatively regulate the expression of NF- $\kappa$ B target genes, by recruiting co-activators or co-repressors. Operating as dimers (p50/p52) bound to NF- $\kappa$ B elements of gene promoters, they act as transcriptional repressors. However, when p50 or p52 bind to a member that contains a transactivation domain (such as p65 or Rel-B), they behave as a transcriptional activator. The Rel-A, Rel-B and c-Rel members contain C-terminal transactivation domains, which are not present in p50 and p52 [109, 115].

The NF- $\kappa$ B dimers are usually sequestered in the cytoplasm in inactive form by NF- $\kappa$ B inhibitory molecules (I $\kappa$ B). Activation of NF- $\kappa$ B involves the phosphorylation and proteolysis of I $\kappa$ B proteins, and the concomitant release of nuclear translocation of NF- $\kappa$ B factors. The process described is mediated by the IKK complex, which compresses two catalytic subunits (IKK $\alpha$  and IKK $\beta$ ) and a regulatory subunit IKK $\gamma$  (NEMO). This process was referenced by Bonjardim and colleagues (2005) as a response to dsRNA or viral infection [82]. After its activation, the IKK complex is phosphorylated by upstream signals, which leads to its degradation. There is a broad variety of potential combinations for the NF- $\kappa$ B complex, but the most common form is represented by the integration of p65, p50 and I $\kappa$ B $\alpha$  mainly regulated by intracellular compartmentalization. There are other subunits that include p52, c-Rel, Rel-B, I $\kappa$ B $\beta$ , I $\kappa$ B $\epsilon$  and I $\kappa$ B $\xi$ . The I $\kappa$ B subunit works by masking nuclear translocation signals in the F- $\kappa$ B complex and maintains the inactive state. When stimulated by molecules such as TNF- $\alpha$ , calcium or by cellular stress, the IKK complex is activated by a series of intermediate steps that lead to the phosphorylation of I $\kappa$ B then end in ubiquitination and degradation [73, 110, 112, 115, 116]. The range of stimuli that induce NF- $\kappa$ B extends to the physical, physiological and oxidative stress, and the other functions include regulation, survival, differentiation and cell proliferation. Consequently, the deregulation of its activity is linked to inflammatory and metabolic disorders, autoimmune diseases and cancer [108]. The regulation of NF- $\kappa$ B after stimulation can occur through the activation of two routes: the canonical and the non-canonical pathways. Its stimulation may require distinct IKK complexes, induce characteristic NF- $\kappa$ B complexes, and focus on different target genes. There are also additional NF- $\kappa$ B stimulation routes, known as atypical IKK $\beta$  pathways [56, 109, 110, 115]. Many of the biological functions of TRAF are mediated by NF- $\kappa$ B, whereby TRAF can regulate positively or negatively its canonical and non-canonical signaling pathways [108].

The canonical pathway is induced by most physiological stimuli such as signals from cytokine receptors, such as: TNFR, IL-1R, antigen receptors and PRRs, including TLR4. On this pathway, defined as dependent on IKK $\beta$  and NEMO, complexes such as p50/RelA-I $\kappa$ B $\alpha$  are induced by an IKK complex that depends upon IKK $\alpha$ , IKK $\beta$  and NEMO. In its activation process, NF- $\kappa$ B is a transcription

factor in the cytoplasm of resting cells through the association with inhibitory proteins of NF- $\kappa$ B (I $\kappa$ B). After the cell stimulation process, I $\kappa$ B is phosphorylated in two critical residues conserved by the IKK complex, leading to ubiquitination and succeeding degradation. The degradation of I $\kappa$ B leads to the synthesis of cytokines and chemokines important for host defense in the initial steps of inflammatory response. The stimulatory signaling in this pathway can be mediated by TLR, IL-1R, TNFR and antigen receptors. The typical stimulating signaling molecules are: TNF $\alpha$  (*Tumor necrosis factor alpha*), LPS (*lipopolysaccharide*), and IL-1 $\beta$  (*interleukin-1 $\beta$* ). The stimulation through these receptors leads to the activation of the IKK complex, which in turn phosphorylates I $\kappa$ B $\alpha$  mainly mediated by IKK $\beta$  [56, 73, 108, 109, 115, 327].

The non-canonical pathway depends upon only IKK $\alpha$  to induce partial degradation of p100/RelB to p52/RelB. This pathway emerges from distinct factors that lead to the activation of NF- $\kappa$ B, inducing the NIK kinase, which phosphorylates and predominantly activates IKK $\alpha$ . Enzymatic activity induces phosphorylation of p100 that generates its ubiquitination and partial degradation of p52. The mechanisms associated with the activation of the non-canonical pathways are independent of the activity of IKK $\beta$  and NEMO, and are tightly regulated by sequences located in the C-terminal region of p100 [56, 73, 109, 115]. Non-canonical signaling is induced by specific members of the cytokine TNF family, such as ligand CD40, BAFF and lymphotoxin- $\beta$  [108].

It has been suggested that certain nuclear events are required for the activation of atypical NF- $\kappa$ B pathways. Some of them are genotoxic agents that induce two independent signaling pathways, such as SUMO-1(*small ubiquitin-like modifier 1*) and ATM (*ataxia telangiectasia mutant*). The modification of SUMO-1 acts as a regulatory mechanism. ATM activation corresponds to signal transduction kinases that mediate certain forms of DNA damage. However, it is unknown how ATM kinase causes inactivation of the cytoplasmic IKK complex to mediate NF- $\kappa$ B activation [109, 110].

The activation of NF- $\kappa$ B occurs by the release of I $\kappa$ B molecules or by the inhibitory dissociation of the ankyrin repeat domains of p100 and p105. The termination of transcriptional activity is achieved mainly because NF- $\kappa$ B up-regulates its own inhibitors I $\kappa$ B and also by some negative regulators, such as deubiquitinases A20 and CYLD. In acute inflammation, it usually ends with the complete inactivation of NF- $\kappa$ B during the negative regulation cycles. Under conditions of chronic inflammation, the persistence of stimuli that activate NF- $\kappa$ B seems to self-execute the inhibitory feedback loop that leads to an intense NF- $\kappa$ B activity. Other kinases that have links to NF- $\kappa$ B correspond to several members of the MAPK family, including JNK (*Jun-N-terminal kinase*) and p38, which are triggered by stimuli that activate NF- $\kappa$ B (just like with TNF- $\alpha$ ). The p38 protein and related kinases are known to be cofactors in the activation of NF- $\kappa$ B, while there are others that counteract the relationship between NF- $\kappa$ B and JNK. This indicates the signaling on NF- $\kappa$ B often depends on the type of cell or microenvironment [109]. NF- $\kappa$ B does not exist in isolation. Since its activation depends on degradation of I $\kappa$ B, the IKK complex is

the gateway for signaling, and that is why it represents a fundamental node in its interaction process [108].

**NF-kB and TLRs** Among the main activators of NF-kB are the TLRs, and these are the frontal receptors that provoke an inflammatory response by microbial infection [60]. The importance of the TLR signaling pathway dependent on NF-kB has been specifically demonstrated in weak mice in IRAK4, one of the kinases required in TLR signaling to activate IKK and induce NF-kB function [56]. Ligands that bind to TLRs lead to the recruitment of fundamental adaptor proteins. Molecules containing TIR domains include MyD88 (which is used by almost all TLRs), along with TIRAP, TRIF and TRAM. These adaptors activate a series of signal transduction molecules that include: IRAKs, TRAF6, TAK1, and the inhibitor of the I $\kappa$ B complex. These events lead to the activation of MAP kinases and the nuclear translocation of the transcription factor NF-kB, key regulators of several inflammatory response pathways [80]. All TLRs have the ability to initiate inflammation processes, inducing the production of principal NF-kB targets to bind discrete nucleotide sequences in upstream genes regions. These events generate mRNA expression and the production of proinflammatory cytokines, such as TNF- $\alpha$  and IL-6, thus regulating their expression and achieving a measure of integration between the TLRs and the NF-kB-dependent pathway [56, 73]. The N-terminal and C-terminal regions of TRIF can independently activate an NF-kB response promoter. In contrast, only the N-terminal region that is directly associated with TRAF6 and TBK1, leads respectively to the activation of NF-kB and IFN- $\beta$  gene [73]. RIP1 is also involved in NF-kB activation via the TLR3-TRIF pathway [108].

**NF-kB and cancer** The contribution of NF-kB in the initiation and progression to cancer is multiple and complex [109]. The constitutive activation of NF-kB has been found in numerous tumors including breast cancer, melanoma and squamous cell carcinoma [111]. The immune defense of NF-kB against cancer cells is usually not sufficiently fine-tuned to eliminate all aberrant cells. This event result in an equilibrium phase change, which is often followed by an escape phase of the cancer cells, in which they surpass the immune system. These phases of equilibrium and escape seem to be characterized by a chronic inflammatory condition, regularly at moderately elevated levels of NF-kB activity [109].

NF-kB is associated with the control of the initiation and progression of human cancer, and with a broad variety of malignancies that indicate a crucial role in oncogenic conversion and in facilitating late stage tumor properties such as metastasis. The NF-kB and IKK proteins, activated downstream by several oncoproteins, have been related to oncogenesis. These proteins can activate target genes in a variety of solid tumors and hematologic malignancies. Bassères and colleagues (2006) emphasize the existence of an enormous amount of data that relate the transcription factor NF-kB with a variety of oncogenic mechanisms, and highlight the importance of its role in the effectiveness of therapies in cancer modulation [115]. A key role has also been shown for the NF-kB signaling pathway in the initiation of tumor progression,

metastasis and chemo-resistance, mediated by the production of a vast variety of proinflammatory cytokines, chemokines, growth factors, anti-apoptotic proteins and collagens [60, 109, 115]. Among the mechanisms related to cancer, known to be affected in one way or another by the NF- $\kappa$ B pathway, the following have been mentioned: self-sufficiency and loss of growth inhibitory mechanism, suppression of apoptotic borders, improved angiogenic properties with ability to invade the local tissue and produce metastasis in different sites [115]. In addition to its role in survival and immune response in the presence of cancer cells, NF- $\kappa$ B can be activated in cancer stem cells (CSC). These cells can promote a proinflammatory environment, inhibit apoptosis and stimulate cell proliferation. It is believed that CSCs comprise a subpopulation of cancer cells that mediate tumor growth and resistance to chemotherapy [109].

Inflammation in general and NF- $\kappa$ B in particular, perform a double-edged role in cancer [60, 109, 115]. On one hand, NF- $\kappa$ B participates in the defense of the immune system, counting on the ability to eliminate transformed cells. Under this role, it usually upregulates anti-apoptotic genes that provide mechanisms for cell survival, which in turn allows triggering inflammatory responses and induce cytokines such as: TNF- $\alpha$ , IL-1, IL-6 and IL-8. These cytokines allow it to regulate the immune response, providing the ability to control processes of apoptosis and cell proliferation [109, 112]. Under its other role, NF- $\kappa$ B is active in several types of cancer and can perform functions that contribute to tumor progression, through the control of vascularization of tumors via the upregulation of VEGF (vascular endothelial growth factor) and its receptors [60, 109]. In addition, NF- $\kappa$ B signaling contributes to the progression of cancer, through the upregulation of MMPs (*matrix metalloproteinase*) by producing the relaxation of the extracellular matrix to provoke an evasion of cancer cells [109].

A tumor can establish an intense NF- $\kappa$ B activity, either due to intrinsic or extrinsic factors. In the first instance, its enhanced activity can be directly induced by mutations of NF- $\kappa$ B genes and/or oncogenes that activate the signaling pathway. Nonetheless, direct mutations in the NF- $\kappa$ B signaling pathway in solid tumors represent rare events. In the second instance, a tumor can reach high NF- $\kappa$ B activity through the increasing release of cytokines by macrophages in the tumor microenvironment. However, there is a mystery in the unwanted cross between solid tumors and neighborhood macrophages. While more tumors are characterized by high levels of cytokines released by macrophages (M1), classically activated such as: TNF- $\alpha$  and IL-1 $\beta$ ; the macrophages in the tumor microenvironment seem to change to the M2 phenotype, activated alternatively. M2 appears to be the predominant form of tumor-associated macrophages (TAMs), rather than those released by pro-inflammatory cytokines. NF- $\kappa$ B and IKK $\beta$  apparently polarize macrophages towards the M2 phenotype, which allows tolerating and promoting the tumor instead of attacking it [109]. The role of NF- $\kappa$ B in the control of oncogenesis based on inflammation, suggest that inhibitors of this factor serve as key components in the prevention of several cancers. This includes components that have already been approved for use in therapy, others that have undergone clinical trials and others that are shown as therapeutic promises. Therefore, it has been suggested

that blockers of the NF- $\kappa$ B signal may equally serve to arrest cell growth [115]. One of the challenges facing NF- $\kappa$ B in immunity is based on being able to differentiate situations in which it must assume a pro-oncogenic or anti-apoptotic role; in circumstances where it can intervene as a tumor suppressor or a pro-apoptotic factor; and in being able to identify the upstream regulatory pathway, under role that operates in an anti-apoptotic or proliferative way, as a therapy or adjuvant for the control of cancer [109, 115].

## 2.4 Viruses

Providing a precise and differentiated definition of the current notion of viruses seems to be fundamental to enter context, but at the same time it ends up being “something extremely difficult”, as expressed by Pradeu and colleagues (2016) [117].

In theory, each virus description could contain at least a single symbol that distinguishes it, or a set of characters that simply or in combination describe it. The initial description of a virus is often based on phenotypic characters such as: symptoms, host class, type of vector and virion form. Subsequently the details of their genetic refine such characterization, although the initial description can be done by sequencing fragments of the obtained gene, for which mixtures of group-specific primer are used [118]. There is equally an orthodox point of view against which many virologist agree, which describes viruses as sub-cellular genetic parasites which lack functional autonomy. These viruses do not self-replicate or reproduce on their own, but passively rather than actively replicate, through the metabolic activities of the infected cells. Virions possess intrinsic properties such as: size, mass, chemical composition and sequences of capsid proteins and nucleic acids. Additionally, viruses possess relational properties that arise by virtue of interactions with other objects such as a host or a vector, which are updated only during the processes of transmission and infection [119].

According to ICTV (*International Committee on Taxonomy of Viruses*), “viruses are real physical entities produced by biological evolution and genetic, whereas virus species and higher taxa are abstract concepts produced by rational thought and logic. The virus/species relationship thus represents the front line of the interface between biology and logic.” Under the lens of ICTV, any virus can be classified into one and only one species [120]. Unlike this concept, Morgan GJ (2016) proposes another approach that he himself named as “radical pluralism.” The author suggests that viruses whose evolution has been driven by horizontal genetic transfer would be better classified under a non-hierarchical system. Under this system, viruses are allowed to be members of more than one category, making it easier to generate predictions and help focus on new biological entities that are easily omitted in a hierarchical classification system [121]. Another approach is presented by Calisher CH (2016), who considers that viruses are specific (real) entities, which exhibit characteristics that species do not have. These viruses have the ability to replicate and can infect cells. The biological characteristics of viruses can be approximated to a point

where the difference between species (concepts) and their members (viruses with characteristics) is confused. A genomic sequence of nucleic acids is clearly sufficient to allow a taxonomic location, although the sequence itself represents a sequence (a chemical), not a virus (a biological entity). The taxonomic location provides a basis for comprehending the evolution of the virus (reflecting phylogeny). In addition, phenotypic attributes can be used as an adjunct but not as a substitute for genotypic characteristics [122].

Appreciating the value of all the previous definitions and many others that the authors have presented in their publications, in our artificial life environment we works with a classic interpretation in its expression and complex in its content. Morgan GJ (2016) confirms that viruses clearly possess genetic material, adapt to their environments, and evolve. In addition, they interact and co-evolve with their biological hosts [121].

#### 2.4.1 Viral Classification

The International Committee on Taxonomy of Virus (ICTV) was created as a committee of the virology division of International Union of Microbiological Societies (IUMS). Its objectives as an organization are aimed at developing internationally accepted information regarding the following topics: Taxonomy for viruses, names for their taxa, to communicate taxonomic decisions to the international community of virologists; and to maintain an index of virus names. The taxonomy of viruses differs from other types of biological classification because ICTV not only regulates a code nomenclature but also considers and approves the creation of viral taxa. Currently, this structure corresponds to the following hierarchical levels: order, family, subfamily, gender and specie. Denoting “order” as the most elevated taxonomic level, and “specie” as the lowest level. Up to 2018 there are ten categories defined at the “order” level, each of which groups a distinct number of virus families. At Order level there are the following groupings: Bunyavirales (9 families), Caudovirales (4 families), Herpesvirales (3 families), Ligamenvirales (2 families), Mononegavirales (9 families), Nidovirales (4 families), Ortervirales (5 families), Picornavirales (7 families), Tymovirales (5 families), and 86 families not assigned to a specific order [120].

In humans, the viruses that cause tumors belong to a variety of families, among which are RNA (*Retroviridae and Flaviviridae*) and DNA (*Hepadnaviridae, Herpesviridae, and Papillomaviridae*) virus families [123]. On the basis of their genetic constructs, both groups encode oncoproteins that have or not the ability to focus regulatory mechanisms common in host cells [124]. Viruses of DNA origin can be integrated into the genome of the host cell. In contrast, RNA viruses are essentially cytoplasmic and cannot be integrated into the genome of the host cell [125]. RNA viruses that cause tumors include retroviruses such as: HTLV-1 (*human T-cell leukemia virus-1*), HIV-1 (*human immunodeficiency virus-1*); and flavivirus like HCV (*hepatitis C virus*) [124]. Among the DNA viruses that cause malignancies in their natural hosts are the following: HBV (*Hepatitis B Virus*), HPV (*Human*

*Papillomavirus*), EBV (*Epstein-Barr Virus*), HHV-8 (*Human Herpesvirus 8*) which is equally known as KSHV (*Kaposi Sarcoma-associated Herpesvirus*) [123], and MCPV (*Merkel cell polyomavirus*) [124]. EBV and HHV-8 are non-integrative viruses into the host genome, whereas HBV, HPV and MCPV integrate the host genome and are involved in different carcinogenesis pathways according to the viral integration region [126]. Some studies on DNA viruses, which include: adenoviruses, polyomaviruses and papillomaviruses, have been instrumental in elucidating the molecular mechanisms underlying cellular transformations induced by viruses. Although these viruses are evolutionarily distinct, the similarity evidenced in their transformation functions is striking, emphasizing the mutual need of these viruses to use the host's replication machinery to achieve an efficient viral replication [123].

Both type of viruses, DNA and RNA, pertain a broad range of families that represent the group of tumoral viruses. They encode various types of oncoproteins, which have the ability to focus or not on regulatory mechanisms common in host cells. Acknowledging the way in which viral oncogenes modify the expression of growth-promoting factors has provided novel insight into the fundamental mechanisms of cancer development [124]. There is no obvious molecular rule that allows establishing or eliminating a priori an agent such as a virus that causes tumors in humans. In addition, almost all viruses that cause tumors have close relatives that do not trigger cancer in humans [127].

#### 2.4.2 Viral Life Cycle

Viruses have to compromise with quantities of positive and negative factors present in the target cells for their survival. In the absence of an appropriate interaction among cells, they do not replicate at all. The viral tropism can therefore be determined in each step of replication, beginning with the entry to the cells and ending with the generation of their descendants. There are two primary types of viral tropisms, that is, independent and dependent-tropisms of the receptor. Restriction of viral replication occurs on the surface of the cell (receptor-dependent viral input) and/or intracellularly (receptor-independent post-entry replication) [128]. Most DNA viruses are known to replicate their genome in the nucleus of the host cell as an inherent part of their viral life cycle [129]. Habitually the "life cycle" of a virus refers to the process of adhesion, entry of the virus into a host cell, replication within the cell and finally the release of progeny particles [125, 130]. Cellular entry represents a fundamental process of the infectivity of any virus. Cell surface receptors are the fundamental molecules in the recognition of target cells that regulate the cellular tropism and peculiarity of the species [131].

The viral life cycle presented by Nowak and colleagues (2001), poses eight key events, these are: (i) adhesion of the virus by itself to the host cell; (ii) invasion of the cell by the virus or its genetic material; (iii) genetic material of the virus keeps uncoated; (iv) production of viral proteins that initiate manipulation of host cells; (v) replication of the viral genome by the machinery of the host cell; (vi) production of additional viral proteins to be used in the capsid and envelope of new viral particles;

(vii) assembly of new viral particles; (viii) release of new particles from host cells [132]. Sequential steps in the reproductive cycle of a virus can differ substantially between specific viruses. These may influence their rates of cell elimination and propagation, therefore affecting their oncolytic potential. The number of genes and the complexity of their regulation differ widely in each virus. For many viruses, the transfer of viral nucleic acids to the nucleus is required to make use of the genome amplification and the transcription machinery of the host cell [125].

During infection, viral nucleic acids that are deposited in the nucleus or cytoplasm of abnormal cells act as potent stimulators of immune response pathways. For a protein to be classified as a viral DNA detector, it must satisfy two requisites. First, physically it must interact with the viral DNA; and the second, it must stimulate an innate immune program which means that it must induce the expression of interferons, chemokines and/or cytokines [129]. In response to a viral infection, inflammasomes detect PAMPs including double-stranded DNA [70]. The detection of nucleic acids is carried out by the PRRs, and these in turn can detect PAMPs and DAMPs [129].

In the case of viruses that cause disease, metaphorically the infectious process is presented as a war between a host and a virus, assuming the virus as such is capable of developing novel strategies and mechanisms to escape the protective immune responses of the host. However, conforming to van Regenmortel and colleagues (2016), the ability of the virus to defeat the immune system is solely due to stochastic mutations that arise from the error-prone activity of the reverse transcriptase of the viral enzyme [132]. Crow and colleagues (2015) consider the effectiveness of the immune response is influenced by the dynamic balance between host defense and the ability of viruses to block these mechanisms. One of the strategies used by viruses is to block the immune signaling through the sequestration of antiviral factors of the host and also via its degradation [129].

Tumor viruses, through their oncoproteins and other regulatory molecules, modulate almost all major signaling pathways including: MAP kinases, JAK-STAT, TGF- $\beta$ , NF- $\kappa$ B, Notch, TNF, Wnt and Hedgehog. It has been suggested that tumor viruses modulate these pathways in order to promote an environment conducive to their replication and to drive host cells to divide and proliferate actively. Viral oncoproteins usually reprogram host cells by sequestering and reusing the regulatory components of the host transcriptional networks. MAP kinases are activated in response to growth factors (ERK) or stress signals (JNK and p38), and are active in cancer conditions [124].

## 2.5 *Integration of a DNA Virus into a Human Host*

Cells always emerge from other cells, and it is these cells more than viruses that produce the virions. From these explanations, virologists distinguish various stages in the replication cycle of a virus. Initially, they define an extracellular infectious state, which corresponds to the virions. Subsequently, they describe a lytic or

replication state in an abnormal cell, which corresponds to an eclipse phase when the virions are no longer present. Ultimately they report a latent or proviral state, when the viral genome has been integrated into the host genome, but without producing virions and without triggering a host's antiviral immune responses. Virus particles, as well as genes are inert components outside the cell. When the viral genome is integrated into an infected cell, it is capable to instruct it to produce virions and viral proteins, although it is the activity of the cell that synthesizes those [132]. In DNA viruses the genetic material can be directly inserted into host genome [123].

Once a virus penetrates a tumor cell, the tumor microenvironment provides sufficient support for viral replication to take place [133]. It is believed that the sequestration of the DNA methylation machinery produced by DNA tumor viruses represent presumably a viral mechanism that promotes viral replication to evade antiviral immunity. Immunosuppression caused by aberrant DNA methylation over time can contribute to the development and progression of cancer associated with DNA tumor viruses. In contrast, viruses can also decrease DNA methylation to regulate host gene expression. Alterations in DNA methylation states of particular genes caused by viruses can have profound effects on cancer development and progression [134].

For DNA viruses that integrate into the host, their viral genome needs to gain access to the nucleus so replication can occur. Once the particles complete the assembly process, they need to be transported out of the core. The nuclear envelope acts as an initial and final barrier, both to the entry and exit of viruses, and they have co-evolved various means to interrupt this obstacle. Once in the nucleus, the viral genomes face a number of challenges to initiate and complete the replication process. Many DNA viruses do not encode their own DNA synthesis enzymes, which is closely associated with alterations in the normal cell cycle progression of host cells. In addition, the viral genomes of entry or replication can be recognized by the host as damaged DNA or "aberrant DNA", and therefore the find and need to overcome the responses that are activated by the host to eliminate this potential damage [135].

Viruses contain a series of complex mechanisms that have evolved to combat and inactivate cellular damage response pathways. Cellular DNA damage response (DDR) is a general term used to describe a series of complex cellular pathways that detect DNA damage [136]. DDR is a signaling cascade that cells mount when they detect the presence of several types of DNA damage to halt the progression of the cell cycle, which allows them to repair the damage. When the damage is severe, the cells initiate an apoptotic pathway dependent on p53 to eliminate the damaged cell. Failure to initiate DDR can lead to genome instability and the development of cancer [135].

There is a complex interrelation between the viral infection and the response pathways and DNA damage repair of the host cell. Viruses have evolved to administer the function of three major DNA damage response pathways, controlled by the three PI3K-related protein kinases, these are: ATM (*ataxia-telangiectasia-mutated*), ATR (*ataxia telangiectasia and Rad3-related*) and DNA-PK (*DNA-dependent*

*protein kinase*). These kinases regulate the control of cell cycle checkpoints, DNA replication, DNA repair and apoptosis in response to genotoxic stress [136]. The ATM and DNA-PK kinases respond mainly to breaks in double-stranded DNA, and contrary to the ATR kinase, none of them is essential for cell survival [137]. When ATM and ATR are in active state, these two kinases can phosphorylate numerous downstream effectors that are involved in signaling and repair processes. The inherent ability of certain viruses allows them to inactivate DDR that has been used in conjunction with DNA damage agents to treat tumors that otherwise turn out to be chemoresistant [135]. The infection caused by a virus is enough to start a DDR, activating some or all of the repair pathways [136].

The ATM kinase is involved in p53-dependent cell cycle checkpoint control G1/S, intra-S phase and G2/M checkpoint control [136]. ATM has a unique role in lymphocyte biology, since repairing programmed DNA damage represent part of the rearrangement of the gene necessary for the formation of a repertoire of highly diverse T-cell receptors. Under stress conditions, ATM is inactive and exists in the form of a dimer, which requires a signal for its activation. The signaling through intermolecular autophosphorylation in the S1981 ATM residue results in the dissociation of this dimer into monomers. The accumulation of damaged DNA and failures to repair the breakdown of dsDNA due to the deficiency of the ATM-dependent DNA repair machinery during chronic viral infection can cause wide implications through the deterioration of various cellular functions. The loss of T-cells mediated by damaged DNA represents a novel mechanism of immune evasion. Counteracting ATM deficiency can restore T-cell competence during viral infection and prevent premature immune aging [138].

ATR kinase regulates DNA replication during S phase, at stalled replication forks and in response to genotoxic stress [136]. ATR kinase orchestrates multiple ramifications of the stress response replication, and turns out to be essential for cell viability. Among other functions it participates in the restoration of dsDNA damage, inter-strand crosslink repair and meiosis, as well as at telomeres and in response to mechanical and osmotic stresses. ATR is compared in functional homology with two other DNA damage response kinases, such as: ATM and DNA-PK [137]. The DNA-PK kinase plays a fundamental role in the repair process of double-stranded DNA damage site, regulating NHEJ (*non-homologous end-joining*). DNA-PK consists of a large catalytic subunit called DNA-PKcs, and two regulatory subunits named Ku70 and Ku86 [136].

**ATM signaling pathway** ATM kinase normally exists as an inactive homodimer, but is activated when it responds to double-stranded DNA breaks, thus forming active ATM monomers. These monomers act as transducers and they are recruited with the MRN complex (MRE11-RAD50-NBS1) at the breaking sites. Then H2AX phosphorylated by ATM recruits the substrate MDC1, and they are located in damaged DNA sites. The complex formed by MRN, HEAT and MDC1 (acting as sensors), recruit the ligases ubiquitin RNF8 and RNF168 (acting as activators). ATM promotes the recruitment of repair proteins, such as: 53BP1, BRCA1 and BLM

(acting as adaptors), in damaged DNA sites. ATM also regulates checkpoints of the cell cycle through the activation of CHK2 and p53 (acting as effectors) [136].

**ATR signaling pathway** Several types of damage and stress active ATR, and many of them have been tracked by a common DNA structure formed in the replication fork that ATR can recognize. Several types of damage and stress can activate ATR, and many of them are tracked by a common DNA structure formed in the replication fork that ATR can recognize. At least partially, this structure contains single-stranded DNA (ssDNA) [137]. ATR is recruited at replication sites or sites of damage through the ATRIP protein (acting as a transducer), which binds directly to RPA70. Then the RAD9-RAD1-HUS1 (9-1-1) complex is associated with dsDNA junctions adjacent to ssDNA loaded with RPA, which facilitates the recruitment of several proteins, such as: TOPBP1, Claspin, Tipin, Timeless, and hnRNPUL1 (acting as adaptors). Consequently, all of them together promote the regulation of the cell cycle through the activation of CHK1, which regulates the G2/M checkpoint (acting as an effector) [136].

**DNA-PK signaling pathway** At the DDR response of the DNA-PK kinase, the Ku regulatory complex (Ku70, Ku86) (acting as sensors) recognizes and binds to the damage site. Subsequently it recruits and stabilizes the interaction of the DNA-PKcs catalytic subunit (acting as mediators) with DNA. Later, two catalytic molecules (DNA-PKcs) are bound in concert at the end of the DNA forming a synaptic complex, and then they recruit the DNA lineage IV-XRCC4 complex (acting as activators) to rejoin the broken DNA ends [136].

**Activation of the ATR pathway in bacteria** The blocking of the polymerase on the lagging strand generates a single-stranded DNA space that binds to the RPA (*replication protein A*) protein, providing a platform for ATR activation. The 5' – ended ssDNA-dsDNA junction formed at the Okasaki fragment adjacent to this ssDNA serves as the loading point for the 9-1-1 clamp complex, which is loaded onto the DNA by the RFC2-5 (*RAD17-replication factor C subunits 2-5*) clamp loader. The 9-1-1 complex, assisted by RHINO (*RAD9-HUS1-RAD1-interacting nuclear orphan*) and MRN complex, recruits the activator TOPBP1 (*topoisomerase II binding protein 1*), thus allowing the stimulation of ATR and the phosphorylation of specific downstream effectors, including CHK1 (*checkpoint kinase 1*). ETAA1 (*ewing's tumour-associated antigen 1*) joins RPA to activate ATR on a parallel pathway [137].

**ATR activators in vertebrates** Each of the regulators Dpb11 (*DNA replication regulator DPB11*)), Dbc1 (*deleted in breast cancer 1*) and Dna2 (*DNA replication ATP-dependent helicase/nuclease DNA2*) has an ADD domain, which contains fundamental hydrophobic amino acids that are necessary for the activation of the entry point to mitosis. In humans, the ATR activator TOPBP1 is a homolog of Dpb11 whereas ETAA1 is not related to any yeast protein, and unlike other

activators it contains two motifs (RPA70 and RPA32C) that interact with two RPA domains. The kinase domain of ATR is followed by two motifs that are needed for ATR activation: the PRD (*PIKK-regulatory domain*) and the FATC (*FAT-carboxy-terminal domain*), which may directly contact the AAD of TOPBP1 or ETAA1. The ATRIP () protein contains an amino-terminal interaction domain RPA70N, a dimerization domain CC (coiled-coil), a motif that is needed for the AAD interaction and ATR activation, and a carboxy-terminal region that interacts with ATR [137].

Adenoviruses inhibit the DNA-damage response by promoting the rapid degradation of key components, but they are equally capable of affecting the location of several DDR proteins. During adenovirus infection the ATR pathways are selectively regulated, a number of DDR proteins are recruited at the viral replication centers, and they can also promote their relocation to other sites. Polyomavirus seem to use the ATM pathway for optimal replication of the virus. However, the relationship between viruses of the polyomaviridae family and the damage response seems to be more complex than in the case of adenoviruses, since the behavior of the ATM pathway probably differs among several viruses of the same family. In the case of papillomavirus, it seems that HPV infection allows the activation or selective repression of DDR pathways to promote the replication of the virus. It could be thought that DDR proteins in the ATM and ATR pathways are activated selectively, and that they are not simply inevitable consequences of the infection. Nevertheless, specifically in response to the expression of HPV proteins and the production of viral DNA during infection, these pathways facilitate both the replication of the vegetative virus in the basal layer and the amplification of the viral genome in the suprabasal layers [136].

There are several double-stranded DNA viruses known to be etiological agents of cancer [139]. Although these viruses share some common characteristics during their cycle of replication and DNA-damage response, each of them possesses unique attributes. Considering our interest in dsDNA viruses that are integrated into the host, we review below information related to HBV, HPB and MCPyV.

### 2.5.1 Hepatitis B Virus

The criteria for demarcation of the species for the genus to which the Hepatitis B Virus (HBV) belongs include: nucleotide sequence diversity, differences in the host class and oncogenicity. Within the ICTV classification, HBV corresponds to the following assignment. (i) Order: unassigned; (ii) family: Hepadnaviridae; (iii) sub-families: does not report; (iv) genera: three, among which one is unassigned: (v) species: 14. Among the two genera assigned, the first is called Avihepadnavirus, which groups three species. The second genus is called Orthohepadnavirus and groups eight species. The unassigned genus groups three species. Specifically, HBV belongs to the genus named as Orthohepadnavirus [120].

From the morphological point of view, the hepadnaviruses are spherical occasionally pleomorphic, with a diameter between 42 and 50 nm. Its envelope contains the surface proteins and surrounds an icosahedral nucleocapsid core that is composed of a major protein species, the core protein. The nucleocapsid encloses the viral genome (DNA), the viral DNA polymerase and associated cellular proteins, including protein kinase and chaperones that appear to play role in the initiation of viral DNA synthesis. The virions of HBV have a diameter between 40–45 nm with an internal nucleocapsid of 32–36 nm. Typically, its subviral particles are spherical (16–25 nm diameter) and filamentous (20 nm diameter and variable in length) [120].

HBV is a double-stranded DNA virus organized in a covalently closed circular DNA (cccDNA) within the nucleus of the cell [126]. The double-stranded genome of HBV is about 3 kbp in length, as well as an endogenous DNA polymerase activity that allows it to incorporate nucleotides into the viral DNA genome [140].

At the protein level, HBV itself encodes seven proteins: S, M, L, preCore, core, pol and HBx. Within this set, the S, M and L are surface proteins. The cover or surface proteins are partially glycosylated and establish a C-terminal nested set. The L protein has about 400 amino acids (aa); S protein possesses 226aa residues that function as a main envelope protein; and the M protein has about 271aa. The other four proteins include: a capsid protein (core), a secreted protein (preCore, secreted by HBeAg), a protein polymerase (pol), and non-secreted protein (HBx). The core protein has about 180aa and it is a subunit of the viral nucleocapsid. The preCore protein, which corresponds to a core protein with an N-terminal peptide signal, is proteolytically processed at its N and C terminal before the secretion of infected cells. The pol protein encodes the enzymes necessary for the synthesis of viral DNA that takes place in nucleocapsids located in the cytoplasm of infected hepatocytes. The HBx protein possesses 150aa and its fundamental function in the life cycle of the virus is uncertain [120, 140]. Nevertheless, some authors consider that HBx plays an essential role during HBV replication, which provides it the ability to modulate factors and cascades of hepatocytes signaling linked to mechanisms underlying cellular transformation [141]. It is also reported that HBx can potentiate HBV replication by interacting with some epigenetic modification genes, miRNAs, transcription factors, chromatin-modifying enzymes, and components of the transcriptional machinery [142]. In addition, it has been pointed out that HBx is required for transcription from cccDNA in HepaRG cells and most likely in HepG2 cells infected with HBV [140].

Regarding the replication and organization of the genome, the hepadnavirus contains the following: main's ORFs: precore/core (preC/C), polymerase (P), surface (preS/S); a fourth ORF; and the X gene. Replication can be considered in two stages: an incoming or afferent arm, in which the input viral genome enters the nucleus and is covered by cccDNA; and an outgoing or efferent arm, in which RNA transcriber from the cccDNA are encapsidated and reverse transcribed within the core particles in the cytoplasm, and the resulting genomic DNA is transported to the nucleus or enveloped and secreted [120]. All the members of the hepadnavirus use a genome replication strategy in which the virus replicates its DNA genome by reverse transcription of an RNA intermediary that uses the reverse transcriptase

activity of the viral polymerase [141]. The fate of DNA containing nucleocapsids depends on the stage of infection of a hepatocyte. When the concentration of the envelope proteins is poor, the nucleocapsids enter a retrograde transport and deliver their DNA charge to the nucleus of the cell, which leads to the amplification of the copy number of the cccDNA. Then in infection, the nucleocapsids bind to the envelope proteins and mature between infectious virions that are secreted into the bloodstream. Notwithstanding the foregoing, some mechanisms are not all completely understood, among them: exact HBx action mechanism, nucleocapsid clearance, cccDNA synthesis and cccDNA elimination [140].

**Infection caused by HBV** Productive HBV infections trigger inflammation and continuous necrosis mediated by the immune response against infected hepatocytes [143]. Virions containing infected DNA bind their target cell through the interaction of the protein with cellular receptors that are not yet fully characterized. The nucleocapsid presumably is delivered to the cytoplasm and transported to the nuclear pore where the genome is released into the nucleoplasm. Particularly the genus Orthohepadnavirus, to which HBV belongs, infects mammals, with a narrow host range for each species of the viruses. The only known natural hosts of HBV are humans and more enormous monkeys (chimpanzees, gorillas, orangutans and gibbons) [120]. Hepatocytes represent the primary targets of an HBV infection. A chronic HBV infection is the main, although not the only, cause for the development of hepatocellular cancer (HCC). As a result of the host's immune response to infection, HBV can cause acute or chronic infection and some complex diseases. The molecular mechanisms that link chronic HBV infection to the development of HCC are incompletely understood, but are typically associated with prolonged periods of persistent chronic infection. This persistence typically includes the development of cirrhosis associated with HBV prior to the progression of HCC [141]. As soon as infection occurs, the HBV DNA is converted to a covalently closed circular DNA (cccDNA) that accumulates in the nucleus of abnormal cells as a stable episome organized into nucleosomal structures. The pregenomic RNA (pgRNA) of HBV is encapsidated between particles of the cytoplasmic nucleus where it is reverse transcribed by means of the viral polymerase to produce the first strand of DNA HBV and sustain the viral replication inside the cell. The enzymes that modify the chromatin, the cellular transcription factors and the viral proteins HBx and HBc are recruited on the cccDNA minichromosome to regulate its transcription and finally its viral replication [143].

It is believed that the regulation of the cell cycle mediated by HBV is largely carried out through the activity of the HBx protein. It has been shown in primary hepatocytes that HBx alters cell cycle regulators, including decreased expression of p15 and p16, decreased DNA synthesis, and increased expression of p21, p22, p27 and cyclins D1 and E. This regulation of the cell cycle mediated by HBx could produce a long-term impact on the physiology of hepatocytes, altering their proliferation pathways and contributing to the development of HCC and diseases associated with HBV [141]. The oncoprotein HBx upregulates expression of the enzymes

DNMT1 and DNMT3A, which leads to promoter methylation and transcriptional repression of several tumor suppressor genes. HBx activates the host cell cycle by upregulating DNMT1 through positive feedback mechanisms. HBx represses the expression of the cyclin-dependent kinase (CDK) inhibitor p16 by methylation of the promoter mediated by DNMT1. The deregulation of p16 expression leads to the activation of the cell cycle through the inhibition of pRb and upregulation of E2F1. This produces increased levels of DNMT1 and creates a positive feedback loop that then reduces pRb expression by methylation of the p16 promoter. HBx also promotes cell cycle progression through hypermethylation of other CDK inhibitors, such as: p21 and p27 [134]. Additionally, it is reported that there is a positive cycle between HBx and the androgen receptor (AR). HBx increases the activity of the c-Ser kinase, which improves the transcriptional activity of the AR gene, thus increasing its mRNA level, and therefore allowing the androgen signaling pathway to increase the transcription and replication of HBV genes [142].

### 2.5.2 Human Papillomavirus

The demarcation of papillomavirus (PVs) by phenotypic criteria becomes difficult for a variety of reasons. One of these reasons is that this virus does not trigger consistent humoral immune responses, either in infected humans or in other mammalian individuals, and therefore it is not possible to develop a taxonomy based on serology. The absence of reliable cell culture systems and host laboratory animals represent other difficulties. Despite these limitations, two main pillars for PVs taxonomy have emerged. The first pillar shows that all known PVs are strictly specific to the host species, and this restriction needs to be reflected in the taxonomy. The second pillar states DNA sequence comparisons lead to refined phylogenetic studies, which shows that all PVs genomes are monophyletic in origin, evolve more slowly than virtually any other group of viruses and they do not recombine. The topology of phylogenetic trees is an indispensable criterion for the taxonomic evaluation of this family of viruses [120].

Conforming to ICTV, the human papillomavirus (HPV) corresponds to the following assignment. (i) Order: unassigned; (ii) family: Papillomaviridae; (iii) subfamilies: 2; (iv) genera: 53, among which one is unassigned; (v) species: 133. Within the two subfamilies, the first is called Firstpapillomavirinae which has 52 genera, which in turn group 132 species. The second subfamily is called Secondpapillomavirinae, which has one genus and it groups the members of unique species. Among the genera with the most significant number of species are: the genus *Gammapapillomavirus* with 27 species and the genus *Alphapapillomavirus* with 14 species. Among the remaining genera, each one groups between one and seven species. Particularly, HPV is classified within the genus *Alphapapillomavirus* [120].

From the morphological point of view, HPV is a double-stranded DNA virus [126]. The HPV virions are naked. They have a diameter of 55 nm, and their icosahedral capsid is made up of 72 capsomers, of which 12 are pentavalent and 60

hexavalent [120]. At the proteins level, the genome of the virus encodes between eight and ten proteins with sizes ranging from seven to 73 kDa. The L1 and L2 proteins participate in the conformation of the capsid. The E1 and E2 proteins are involved in replication, and among these, particularly E2 is involved in intragenomic regulation. The E5, E6 and E7 proteins induce cellular DNA replication. The E4 protein may represent a late function and binds to specific structures of the cytoskeleton [120].

The classification of HPV by “type” is based on the nucleotide sequence of the viral gene L1 that codes for the major protein. When the complete genome has been clones and the DNA sequence of L1 differs by more than 10% of the type, with respect to the closest known PV, it is recognized as a new type. If the difference in homologation varies between 2% and 10%, a new subtype is recognized. When the difference is less than 2%, it corresponds to a new variant. Regarding the complete nucleotide sequence of L1 ORF within the same species, the “types” share a similarity that varies between 71% and 89%; the “species” share similarity between 60% and 70%; the “genera” share less than 60%; and the “families” share between 23% and 53% [144–146]. At the genotypic level, the differences found between the various types of PVs are primarily based on the amino acids that conform the L1 protein, and it is these characteristics of the protein that allow the virus to be treated as high or low risk. Based on data from 11 case-control studies conducted in different countries, the association between cervical cancer and PVs infection was explored. Considering their causal association with cervical cancer, 15 viral types of “high-risk” HPV were identified (16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73 and 82); three types of probable high-risk HPV (26, 53 and 66); and 12 types of “low-risk” HPV (6, 11, 40, 42, 43, 44, 54, 61, 70, 72, 81 and CP6108) [18, 147]. There are more than 100 HPV types, but only about 13 cause cancer [139]. HPV types 16 and 18 cause about 70% of all cervical cancers and almost other HPV-related cancers [18].

The variants are considered markers of specific genomes of HPV and are used in epidemiological and etiological studies to investigate the spreading of HPV in populations. To establish the variants, the studies were based on an analysis of nucleotide sequences, specifically on the LCR region. Isolates of the same type are closely related to each other on the basis of the nucleotide sequence and represent a phylogenetic group. New types can be recognized and studied after the cloning of partial genomic sequences, but there is an agreement in which it is established that only a new taxonomic variant will be formally recognized if the whole sequence is cloned [148, 149]. Five phylogenetic branches are identified among 301 HPV isolates that were obtained from 25 different ethnic groups and geographical locations in Europe, Asia, Africa, North and South America. The phylogenetic branches were designated as follows: E (European), As (Asian), AA (Asian-American), Af1 (African-1) and Af2 (African-2). The reference genome of HPV16 is a German isolation and corresponds to a member of the “E” lineage. In the LCR-based analysis, the European variant is more associated with the presence of persistent infections, while the African and Asian-American variants are more associated with the development of cervical cancer [150, 151]. In Europe and North America the class “E” variant pre-

dominates. In Africa, 92% of the variants “Af1” and “Af2” predominate. The variant “As” is primarily found in Southeast Asia. The “AA” variant groups Central America, South America and Europe, but within Europe this variant is presented only in Spain. The geographical distribution of HPV types and specific variants may be influenced by several factors including the co-evolution of HPVs with human races, human migration patterns and optimal measurements of the virus, as in the case of transmissibility. It has been shown that HPVs are extremely ancient viruses that have co-evolved with their hosts from the three largest human races (African, Caucasian, East Asia), and also that some specific variants of HPV16 are associated with viral persistence and the development of high-grade cervical lesions [152]. At a more recent date and based on a total of 953 isolates of E6/LCR sequences from around the world, with a considerable proportion of isolates from Africa and Asia, Cornet and colleagues (2012) produce an update of the HPV16 phylogenetic tree that clearly identifies nine sublineages: EUR, As, AA1, AA2, NA, AFR1, AFR1b, AFR2a, and AFR2b. This updated confirmed two important events: (i) it was confirmed that only E6 does not allow to distinguish the three sublineages AA1, AA2 and NA. The inability of E6 to differentiate these three sublineages is important in epidemiological studies, since the AA1 sublineage is associated with CIN3 or with higher-risk CIN3+ based on whole-genome analysis; and (ii) it is confirmed that LCR contains much more phylogenetic information than E6, and that it is able to distinguish all nine sublineages proposed [153].

As for the organization and replication of the genome, the virions that bind to cellular receptors are engulfed by the cell, and the DNA is discovered and transported to the nucleus. During productive infection, the transcription of the viral genome is divided into early and late stages. In the early stage, regulatory proteins are encoded that can exhibit transactivation properties which are required for DNA replication. Viral DNA replication is initiated in the nucleus before late events begin. Translation of late transcripts produces structural proteins that are involved in capsid assembly. Several of the viral proteins contain sequences, called nuclear localization signals, which facilitates the transport of proteins to the nucleus of host cells where virion maturation occurs. Virions are released by lysis of the virus-producing cells. Papillomavirus are extraordinarily host species-specific and tissue-restricted. Terminal differentiation is required by all known HPVs for replication and virion production [120].

**Infection caused by HPV** HPV infection seems to occur mainly through microlesions of cell that proliferate in the basal layer. The spread of the virus occurs through the release of virions from the surface of warts and papillomatous lesions, which frequently contain considerable amounts of viral particle within the differentiated superficial layers [120].

Viral replication can occur in two stages: non-productive and productive. The non-productive stage occurs in the epithelium of the basal layer where the virions penetrate the epithelium and infect the located stem cells in the layer of the basal epithelial cells. In the basal compartment of the epithelium, the virus establishes itself with a low copy number (approximately 100), using the two early viral proteins

E1 and E2. In addition, using the cellular DNA replication machinery of the host to synthesize it, on average once per cell cycle bidirectionally [147]. Replication starts at the only binding site of E1. This site is located in the 3' segment of long control region (LCR). The E1 and E2 proteins interact and form a complex in the viral origin of replication (Ori) and cooperatively recruit the cellular replication machinery for the viral genome. The E1 protein is an ATP-dependent DNA helicase that interacts with an AT-rich binding site at the Ori site, where it interacts with the E2 protein and some partners. The E2 multifunctional protein controls the transcription and replication of the viral genome through interaction with the E1 protein and some cellular partners. The E2 protein binds to DNA binding sites bordering the E1 binding site, and N-terminus of E2 interacts with E1 and aids in the recruitment of the origin of replications [154]. The functions of E2 are mediated by their junctions to numerous sites in the long control region. HPV16 includes four of these sites (BS1-BS4). Two of these sites (BS1, BS2) are upstream of the TATA box, separated from each other, and from the TATA box it is separated by three or four base pairs. Of the other two sites, BS3 is next to the site linked to DNA E1, and is involved in the regulation of viral DNA replication; and BS4 is found 300–400 base pairs upstream. The binding of E2 to BS1-BS2 causes transcriptional repression, whereas the binding of E2 to BS3-BS4 can lead to activation. However, the precise effects of E2 on transcription, activation or repression, vary depending on E2, LCR and host cells [155]. During variable periods of time, this reduced number is maintained in cells initially infected, but still competent and capable of replication. The expression of the viral proteins E6 and E7 delays the arrest of the cell cycle and its differentiation. These are observed as epithelial cells that are directed towards the upper part of the membrane. Subsequently they become mature keratinocytes, producing a thickening of the skin (wart), characteristic of some infections caused by PVs [147]. The E7 protein of HPV16 can stimulate DNA synthesis and cell proliferation which allows it to improve the proliferation of keratinocytes and reactivate the DNA replication machinery in differentiated, non-cyclic, suprabasal cells. E7 also induces cell proliferation in cells that have already migrated from the basal layer, suggesting that differentiation occurs, but in a modified state [156]. The HPV E6 and E7 proteins also improve the methylation promoter by deregulating DNMT1 expression, through the degradation of p53 and a direct interaction with the DNMT1 protein, respectively. It is suggested that high-risk HPV E6 interfere with the expression of IFN-I to promote HPV persistence in host cells. Deregulation of CXCL14 by E7-induced methylation of HPV to evade host immunity contributes to the suppression of antitumor immune responses during HPV persistence and progression to cancer [134].

The productive stage occurs in the suprabasal layer of the epithelium. At this site the virus, in a rolling circle mode of replication, amplifies its DNA to a high copy number, and synthesizes the capsid proteins L1 and L2. These proteins are essential in the replication phase, and they are responsible for the virus assembly to occur. Subsequently, the virions are released into the environment since the topmost layer of the epithelium is eliminated. During viral persistence, the immune system maintains the infection in this state, which in turn prevents the basal cells from accessing

the suprabasal layer and lose the ability to divide. At that moment the terminal differentiation phase begins. The PV is replicated in this behavior. When this occurs, the uncoded late viral proteins, L1 and L2 are assembled in the cell nucleus. The mature virions are then assembled and released from the epithelium within the superficial cells, taking advantage of the disintegration of the cells that are produced as a consequence of the natural turnover in the superficial layers of the epithelium [147, 157, 158].

### 2.5.3 Merkel Cell Polyomavirus

The criteria for demarcation of the included species in the genus Alphapolyomavirus, to which the Merkel cell polyomavirus (MCPyV or MCV) belongs, are the following: sufficient information about the natural host, the genetic distance observed from a member of the most closely related species, must be greater than 15% for the coding sequence of LTag (large T antigen); and when two polyomaviruses exhibit observed genetic distance of less than 15%, biological properties may be important additional criteria. According to the ICTV classification, this virus is within the allocation that follows. (i) Order: unassigned; (ii) family: Polyomaviridae; (iii) subfamilies: does not report; (iv) genera: 5, among which one is not assigned; (v) species: 88. Among the five genera, the first is called Alphapolyomavirus and groups 38 species. The second genus is called Betapolyomavirus and groups 32 species. The third genus is called Deltapolyomavirus and groups four species. The fourth genus is called Gammapolyomavirus and groups nine species. The unassigned genus groups five species. Explicitly, the MCPyV belongs to the genus Alphapolyomavirus and to the species called Human Polyomavirus 5 (HPyV5) [120, 159].

From the morphological point of view, polyomaviruses (PyVs) are a family of small naked viruses with double-stranded DNA (dsDNA) genomes of approximately 5000 base pairs (bp). The phylogenetic relationship between polyomaviruses is based on the sequence of the long tumor antigen of the viral protein, based on which the four classified genera are delimited (Alpha-, Beta, Gamma, and Delta-polyomavirus). Members of the genus Alphapolyomavirus can infect mammals, birds and fish. Some of its members are known human and veterinary pathogens causing symptomatic infection or cancer in their natural hosts. The mature virions of MCPyV measure in diameter between approximately 40–45 nm. They are made up of 88% proteins and 12% DNA. The virions are naked, the capsid has right icosahedral symmetry ( $T = 7$ ) and is constructed of 72 pentameric capsomeres. Each pentamer is made up of five VP1. The capsomeres are interconnected by the C-terminal arm of VP1. Each virion contains a single copy of circular dsDNA. MCPyV (specie: Human Polyomavirus 5) has the longest genome (5387 bp) [120]. During the normal life cycle, the polyomavirus genome is maintained as a circular double-stranded DNA episome [160].

At the protein level, the polyomavirus genome encodes two regulatory proteins: LTag (*large tumor antigen*) and STAg (*small tumor antigen*), which are expressed

early during infection; and three capsid proteins: a major protein VP1 and two minor proteins inside the capsid: VP2 and VP3, which are expressed after the start of viral DNA replication and are therefore called late proteins. However, some polyomaviruses produce additional early and late proteins [120, 161, 162]. The VP1 protein determines the antigenicity and specificity of the receptor and therefore has a significant impact on adhesion, tissue tropism and pathogenesis of PyV [161]. The LT antigens of polyomaviruses contain a number of important motifs and common domains that facilitate the viral life cycle [162].

In terms of genome replication and organization, MCPyV is a small circular double-stranded DNA virus that belongs to the polyomavirus family and the HPyV5 specie [120]. The viral genome is divided into three regions: a non-coding regulatory region (NCRR), and two transcriptional regions that include the early coding region and the late coding region [160]. The NCRR region encompasses an origin of DNA replication and transcriptional promoters and enhancers. The transcription of the early region results in a single mRNA precursor from which different transcriptors are generated through alternative splices. The principal translational products, generated from these alternative mRNAs, are the regulatory proteins LT and ST. The processes of transcription of early and late genes are directed by NCRR, a region that is located between the early and late regions. The early promoter includes a TATA box, whereas the late promoter lacks such a motif. The NCRR region contains multiple binding sites for cellular transcription factors [120].

In MCPyV, the NCRR region contains a minimum origin of 71 bp of replication (viral origin of replication) which has a stretch rich in AT, an LT binding domain and an early promoter region (transcriptional regulatory elements). The early region expresses three T antigens: long T antigen (LT), small T antigen (ST) and 57kT antigen, which share a short common amino terminal [163]. Moreover, it contains an additional product, from an alternative frame to the open reading frame LT known as ALTO (*alternative tumor antigen*) [162]. The late region encodes the capsid proteins VP1 and VP2. The late region of the MCPyV genome contains an open reading frame (ORF) for the major capsid protein VP1 and minor capsid protein VP2 [163]. Specifically in MCPyV, the VP3 protein is absent [120, 161].

**Infection caused by MCPyV** Currently, among polyomaviruses (PyVs), one of the best studied is MCPyV [126]. Typically, PyVs induce latent infection without evidencing disease, but in the context of immunosuppression this virus can produce tumors after the integration of viral DNA between the genome of the host [164]. MCPyV infection occurs at an early age and establishes a persistent infection. Viral transcription of MCPyV is present in physiological conditions of human skin along with transcriptors of other polyomaviruses, but viral integration is only found in malignant tissues. So far no preferential insertion site within the human genome has been shown. However, it is recognized an LT antigen binds and inactivates the tumor suppressor genes Rb and p53 in the polyomavirus family [126]. The LT antigens, which play a coordinated role in viral transcription and replication, are commonly known for their ability to interrupt cellular pathways involved in signaling and cell cycle regulation [165]. The LT antigen regulates the life cycle

of the virus and the cell cycle of the host cell. The cell cycle regulation of the host cell is achieved through interaction with the p53 tumor suppressor gene and elements of the retinoblastoma (Rb) gene family [164]. LT promotes entry into the cell cycle through the inactivation of Rb [165]. The J-domain that carries the N-terminal of LT, binds to heat-shock proteins; a motif for binding to Rb, which makes it possible to inhibit members of the Rb family; and the binding domains at the origin of the C-terminal and helicase/ATPase, all together are required for efficient viral DNA replication [160]. The ST antigen is able to stimulate cell proliferation through the activation of several cellular pathways [164]. ST mRNA shares a common exon with LT, but reads through a splice junction present in the LT mRNA to generate a short alternative reading frame protein (18 kDa) that possesses unique cellular targeting features [166].

The entry of PyV is initiated by the major capsid protein VP1, which binds to cellular receptors to promote internalization. During viral entry, MCPyV uses sulfated carbohydrates termed *glycosaminoglycans*, as primary binding receptors; and *sialylated glycans*, as post-binding secondary co-receptors for gene transduction [161, 163]. After the entry of the virus into the cell, the decapsidation of the virion in the cytosol begins. The partial coating exposes the nuclear localization signals in VP2 and helps guide the particles into the nucleus, where then the clearance occurs. The viral genome remains episomal. At an early stage of infection, transcription occurs from one strand, an in, one direction resulting in mRNA encoding LT, ST and early alternative proteins. LT self-regulates its own transcription and is only viral protein required for replication. ST also contain the J-domain and can bind and inactivate the cellular protein PP2A (*protein phosphatase 2A*), which leads to the activation of cyclin-D1 and cyclin-A. Some of the early alternative proteins can bind to members of the retinoblastoma family. The ALTO protein of MCPyV is not required for replication [120].

After entry, PyV moves from the cytoplasm to the nucleus, the latter being the place where the uncoated genome is accessible to the replication machinery of the host cell. The virion assembly occurs in the nucleus, although little is known about the process of viral particle outflow [161]. In the origin of MCPyV replication there are more pentanucleotides sequences and their pentameric elements exist in greater proximity than in any other PyV. It is suggested that this greater proximity allows intermolecular OBD-OBD interaction that occurs in LT proteins that bind to the origin of replication. Nevertheless, the replication in the MCPyV viral origin and the LT seeding processes can be very different and more complex than in other PyVs [163]. After viral replication begins, the late genes are transcribed from the opposite strand in a direction opposite to that of the early genes. The proteins of the capsid are synthesized in the cytoplasm and transported to the nucleus, where the assembly with the viral genome among viral particles occurs. Viruses are released by mechanisms dependent and independent of lysis. For MCPyV, it has been suggested that dermal fibroblasts are the genuine host cells where infectious virus particles are produced. The infection of non-permissive Merkel cells that do not sustain the viral life cycle results in the integration and transformation of the MCPyV

genome [120]. Immune dysregulation by viral proteins allows MCPyV to establish a persistent infection [160], and persistence of long-life is possibly reached by very low levels of viral replication [161].

## 2.6 Cancers Associated with Viruses

The theory that cancers could be caused by an infectious agent was propagated at the beginning of the nineteenth century [167]. From then until now, according to the information published by the International Agency for Research on Cancer (IARC), 11 infectious agents have been identified that classify as carcinogenic agents in humans (referred to as Group1). Currently, there is well-established evidence of a causal link among these infectious agents and the type of cancer associated with them. This group include: an agent that originates in bacteria, seven agents that originate in viruses, and three agents that originate in parasites. The agent of bacterial origin is *Helicobacter Pylori*, which is associated with the gastric carcinoma (both cardiac and non-cardiac) and gastric non-Hodgkin lymphoma. The three infectious agents that originate in parasites include: *Opisthorchis viverrini* and *Clonorchis sinensis*, which are associated with cancer of the bile ducts; and *Schistosoma haematobium*, which is associated with bladder cancer [2]. The seven agents of viral origin are integrated by: (i) Hepatitis B Virus (HBV), which is associated with liver cancer; (ii) Hepatitis C Virus (HCV), which is associated with liver cancer and non-Hodgkin's lymphoma; (iii) Human Papillomavirus (HPV) high risk (types 16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58 and 59), associated with the carcinoma of the oral and oropharyngeal cavity, including tonsils and base of the tongue, larynx, anus, cervix, vulva, vagina and penis; (iv) Epstein-Barr Virus (EBV), is associated with Hodgkin's lymphoma, Burkitt's lymphoma and nasopharyngeal carcinoma; (v) Human Herpesvirus-8 (HHV-8), which is associated with the Kaposi sarcoma; (vi) Human T-cell Lymphotropic Virus type 1 (HTLV-1), which is associated with leukemia and adult T-cell lymphoma [2]. In 2008, a new human polyomavirus known as Merkel cell polyomavirus was reported; (vii) MCPyV is associated with a rare tumor named Merkel cell sarcoma [168], which can then evolve into the development of a skin neuroendocrine carcinoma known as Merkel Cell Carcinoma (MCC) [169–171], considered one of the most aggressive skin cancers [160]. Among agents of viral origin that are associated with human cancers, three of them are not integrated into the host (HCV, EBV, HHV-8), and the remaining four are integrated into the host's genome (HBV, HPB, MCPyV, HTLV) [126].

Moore and colleagues (2010) refer to the two categories in which infectious cancer agents have also been classified, these are: direct carcinogens and indirect carcinogens. In the first classification, the infectious agents express intracellular viral oncogenes that directly contribute to the proliferation of cancer cells. In the second classification, chronic inflammation induced by the virus initiates growth of cancer cells rather than a viral oncogene, presumably causing cancer through infection and chronic inflammation which eventually leads to the carcinogenic mutation

in host cells. By definition, a direct viral carcinogen is present in each cancer cell and expresses at least one transcriptionist to maintain the transformed tumor cell phenotype, as occurs with cancers related to HPV, MCPyV, EBV and KSH. However, viral agents such as: HBV, HCV and HTLV-1 do not fit appropriately into these categories [127]. Another way of looking at infectious carcinogens is to establish whether a tumor virus is a rare or recurrent infection in humans. Most tumor viruses are not common infections, and therefore, patterns of tumor occurrence reflect patterns of virus infection. At least in North American and European populations, infections caused by HBV, HCV, HTLV-1, HPV and KSHV viruses are generally not considered common infections [172].

DNA tumor viruses usually manipulate two major cellular processes that include signal transduction and cell cycle. The continuous activation of cascades of signal transduction as well as the interruption of cell cycle regulation caused by this type of virus results in uncontrolled cell proliferation. Tumor viruses change cells by integrating their genetic material into the genomic DNA of the host cell, although the mechanism of insertion may differ depending on whether the viral genome is DNA or RNA. The process of transforming a conventional cell into a cancer cell involves three key events, these are: initiation, promotion and progression. The presence of the viral genome within the cell is essential for cell transformation. The transformed cells exhibit increased cell division that may favor viral propagation. After the host cells are transformed, they acquire the ability to invade the tissue and then cause metastasis in distant locations [124]. The progression of a tumor is typically considered a stochastic dynamic process [133]. Although oncogenic mechanisms have been poorly understood, it has been observed that tumors caused by viruses that integrate into the host's genome continue to be unique in their natural history, prognosis and response to treatment [126].

Related cancers to immunosuppression may have been induced by tumor viruses [127]. The host defense mechanisms interpret a fundamental role in the modulation of viral carcinogenesis [123]. Virus-mediated carcinogenesis is a multi-step process that involves a series of various complementary events in order to transform a conventional cell into a cancer cell [124]. Virus oncogenes that produce tumors in humans are viral genes, rather than mutated versions of cellular genes that were accidentally assimilated during the viral replication cycle. These oncogenes perform fundamental roles in the life cycle of the virus, and its oncogenic potential represents a manifestation of those activities [123]. The direct link between viral insertion and malignant transformation can demonstrate the result of the impact produced by these integration events on oncogenic proteins [126]. Some viral proteins have the ability to induce the secretions of enzymes that favor cells migration, decreasing cell-to-cell adhesion and allowing progression to cancer. Viral proteins can alter cellular signaling pathways by simulating the action of their cellular homologs, such as receptors or adapter proteins, which mediate in the perpetual activation of these pathways. Despite variation in insertion sites among different DNA tumor viruses, the integration event adversely affects the human genome by means of deletions, duplications, chromosomal translocations, transcriptional enhancement driven by viral promoters, insertional mutagenesis and induction of genomic insta-

bility. Viral integration may be associated with the incorporation of the viral genome into the host and/or the continuous expression of some of the viral proteins, which alters the genetics and physiology of the host cell [124].

Among the infectious agents of viral origin, previously referred, this article focuses on a group of three DNA viruses (HBV, HPV and MCPyV) that are integrated into the genome of the host and that are considered etiological agents of human cancer. Although these viruses coincide in being double-stranded DNA viruses and possess the characteristic of integrating into the human genome, this does not imply that their life cycle and the mortality they produce in the population are similar. Regarding mortality rates for cancers associated with these three types of viruses, Globocan (2012) predicts that by the year 2030 deaths in both sexes of all ages, around the world will include the distribution that follows: Death caused by: (i) Liver cancer: 1,165,249 people; (ii) non-Hodgkin lymphoma: 308,335 people; (iii) Lip and oral cavity cancer: 220,627 people; (iv) Cervical cancer: 382,970 women; (v) Hodgkin lymphoma: 35,710 people; (vi) Nasopharyngeal carcinoma: 74,675 people; (vii) Kaposi sarcoma: 37,604 people; (viii) Leukemia: 388,640 people; and (ix) Skin melanoma: 87,725 people [1].

### 2.6.1 Cancers Associated with HBV

Hepatocellular carcinoma (HCC) arises after a prolonged liver cirrhosis from death and chronic regeneration of cells induced by viruses. Both HBV and HCV are clonally integrated into the genome of tumor cells in almost all cancers related to these viruses, but it is not clear whether their persistent expression is required for HCC cell proliferation [127]. The precise biology associated with HCC has not yet been completely elucidated [173].

Chronic HBV infection can cause cirrhosis and liver cancer and is also considered a risk factor for the development of non-Hodgkin's lymphoma [139]. HBV contributes to the development of HCC [143]. According to Seeger and colleagues (2015) the pathogenic process that leads to the development of HCC is involved in at least three stages: initiation, promotion and progression. In the initiation stage, the fixation of one or more mutations in the host DNA is carried out via cell division. The promotion stage could correspond to the clonal expansion of mutated hepatocytes, which demonstrate an elevated risk of the oncogenic transformation. The progression stage includes the steps by which members of these clonal populations evolve, until they become cancer cells [140]. Levrero and colleagues (2016) present three different mechanisms by which HBV can promote the development of HCC: (i) classical insertion mutagenesis, similar to a retrovirus with the integration of viral DNA among cancer genes in the host, such as: TERT (*telomerase reverse transcriptase*), CCNE1 (*cyclin E1*) and MLL4 (*mixed-lineage leukemia protein 4*); (ii) the promotion of genomic instability as a result of the integration of viral DNA into the host genome and the activity of viral proteins; and (iii) the capacity of viral proteins of wild type, mutated and truncated (HBx, HBc and preS) to affect cellular

functions, to activate oncogenic pathways and sensitize liver cells to mutagens [143]. HBV can lead to oncogenesis induced by chronic inflammation processes and through molecular mechanisms, such as the activation of signaling pathways NF- $\kappa$ B, MAPK and AKT mediated by HBx oncoproteins [126]. It is also suggested that cell cycle manipulation by HBx through DNA methylation can contribute to the development of HCC [134].

HBx (Q156X2\_HBV, Uniprot) is a multifunctional protein that performs a key role in the silencing of the host antiviral defenses and promotes viral transcription. Yue and colleagues (2016) show that HBx first induces DNA damage and then disrupts the metabolism of nucleic acid. Abnormalities of metabolism block DNA repair and induce the emergence of HCC [174]. HBx itself can inhibit the p53 tumor suppressor gene and the DDB1 (*DNA damage-binding protein 1*) DNA repair protein, and thus activate several oncogenes. HBx can also enhance the interaction with genes of genetic modification, miRNAs, transcription factors, chromatin-modifying enzymes, and components of the transcriptional machinery [142]. The HBx integration process occurs in actively transcribed regions, preferentially within the MLL4 and TERT genes. MLL4 represent an epigenetic regulator that acts as a transcriptional activator. TERT (telomerase reverse transcriptase) is a gene that codes for reverse transcriptase of human telomerase, which plays a fundamental role in cell immortalization. Viral integration in TERT can induce the activation of mutations, while viral integration within the TERT promoter can induce over-expression of human telomerase reverse transcriptase. However, it is not currently acknowledged if the integration events are selective or if the recurrent integration sites only result from a selective bias. Nevertheless, alterations of TERT and MLL in non-viral HCC have also been reported, which indicates that viral integration is not always necessary for alterations of specific genes [126]. The most prevalent mutations are in the TERT promoter, which can lead to telomerase reactivation, allowing cells to avoid programmed cells death and acquire malignant potential. Also involved is the tumor suppressor gene TP53 and other oncogenic pathways that include the signaling of: Wnt/b-catenin, p53/cell-cycle, PI3K-AKT-mTOR, MAPK, JAK/STAT7, Rb. Additionally estrogen regulates the estrogen receptor which interacts the binding of HNF-4 $\alpha$  (*hepatocyte nuclear factor*) with HBV enhancer, thereby reducing the level of IL-6 and the activity of NF- $\kappa$ B, STAT3 (*signal transducer and activator of transcription 3*) and C/EBP $\beta$  (*CCAAT/enhancer-binding protein  $\beta$* ). Likewise, the ligand-activated androgen receptor (AR) increases cell cycle-related kinase transcription and activates the  $\beta$ -catenin signaling pathway, which is identified as a classic oncogenic pathway in the development of HCC [142]. HBx also induces methylation of the IL-4 receptor (IL-4R), which leads to deregulation of its expression. IL-4 suppresses the growth of the host cell and induces apoptosis. Replication and viral gene expression are inhibited by IL-4. Therefore, deregulation of IL-4R expression mediated by promoter methylation induced by HBx represents a mechanism of immune evasion of HBV. Considering that IL-4 mediated signaling is pro-apoptotic, the deregulation of IL-4R expression in HBV by means of the methylation promoter may contribute to cell proliferation and the development of HCC [134].

## 2.6.2 Cancers Associated with HPV

Persistent infection with HPV can cause six different types of cancer, in addition to genital warts. Persistent HPV infection is responsible for all cervical cancers, 90% for anal cancers, about 70% for oropharyngeal cancers, and between 60% and 70% for vaginal, vulvar and penile cancers in the United States [139]. HPV causes the most common sexually transmitted infection in the world. Persistent infection by oncogenic types of HPV is one of the prerequisites for the development of cervical cancer. HPV type 16 (HPV16) is the most typical etiologic agent in invasive cervical cancer and represents about 70% of all cervical cancers in the world. HPV16 is also associated with anal, penile, vaginal, vulvar and oropharyngeal cancers. More than 80% of cases of cervical cancers occur in less developed regions and mortality rates in these countries are 18 times higher than in developed countries [175].

The immune system performs a key role in the battle of organisms against cancer. According to the modern concept of immunoediting there are three phases of interaction between the host and the tumor cells: Elimination, Balance and Escape. The Elimination phase consists in the recognition of a growing tumor due to the fact that cancer cells express their own antigens, which causes the immune cells to recognize them and provoke a subsequent infiltration to the tumor by means of immune effector cell as happens with NK cells. These cells produce multiple cytokines, including interleukins (ILs), which contribute to the death of cancer cells that inhibit angiogenesis. Subsequently the cancer cells that survive the phase of elimination then enter the equilibrium phase where the lymphocytes exert a selection pressure on tumor cells, which are genetically unstable and mutate rapidly. The Equilibrium phase results in the immune selection of tumor cells with decreased immunogenicity. These cells become those that likely survive in an immunocompetent host, which could explain the paradox of the occurrence of cancer in immunologically intact patients. In the final phase, tumor cells that have acquired resistance to the immune system continue to grow and expand uncontrollably, therefore leading to a condition of malignancy [176, 177].

Oncogenic viruses that induce cancer can do it directly and indirectly, through the activation of inflammatory signaling pathways, and the production of cytokines that stimulate the growth of malignant cells and that manage to inhibit the process of cellular apoptosis. Oncogenic viruses induce inflammation through different mechanisms to promote the initiation and progression of cancer. In the context of cancer, chronic inflammation is linked to tumorigenesis, increased cell proliferation, survival, invasion, angiogenesis and metastasis. In the context of cancer, chronic inflammation is linked to tumorigenesis, increased cell proliferation, survival, invasion, angiogenesis and metastasis. Tumorigenesis typically follows the accumulation of mutations in gene coding regions that result in uncontrolled cell growth [178]. The modulation of the pro-inflammatory response can constitute a crucial step during carcinogenesis by HPV, a process that includes the establishment of infection, persistence, progression, angiogenesis and metastasis [76].

Sometimes high-risk HPV infections result in viral episome integration into host DNA. Virus-infected cells acquire an extended lifespan that allows them to preserve the ability to proliferate and tend to develop and perpetuate mutations in the DNA germination line, which is attributable to the actions of E6 and E7 proteins [157, 158]. Exposure to high-risk HPV causes an initial infection of the squamous epithelium in the transformation zone. This event is followed by persistent infection, integration of the viral genome into the host genome, genomic alterations, immortalization and transformation of epithelial cells [179]. In the context of cervical cancer due to HPV, the viral life cycle is disturbed by the progressive histopathological changes that arise in the cervical epithelium, which influences the loss of terminal differentiation; and in addition, by the integration into the circular viral DNA genome and the host DNA genome, which together leads to the selective expression of E6 and E7 proteins. These integration sites are randomly distributed in the complete genome with a predilection for fragile sites [180]. The initiation of cancer during viral infection can be based both on the ability of the virus to directly contribute to oncogenesis, and on the capacity of viral oncogenes to inhibit apoptosis. Viral carcinogenesis can also occur indirectly if chronic infection leads to inflammation [178].

HPV have a predilection for cutaneous or mucous epithelial surfaces, and this is one of the reasons why it directly infects cervical keratinocytes and interferes with TLR signaling. HPVs are separated from the immune system, since they first infect the basal cells of the squamous epithelium and do not include a viremic phase. The limited innate immune response, the reduced levels of viral gene expression in the basal epithelium and the absence of cytopathic effects, generally result in a delayed adaptive immunity against the initial HPV infection favoring the establishment of the persistence of the viral infection [57]. During the early stages of HPV infection, the innate immune system creates a proinflammatory microenvironment by recruiting innate immune cells, to eliminate infected cells and to initiate an effective acquired immune response. However, HPV exhibits a broad range of strategies to evade immune surveillance, in this way generating an anti-inflammatory microenvironment [55, 76]. Although it is a principal causative agent of cervical intraepithelial neoplasia (CIN) and cancer, it is believed that alone HPV viral infection is not enough for the establishment of malignity and progression of the disease [57]. HPV infection experiences an inactive period or almost 10 years, between the initial infection and the development of cervical cancer, and it is believed that chronic HPV infection can stimulate the tolerance of the immune system [55]. Most high-risk HPV infection in immunocompetent individuals are cleared, although this work can take up to 4 years, and only a minority of apparently immunocompetent individuals will develop the persistent infection [181]. The location of many cervical cancers in the transformation zone indicates that this area may have increased their susceptibility to cervical cancer and produce inefficient immune responses. Although HPV infection does not easily stimulate an immune response, many HPV infections are cleared by immune mechanisms, before these infections become malignant [55].

Cervical cancer is a malignant neoplasm that arises from cells of the cervix, which are mostly infected with high-risk HPV. The early proteins E6 and E7 play a fundamental role in the initiation and progression of cervical cancer [182]. The induction of cervical carcinogenesis due to high-risk HPV infection represents a multistep process that involves the transformation of the normal cervical epithelium into a pre-neoplastic stage. This process includes persistent infection and several stages of cervical intraepithelial lesions (CINs) that eventually lead to the development of invasive cervical cancer [60, 183].

Viral infections are normally detected by antigen-presenting cells (APCs), but their mechanisms do not appear to be fully functional, which may be due to the low expression of MHC class-I molecules on the cell surface [55]. When severity is reported in the disease and there is presence of HPV in the epithelium, elevated levels of TLRs expression in the diseased stroma are produced, which induces an increase in IFNs responses. This upregulation can contribute to the persistence of HPV infection and cultivate the progression of the lesion. Although TLRs are not being induced in late carcinogenesis in the cervical epithelium, the presence of HPV triggers the activation of TLR pathways in the cervical stroma, causing innate immune responses [53].

The TLRs bind to microbial compounds to trigger innate and adaptive immunity, in the process of defense against immunological danger. TLRs play an essential role in the detection and initiation of antiviral immune responses [57]. Efficient innate immune responses depend largely on the body's ability to detect an aggressor pathogen. TLRs are present in diverse types of cells and have the ability to detect and bind to distinct ligands associated with pathogens, signaling their presence and initiating and directing an immune response against the invading pathogen [77]. The most essential role of TLRs in the defense of the host is the regulation of innate and adaptive immune responses in epithelial cells. These receptors play an essential role in the functions of the innate immune system, through the recognition of discrete exogenous PAMPs and endogenous ligands of DAMPs, to induce an innate immune response [53]. The activation of TLRs leads to the expression of effector molecules, such as proinflammatory cytokines and cytokines that mediate immunity [77]. The ability of TLRs to differentiate pathogens is recognized as the first line of defense in the detection process within the innate immune system. The damaged TLR signaling can lead to susceptibility to infection, but often if a TLR type is removed, other can compensate for its deficiency and still mount a challenge against the invading antigens. TLRs seem to interact with HPV16, and it is indicated HPV can modulate TLR expression by interfering with its signaling pathways, leading to persistent viral infection and eventually to the development of carcinogenesis. Despite the increase induced by TLR expression, HPV infection causes a decrease in the production of cytokines, and this behavior is attributed to the viral oncoproteins E6 and E7 [53, 55, 57]. Several TLRs are involved in viral recognition and are identified as follows: TLR1, TLR2, TLR3, TLR4, TLR6, TLR7, TLR8 and TLR9. It is believed that one or several TLRs of viral recognition, acting as one of the primary points of the pathogen recognition immune system, are important in shaping the host response

to cervical HPV infection [77]. The expression TLR1-TLR9 is observed in epithelial cells of the human female reproductive tract. However, distinct portions of the reproductive tract seem to express different TLRs. The cervix shows high levels of TLR2, TLR3, TLR4, TLR6; and reduced levels of TLR1, TLR4, TLR7, TLR8 and TLR9. The expression TLR seems to increase with the expression of TNF- $\alpha$  [55]. The TLRs establish a suitable microenvironment for the growth of tumor cells, which allows them to evade immune cells, infiltrate, metastasize and undergo malignant progression [78].

In the carcinoma in the epithelium, significant decrease in TLR1 expression has been observed, and it is thought it may represent the only TLR pathway upstream affected in keratinocytes with late carcinogenesis. It is also believed the increase in the expression of TLR1 in the malignant stroma is due to the increased infiltration of monocytes linked to DAMPs, from dying cancer cells as potential ligands. During cervical carcinogenesis and when the cervical epithelium is still undergoing differentiation, which typically occurs in parallel to the HPV life cycle, the TLR3 receptor triggers antiviral response by means of the regulatory factor interferon 3 (IRF-3) increasing the expression in the dysplastic epithelium. It is believed that this upregulation may represent an attempt by infected cells to initiate immune responses against HPV infection [53]. The TLRs are not only expressed by immune cells but also by the tumor cells, generating conditions that facilitate the occurrence and development of the tumor. TLRs can promote an environment that leads to carcinogenesis through inflammatory responses, angiogenesis and cell death. Meantime, HPV can invade innate immunity by avoiding TLR recognition. As a result of this interaction, an ideal microenvironment for tumor growth can be established, which allows tumor cells to evade immune cells, infiltrate them, conduct them to metastasis and subject them to malignant progression [184]. In this same context, a decrease in the expression of TLR4 during progression to CIN has also been observed, and this event seems to be associated with the expression of a critical marker of HPV integration among infected cells [57]. After being activated by its ligands, TLR4 induces immunosuppressive factors such as: TGF- $\beta$  and VEGF; and stimulates the production of interleukin-8 (IL-8), which generate resistance of tumor cells to TNF- $\alpha$ , thus inducing apoptosis and immune evasion of cancer cells [78]. Aggarwal and colleagues (2015) observed a significant upregulation of the relative expression of TLR1 genes and deregulation of TLR3, TLR4 and TLR5 in patients with cervical cancer, compared to their controls [184]. TLR5 and TLR9 may interpret a significant role in the tumor progression of cervical neoplasia and may represent a useful marker for the malignant transformation of cervical squamous cells [60]. In addition, the expression TLR5 and TLR9 are increased according to the progression of the histopathological grade CIN; this means that their levels increase as they go from low-grade CIN to high-grade CIN, and particularly high-levels of expression in invasive squamous cell carcinoma are observed [57]. Similarly, TLR4, TLR7 and TLR9 report high levels of expression in a variety of solid tumors and promote tumor development. These receptors are expressed in stromal cells and also in cancer cells [78].

In cancerous immunity, the functional role of plasmacytoid dendritic cells (pDCs) is believed to depend on the balance between immune responses and tolerance that is initiated in the tumor microenvironment and/or in lymphoid organs. Soluble factors produced in the tumor microenvironment contribute to the functional damage of pDCs associated with the tumor, through specific interference with TLR9, which alters the expression of its own receptor or its downstream signaling pathway [88]. The demodulation of TLR9 by the E6 and E7 proteins of HPV16 may be important for the selection and proliferation of cancer cells by mechanisms that are still unknown [57]. TLR9 is upregulated during some stages of tumor development, increases gradually with progression to cervical carcinoma, and is also linked to other forms of cancer. TLR9 is upregulated during some stages of tumor development, and its expression levels increase gradually with progression to cervical carcinoma; however, it is also linked to other forms of cancer [55]. Although the expression levels of TLR9 are considered low in normal cervical epithelium, the increase of their levels of expression on the tissue is associated with the progression of epithelium persistently infected with HPV, which then leads to the development of cervical cancer [185]. TLR9 can also induce the development of tumor cells, increase their adhesion and evasion, promote their growth and inhibit their apoptosis process [78].

In addition to the promotion of tumor cells, TLRs also provoke malignant biological behavior through the induction of the expression of anti-apoptotic proteins via the activation of NF- $\kappa$ B and upregulation of lipid mediators of tissue remodeling, associated with inflammation. These events contribute to cancer formation and progression such as occurs with: COX-2 (*cyclooxygenase-2*) and PGE2 (*prostaglandin E2*) [78]. In its interaction with HPV, the expression of COX-2 which is involved in the production of proinflammatory prostaglandins, upregulates its expression levels during the development of cervical cancer [185]. The E5, E6 and E7 proteins of HPV16 are shown to induce COX-2 and PGE2 expression [178]. Several soluble factors produced in the tumor microenvironment, such as: PGE2, IL-10 and TGF- $\beta$ , show that they damage the production of IFN type I by pDCs in response to CpG. The inhibition of IFN-I by pDCs in the tumor microenvironment may confer a selective advantage for tumors [88].

In cervical cancer associated with HPV infection, the anti-inflammatory and immunosuppressive cytokines are expressed in the cervical microenvironment. These cytokines determine the persistence of HPV and tumor progression by disrupting the mechanisms of cellular immune surveillance, whether the change is made as a secondary effect induced by the tumor cells or due to the persistence of the viral infection itself [186]. HPV proteins promote the expression of mediators of proinflammatory proteins [185]. The secretion of cytokines by infected cells and underlying tissues promotes an inflammatory environment that can stimulate the development of cancer, mainly mediated by IL-1 $\beta$ , TNF- $\alpha$  and IL-6. Inflammatory stimuli and direct effects of the virus activate signaling pathways responsible for the development of cancer, including NF- $\kappa$ B, STAT3 and the MAPK complex [178]. Progression to cervical cancer is associated with changes of cytokines, Th1 type to Th2 type, induced by two epitopes from the E6 oncoprotein and an increase in the

expression of IL-10 [187]. In the cervical carcinomas, the expression of several immunoregulatory cytokines is also altered. The cytotoxic T-cells (CTLs) and helper T-cells (Ths) activity, specific for E6 and E7 of HPV16, are found in the blood of patients with pre-malignant cervical neoplasia and in healthy controls. This suggests that spontaneous regression of CIN lesions may be associated with immune responses mediated by cells that act against HPV E6 and E7 oncoproteins [188]. The HPV E6 and E7 proteins regulate the expression of TGF- $\beta$  in cervical cancer. The HPV E2 protein binds to the regulatory gene IL-10 and induces high epithelial cell-promoting activity [186].

It is suggested that IL-10 is responsible for viral persistence. The increased expression of IL-10 can induce viral persistence, the transformation of cervical epithelial cells, and consequently the development of cancer. The regulation of the IL-10 gene in cervical cancer can be mediated by HPV proteins during the various phases of the viral transformation process. Because of the presence of this cytokine, it allows the virus to contribute to the development of a local immunosuppressive state. Induction of IL-10 production by the host during chronic infection appears to be one of the key viral media to alter the class of antiviral immune responses and induce widespread immunosuppression [179, 186]. In cervical cancer lesions certain levels of immunosuppressive cytokines are produced to promote carcinogenesis. The cytokines IL-4 and IL-10 are upregulated in patients with CIN lesions. The microenvironment of low-grade cervical lesions is predominantly anti-inflammatory, and it is modified in a way that favors HPV infection. Elevated levels of IL-10 are found in HPV-positive patients, compared to HPV-negative patients. Nevertheless, IL-10 is not always found in low-grade lesions, although there is evidence of over-expression of other anti-inflammatory cytokines like TGF- $\beta$ . Once HPV infection has been established, the expression of the E6 protein leads to upregulation of IL-17, which may establish a significant step for tumor development and progression. IL-17 promotes angiogenesis (increase of new blood vessels that the tumor needs to grow) and tumor growth. In addition IL-8 is involved in angiogenesis, metastasis and in the signalling pathways of IL-17. The upregulation of IL-8 is correlated with the expression of metalloproteases MMP-2 and MMP-9, which are involved in angiogenic mechanisms [76]. In cervical tumors, high expression levels of IL-4, IL-10 and TGF- $\beta$  are reported. In biopsies of patients with premalignant lesions and cervical cancer, high-level of cytokines expression are shown, such as: IL-4, IL-10 and TGF- $\beta$ , which can induce a local state of immunosuppression [186]. Cancer cells transfected to express the HPV E2 protein, induce high levels of IL-10 mRNA in cell infected by HPV. The cytokine IL-10 reports high-levels of expression in tumor cells, and its expression is directly proportional to the development of cervical cancer by positive HPV, which suggests a fundamental role of HPV proteins in the expression of IL-10 [174, 186].

Interferons type-I (IFN-I) are strong anti-tumor factors that negatively regulates the process of tumor immunosurveillance. Elevated concentrations of IFN-I are significantly correlated with more significant tumor-specific T-cell responses. Elevated concentrations of IFN-I are positively correlated with stronger tumor-specific T-cell responses [88]. The deregulation of Interferon Gamma (IFN- $\gamma$ ) in samples of

premalignant epithelium is attributed to the interference of the HPV E6 protein with the IFN pathway, blocking IRF3. It is also suggested that the deregulation of TLR9 by means of the HPV16 E6 protein, the methylation of the IFN- $\kappa$  promoter and the deregulation of IFN- $\kappa$ , can represent mechanisms that deregulate the IFN- $\kappa$  itself in the epithelium [53]. IFN- $\gamma$  secreted by CD8+ T-cells but activated by NK cells, NKT and DCs, makes it a key intermediate mediator of effector function of CD8+ T-cells that improves the presentation of antigens as well as the polarization of the immune response by Th1 type cells. The damaged antigenic presentation may be one of the reasons why the HPV infection is slow to clear when there is adequate cellular immunity [181]. Treg cells specific for E6 and E7 antigens are identified in cervical cancer and high-grade squamous intraepithelial lesions, infiltrating lymphocytes. The infiltration of the tumor with Treg foxp3+ cells is associated with little survival in cervical cancer; and a reduced rate between Treg cells and CD8+ T-cells is strongly associated with a poor prognosis [187].

NF- $\kappa$ B represents a master regulator of cell differentiation induced by local pro-inflammatory signals, it is upregulated in epithelial cells in cervical cancer, and it is associated with a poor prognosis. Its expression is upregulated by the E6 and E7 proteins of HPV16 [185]. The E6 and E7 proteins of HPV16 inhibit NF- $\kappa$ B activity, when they are induced by TNF- $\alpha$  in cells of the cervical transformation zone, where the majority of cervical cancers develop. A decreased activation of NF- $\kappa$ B may represent a mechanism by which the virus interferes with the host immune response and promotes the persistence of the infection. Vandermark and colleagues (2012) show that the deregulation of NF- $\kappa$ B stimulates the growth and immortalization of cervical epithelial cells [189].

### 2.6.3 Cancers Associated with MCPyV

Merkel cell polyomavirus (MPCyV) is the unique human polyomavirus that includes a robust collection of scientific evidence that supports its classification as a causative agent of a human malignancy known as Merkel Cell Carcinoma (MCC) [163, 190]. MCC is a highly malignant skin cancer characterized by metastasis and low survival [191]. Because it was classified as a malignancy of cutaneous neuroendocrine cells, it was believed that MCC originated in the transformation of Merkel cells [163]. Merkel cells are found in the skin and mucosa of all vertebrates and continue to be the only neuroendocrine cells in the skin with both neuroendocrine and epithelial profiles [164]. It has been suggested that Merkel cells cannot be the cells of origin of MCC, because MCC expresses PAX5 (*paired box gene 5*) and TdT (*terminal deoxynucleotidyl transferase*), and these genes are only expressed in pro/pre-B cells under certain physiological circumstances. Therefore, it is speculated MCC tumors can arise from pro/pre-B cells, and that the true origin of MCC cells are early B-cells. However, there are many unsolved events and contradictory finding that need to be tested experimentally [191]. Liu and colleagues (2016) show human dermal fibroblasts as natural host cells, and they explain that these cells can support the productive (life cycle) infection of MCPyV [160]. Another explanation

suggests that MCC arises from dermal stem cells and/or pluripotent epidermal stem cells [164]. Nevertheless, the true origin of MCC tumors continues being unknown [160, 164].

MCC is one of the most aggressive and lethal skin cancers. It is believed that a combination of immunosuppression, the potential oncogenic pathways triggered by ultraviolet exposure and advanced age represent events that are involved in pathogenesis MCC, although these factors alone do not completely explain the etiology of MCC [163]. However, it is suggested that MCPyV can be virtually associated with all cases of MCC [190]. It is suggested that there are some MCCs that are dependent on viruses and others that are not dependent on viruses. Those belonging to the first group tend to occur in the trunk, produce multiple viral copies, show very few genomic deletions of the host, and lack p53 expression. Virus-independent MCCs tend to occur in areas damaged by ultraviolet rays in the head and neck, appear to produce more genomic deletions from the host, and express p53 [164].

DNA MCPyV is clonally integrated into the tumor genome of MCC with persistent expression of LT and ST antigens [161], so that the tumorigenesis associated with MCPyV can be characterized and mediated by the classic expression of LT and ST antigens [162]. The ST and LT viral antigens of MCPyV-positive are expressed in MCC cells. The two proteins have in common 78 N-terminal amino acids, while in respect to the C-terminal both are different. The LT proteins expressed in tumor cells lack domains essential for viral replication, while the binding motif for the Rb protein is always preserved [190]. Shuda and colleagues (2011) found that LT does not promote cell proliferation. They consider that in the configuration of human MCPyV tumors, the LT protein probably acts in combination with other T-antigens of the virus, immunosuppression and possibly host cell mutations in order to promote MCC growth [166]. One of most striking features of the MCPyV LT antigen is that the C-terminal domains are consistently truncated by polymorphisms associated with tumor, but in N-terminal domains up to the binding motif for RB they are always intact. It is suggested that the C-terminal portion of LT typically produces an antiproliferative effect that needs its deletion in tumorigenesis [162]. The viral integration event seems to take place before the clonal expansion of the proliferation of tumor cells. The clonal integration pattern of MCPyV suggests that viral infection and the non-productive integration of genomic DNA between host cells occur prior to the clonal expansion of the proliferation of MCC tumor cells [160, 190]. The ST protein of MCPyV also deregulates the expression of cellular genes associated with NF- $\kappa$ B pathways, which may subvert the host immune response to allow the establishment of persistent MCPyV infection in host cells. The ST antigen is the main viral oncogene that contributes to virally induced cell transformation [160]. The ST-antigen of MCPyV has an interaction domain PP2A (*protein phosphatase 2*) in its C-terminal region [166]. However, the explanation of how the MCPyV infection leads to MCC is not yet fully understood, and many aspects of its infectious life cycle continue to be unknown [160].

In tumors, not solely is MCPyV clonally integrated into the genome of the host cell, but it equally contains various mutations at the 3' end DNA strand of the T-antigen gene. These mutations eliminate the viral helicase activity LT, but are downstream of the open reading frame of ST. Shuda and colleagues (2011) show that the ST protein of MCPyV is detected in human MCC tumors more commonly than the LT protein [166]. LT protein isolated from tumors typically contains a truncated form of LT antigen that is functionally incapable of supporting viral replication [161]. The virus within the tumor cells has tumor-specific LT mutations. These mutations prevent the integrated virus from replicating independently of the tumor cell. The integration of the viral DNA among the tumor DNA also makes it non-transmissible [164]. In order for the tumor cells to survive, the persistent expression of MCPyV T-antigens is required from the integrated viral genome. In addition, it has been shown that deletion of the C-terminal LT is necessary for the oncogenic progression of cancers associated with MCPyV [160].

Deciphering the interactions that arise between MCPyV and the human immune system becomes relevant to achieve a proper understanding of how the processes associated with the development of MCC tumors induced by this virus are produced. As with other DNA viruses, MCPyV is equally capable of escaping immune destruction and progressing to the development of MCC. It has been observed some of the pathways used by MCPyV to evade immune system responses include deregulating TLR9 and causing the deregulation of MHC-I, which affects the presentation of peptides from intracellular proteins before CD8+ T-cells [160]. Polyomaviruses that focus on the PI3K-Akt-mTOR signaling pathway can equally contribute to carcinogenesis. The importance of *cap*-dependent translation in tumorigenesis regulated by PI3K-Akt-mTOR signaling is emphasized. A key regulator of this process is the 4E-BP1 (*4E-binding protein 1*) factor, which when bound to eIF4E (*eukaryotic translation initiation factor 4E*) is inhibited at the moment when 4E-BP1 is phosphorylated by the kinase mTOR (*mammalian target of rapamycin*). Once the hyperphosphorylation of 4E-BP1 occurs by mTOR, it causes the binding between 4E-BP1 and eIF4E to be released, allowing eIF4E to be free to form the *cap* assembly and initiate *cap*-dependent translation. The hyperphosphorylation of 4E-BP1 induced by mTOR promotes mitogenesis and cell proliferation. Shuda and colleagues (2011) observe that ST of MCPyV focuses on *cap*-dependent translation through a mechanism that contributes to the tumorigenesis of Merkel cells. They conclude the *cap*-dependent translation through hyperphosphorylation of 4E-BP1 may contribute to the carcinogenesis of Merkel cells [166].

Most MCCs harbor copies of the MCPyV genome, clonally integrated with ST and previously truncated versions of LT. It has been shown that the growth of MCC tumor cells, induced by MCPyV positive, depends on the interaction of LT with Rb (retinoblastoma); whereas ST promotes cell proliferation by deregulation of the mTOR signaling pathway, mediated by the inactivation of 4E-BP1 [192]. It has also been shown that ST expression results in increased hyperphosphorylation of 4E-BP1, independent of mTORC1 signaling. Although the precise mechanism explaining the ST-mediated hyperphosphorylation of 4E-BP1 is not yet defined, it is recognized to be independent of the interaction between PP2A and Fbw7 (*F-box*

and WD repeat domain containing 7) [162]. Akt-mTOR activation is common for other tumor viruses that focus on upstream components of this signaling cascade and are extremely sensitive to mTOR inhibitors. In contrast, ST of MCPyV is activated downstream of mTOR pathway. Therefore, it is established that although ST of MCPyV is necessary for MCC, this viral factor alone is not enough to cause the tumor. ST of MCPyV can focus on 4E-BP1 to increase replication and viral transmission, but in doing so, it places the infected cells at risk of the carcinogenic transformation [166]. Cook and colleagues (2016) consider that several of the oncogenic pathways including bcl-2, Wnt signaling and MAP kinase signaling pathways, report little involvement in the pathogenesis of MCC [164]. Alternatively, it is suggested that the induction of MMP (*matrix metalloproteinase*) genes via the WNT/β-catenin signaling pathway together with other growth factors, stimulate MCPyV infection and in this manner can lead to MCC tumorigenesis [160].

## 2.7 Cancer Immunotherapy

Conforming to “American Cancer Society” (ACS), the immunotherapy in cancer defines a treatment that uses parts of the immune system to counteract the disease which can be done either by stimulating your own immune system to work harder or more intelligently, and therefore attack cancer cells; or by supplying some components of the immune systems like man-made proteins [91]. Wang and colleagues (2017) define as cancer immunotherapy targets the activation and expansion of the population of cancer-specific T-cells, which attempt to eliminate cancer cells by recognizing target antigens that are expressed on them [193]. One of the objectives of cancer vaccines, alone or in combination with other immune modulators, is to induce protective immunity or therapeutic antitumor immunity that can stimulate the benefits of current active immune therapies but with less toxicity and lower cost [194]. This category includes both vaccines that help to prevent cancer and vaccines that help treat cancer. In the latter, the specific goal is to help treat cancer or help prevent it from coming back after other treatments [18].

According to Ophir and colleagues (2016), cancer immunotherapy focuses on utilizing the immune system to reject and/or prevent its recurrence. They highlight three different approaches to immunotherapy. (i) Passive immunotherapy: it is associated with the administration of antibodies generated exogenously or by immune cells transferred adoptively, in order to mediate an anti-cancer immune response; (ii) Active immunotherapy: it is linked to vaccination, attempt to activate endogenous immune cells to recognize specific tumor-associated antigens (TAAs) and eliminate cancer cells; (iii) Immunomodulatory agents: it focus on improving the immune response to increase anti-cancer immunity [195]. In contrast to the active approach, cancer immunotherapy is based on tumor-specific antigens and tumor-specific T-lymphocytes which can recognize tumor antigens that are presented by antigen-presenting cells (APCs). Therefore, it is fundamental to identify tumor antigens for the development of optimal cancer vaccines. Tumor antigens can be classi-

fied into two groups: tumor shared antigens and tumor specific antigens. Shared antigens (also known as tumor-associated antigens, TAAs) mainly include tissue differentiation antigens, viral antigens, and germline tumor antigens. The specific antigens are personalized antigens or neoantigens that are the product of non-synonymous and more specific somatic mutations for tumor cells compared to shared antigens. In cancer immunotherapy, specific antigens stand out because they are unique in various types of cancer and are possible targets to be used in personalized vaccines against cancer [196].

The recognition of the different immunosuppression pathways has provided the basis for the emergence of several immunotherapeutic approaches, directed towards the multiple mechanisms that tumors evolve to suppress the responses of the immune system [197]. Cancer immunotherapy has predominantly focused on the restoration of the host anti-tumor response. The majority of oncological therapies are oriented towards the elimination of tumor cells and the reduction of tumor burden. In general, in the face of the selection of therapies, little attention has been paid to the interactions between the tumor and the host's immune system. The principal obstacle facing successful immunotherapy corresponds to the ability of tumors to disarm host defenses, through evolved mechanisms, in addition to the fact that each tumor creates a single microenvironment and produces a unique immune signal [198].

Conforming to the ACS, four principal types of immunotherapy are currently being used to treat cancer. These are: monoclonal antibodies (mAb), immune checkpoint inhibitors, cancer vaccines, and other non-specific therapies [91]. In general, the available therapies focus on immunosuppressive mechanisms directed by the tumor, in an attempt to counteract its evasion. According to Whiteside and colleagues (2010), these therapies are classified into seven distinct categories. These are: (i) Monoclonal antibody therapy (mAb): this therapy involves the use of mAb specific for TAAs in order to eliminate tumor cells that express antigens, to promote the formation of antigen-antibody (Ag-Ab) complexes strongly immunogenic and to improve the development of anti-tumor responses, both cellular and humoral; (ii) Cytokine therapy: this therapy aims to induce cytokine-mediated protection of immune T-cells, activated from apoptosis, remodeling of the proinflammatory tumor microenvironment and upregulation based on the amplification of functions on immune cells; (iii) Vaccination/adjuvant therapy: this therapy tries to deliver adjuvants, usually in combination with therapeutic anti-tumor vaccines for cancer patients, focused on the activation of anti-tumor responses and the development of long-term immunological memory; (iv) Cell therapy: this therapy is based on the stimulation of T-cells, alone or together, with adoptive transfer of T-cells constructed *in-vitro* to increase anti-tumor effector functions that provide for the *in-vivo* survival of these cells; (v) Cell depletion/neutralization therapy: this therapy aims to eliminate Treg cells and/or MDSC (*myeloid-derived suppressor cells*), and block the soluble factors produced by these cells; (vi) Molecular therapy: this therapy uses small molecules to block suppressive signaling; and (vii) Cancer stem cell therapy: this therapy aims to identify and eliminate cancer stem cells [198]. Additionally, another approach proposes the development of combined strategies. This approach

presents three different categories that aim to achieve therapeutic benefits for an enormous population of patients with a variety of tumors. These are: (i) Modalities that improve the presentation of antigens, such as: radiation, cryotherapy, chemotherapy, focused agents, vaccines, TLR agonists, interferons type-I, and oncolytic viruses; (ii) Additional agents focused on reversing T-cell dysfunction, like check-point inhibitors, where the combination of anti-CTLA-4 therapies is explored with other co-stimulatory receptor agents or T-cell co-inhibitors; and (iii) Agents focused on compensation of immune inhibitory mechanisms in response to therapy, such as: Treg cells, IDO (*indoleamine-2, 3-dioxygenase*) and MDSCs [197].

To develop monoclonal antibodies it is first proposed to identify the correct antigen to attack, but in cancer this is not always easy, because mAbs can be useful for some types of cancer and not for others [91]. Anticancer antibodies are considered promising medications, because they have a favorable toxicity profile and can participate in different effector mechanisms of the host. The antibodies activate the cell-mediated cytotoxicity through the activation of *Fc* receptor ligands expressed on several immune cells and forming immunogenic Ag-Ab complexes, which are processed by APCs, favoring the development of Th1-type immune responses. The binding of immune complexes *Fc* receptors on APCs result in phagocytosis and presentation of peptides on MHC class-I and class-II molecules. This cross-presentation form elicits effective peptide-specific T-cell responses. All antibodies on immunosuppressive elements produced or expressed by human tumors, either with antibodies that try to inhibit signals induced by tumor-associated inhibitory ligands or antibodies that attempt to neutralize immunosuppressive cytokines, such as: IL-10 or TGF- $\beta$ 1, or antibodies designed to remove suppressor cells. The antibodies can promote the development of tumor-specific CTLs, by means of several antitumor mechanisms which lead to the release of TAAs in biologically active form, in that way biasing the APCs towards a Th1-type response. Therefore, being able to achieve the appropriate balance in the therapy that leads to the immune activation, represents a challenge to maintain contracted the upregulation of immune responses against the immune suppression induced by the tumor. The tumor tends to dominate the microenvironment, making it hostile to immune effector cells, so the elimination of effects mediated by Treg, MDSC and inhibitory cytokines, seems to be a cancer immunotherapy of outstanding importance [198]. Currently, several types of monoclonal antibodies are in use, among which are: (i) Naked mAbs, in which there is no medication or radioactive material attached to them; (ii) Conjugated mAbs, which together with a chemotherapy drug or radioactive particle, seek to be used as a device for cancer cells to capture some of their substances; (iii) Bi-specific mAbs, which are part of diverse mAbs, try to bind to two different proteins at the same time [91].

Immune check inhibitors to treat cancer use certain molecules on determined immune cells that need to be activated to initiate an immune response. Currently, treatments are available focus on CTLA-4, PD-1 or PD-L1 [91]. Antibodies targeting co-stimulatory receptors and T-cell co-inhibitors have demonstrated profound clinical benefits seen in patients with various type of cancer. Antibodies that block human CTLA-4, a molecule that is upregulated during the early activation of T-cell,

report durable responses in advanced melanoma. Other inhibitory receptors, such as: PD-1, LAG-3 and TIM-3, are shown to be active during the late stages of T-cell activation and regulated within the context of the tissue microenvironment. The PD-1/PD-L1 and CTLA-4 checkpoint proteins are shown as targets that promise to be attractive for cancer therapy [197], including among others, treatments for skin melanoma and Merkel cell carcinoma [91].

In vaccination approaches have been undertaken that combine immunomodulatory antibodies with activators of innate immune response, as observed with TLR agonists. Intratumoral therapy with TLR agonist is being used as components of antitumor vaccines based on peptides; in addition, it is part of the design of vaccines based on DCs, to be used as monotherapy for the activation of innate and adaptive immune responses; and in combination with other categories of adjuvants or cytokines-adjuvants, to be applied as therapy in cancer and viral infections. Nevertheless, it is suggested to be careful because the TLRs are expressed not only by immune cells but also by tumor cells. When the signal of the respective ligand is present, the TLRs can promote tumor growth and increase the resistance of tumor cells to drug and immune interventions, thus achieving the opposite effect. Therefore, it is emphasized that the TLR adjuvant must be carefully dosed, clearly understanding its effects on cellular networks [197, 198]. Oncolytic viruses are also used as an *in-situ* vaccination strategy, which can infect, alert the immune system and kill cancer cells [91]. These oncolytic viruses are explored to enhance the efficacy of immunomodulatory antibody therapy against multiple cancer antigens and elicit the release of PAMPs and DAMPs, necessary for efficient APC maturation and antigens presentation [197, 198]. GM-CSF (*granulocyte macrophage colony-stimulating factor*) is used as an adjuvant in antitumor vaccines, designed to restore immune responses mitigated by the tumor. Although GM-CSF cannot be an effective adjuvant, it does have the ability to mitigate the antitumor response induced by the vaccine [198].

The category of adjuvants and immunotherapies that are not cancer-specific do not focus exclusively on cancer cells but rather try to stimulate the immune system in a more general way, and this can sometimes lead to better immune responses against cancer. Some of the non-specific immunotherapies are given by themselves as cancer treatments and others are used as adjuvants to stimulate the immune system. This category includes the cytokines and chemokines, interleukins, interferons and other drugs that are not found naturally in the body but that try to stimulate the immune system [91]. Some agents focus on co-stimulatory receptor targets of the tumor necrosis factor receptor superfamily (TNFRSF), such as: GITR, 4-1BB, OX40 and CD40 [197]. Among the drugs that upregulate the stimulating activities of APCs are the agonist antibodies for CD40 and/or CD40L. Particularly CD40, member of the TNF receptor superfamily, is a co-stimulatory protein expressed on APCs and targets agonist antibodies. The signaling pathway of this receptor improves HLA expression and the production of IL-12 through APCs that lead to T-cell activation, direct tumor inhibition, interference with angiogenesis and activation of NK-cells, B-cells and macrophages. CD40 agonists demonstrate enormous therapeutic potential when used in conjunction with other oncological therapies

[198]. Elevated doses of antibodies focused on 4-1BB have reported hepatotoxicity, while at low doses they do not show significant toxicity. In the evaluation of antibodies that focus on several immunomodulators, two categories are included: agonist antibodies that focus on active receptors of the T-cell surface and antibodies that block inhibitory receptors on T-cells and NK-cells [197].

Another category of drugs includes an enormous number of inhibitors used to attenuate the activities of immunoinhibitory enzymes, antagonist cytokines and small molecules that modulate immune cell functions in the tumor microenvironment. Immunotherapy transfers adoptive T-cells or delivers exogenous cytokines that are designed to activate, mobilize and upregulate the host's innate and adaptive immune responses. Other cytokines oriented to T-cells growth factors, seen as promising adjuvants or stimulators of antitumor cells immune activities are: IL-15 and IL-7. The cytokine IL-15 that inhibits cell death induced by T-cell antigens, reverses the T-cell anergy promoted by tumor-derived factors, boost the differentiation of DCs *in-vitro* and improves the activity of NK-cells. Unlike IL-2, the IL-15 does not support Treg activity. The cytokine IL-7 administered to patients, mainly involved in the survival of new T-cells, induces the increase of peripheral T-cells CD4+ and CD8+ without apparent toxicity. The best-known cytokine that activates T-cells is IL-2. The systemic administration of elevated doses of IL-2 has been approved for melanoma therapy, and this therapy has been made less toxic and safer for patients with cancer. Another cytokine seen as potent vaccine adjuvant and polarization cytokine in Th1 responses is IL-12. In its role as an immune adjuvant, it promotes the release of IFN- $\gamma$  from immune cells that express its IL-12R receptor. In its polarizing role, it is capable to induce Th1 polarization and the proliferation of effector T-cells producing IFN- $\gamma$ . The cytokine TNF- $\alpha$  performs a fundamental role in the inflammation induced by the tumor. The tumors constitutively produce TNF- $\alpha$ , and this cytokine derived from the tumor has been shown to improve tumor proliferation in experimental animals. Neutralization of TNF with anti-TNF antibodies and other TNF antagonists are introduced as potential therapies for patients with advanced malignancy [91, 198]. Inhibition of the proinflammatory cytokine TNF- $\alpha$  is used for the treatment of chronic inflammatory diseases mediated by immunity. However, it is indicated that blocking this cytokine increases the risk of acquiring or reactivating certain viral infection. Handisurya and colleagues (2016) investigate a possible association between TNF- $\alpha$  inhibition and anogenital HPV infection. In their study they observed a trend of higher frequency in anogenital lesions induced by HPV, an increased prevalence of mucosal HPV DNA, and a significantly higher seropositivity for high-risk HPV16 in patients with psoriasis [199]. Werberich and colleagues (2016) indicates that there may be a relationship between immunosuppression induced by TNF- $\alpha$  treatment and cancer development, considering that the use of this inhibitor decreases host defenses against viral infections [200]. Inhibition of TGF- $\beta$  signaling is used as an emerging strategy for cancer therapy and is mainly considered to normalize homeostasis of the tumor microenvironment by deregulation of stromal stimulation, which results from excess production of TGF- $\beta$  from tumor and its related tissues, with a direct impact on cancer cells [201].

In the early stage of viral infection, cells infected by the virus produce IFN- $\alpha$  and IFN- $\beta$ , and eventually undergo p53-dependent apoptosis. These virus-induced interferons can act in their surroundings on uninfected cells to help with antiviral defense, by inducing cellular genes that inhibit virus replication, and also by inducing p53 to prime cells for enhanced apoptosis. The cooperation of p53, IFN- $\alpha$  and IFN- $\beta$ , is shown as a link between tumor suppression and antiviral defense of the host [202]. Especially the cytokine IFN- $\alpha$ 2 demonstrates a significant impact on the modification of the tumor environment, both in the STAT signaling on tumor cells, as in immune cells and in the polarization of immune responses that aim to improve the antitumor reactivity [198]. The induction of p53 by IFN- $\alpha$  and IFN- $\beta$  suggests that cells treated by IFN are more susceptible to p53-dependent apoptosis in response DNA damaging agents, such as therapeutic agents used in cancer therapies. These interferons (IFN- $\alpha$ , IFN- $\beta$ ) are useful in the treatment of some types of human cancer, including cervical cancer associated with HPV. Although the molecular basis of the interferon action in the treatment of cancer is not completely clear, it is suggested that in combination therapies with chemotherapeutic drugs the use of lower doses that produce more moderate secondary toxic effects could be allowed. In the defense process by means of the activation of p53, the timely induction of apoptosis of virus-infected cells would be beneficial for the host by limiting the replication of the virus. The production of the virus could be suppressed if the abnormal cells undergo rapid apoptosis before the onset of cell disintegration. The apoptotic cells can be rapidly engulfed by phagocytic cells *in-vivo*, and this event in turn contributes to the inhibition of the spread of the virus. In this type of response, p53 is not absolutely dependent, but it can be improved with IFN- $\alpha$  and IFN- $\beta$  [202]. Taking into account the autocrine signaling characteristics rather than paracrine IFN- $\kappa$ , immunotherapeutic approaches involving the focal administration of IFN- $\kappa$  could be an alternative treatment option for HPV related diseases, such as: cervical, vulvar and anal; as well as cancer of the head and neck [203].

Radiation stimulates antitumor immune responses and is thought to occur through several mechanisms including the release of tumor antigens, cross-presentation of DCs and the induction of proinflammatory cytokines and chemokines, which mediate the recruitment of T-cells and DCs. Treatment with chemotherapy can produce an increase in proinflammatory cytokines in serum, proinflammatory changes in the tumor microenvironment and the induction of tumor-specific immune responses. Conventional chemotherapies show the release of antigens and DAMPs, thus triggering death by immunogenic cells. In addition to the effects on cancer cells, the chemotherapeutic agents exert various effects on the host immune system. The combination of chemotherapy and immunotherapy needs careful estimates to determine the appropriate dose and sequencing of these agents, in order to establish if such combinations could be more active and/or tolerable in specific type of cancers [197]. In immunotherapy based on drugs focused on cancer-induced immune suppression, there are medications that are already available in the clinic or are in the development stage for clinical use in the near future. To this category belong the antibodies focused on tumors, the antibodies that prevent the

accumulation of Treg and MDSC, the cytokines that counteract the immune suppression, the immune adjuvants or vaccines that induce antitumor effector cells, the T-cell stimulators, the metabolic inhibitors of tumor-derived factors, adoptively transferred T-cells and the elimination of cancer stem cells (CSC). Among the therapies involving cancer stem cells (CSC) there is the possibility of a future approach that includes CSCs with mAbs and specific T-cells, alone or in combination with chemotherapy. It is believed that CSCs perform a role in the promotion of tumor progression, and although they cannot directly mediate tumor-induced suppression, they have the ability to provide tumors with a pathway to acquire resistance to therapy and survive [198].

According to Cancer Research UK (2018), cryotherapy is defined as a treatment that uses extreme cold to destroy cancer cells [204]. Cryotherapy is equally known as cryosurgery or cryoablation, a treatment that uses liquid nitrogen (or argon gas) to destroy abnormal tissue. This technique treats external tumors (on skin) and internal tumors [91]. Cryotherapy involves another strategy of local removal that tries to induce localized immune response and release of antigens. Although the precise mechanism of immune activation is not yet clearly understood, it is believed the pathways used are the same as those of radiation, such as antigens and DAMPs released by the activation and cross-presentation of DCs [197]. Modern cryosurgery is used exclusively to treat spectra of tumors and cancers with various indications, from benign adenomas and precancerous lesions, to cancerous and low-grade lesions or early localized solid tumors. Some indications for the use of cryotherapy in the treatment of early stage cancers include the eradication of small retinoblastomas, basal and squamous cell skin cancers, cervical intraepithelial neoplasia and low-grade bone tumors. Additionally, some research studies the benefits that cryosurgery could have on cancers of the breast, liver, prostate, colon, kidney, pancreas and esophagus [205].

Another form of experimental immunotherapy is the adoptive transfer of T-cells, which is based on the *in-vitro* manipulation and expansion of autologous reactive T-cells of cancer, which can be obtained from peripheral blood and body fluids. After *in-vitro* expansion, these T-cells are reinfused into cancer patients with the expectation that they will eliminate the tumor cells and that they will generate anti-tumor immunity and memory. Tumor infiltrating lymphocytes (TIL), considered to be enriched with T-cells reactive in tumors, are successfully expanded in cultures and delivered to patients with metastatic disease especially melanoma. In the context of the adoptive transfer of *ex-vivo* cultures or genetically engineered T-cells, recombinant IL-2 is typically employed to provide transferred T-cells as a factor capable of supporting their functions and survival. The interleukins IL-15 and IL-7, which are considered survival cytokines, emerge as new adjuvants to improve and seem to significantly increase the efficacy of the treatment. Two other strategies in adoptive T-cell therapy consider: (i) transfected T-cell receptors (TCR), which are selected for recognition of tumor antigens; and (ii) genetic modifications of T-cells, to express or fuse chimeric receptors; which combine domains bound to the B-cell receptor antigens with the signaling or TCR components. Nevertheless, these strate-

gies have reported unexpected toxicities due to the expression of the target antigen on vital organs like the liver [198].

Several treatment options currently available, both approved and experimental allows creating an infinite grid of therapeutic possibilities. Each of individual options needs to be evaluated carefully, as well as some combinations that could be not only ineffective but even worse, dangerous for the patient. Because of the complexity of the interactions among the tumor, the immune system and the distinct therapeutic agents, it is unlikely to obtain a unique formula for the combination of immunotherapies. The response to a specific immunotherapy is influenced by multiple factors, which include the histological type of the tumor, the variability among patients due to the genetic differences of their tumors and the heterogeneity observed within the same patient [197].

### 2.7.1 Oncolytic Virotherapy

Each virus involves a specific cellular tropism that determines the tissues that will be preferentially infected, and therefore the disease that it will cause can be established [206]. Typically, viruses have a lytic replication cycle that leads to the death of infected cells [125]. Pathogens injected directly among tumors produce an inflammatory reaction that can promote infiltration of immune cells in an immuno-suppressed tumoral microenvironment. The introduction of lytic viruses also has the ability to directly destroy malignant cells, which can result in immunogenic cell death [194]. Oncolytic viruses (OVs) are therapeutically useful anticancer viruses that selectively infect and damage cancer tissues without causing harm to normal tissues [206], and these can be found in various kinds of viruses [125]. The OVs have the ability to infect, replicate and kill malignant cells. In addition, they can indirectly eliminate tumor cells by destroying the tumor vasculature [207].

Oncolytic vaccine therapy consists of an *in-situ* administration of genetically engineered viruses, which target tumor cells while controlling host tissue. The destruction of cancer cells within an inflammatory environment configures the stage in which tumor-specific antigens are recognized and memory is established, in order to combat the progression of the disease [194]. Mechanisms based on oncolytic viruses can consider use the various strategies. Lin and colleagues (2018) analyze four strategies that focus on the mechanisms of transduction, transcription, translation and proapoptosis. (i) The mechanism that focuses on transduction considers that tumor cells have high and specific surface expression of the oncolytic viral primary receptor. This receptor is able to bind to proteins on the viral surface, through the receptor-ligand pathway, which is fundamental for transduction. The first step corresponds to the absorption, followed by the combination of the surface adhesion protein with the target cell surface receptor, which initiates the endocytic signaling pathways. (ii) The mechanism that focuses on transcription considers that achieving selective viral tumor replication consists of altering the control function of genetic transcription. This mechanism is essential in viral replication, to a tissue- or tumor-specific promoter. (iii) The mechanism that focuses on trans-

lation considers that signaling pathways regulated by IFN form part of the mechanism that causes the loss of antiviral response in most cancer cells. The deficiency of tumor antiviral activity makes these cells more susceptible to an infection compared to normal cells, which results in a survival advantage for the viruses within the tumor cells. (iv) The proapoptotic mechanisms consider the p53 tumor suppressor as a key function to induce apoptosis through the arrest of the cell cycle. Viruses have evolved the ability to inhibit apoptosis of host cells, in order to perform viral replication [133].

Oncolytic virotherapy (OVT) represents a promising approach to treat cancer. Some effective ways to perform oncological therapies of this kind consider individual or combined cell lysis, apoptosis and innate or adaptive immune responses [125]. It is believed that OVs exert their antineoplastic effects through multiple pathways. Although the mechanisms of oncolysis can vary from virus to virus, there are some common mechanisms that they use to achieve an antineoplastic effect. OVs self-perpetuate, and theoretically, in a single dose they could eliminate all cancer cells unless the virus itself is cleared by the immune system before the cancer cells are eliminated [207].

The oncolytic virus vaccine has been genetically designed and exerts selective antitumor effects through the induction of dead immunogenic cells, which induces the release of natural tumor antigen and molecular patterns associated with pathogens [196]. Several genetically engineered OVs are available and the combination diversity of OVT regimens has grown rapidly in recent years, especially with emphasis on the development of personalized cancer therapy [125]. In oncolytic virotherapy, viral genomes are designed to improve their antitumor specificity [206]. To this end, a variety of genes are inserted into distinct oncolytic viruses, from immune stimulatory genes to proapoptotic genes. The insertion of genes between oncolytic viruses can result in specific overexpression in target cells. Some of the immune stimulating genes are: IL-2, IL-4, IL-12, IL-15, IL-18, IFN- $\alpha$ , IFN- $\beta$  and GM-CSF [207, 208]. Some of the proapoptotic genes that are part of the oncolytic viral vectors are: TNF, p53 and TRAIL (*tumor necrosis factor-related apoptosis induced ligand*) [207]. The coding of genes for TLRs, heat shock proteins, GM-CSF and tumor antigens have been introduced to improve the antitumor immunity and to overcome the immunosuppressive barriers of the tumor microenvironment [194]. Other OVs have also been designed to express enzymes that modify drugs, which convert at the site of viral replication of non-toxic prodrugs into toxic forms, thus causing the selective tumor cells death [133].

The idea of an oncolytic vaccine, combining the destruction of the tumor mediated by the virus with immune recognition of tumor antigens, is attractive but demands careful orchestration since the activated immune system can prematurely suppress the replication of the therapeutic virus [206]. This strategy is also considered a promising therapy when used as a complement to other conventional therapies, although its efficacy continues being evaluated [133]. Oncolytic viruses hold out the promise of improving cancer treatment, since these can eliminate both differentiated cancer cells and cancer stem cells and minimize the risk of relapse of the disease [207]. Another reason why it may be a promising antitumor modality is that

OVs are able to transfer and amplify therapeutic genes, while simultaneously preventing immunosurveillance of infected host cells [133].

Many of the solutions developed have been analyzed in systems of artificial models, although these are not so robust to reveal the consequences, in favor and against of a certain modification, for all aspects of the overall treatment paradigm [206]. Producing quantitative models that possess the ability to simulate an effective OVT can mean a significant reduction in time and effort in the search for optimal OVTs for subpopulations of cancer patients [125].

### 2.7.2 Therapeutic Vaccines Against Cancer

Understanding the function of viral proteins has paved the way for the development of antiviral drugs that focus on enzymatic activities of the virus. However, Law and colleagues (2013) consider that many of these medications do not work properly. They argue the virus-based approach has not proven adequate to decipher the complex and multifaceted interactions between the host and the virus, which are critical in viral recognition, immune signaling and disease outcome [130]. The principal objective for the improvement of the immune response in cancer vaccines is to discover strategies that can deliver antigens to the DCs population, more specifically, and induce subsequent activation of durable T-cell immune responses against cancer antigens; and at the same time, be able to reverse the immunosuppressive network of the tumor microenvironment [209]. Traditional vaccines that focus on viruses that can cause certain cancers help protect against some types of cancer, but don not focus directly on cancer cells. Vaccines for the treatment of cancer are different from vaccines that work against viruses. Treatment vaccines are intended to force the immune system to prepare an attack against cancer cells in the body. Some vaccines for the treatment of cancer are designed from cancer cells, part of cells or pure antigens. These vaccines are often combined with other substances called adjuvants, which help to strengthen the immune response much more [18].

Ophir and colleagues (2016) summarize the mechanisms that lead to an effective anticancer immune response, establishing three stages. Stage-I: Dendritic cells (DCs) act as key players especially through their ability to capture, to process and to present antigens, which allows them to activate T-cells. To initiate immunity, DCs can capture TAAs derived from a vaccine. Stage-II: The delivery of an appropriate activation signal by the DCs allows their differentiation and migration to the lymph nodes. In this site, mature DCs present TAAs that can activate and expand tumor-specific T-cells in sufficient quantities to generate an immune response with therapeutic implications. Stage-III: Cancer-specific T-cells leave the lymph nodes, infiltrate the tumor microenvironment and perform their effector function, thereby attempting to eradicate the tumor [195]. The active ingredients of cancer are made up of four essential components: tumor antigens, formulations, immune adjuvants and means of delivery [210]. For the successful development of therapeutic vaccines against cancer, Melief and colleagues (2015) propose a guide that includes the

following aspects: (i) sufficient antigenic concentration must be achieved on DCs; (ii) an effective route of administration must be used; (iii) DCs can be activated with adjuvants; and finally, (iv) the therapeutic cancer vaccines should not be expected to act as a monotherapy [211]. The most prominent types of therapeutic cancer vaccines based on tumor antigens include: dendritic cell vaccines, peptide vaccines and genetic vaccines [196].

**Dendritic cell vaccines** The dendritic cells (DCs) represent an essential component of vaccination through their ability to capture, process and present antigens to the T-cells. While immature DCs in peripheral tissues efficiently capture antigens, the presentation of antigens typically results in immune tolerance due to the lack of co-stimulatory molecules. The DCs can be exploited for cancer vaccination through various means including: (i) vaccines based on nucleic acids and peptide/protein not focused, captured by *in-vivo* DCs; (ii) Vaccines composed of antigens directly coupled to anti-DC antibodies; and (iii) vaccines composed of DCs generated *ex-vivo*, which are loaded with antigens [212]. DC-based vaccines operate as exogenous APCs to overcome tumor suppression and to stimulate T-cell responses specific for tumor antigen that attempt to attack the tumor [196]. Most DC vaccines cannot migrate to the lymph nodes to stimulate T-cells, nor can they function directly on their own to prime immunity. Instead, a cross-presentation by resident DCs in lymph nodes is required [213]. DCs serve as the most efficient antigen-presenting cells, which sensitize T-cells restricted to MHC molecules to initiate immune responses. The reason why DCs can be used as one of the cancer vaccine strategies is because the tumor has the ability to suppress maturity and promote apoptosis of endogenous DCs [196]. DCs also have the ability to present non-conventional antigens. New classes of antigens include phosphopeptides, which are synthesized as a result of deregulated phosphorylation, citrullinated peptides and antigens derived from non-coding DNAs that are expressed in cancer cells [213].

Vaccines based on DCs have not achieved the expected success. The limited effectiveness of the immune responses induced by DCs vaccines against the tumor, obeys several mechanisms. It is believed the principal obstacle is generated by the presence of regulatory T-cells (Treg) and the suppressive pathways established by the tumors [214]. Many of the molecular changes that allow a tumor grows and infiltrates its environment also produce an immune-suppressive medium that counteracts tumor rejection and prevents complete activation of DCs [215]. Other mechanisms that decrease the effectiveness of DCs vaccines involve the reduced amount of DCs in the tumor site, restricted access of DCs to the tumor antigen, the limited capacity of tumor cells to activate intratumoral DCs and the secretion of cytokines by of tumor cells that inhibit the maturation of DCs. Although it is attractive option for immunotherapy in cancer, the complexity of the dendritic cells system requires manipulation to achieve protection or therapeutic immunity [214].

**Peptide-based vaccines** The general principle of a peptide vaccine consists to activate and expand T-cells specific to tumor-associated antigens [216]. Peptide-based

vaccines have been widely researched due to their good safety profile, ease of manufacture and quality control. However, its antitumor efficacy has been questioned, it has been attributed an inefficient co-delivery of the antigenic peptides and adjuvant in the lymph nodes and the consequent development of immune tolerance [217].

The peptides are easily synthesized and stored. They may represent short epitopes or longer T-cell peptides, and may be delivered as single peptides or mixtures of multiple peptides [194]. Long peptides typically contain restricted peptides to MHC class-II and therefore possess the potential to stimulate the two classes of CD4+ and CD8+ T-cells [210]. It is believed that professional APCs as DCs, are necessary to present long peptides to new T-cells, which can generate productive responses; whereas non-professional APCs (lacking co-stimulatory molecules) may present short peptides, but induce T-cell anergy [216]. The short epitope peptides bind to MHC class-I molecules on the surface of immature CTLs thus bypassing antigen processing and directly activating effector cells. These peptides can induce peptide-specific CD8+ T-cell responses. Some of the limitations in using short peptides in vaccines include: rapid degradation by serum or tissue peptidases, temporary immune responses of varying magnitude, restricted applicability for expression of MHC class-I allele and limited variety of antigen. Long peptides require internalization in professional cells presenting antigens (as DCs) for processing and presentation of antigens restricted to MHC. These peptides may also contain more epitopes allowing presentation through a considerable number of MHC alleles, or otherwise, they may include a number of own epitopes that predispose the host to additional self-reactivity [194].

In peptide-based vaccines, peptides TAAs can be recognized by T-cells, followed by an active immune response of the host immune system which has the ability to destroy tumor tissues and that is why they are used to make peptide vaccines against cancer. Considering that some natural TAAs peptides have low immunogenicity, they can be improved by being modified in order to increase the T-cell response. The single peptide vaccine has limited use because some TAA peptides can be lost or present at distinct stages with tumor progression. For this reason, multi-peptide vaccines have been developed that are based on focusing multiple TAA epitopes. However, some natural TAA peptides have a weak immunogenicity, with which the need arises to modify them to improve their immunogenicity [196]. Considering that by nature TAAs are not very immunogenic, an immunostimulatory adjuvant is essential to induce the generation of an effective immune response [218]. Considering that the immunogenicity of a peptide depends on its length and formulations, an appropriate selection and in some cases modification of the immunogen is necessary. The modification of the amino acid sequence can be considered as a strategy to improve the immunogenicity of the peptide. The length of the peptide may also be meaningful to achieve an effective CD8 response. It has been shown that adding few amino acids to the CD8 epitope, the capacity and duration of cross-presentation by APCs can be increased [216].

Some specialized variants of peptide-based and protein-based vaccines are vaccines based on anti-idiotype antibodies and the complex peptide-protein heat shock

(HSPPC) vaccines. Anti-idiotype vaccines contain idiotypes directed against an antibody of another antibody that recognizes a particular tumor antigen. The anti-idiotype antibody is used as a substitute antigen for vaccination rather than a passive antibody therapy. The heat shock proteins function as intracellular chaperones and can bind and present tumor antigens on APCs, through MHC class-I and class-II molecules leading to the activation of antitumor T-cells. Protein-based vaccines are composed of recombinant proteins or purified with peptides that are the antigen format most commonly tested in vaccination [210].

**Genetic vaccines** Part of the strategy associated with genetic vaccines is based on the transfection of somatic cells (myocytes, keratinocytes) or DCs that infiltrate the muscle or skin as part of an inflammatory response to vaccination, which consequently results in cross-priming or direct antigen presentation. Vaccines based on viruses, RNA vaccines and DNA vaccines form part of this classification [218]. Virus-based vaccines use viruses as tumor antigen vehicles to drive antigen-specific immune responses. The adenoviruses represent another viral system for virus-based vaccine [196]. The reason for using viruses as immunization vehicles is based on the phenomenon that causes viral infection to result in the presentation of virus-specific peptides, restricted to MHC class-I and class-II molecules on infected-cells. Viral vectors with low intrinsic immunogenicity are designed to encode TAAs, alone or in combination with immunomodulatory molecules [218]. RNA vaccines represent the alternative method of DNA vaccines. These inject the messenger RNA (mRNA) directly into the host to drive the immune response, which is considered safer than the DNA vaccine because it does not enter the nucleus [196]. The RNA vaccination is typically carried out together with other agents to obtain adjuvant or stabilizing effects [218]. Due to the instability and short half-life of mRNA, the development of RNA vaccines has been combined with some agents like protamines or liposomes to increase the immune response. Genetic vaccines use viral DNA vectors or plasmids carrying the expression to deliver the coding region of antigens or antigen fragments for vaccination purposes. The genetic and transferred material to the DCs or somatic cells, results first in the expression and then in the presentation of the antigens [196]. The version that is based on bacterial plasmids is constructed to function as a launch system and encodes vaccine antigens driven by efficient eukaryotic promoters. This type of vaccine allows delivering and expressing tumor antigens to be presented to both complexes: MHC class-I and class-II, which enables the stimulation of both CD4 and CD8 T-cells [218, 219]. This event allows DNA vaccines to trigger innate immune responses, and depending on their design and delivery sites, they can also stimulate humoral and cellular immune responses specific for the antigen. DNA vaccines can trigger the activation of helper T-cells (Th1 or Th2), thus being able to polarize the resulting immune response [220].

DNA vaccines depend on the presentation of tumor antigens by DCs and myocytes in which the DNA is delivered. These vaccines are introduced either as tumor antigens expressing naked DNA or by means of a viral vector, and through an intra-muscular, subcutaneous or oral route. DNA vaccines are easy to synthesize and

store [194]. In order to improve the potency of DNA vaccines, several strategies have been investigated, among which are included: plasmids that encode antigens to promote the expression and presentation of antigens, immunomodulatory molecules that attempt to stimulate the immune system or reduce immunosuppression [220], improvement in the design of vectors, codon antigen optimization, use of molecular adjuvants and traditional adjuvants, electroporation, co-expression of molecular adjuvants and strategies to boost and improve the response [219]. The plasmids used in DNA vaccines are of bacterial origin and appear to serve as a ligand that stimulates Toll-like receptors. The plasmid design produces improved epitopes which more effectively activate lymphocytes co-administer DNA encoding immunomodulatory and immunostimulatory molecules, the use of impulse primer strategies and approaches to break immunosuppressive networks in the tumor microenvironment. The administration of molecules to modulate immune cells can improve the potency of DNA vaccines against self-antigens. The blocking of CTLA-4 with anti-CTLA-4 antibodies after vaccination, and DNA plasmids encoding p53 and gp100 together with DNA encoding CD40L have been used under this strategy. The results have shown that immunomodulatory molecules can be incorporated among DNA vaccines to improve the potency of the vaccine. These vaccines have demonstrated potential as an effective immunotherapeutic strategy against cancer. DNA vaccines represent a platform for cancer immunotherapy with potential for mass application, these are very safe and stable, these are easy to manufacture and low production cost and can be easily stored and transported [220].

A crucial advantage of genetic vaccines is the easy delivery of multiple antigens in an immunization and the activation of the two type of immunity [218]. DNA vaccines turn out to be more cost-effective than other therapeutic vaccines, such as those that focus on recombinant proteins, tumor cells or viral vectors [220]. This is primarily justified to the fact that its cloning between plasmid vectors involves an only step. Also because the expression of an antigen gene driven by eukaryotic promoters and endogenous translational modification result in new protein structures that ensure *in-vivo* processing and immune presentation, and such events may involve reduction in cost and time of production [219]. Regarding their disadvantages, DNA vaccines report limited success in producing therapeutic effects against many cancers due to their poor immunogenicity, to the mechanisms of central and peripheral immune tolerance that limit their effectiveness and to the obstruction of therapeutic effects because of the fact that tumors can induce mutation or loss of immunodominant epitopes. Naked DNA does not easily spread *in-vivo* from cell to cell [220]. The problem of immune tolerance is yet a severe difficulty for DNA vaccination, since most tumor antigens are self-antigens that cannot trigger potent immune responses, thus producing inadequate immunogenicity [219, 220]. Another relevant problem of DNA vaccines corresponds to the failure to induce efficient immune responses, which is primarily due to the poor localization into the resident cells and the lack of co-stimulatory molecules in local APCs. The combination of DNA vaccines with other immunotherapy methods can improve the clinical efficacy of DNA vaccines [196].

DNA vaccines face challenges in trying to induce potent antigen-specific cellular immune responses as a result of immune tolerance against endogenous self-antigens in tumors [220]. To confront this tremendous challenge, the strategy to follow involves possessing a robust understanding of the immune pathways and the effects of vaccination accompanied by a broad array of immune receptors, the signaling molecules, cytokines, and transcription factors. All these elements together, are in the process of being tested as DNA vaccine adjuvants [219]. Vaccines based on nucleic acids, DNA and miRNAs, are being developed as a means to encode antigenic proteins and provide adjuvant function [210]. In DNA vaccines for humans, the safety profile is universally accepted and the principal interest in clinical trials is to demonstrate its efficacy [220].

### 2.7.3 Therapeutic Vaccines Against Cancers of Viral Origin

The challenges in the development of highly effective therapeutic vaccines against cancer derive from several sources including: the reduced immunogenicity of cancer cells, the immunosuppressive environments (micro and macro) induced by malignant cells, the compromised immune system of treated patients and the selection of patients to whom the vaccine will be administered that tries to eliminate large tumors that are already established [195].

The therapeutic vaccines against cancer use tumor antigens to stimulate the immune system of patients suffering from the disease [196]. Some cytokines are used in cancer immunotherapy or as adjuvants. In the first case, the cytokines do not focus exclusively on cancer cells, but try to stimulate the immune system in a more general way; however, in some cases it can lead to a much more appropriate immune response against cancer cells. In the second case, the cytokines are used in conjunction with a principal treatment, usually a vaccine to enhance the immune system. In the second case, the cytokines are used in conjunction with a particular treatment that is typically a vaccine, which is to reinforce the immune system. Interleukins and interferons are some groups of cytokines that act as chemical signals among cells. Interleukin-2 (IL-2) for example, can be used as a medicine for the treatment of some cancers or it can be combined with chemotherapy or with other cytokines, as is currently used with interferon alpha (IFN- $\alpha$ ). These treatments allow being more effective in counteracting some cancers, but the side effects of the combination therapy also increase. Interferons improve the body resist viral infections and cancers. IFN- $\alpha$  reinforces the ability of certain immune cells to attack cancer cells, slow down the growth of cancer cells and the blood veins that the tumor needs to grow. Some of the cytokines currently being studied for use against cancer and/or as adjuvants are IL-7, IL-12 and IL-21 [18, 174, 221, 222].

There are other medications that are not found naturally into the body, but they can reinforce the response of the immune system in a non-specific way. These drugs are called inhibitors, and they try to focus on molecules such as: PD-1, PD-L1 and CTLA-4, which normally help the immune system to be under control [223, 224].

These proteins help prevent the immune system from damaging conventional cells. This type of medication helps to strengthen the response of the immune system against some cancers. There are also other immunotherapies and other approaches that are currently being investigated to treat cancer, in order to obtain safer and more effective results. Such are the cases related to the development of new monoclonal antibodies, tumor cell vaccines, antigen vaccines, dendritic cell vaccines, vector-based vaccine, oncolytic viruses, T-cell therapies, infiltrating tumor lymphocytes and IL-2 [18, 225].

**Therapeutic vaccination against cancers that originate in HBV** Several strategies of therapeutic vaccines against persistent HBV infections and their sequelae have been developed. These vaccination platforms include: recombinant HBV proteins, DNA vaccines, recombinant virus vaccines and subviral particles as well as immune complexes of HBV surface antigens. Significantly, HBV does not contain oncogenic proteins that need to remain expressed in transformed cells, but they can cause HCC through indirect mechanisms like inflammatory events. For this reason it is necessary to focus the persistent viral infection before the malignant transformation, because HCC cannot express viral proteins [211]. Some of the strategies to control the infection and progression of HCC through antiviral therapies focus on the inhibition of viral replication by reducing the cccDNA load, blocking the polymerase and immunomodulation by IFN- $\alpha$ . Currently, the strategies available to treat patients with advanced HCC are surgical resection and liver transplantation. However, several immunotherapeutic treatments are available, including: (i) chemotherapy treatment, whereby small molecule drugs are used that focus on several signaling pathways. This modality is used after a surgical resection; (ii) antibody therapy; (iii) adjuvant therapy, mainly based on IFN; (iv) peptide vaccines through the use of the synthetic peptide SP94, the carcinoembryonic antigen GPC3 (*glypican-3*) that can act as an anti-cancer immunotherapeutic target and the immunotherapeutic antigen MRP3 (*multidrug resistance-associated protein 3*) that is used as an HCC target; (v) miRNA-based therapies, which correspond to drugs based on nucleic acid siRNA (*small interfering RNA*) and miRNAs (*microRNAs*). Some of the miRNAs used for HCC control include the following: miRNA-7, -21, -29, -34a, -122, -221 and miRNA-224 [226]. Cytokine-based immunotherapy has several studies under development that seek to identify the appropriate cytokines to be used as markers with diagnostic, prognostic and therapeutic approaches. Under a therapeutic approach, the cytokines IFN- $\alpha$  and IL-12 are studied. Some combined immunotherapies are also being studied, including: hTERT+IL-12, cytokines M1 (IL-1 $\alpha$ , IL-12/IL-23) together with TNF- $\alpha$ . Other immunotherapy approaches based on cell populations consider: immunotherapy based on NK-cells, DCs and T-cells. In addition, some therapies that focus on checkpoint inhibitors are under study, such as: PD-1, PD-L1 and CTLA-4; immunotherapies based on antibodies, which include the Trop-2 and Licartin; and other therapies that focus on growth factors like IGF (*insulin-like growth factor*) [227].

**Therapeutic vaccination against cancers that originate in HPV** Several therapeutic vaccines have been developed that include live vectors, proteins or peptides,

nucleic acids and cell-based vaccines. Most of these vaccines focus on the HPV E6 and E7 oncoproteins in order to deliver E6 and E7 antigens en various form to the APCs. Consequently, APCs try activating CD4+ or CD8+ cytotoxic T-cells specific for the HPV antigen. In vaccine therapies in cancers associated with HPV, there are certain advantages but there are also some limitations. Vaccines based on live vectors tend to report a potential safety risk, particularly in immunocompromised individuals, and efficacy after numerous immunizations using the same vector is limited. RNA vaccines focusing on HPV antigens and diseases associated with HPV have not yet been explored in the clinical context. DC-based vaccines are technically expensive, and for this reason they turn out to be ineligible for large-scale production. Furthermore, according to the culture techniques used, they can lead to a vaccine with inconsistent quality and lacking standardized criteria at the time of evaluation. Additionally, the most appropriate supply route for DC-based vaccine has not yet been determined. Vaccines based on tumor cells focused on HPV have not yet been developed and tested in clinical studies. Mainly because of its nature and the potential risks involved its production in mode and time turns out to be significantly expensive [68]. The effects of vaccination after the establishment of persistent HPV are currently unknown. It has not yet been determined whether vaccination against HPV antigens, such as E5 and E7, will represent an effective strategy to prevent cancer on individuals with persistent infection. It is also unclear which patients can benefit from certain immunotherapeutic approaches, considering that some specific group can respond to a certain therapy but others cannot, probably due to the HLA subtype or other immunomodulatory proteins [228].

**Therapeutic vaccination against cancers that originate in MCPyV** Although MCC is an extremely aggressive carcinoma, there is currently no approved therapeutic regimen available to treat advanced stages of this cancer [160]. However, the ST Oncoprotein of MCPyV in humans possesses the potential to be used as a diagnostic marker and a therapeutic target for MCC [166]. Based on animal studies, it was reported that a vaccine using MCV VLPs (*MCV virus-like particles*) can lead to *in-vitro* production of anti-MCV antibodies. Although these reports are inconclusive, it is suggested that anti-MCV particle vaccines could play an essential role in the future therapeutic advancement for the treatment of MCC. However, it is necessary to evaluate the usefulness of new immunomodulatory and molecular therapies that focus on MCC treatment [164]. Immunotherapy in MCC is of particular interest because positive MCPyV tumors express viral oncoproteins, and negative MCPyV tumors support a high mutational load associated with the production of neoantigens. Both cases are considered important immunotherapeutic targets. Currently, immunotherapies are being investigated that oppose the local immune evasion invoked by MCPyV. Inhibitions with PD-1 or blockade of PD-L1 revitalize the T-cells and activate the antitumor response [229]. Some studies show successful use of antibodies focused on PD-1 and others focused on anti-PD-L1 [230]. Other additional immunotherapies that are being investigated in clinical trials focus on: adoptive T-cells transfer, intratumoral IFN, IL-12 and TLR4 agonist [229, 230].

## 2.7.4 Personalized Immunotherapy

Diseases that originate in infectious processes caused by viruses are particularly challenging from the point of view of personalized immunotherapy. A thorough understanding of both the host and the pathogen, including its characteristics and interactions, is required to try to determine how these processes result in each individual [231]. Three components are essential in the composition of tumor immunotherapy. These are: the antigen against which immune responses are induced, the adjuvant that acts as a warning signal to alert the immune system, and delivery system that is responsible for providing the vaccine antigens and adjuvants in DCs [209]. The selection of an appropriate tumor antigen is one of the principal obstacles in the development of cancer vaccines. Silva and colleagues (2013) believe focusing on anti-cancer vaccines should consider some general criteria: (i) include peptide sequences that bind to MHC class-I; (ii) be processed by tumor cells and become available to bind MHC class-I molecules; (iii) be recognized by the T-cell repertoire in a form restricted to MHC class-I; and (iv) lead to expansion of CTL precursors that express specific T-cell receptors. An alternative strategy to overcome the problem inherent to the use of universal antigens corresponds to the development of personalized vaccination [232]. Some of the personalized immunotherapy strategies use vaccination with: defined neoantigens, complete autologous tumors, delivery systems of nanoparticles and DCs generated *ex-vivo* [195]. This strategy considers several delivery formats, including: long peptides, messenger RNA, DNA plasmids, viral vectors, engineered bacteria and *ex-vivo* antigen-loaded DCs [233].

**Personalized vaccines with defined neoantigens** Tumor antigens are involved in tumor development and progression, but from an immunological perspective they are grouped into two differentiated classes. One class corresponds to the self-tumor antigens, also known as tumor-associated antigens (TAAs) or tumor-specific antigens (TSAs), which are derived from normal proteins with elevated expression in the tumor but with limited expression in normal tissues [234]. TAAs are overexpressed in malignant cells but are also present in normal cells at reduced levels of expression. Vaccines that focus on TAAs have achieved limited success because TAAs are normal host proteins, and therefore are subject to both central and peripheral tolerance mechanisms [235]. The other class corresponds to non-self tumor antigens, which include both neoantigens and viral antigens [234]. The immune system has the ability to distinguish the self from the non-self and this provides it the ability to recognize and focus non-self antigens on cancer cells to exercise its control [235]. Oncogenic viral antigens have been identified in virus-induced cancers, such as: HPV, HBV and MCPyV. Given that these antigens are foreign to the body and are expressed only by cancer cells, they are extremely suitable for use in a cancer vaccine [210]. The human neoantigens can be recognized and/or focused by the immune system. Personalized vaccines that focus on neoantigens can successfully induce CD4 and CD8 T-cells responses [234].

Neoantigens are tumor-specific antigens that can arise from three classes of mutations. These are: (i) transient mutations, which result in genomic instability

within the tumor; (ii) conductive mutations, which help to growth of cancer and the resistance of focused therapies; and (iii) other genetic modifications. All these mutations can alter the amino acid coding sequences. In human tumors that originate from viruses, the proteins encoded by the viral open reading frame (ORF) correspond to another type of neoantigens [235]. Neoantigens result from a considerable quantity of somatic mutations that are found in human cancer cells and therefore are totally tumor-specific. Mostly part, neoantigens are patient-specific since the particular mutations found in any pair of tumors are usually different. Regarding vaccination with defined neoantigens, the central tolerance does not represent a concern and therefore the affinity of T-cells for neoantigens can be significantly superior. In addition, focusing on tumor-specific neoantigens may not induce autoimmune toxicity in healthy tissues [195]. Considering that tumor neoantigens are generated as products of somatic mutations, and therefore they are not only tumor specific but also highly immunogenic based on the absence of central tolerance [210]. Sahin and colleagues (2018) hypothesize that only tumors with a high mutational load, and therefore with a greater diversity of T-cell specificities reactive to cancer that occurs spontaneously, may qualify for immunotherapy based on neoantigens [233]. Although neoantigens are attractive targets for cancer immunotherapies, many neoantigens arise from single mutations and are not shared among different patients. Therefore, immunotherapy directed by neoantigens probably needs to be personalized [235]. When normal and tumor DNAs are compared using complete exome sequencing, somatic mutations can be identified. The use of exome sequencing data and transcriptome data from the sequencing of the tumor RNA, allows predicting the neoantigens and their possible affinities towards the MHC/HLA molecules, which is achieved through the use of computational algorithms [196]. Yarchoan and colleagues (2017) briefly explain this process, which begins with the extraction of host DNA, tumor cells and the decomposition of DNA into fragments. The encoded fragments of the DNA are sequestered with artificial DNA or RNA baits that are complementary to the focused DNA, and the uncoded sequences are discarded. Coding sequences are then amplified, sequenced and analyzed using a referent genome. Subsequently, DNA mutations can be constructed between an altered amino acid sequence by coincidence with each DNA codon and its corresponding amino acids, thus providing the initial data required to prioritize tumor neoantigens [235]. The exhaustive mutational analysis is carried out through the complete sequencing of the exome. Then neopeptides encoded by somatic mutations in the tumor are selected, those that have the most significant probability of being presented by the MHC molecules of the individual on the basis of affinity predictions [210].

Antigen-specific vaccines differ from whole-cell vaccines in that they scarcely contain the antigenic portions of the tumor cells that are necessary to boost an immune response [210]. Based on current knowledge, vaccines cannot be designed to target shared neoantigens for a huge group of patients. Therefore, most attempts at active immunotherapy in cancer patients have focused on their self-antigens. The design of self-antigen vaccines to be used in cancer clinical trials is relatively simple and can be applied in a considerable number of patients. Therefore, self-antigens

can be shared among several patients, some even among patients with diverse types of cancers [195].

Some of the platforms used to focus on neoantigens include DNA vaccines, peptide-based vaccines and *ex-vivo* dendritic cells [234]. Although several strategies have been proposed to focus on personalized neoantigens, the two best-established strategies correspond to: T-cell products specific for neoantigens for which adoptive cell therapy is used, and personalized vaccines encoding predicted neoantigens. However, the design of antibodies directed to personalized tumor neoantigens is impractical and significantly expensive [235]. It has been suggested that an effective cancer vaccine needs to target multiple neoantigens and should be customized for tumors from a sole individual [210].

In personalized cancer vaccines, tumor tissue is required to identify and prioritize candidate neoantigens. Considering that thousands of mutations may be present in an individual with cancer, the use of algorithms that allow prioritizing mutations that can be focused by the immune system is required. The first step to prioritize neoantigens is to confirm the expression of mutant genes in the tumor, which is achieved by sequencing techniques such as: RNA-Seq and RT-PCR, using cDNA from tumor RNA. The second step consist to estimate the binding potential with the appropriate MHC/HLA alleles, for which computational algorithms are used that quantify the predicted binding affinity of candidate epitopes for MHC/HLA class-I alleles [234], which are supported by technologies like a massive parallel sequencing (MPS) and epitope prediction algorithms [235]. Through these techniques it becomes possible to discover specific therapeutic neoantigens potentials of the patient. Mass spectrometry, through the analysis of immunopeptides, also allows to directly identifying the presented neoantigens. This technique is usually more accurate than prediction technologies [196]. Nevertheless, the use of MHC/HLA binding predictions currently presents some challenges. Prediction algorithms need to identify certain types of neoantigens, including those that form MHC class-II antigens, as well as proteins that arise from the fusion of genes and neoantigens that are generated from errors in translation [235]. Through tumor sequencing together with the prediction of MHC binding epitopes, it is possible to identify candidate tumor neoantigens on per-patient basis. On this approach, evidence of its immunogenicity and use for the development of personalized vaccines has been found. Part of this evidence includes: findings of an elevated load of neoantigens associated with stronger T-cell responses and better clinical outcomes in patients with cancer; T-cell populations specific for neoantigens that are expanded in the effective establishment of antitumor immunity; and antigen-specific T-cells that are cytolytic for tumor cells that display mutated peptides and that may contribute to partial or complete tumor regression [210].

The tumor neoantigens approach can be more specific, more effective and less toxic than other cancer immunotherapy approaches. However, various challenges are in force (cost and tumor heterogeneity) and several questions still remain unanswered. Among the pending questions, there is a need to know how to trigger the traffic of reactive T-cells towards the tumor-site, to enable its long-term persistence and achieve tumor destruction [193]. Yarchoan and colleagues (2017) consider that

the best approaches to predict neoantigens are not yet known [235]. Nonetheless, it has been shown that tumor neoantigens are key targets for therapeutic vaccines, for adoptive cells transfer (ACT) and immune checkpoint blockade (CPB) [210]. To destroy established cancers in the presence of immune checkpoint inhibitor (as CTLA-4 and PD-1), T-cells can recognize antigens deployed by MHC/HLA on tumor cells. It has been documented that T-cells specific for neoantigens expressing PD-1 have been identified in the peripheral blood of patients with a melanoma, and are correlated with the activity of PD-1 inhibitors [235]. Neoantigen-specific T-cells may play a role in tumor regression observed in checkpoint blockade therapy, which depend on the presence of tumor-infiltrating T-cells at tumor sites. Neoantigens derived from mutations can play fundamental roles in the induction of tumor-specific immune responses in checkpoint blockade immunotherapy. Neoantigens restricted to MHC class-II have also been employed to generate therapeutic immunity against cancers expressing neoantigens, using peptide-based or RNA-based vaccine strategies [193]. It has been shown that vaccination based on neoantigens is feasible, safe and capable of inducing broad and robust T-cell responses specific to neoepitopes in patients with a melanoma [210].

**Personalized vaccines with whole autologous tumor** Autologous tumor cells constitute an obvious source of tumor-associated antigens (TAA) for vaccination purposes [236] and can be modified to confer greater immunostimulatory characteristics. Generally, these cells are irradiated and combined with an immunostimulatory adjuvant and then administered to the individual from whom these were isolated [218]. Irradiated tumor cells for autologous vaccines are genetically engineered to release cytokines or chemokines [237]. Vaccination with whole tumor antigens is based on its abundant source of antigens, formed by epitopes of both types of T-helper cells both CD8+ and CD4+, possibly a necessary condition to ensure tumor localization of low affinity CD8 cells. Whole -tumor vaccines could also greatly decrease the chance of tumor escape compared to single-epitope vaccines [238]. The use of whole tumor cells in cancer immunotherapy represents a promising approach that obviates some of the challenges in defining specific antigens for the development of the vaccine [237].

Tumors are recognized by T-cells through tumor-associated antigens (TAAs). TAAs can be non-mutated self-antigens and mutated neoantigens. The expression of the self-antigens may be restricted to certain germ lines, be overexpressed in cancer cells or be tissue-specific. The tumor cells used for vaccination can be autologous or allogeneic. The presentation of tumor antigen *in-vivo* in both types of vaccines depends on the uptake and cross-presentation of endogenous DCs. While autologous tumor cells can more accurately represent the antigenic profile of a specific patient, the complexities of making a vaccine adapted to the patients has led to the application of vaccines based on allogeneic tumor cells that can be prepared with a standard approach in great quantities. Among the repertoire of patient-specific TAAs in autologous tumor vaccines, private neoantigens are included, and in this way these can represent a true personalized approach to active immunotherapy [195]. Vaccination based on whole tumor cells offers the benefit of present-

ing tumor-specific antigen and/or patient to elicit immunity against a broad spectrum of tumor-associated antigens [237]. Autologous tumor vaccines are more complex to prepare than allogeneic vaccines, but offer the potential benefit of including unique mutant neoantigens, which may be important antigens of tumor rejection [194].

In active immunotherapy, using complete autologous tumors, the identification of defined neoantigens may be unnecessary [195]. In theory, they have the advantage of incorporating the full range of neoantigens that result from the somatic mutation, without having to identify these neoantigens directly [211]. One of the principal advantages of whole tumor cell vaccines corresponds to their potential to present the complete spectrum of antigens associated with the tumor to the patient's immune system. However, the preparation of autologous tumor cell vaccines requires having enough tumor specimens, which limits this technology to only certain types of tumor or stadiums [218]. Another disadvantage is it is unknown how effective the processing of neoantigens epitopes from the lysate of tumor cells results for DC presentation [211]. Considering that tumor cells are tolerogenic in nature, in order to produce immunogenic vaccines, improved methods of tumor preparation and delivery are required together with an adequate immune impulse [195]. Despite the promise given to the vaccination of whole tumor cells, this approach typically requires *ex-vivo* genetic manipulation of the tumor cells, which leads to an elevated cost and to face important concerns regarding its regulation. A limitation of this approach corresponds to the fact of being able to isolate enough tumor cells for vaccines of autologous tumor cells, given that a substantial number of these cells die or lose their function during *ex-vivo* manipulation, so it is suggested to evaluate in the future if this system can reduce the amount of tumor cells needed to achieve an effective vaccination [237].

**Personalized vaccines with nanoparticles delivery system** The development of nanoparticles (NPs) has been applied in biomedical fields of several scientific disciplines, including the detection and diagnosis of diseases, delivery of pharmaceuticals drugs and immunotherapy based on vaccines [239]. Nanoparticles (NPs) are solid colloidal particles, formed by macromolecular materials that can be used therapeutically as adjuvants in vaccines or as carriers in medicines [240].

The field of nanotechnology has allowed the emergence of nanomaterials with potential biomedical application [241]. The family of nanocarriers includes polymeric nanoparticles, lipid-based carriers, dendrimers, carbon nanotubes and fold nanoparticles, which are used as delivery systems [240]. It has been observed that proteins encapsulated in poly (*lactic-co-glycolic acid*) (PLGA) nanoparticles are able to activate cytotoxic CD8+ T-cells and induce potent antitumor activity. It has also been shown the biodistribution of gold NPs has been associated with several types of cells in the immune system, including B-cells, granulocytes, DCs and T-cells [239]. These particles allow focusing on DCs to exploit their capacity to capture and to facilitate the co-delivery of antigens together with adjuvants for the same target cells, which attempts to promote strong DC maturation [232, 241]. Apart from the increased uptake by DCs, the nanocarriers can also provide protec-

tion for the encapsulated agents including proteins, peptides and nucleic acids which otherwise would suffer immediate degradation when administered. TLR ligands equally benefit from encapsulation by nanocarriers, since some molecules between which bacterial and viral nucleic acids are found are easily degradable targets [232]. Nanomaterials for vaccine delivery are designed to improve the uptake of the antigen by APCs and/or to obtain a controlled or sustained release in the presentation of the antigens [242].

The nanoparticles can be designed either to avoid recognition of the immune system or specifically to inhibit or improve immune responses. Some of these nanoparticles are able to enter APCs by distinct routes, modulating in this way the immune response to the antigen. This event proves to be fundamental for the induction of protective Th1 immune responses against intracellular pathogens [240]. Some formulations of nanoparticles can considerably increase the location of these drugs, either in target lymphoid tissues or within immune cells, and consequently enhance their potency as well as their safety [209]. Typically, the nanocarriers for vaccine delivery can be divided into three groups: (i) nanocarriers that passively focus on APCs, where bound or encapsulated antigens can be internalized more efficiently, can be protected from proteolytic degradation and can be given a particular form; (ii) nanocarriers that actively focus on APCs, where antibodies or specific ligands allow the vaccine material to interact with specific DC membrane receptors, which then result in receptor-mediated endocytosis; and (iii) cytosolic delivery and intelligent nanocarriers, where the delivery of the cytosolic target is executed once the intracellular target is in the cytoplasm, especially for DNA vaccines that require expression in the cytosol for the production of antigens [242]. These carriers must be designed as low toxicity systems with physical and chemical structures suitable for vaccine delivery. For this purpose it is necessary to consider the properties of nanoparticles in terms of composition, size and surface characteristics to manufacture cancer vaccine delivery carriers [209]. The size and surface properties of the nanoparticles represent the predominating factors that control their behavior in biological barrier transport, tissue and cellular uptake and the induction of immune responses [242]. The technique of physical-chemical modification of a surface or interface at the molecular level applied to nanoparticles, can be used to increase the time of permanence in the blood, reduce the non-specific distribution, and in some cases, tissues target or specific cell surface antigens [232].

The efficiency of vaccination with a nanoparticles delivery system may require the uptake and presentation of tumor material through *in-vivo* DCs. However, the vaccine based on tumor cells can fail to deliver appropriate maturation signals for DCs, therefore in this way it can lead to T-cell tolerance. To overcome this obstacle, biodegradable particles have been developed that are used to deliver TAAs to *in-vivo* DCs. TLR ligands are potent activators of DCs, and their encapsulation can defend them from degradation and increase their targeting to the DCs [195]. The nanocarriers can provide direct intracellular access, facilitating the intracellular commitment of TLR3, TLR7, TLR8 and TLR9, together with their ligands, consequently improving their efficacy as vaccine adjuvants. The introduction of TLR ligands or ligands for other PRRs on the surface of nanoparticles becomes an attrac-

tive option to considerably improve the immunogenicity of nanoparticles vaccines, via supplying an intrinsic damage signal [232]. However, some TLR ligands can mediate unwanted immune responses in non-target sites. Although this approach demands further work, it maintains immense potential for efficient delivery of personalized active immunotherapy in cancer patients [195].

The use of nanoparticles offers several advantages over the administration of the antigen in cancer immunotherapy. Saleh and colleagues (2016) report some of these advantages: (i) protection of enzymatic degradation, considering that the antigens are susceptible to digestive enzymes in blood and interstitial fluids; (ii) improvement in absorption in the targeted tumor tissue, either because the permeability and retention effect are improved, or because the active pathway is focused with the use of ligands and the ability to control the pharmacokinetics and the profile of tissue distribution; (iii) enhancement of cellular uptake of DCs to trigger a strong immunostimulatory cascade; (iv) delivery systems designed to initiate immunogenic cell death or to target immune checkpoint molecules that can drive antitumor responses and reverse immune suppression; (v) the simplicity of design and use, coupled with the feature of multifunctionality. All these advantages make nanoparticles an attractive carrier system for tumor vaccines and immunotherapy [209].

Nanoparticles represent a promising approach for efficient delivery of vaccines that are directed towards the lymph nodes and several of them have been developed pre-clinically, although the clinical translation of these nanoformulas have been hampered by limitations in their safety, suitability for their reproducibility in large-scale manufacturing and for its *in-vivo* integrity [243]. Many nanoparticles appear to show some toxicity in several cell types, and it is suggested that the adoptive toxicity of such particles should be eliminated [240]. Concerns about the potential toxicity of nanoparticles have focused predominantly on their biological fates and the resulting biological consequences, especially on materials that are at severe risk of accumulation because they are undegradable. The effects that may occur in non-target locations and the unpredictable toxicity that could be caused to susceptible populations, from unintentional exposure, equally represent cause for concern [242].

Nanomaterials by themselves produce intrinsic immunomodulatory activities that make them potentially applicable as adjuvants. Immunomodulatory activities include inflammasome activation, recruitment of immune cells and triggering of the complement system [242]. Nanovaccines are also explored as a means of delivering vaccines. These can efficiently co-deliver adjuvants/antigens locally administered into the lymph nodes via lymph. Although the pharmacological behaviors of almost all nanovaccines have been studied, either invasively or semi-quantitatively, several aspects have hampered the clinical translation of most synthetic nanovaccines. These obstacles are principally found in their large-scale manufacturing processes, quality control, formulation and safety [243]. Within the development of nano-adjuvant vaccines, there are some challenges that need to be overcome. Zhu and colleagues (2014) define some of these challenges, among which are: (i) the need to establish a way to optimize biological behaviors and minimize the potential risks linked to nanomaterials; (ii) the need to carefully control the extent of immune responses induced by nanomaterials, in order to achieve the desired adjuvant effects

and no toxic responses; and (iii) the need to have effective monitoring techniques or methodologies, particularly *in-situ* quantitative methods and in real time, that help in the characterization of biological behavior between the physicochemical properties of NPs and the triggering of related immune responses [242].

Several approaches with nanoparticles are being developed with a view to counteracting tumor-mediated immunosuppression and improving the immunotherapeutic result. Some of these strategies include: silencing of STAT3 in DCs, combination of delivery of inhibitors TGF- $\beta$  and IL-2 by liposomal polymer gels, intelligent multifunctional nanoparticles focused on TAMs (*tumor-associated macrophages*), nanoparticles for delivery of TLR ligands, and blocking of PD-1/PD-L1 pathway. Explicitly, the multifunctional nanoparticles possess an enormous capacity to achieve multiple purposes simultaneously; including the co-delivery of multiple components including TAAs coupled with adjuvants and immune enhancers and focused specific delivery by modifying the surface of the nanoparticles. In addition, they allow manipulation of the immune system through the promotion of effector immune cells and improve cancer responses. In addition, these allow manipulation of the immune system through the promotion of effector immune cells and improve cancer responses [209]. Yang and colleagues (2016) study the kinetics of cellular biodistribution of polymeric NPs by evaluating them in various immune organs, including blood, bone marrow, spleen and lymph nodes. They suggest the understanding of the kinetics of the biodistribution of polymeric NPs in the immune system is of essential importance in the development of immunotherapies based on targeted delivery of nanoparticles, and also identify potential target cells for delivery of antigens [239]. Kuai and colleagues (2017) propose an alternative strategy that they consider may be applicable to personalized therapies with a broad range of bioactive molecules and images formation agents. They propose preformed nano-carriers mixed with adjuvants and antigenic peptides, including tumor-specific mutant neoepitopes, to produce personalized cancer vaccines. In order to achieve this, they develop a nanovaccine system ideally adapted for individualized neoepitope vaccination based on high density synthetic lipoprotein nanodiscs (sHDL). They manage to demonstrate the power of their proposal to generate broad spectrum T-cell responses with remarkable therapeutic efficacy when combined with inhibitors of immune checkpoints [217].

In summary, the predominant factors of nanoparticles to control their behaviors in the transport of biological barriers, tissue, cellular uptake, and induction of immune responses are the properties of size and surface. However, it still remains unclear what is the optimal size range in the vaccination delivery system. In contrast, it has become clear that nanoparticles turn out to be the most efficient in the crossing of biological barriers and circulation in the blood, in addition to possessing a prolonged average life [242].

**Personalized vaccines based on ex-vivo DCs** Keratinocytes (KCs) are the most abundant epidermal cells and play a fundamental role in innate immunity on human skin. KCs produce a broad range of cytokines, chemokines and antimicrobial peptides and have the ability to express several TLRs. These cells are capable of allowing

ing the elimination of invading pathogens and recruiting immune cells, like DCs, helping in this way to maintain an appropriate balance between reactivity and tolerance of the immune system. In addition, KCs are capable of priming new reactive T-cells in skin [232]. DC-based vaccines can be prepared in both modalities: *in-vivo* or *ex-vivo*. The ex-vivo DCs are generated mainly from progenitor cells of the bone marrow in the presence of GM-CSF (*granulocyte-macrophage colony-stimulating factor*) and the interleukins IL-4 or IL-13 [214].

Three fundamental events related to the optimization of DCs for tumor vaccination include: (i) appropriate selection of tumor antigens; (ii) selection of an appropriate strategy to load tumor antigens into DCs; and (iii) determination of the method of optimal administration of the vaccine to ensure the loaded DCs can migrate to the lymph nodes. All these events together are fundamental to induce immune responses [214]. Current strategies based on dendritic cells use patient's own DCs to generate therapeutic vaccines. DC vaccination requires the induction of DC maturation, in order to trigger potent T-cell effectors and immune responses from memory. The DCs are collected from the patient, matured *ex-vivo* using adjuvants, loaded with tumor antigens and injected back into the patient. After the injection, the DCs present tumor antigens to tumor-specific T-cells, which results in the activation and expansion of T-cells [241].

In relation to personalized vaccines through the use of DCs generated *ex-vivo*, several techniques have been tested to perform the delivery of the primed antigen: (i) dendritic cells with tumor load, means that DCs are boost with known tumor antigens; (ii) fusion of DCs and tumor cells vaccines; (iii) vaccination with DCs transfected with DNA or RNA, containing the gene for the antigen of the protein of interest; (iv) incubation with lysate of tumor cell lines or whole allogeneic or autologous tumors. For the first approach, the DCs can be generated *ex-vivo* from the patient's monocytes, loaded with diverse forms of antigens, then activated and again infected to the patient. By boosting the *ex-vivo* DCs, more opportunities are generated to manipulate the DCs during the impulse and subsequent activation, thus optimizing their immunogenicity. In the second approach, the cell fusion method allows DCs to be exposed to a broad array of TAAs expressed by whole tumor cells. The DCs then process TAAs endogenously and present them through the MHC class-I and class-II pathways in the context of co-stimulatory molecules, which results in the simultaneous activation of CD4+ and CD8+ T-cells. In vaccines using fusion of DCs and tumor cells, this fusion is achieved by the use of a membrane destabilization agent (PEG or TGF), or through electroporation and electrofusion. About the fusion method, the DCs and the cancer cells become hybrid cells that share a unified cytoplasm and preserve the identity of their nucleus, with which all the TAAs are delivered to the DCs to then continue with their process. In the third approach, using DCs driven with whole tumor RNA offers an alternative solution, which harbors many of the advantages of whole tumor DC vaccines. The amplified tumor RNA may be directly introduced to the DCs, generally by electroporation, therefore leading to

the production of tumor proteins that could be processed and presented by the DCs to activate tumor-specific T-cells. In the fourth approach, the tumor lysate includes RNA derived from the tumor, cell lysis and autophagosome. The advantage of using whole tumors or tumor cell lines is that the presentation of the antigen is prolonged and multiple epitopes can be presented on MHC molecules of different haplotype, which allows to possessing the potential to induce CD4+ and CD8+ T-cell responses for a broad spectrum of antigens [195, 214].

Several strategies have been developed to effectively load tumor antigens into DCs, including the introduction of antigenic proteins or peptides, whole tumor cells, and tumor antigens that encoded DNA, mRNA or recombinant virus. For DC vaccines loaded *ex-vivo*, first the DCs are differentiated and activated with cytokines and later loaded with antigens [210]. Currently, therapeutic DC vaccines are principally focused on tumor antigen mRNA transfected DC vaccines [196]. DCs can also be transfected with mRNA encoding TAAs. It has been shown that these DCs possess the ability to induce strong tumor antigen-specific T-cell responses *in-vivo*. It is considered that electroporation is the most efficient method to introduce mRNA into DCs. Bacterial or viral vectors have also been evaluated to target DCs with tumor antigens. These vectors possess many advantages, which include: the ability to insert genes encoding TAAs, remove genes to encode virulence, remove replication factors for safety, and induce DC maturation [213].

Despite the forceful antitumor immune responses that DC vaccination based on *ex-vivo* loaded antigens has shown, this strategy still faces several challenges. Many patients develop T-cell responses again, but this often does not result in a significant improvement in the next clinical response. The *ex-vivo* DC culture procedure is expensive, time-consuming and also depends upon dedicated infrastructures, that which cause difficulties for its wide-spread clinical use. Because it is therapy based on the patient's own cells, it is equally problematic to ensure the quality of the product due it depends strongly on the level of depletion of the patient's immune system [241]. Although the form of vaccination with DCs vaccines may be defined in part, the target tissue to which the resulting T-cells are directed requires several studies to be able to compare the performance of these cells. In addition, other key variables to be able to induce the strongest and longest lasting immune response have not yet been determined. Some of these variables involve: the amount of DCs to be injected, the vaccination program (frequency and doses to be applied) and the injection route (oral, intradermal, subcutaneous, intramuscular, intranodal, intravenous, or even intratumoral) [194, 213]. Although DCs vaccines can produce IL-12, considered a key factor for vaccine efficacy, it is currently unknown how much IL-12 is sufficient and whether the DC vaccine will continue to produce IL-12 after injection or after of the encounter with T-cells. Given the correlation between the efficacy of the vaccine and the secretion of IL-12, it is relevant to know if the expression of IL-12 continues in production after the application of the injection [213].

## 2.7.5 Combined Immunotherapy

Combined immunotherapy in cancer attempts to develop applications from the mixture of multiple modalities that focus on diverse aspects of the tumor and different immune pathways, in order to achieve lasting antitumor effects and more effective therapeutic results [218]. The above, because cancer vaccines used as monotherapies have not achieved the expected success [195, 244]. In this way, the combined strategies are proposed as a key to success for DC vaccines [213]. Fecek and colleagues (2016) group the approaches of combined immunotherapies into two broad categories, immunoregulatory and immunoconditioning strategies. The strategies that are part of the first category are designed to antagonize and/or remove immunosuppressive networks within the tumor microenvironment. Strategies belonging to the second category are designed to condition the tumor microenvironment stimulating it to respond more robustly in DC-based vaccines, thus facilitating the recruitment, optimal functioning and durability of antitumor T-cells induced by vaccines. Two strategies stand out within the immunoregulatory group, these are: (i) therapies that focus on immune checkpoint pathways; and (ii) therapies that attempt to suppress the mechanisms of immunosuppression using inhibitors. Within the immunoconditioning group, the following four strategies stand out: (i) therapies focused on MAPK inhibitors in combination with DC vaccines; (ii) inhibitors Tyrosine kinase in combination with DC vaccines; (iii) HSP90 (*heat shock protein 90*) inhibitors in combination with DC vaccines, which try to conditionally improve tumor cells by loading MHC class-I with antigenic peptides recognized by specific CD8+ T-cells; and (iv) HDAC (*histone deacetylases*) inhibitors that aim to improve the immunogenicity of tumor cells [244].

Other modalities have also been proposed that involve the combination of cancer vaccines with conventional therapies such as: radiation and chemotherapy, to produce synergistic effects [218]. It is reasonable to think the combined use of immunotherapy with innate immune stimulation, chemotherapy or radiation therapy can improve immune responses and clinical outcomes [193]. It is believed that the use of cancer vaccines based on DCs in combination with alternative immunotherapies of cancer or radio/chemotherapies, can define treatment programs to improve their potency in support of antitumor, protective or therapeutic immunity [244]. Koido (2016) suggest that some chemotherapeutic agents contain the potential to increase cancer vaccines. Chemotherapy or irradiation agents can promote intrinsic immunogenicity of whole tumor cells in multiple forms. Some therapies subjected to immunogenic modulation that demonstrate upregulation of several damage signals consider use of HMGB1, HSP70/90, ATP and CRT [245]. Local radiation doesn't merely reduce the tumor but also generates an inflammatory environment, by which the presentation of moribund TAAs released by the tumor in DCs and the consequent priming of T-cells is promoted. Under this approach, some chemotherapeutic agents are able to induce immunogenic death of cancer cells, which results in cross-priming of TAA-specific T-cells and subsequent antitumor immunity [218].

To improve the effectiveness of personalized vaccines, some options are suggested. One of them proposes the combination of next-generation personalized vaccination with other modalities of immunotherapy, which could become the key to achieve significant therapeutic effects in patients with cancer [195]. Combined therapy that includes personalized vaccine in conjunction with checkpoint inhibitors can improve the ability of the immune system to eliminate tumors and promote antigen responses [234]. It has been shown that inhibition of checkpoints using antibodies focusing on CTLA-4, PD-1 and PD-L1 represents a promising option on the path to personalized medicine, and has also provided a proof of concept in which T-cell-based immunotherapy can offer therapeutic benefits for a variety of tumors [196, 218, 234]. Considering that antibodies to block checkpoints increase T-cell responses specific for neoantigens, producing synergy together with personalized cancer vaccines can be substantial. Another option proposes the use of anti-angiogenic therapies, such as blocking VEGF that can overcome the tumor endothelial barrier and improve the orientation of T-cells towards tumors, as well as achieve an alternative effect that promotes immunity by reducing myeloid-derived suppressor cells [195]. Regarding the development of strategies focused on immunosuppressive micro-environment of tumors, the blocking of immune checkpoints combined with DC-tumor fusion cells is also suggested, as an effective treatment for patients with advanced cancer [245].

The success of T-cell-based immunotherapy depends in great measure on the role played by HLA class-I molecules during the presentation of tumor antigens. The reduction or loss of expression of these molecules is associated with immune escape, and to counter this event some proinflammatory cytokines have been used [234]. Active immunotherapy can also be combined with adoptive T-cells transfer to improve therapeutic efficacy. Adoptive transfer of tumor infiltrating lymphocytes (TILs) that recognize numerous TAAs can also enter synergy with whole-tumor personalized vaccine. Such a strategy could prime and expand the in-vivo TILs without having to define their antigen specificity. The personalized vaccination of patients against their own tumor antigens is a promising immunotherapy approach. This strategy can activate a repertoire of tumor-specific T-cells, including high-avidity T-cells specific for private mutated neoantigens, which could mediate antitumor responses with minimal effects on healthy tissues. Nevertheless, further studies are required to evaluate some of its limitations and uses [195].

Of the three signals needed by an effective immune response, the third signal is associated with the use of cytokine-anti-cytokine antibody complexes, which increase the half-life in circulation of these cytokines and in some cases directs their binding to the specific receptors on T-cells. This complex can represent a potent path to improve the effects of peptide vaccines. Although the use of peptide vaccines may not be optimal for treating tumors that express low levels of MHC, some therapies targeting molecules like EGFR and MEK inhibitors, increase MHC expression on tumors and could be combined with peptide vaccines. The combination of peptide vaccines with checkpoint inhibitors can improve the amount and

function of T-cells specific for neoantigens. Because inhibitors alter immune tolerance, these can be beneficial in peptide vaccines along with non-mutated TAAs. The combination of CD4 and CD8 epitopes could represent another strategy to increase the efficacy of the peptide vaccines [216].

The optimization of immunotherapeutic strategies is essential to achieve a robust antitumor immunity, capable of mediating long-lasting objective clinical responses in cancer patients. The use of vaccines based on DCs in combination with protocols and conditioning adjuvants of the tumor microenvironment and immunomodulators are shown as an exceptional promise to improve the potency of cancer therapies [244].

### 2.7.6 Administration Routes

To achieve optimal T-cell priming during the vaccination process, antigens from the vaccine should be protected from degradation and brought into contact with APCs [210]. From the perspective as a system, the medium to be used plays a fundamental role in the development of biological processes, since the interactions are established exclusively through those medium. The chosen medium allows containing the constitutive elements of the biological system, to organize the exchange of material, energy and information and also permits to establish the interaction among their networks [12]. A widely used strategy to increase the immunogenicity of a vaccine is to induce a local inflammatory response. At the vaccination site, the trafficking of APC is generated and conducted to the site where these can capture and process vaccine antigens for presentation of antigens or cross-presentation by T-cells in the draining lymph nodes. For this purpose, emulsions, aluminum salts, water in oil, liposomes, virosomes, among other means are used [210].

The main immunization routes for human vaccines include the following: oral, intranasal, intramuscular, intradermal, intranodal or intralymphatic, subcutaneous, intraperitoneal, intravenous and intratumoral. Although the optimal means of administration has yet been undetermined [214, 242], the preferred routes to administer cancer vaccines correspond to the subcutaneous via and the delivery in lymph nodes, where the tumor antigens can be effectively captured by DCs [196].

The intranodal route has been shown to generate strong T-cell responses and remarkably superior antitumor immunity due to the bioavailability and adjuvant effects inherent to RNA in the lymph node microenvironment [232]. However, the intranodal method proves to be more invasive than other injectable forms, like intravenous and subcutaneous injections. In addition, injection of vaccines into the lymph nodes is technically more complicated, since an incorrect injection could interrupt the architecture of the lymph node [214].

The delivery system of nanoparticles through the intranasal route attempts promoting immunity in the mucosa and lung. The deposition and distribution of nanoparticles in the respiratory tract are governed by diffusion, because of the displacement when they collide with the air molecules rather than the inertial impact, the gravitational assignment or the interception of bulk particles. It has been shown

that once deposited, the nanoparticles seem to easily transfer across barriers to extrapulmonary locations, thus managing to target on different organs. In contrast, large particles are rarely transferred to extrapulmonary sites and they are cleared through mucociliary or phagocytosed movements [242]. With regard to nanoparticles, its development as a delivery system continues representing a fundamental area for future research [240]. It has been suggested that the delivery system of nanoparticles through intradermal injections turns out to be more efficient in overcoming biological barriers compared to microparticles [242].

The skin itself is an attractive organ to be used as an immunization route. The intradermal route offers the potential for safe immune stimulation, because this way prevents direct contact between potent adjuvants and the general circulation [232]. DC vaccines have been administered by intradermal, subcutaneous, intravenous, intranodal or intratumoral routes, although the optimal via administration has yet been unestablished [213]. The route of administration of antigen-loaded DCs affects the migration of DCs towards lymphoid tissues, as well as the magnitude of the antigen-specific CTL response. Intratumoral administration of DC vaccines has shown retention power at the site of injection, with a reduced amount of DCs detected in the lymph nodes, indicating a failure of the vaccines to achieve their objectives [214]. Innovative strategies include the administration of DC vaccines by more than one route, this means intradermal plus intravenous, to induce a systemic response and to be administered directly into the lymph nodes (intranodal) [213]. This last route offers the advantage that DCs do not need to migrate, since they are prepared and are in close proximity with T-cells in the lymph nodes. On DC vaccines it has been shown that the immunization capacity result more effective when using subcutaneous injection than when intravenous injection is used, specifically in the induction of CTLs. This is because the DCs injected subcutaneously accumulate in the lymph nodes, while the DCs injected intravenously are sequestered in the spleen [214]. Regarding nano-adjuvants injected either intravenous or by nanoparticles that pass the tissue barriers at the site of administration, they typically enter directly into circulation [242].

DNA vaccines may be administered through different methods, including syringe injection, gene gun, electroporation, nanoparticles, microneedles, and liposomes [219]. It has been shown that a DNA plasmid vaccine, administered directly into lymphoid tissues, turns out to be more effective for the generation of CTL immune responses than intradermal or intramuscular routes. The approach of antitumor vaccines in lymph nodes through direct intranodal administration or through the use of delivery systems of particles that can travel via the lymphatic vessels and reach the lymph nodes is shown as an attractive treatment option [232]. Kumai and colleagues (2017) believe the route of administration of peptide vaccines represents a fundamental factor in regulating the intensity of the immune response. Although subcutaneous vaccination is effective in inducing antibody responses, it is shown that intravenous injection of peptide-based vaccine reports superior results compared to those operating the subcutaneous route. It is also suggested that the systemic administration of peptides delivers antigens to more lymphoid organs and their response facilitates the recruitment of naïve T-cells [216].

DNA vaccines can be delivered intradermally, which leads to the transfection of epidermal keratinocytes and Langerhans cells. The intramuscular delivery of DNA vaccines results in the transfection of myocytes [196, 220]. Generally, DNA vaccines are introduced intradermally or intramuscularly with most of the vectors that usually end up in the extracellular space. However, many cells at the site of injections are inefficient in uptake of injected DNA vectors, which results in low transfection efficiency. Faced with these deficiencies, strategies like administration with a gene gun (intradermal) and electroporation (intramuscular) have been designed. With regard to the first strategy, the plasmid DNA is coated in heavy metal nanoparticles, usually made of gold and then bombarded between the keratinocytes with the help of compressed helium as an acceleration force. The strategy results in the direct introduction of the antigen between the immature DCs. In the second strategy, the uptake of plasmid DNA in muscle cells is considerably increased by the application of brief electrical pulses that temporarily permeabilize the cell membrane. Electroporation equally serves as a form of adjuvants when it damages the site of application, which leads to inflammation and consequent release of cytokines to finally recruit APCs, such as DCs and macrophages [220]. However, it is unguaranteed that the immunogenicity of the vaccine increases and does not always prove to be viable for many human vaccines. In the case of the DNA vaccine, candidate methods are proposed for more efficient deliveries, among which there are transfer systems in needle-free skin or intrapulmonary or intranasal transference systems. For the first method, fine high-pressure flows are used to focus distinct depths of the skin, which allows the vaccines to be able to transfect the Langerhans cells in the skin. In theory with the second method, DNA vaccination should expose delivered DNA and expressed antigens to a broad surface of epithelial and immune cells, while avoiding the need for needle injection. The most significant challenge in pulmonary delivery corresponds to the need to have specialized delivery methods, and consequently, strategies like the nebulization of surface acoustic waves in this field have been innovated [219].

### 2.7.7 Future Vision

In spite of the vertiginous advance that cancer immunotherapy has evidenced at the moment, several issues persist that imply challenges in constant development. Some of these challenges attempt to establish how to optimize cancer vaccines in terms of antigen selection, antigen form, vaccine adjuvants, combined options [194], and the ability to elucidate the frequency and appropriate intervals in vaccine therapy to ensure its effectiveness [216]. One of the most enormous obstacles experiencing the effective development of cancer vaccines refers to the central tolerance of the host immune system [220].

Regarding strategies in development, molecular mechanisms are currently being explored to enable virus-host integration events, which can lead to the evolution of innovative treatment methods, including therapeutic vaccines to combat the diseases that develop from the persistent infection caused by DNA viruses, similar to

those reported in this article. These include therapeutic vaccines to fight diseases that develop from persistent infection caused by DNA viruses, similar to those reported in this chapter. One approach that is being explored is the differentiation of DCs from human pluripotent stem cells including induced pluripotent stem cells and embryonic stem cells. This innovative approach to DCs possesses potential for large-scale production through the employment of bioreactors, unlike the methods currently used, which depend largely on operational quality standards. This technology offers the potential to improve DC vaccination by establishing an unlimited source of DCs, with the prominent feature of being to transduce the antigen directly, thereby ensuring presentation on MHC and stimulating CD8+ T-cell responses [213]. Cancer immunotherapies that integrate subgroups of primary DCs derived from blood are also being actively studied. Dendritic cells derived from blood do not depend upon prolonged periods of ex-vivo culture and differ biologically from DCs derived from cultured monocytes. It is suggested that in patients, plasmacytoid DC-based vaccines promote more durable antitumor T-cells responses compared to homologous vaccines composed of myeloid DCs. Natural Killer cells also emerging as essential factors that contribute to antitumor immune responses and their activation can be focused by vaccines based on DCs [244]. A promising development of DC immunotherapy is the generation of off-the-shelf vaccines. To avoid the laborious preparation and variable quality associated with ex-vivo prepared cell-based vaccines, the DCs can be directly targeted on triggering their maturation [241].

With regard to approach to personalized cancer vaccines, Hu and colleagues (208) propose four areas in which improvements are expected in the near future, these are: (i) improvements in the MHC/HLA binding algorithms, so as to increase the probability of targeting on neoantigens that are expressed by cancer cells of a patient. This should allow providing more precise algorithms to enhance the prediction of connections to MHC class-II; (ii) prevent immune escape through the development of combination therapies that include personalized vaccines focusing multiple neoepitopes and complementary therapies to reverse immune suppression in the tumor microenvironment; (iii) optimizing the dosage, programming and administering personalized cancer vaccine routes through the development of preclinical models, optimizing immune adjuvants and establishing new approaches for the delivery of the vaccine where the nanoparticles technology takes place; (iv) modifications in the assembly of personalized vaccines that allow to incorporate drastic changes to the standard manufacturing practices, and that in turn permit to reduce costs and production times [210]. The identification of both antigenic-shared and neoantigens-unique cancer-specific will represent the key to the development of immunotherapies for many types of cancer [193].

The medium and long-term future around cancer immunotherapy offers multiple alternatives. The combination of vaccines to focus several DC groups, such as plasmacytoid DCs, will represent a valuable strategy to elicit robust and durable responses of CD4+ and CD8+ T-cells. It is also expected that these vaccines will be combined with checkpoint blocking agents [241]. Fecek and colleagues (2016) speculate that vaccines based on optimized DCs will become a fundamental ingredient in the future of the effective combination of immunotherapies. Therapies that incorporate gamma/delta ( $\gamma/\delta$ ) T-cells licensed as APCs in combination with mono-

clonal antibody therapy may represent an effective cancer treatment strategy in the future. The upcoming development of clinical vaccination protocols implementing subgroups of primary DCs or  $\gamma/\delta$  T-cells from the patient's peripheral blood as APCs remains significant promises of treatment [244]. For his part, Yang and colleagues (2014) suggest that DNA vaccines designs have to be strategically optimized to achieve the desired translational efficiency. Effective strategies are needed to help enhance the potency of DNA vaccines and for this purpose, they suggest focusing on improving antitumor immunity by preventing immune tolerance, breaking immunosuppressive networks in the tumor microenvironment and inducing long-term memory. They also suggest focusing on incorporating immunomodulatory molecules either as adjuvants or encoded in DNA plasmids to promote the improvement of immunogenicity against self-antigens [220].

The future glimpsed by Song and colleagues (2018), indicates that research may be focused on achieving a more appropriate combination of different personalized vaccines against cancer and other therapeutic methods for distinct types of cancer patients and several diseases stages of these same ones patients [196]. The fact that a broad range of viruses and various viral components activate caspase-1, IL-1 $\beta$  and IL-18, demonstrate the potential role of inflammasomes in the viral immune response [54]. This evidence establishes them as potential targets in the therapeutic design that attempts to implement mechanisms of activation or blocking of downstream signaling pathways. It is expected that in the near future, vaccines will be designed based on combination of tumor antigens (patient-specific) and adjuvants co-delivered by functionalized nanomaterials [241].

## 2.8 Artificial Immune System

The biological immune system is considered a complex system and its capabilities have allowed the construction of automated systems that simulate in some way several of its properties, and this has generated a field of research that has adopted the name of artificial immune system (AIS) [4]. In the artificial world, the immune system is frequently seen as a protection against infectious agents such as: viruses, bacteria, fungi and parasites [24].

Characteristics like robustness, adaptability, diversity, scalability, multiple interactions on a variety of times scales, represent some of the properties of the biological immune system, which the artificial immune system has wished to possess. The artificial immune system has tried capturing the dynamics of the biological immune system [246]. Based on this, AIS has allowed the construction of algorithms inspired by immunity applied to problems such as: robotic control [247], intrusion detection in networks [248], fault tolerance [249], bioinformatics [250], machine learning [251], and computer viruses control [252], among other several bio-inspired algorithms.

The main AIS developments have focused on three most relevant theories of immunology: clonal selection, immune networks, and negative selection. These

theories have stimulated the development of works oriented to the mechanisms of learning and memory of the immune system for the generation of detectors and classifiers. Various optimization algorithms, pattern recognitions, computer viruses control [252], static and dynamic learning have been inspired by clonal selection [253]. Design of hierarchical clustering tools, fuzzy systems methods, performance algorithms and predictions have been inspired by immune networks [254]. Many of the intrusion detection algorithms, security mechanisms and antivirus control [255] have been inspired by negative selection among several other approaches.

With an interdisciplinary (biology, mathematics and computing) approach and using modeling inspired by immunology, our interest focuses on the construction of a robust artificial immune system, which additionally allows us to comprehend the underlying complexity inherent in immune system of human beings. AIS is not a rival to its natural counterpart since both exhibit the similar level of complexity or perform the same function, but AIS does not capture essential properties of the biological immune system, which makes it a paradigm of competitive computational intelligence [4]. However, the approximations made in this area of research allow a more precise understanding of the relationship between the immune system and disease and in this case particularly, a broader knowledge of the immune response to infections caused by DNA viruses that are integrated the guest.

## 2.9 *Artificial Life*

Conforming to Liò and colleagues (2015), Artificial Life (AL) represent an interdisciplinary research field that maintains the purpose to achieve a more proper understanding of biological life by artificially synthesizing novel and simple forms of life, as well as reproducing lifelike properties of living systems [256]. Several approaches have been referred to describe artificial life. Initially, AL was described as “the study of man-made systems that exhibit behaviors characteristic of living systems.” [257]. Langton (1989) summarizes the essence of artificial life in the following characteristics: life-like behavior, parts of man-made systems, semi-autonomous entities whose local interactions with others are governed by a set of simple rules, populations rather than individuals, simpler than complex specifications, local rather than global control, bottom-up rather than top-down modeling, emergent rather than predefined behaviors [4, 257]. A more synthetic approach than reductionist was presented in 1994, and it was intended to describe an artificial life from that exhibited natural evolution. Under this approach, the concept of artificial life is presented as an initiative for the understanding of biology, by means of the construction of biological phenomena from artificial components and not as a way of separating the natural life forms among the parts that compose it [258]. Artificial life (ALife) was initially motivated by the need to design biological systems, which brought with it the demand for computing. Representing a multidisciplinary field, ALife involves a variety of topics from various disciplines that include essential elements of biological life and artificial life, origins of life and self-organization,

evolutionary dynamics, replication and development of processes, learning and evolution, emergency, computing and living systems, and simulations systems to investigate artificial life, among many other approaches [4].

Currently, research on artificial life may be classified into 13 subjects that include the origin of life, autonomy, self-organization, adaptation (evolution, development and learning), ecology, artificial societies, behavior, computational biology, artificial chemistry, information, living technology, art and philosophy [259]. The artificial life has motivated the construction of many computational models based on agents of biological systems, and precisely the agent-based modeling as an artificial life mechanism, has been applied to the biological understanding of these systems. The artificial life represents an approach to explore biological systems that try to infer mechanisms from the biological phenomenon adding the elaboration, refinement and generalization of their machinery, to identify dynamic properties of these systems. The essential characteristics of an ALife program include: (i) a population of diverse organisms or individuals, on which their characteristics, behavior, resources in time and space can vary; (ii) interactions that require detection of the local or neighborhood scenario, interactions with the individual and interactions with the environment; (iii) sustain and renewal that demands the acquisition of resources that may be provided by the environment or other agents; (iv) self-reproduction and replacement, where the organism can be transformed through changes in its attributes and behaviors, which may occurs through the introduction of new organisms and substitutions [4].

The field of ALife is intimately connected to agent-based modeling (ABM). Consequently, ABM modeling has grown around the need to model the essentials of artificial life, and in turn many of the aspects of ALife have been incorporated into the development of agent-based models.

### 2.9.1 Agent-Based Models

The most generalized work environment for the modeling and simulation of complex systems are agent-based models (ABMs) [6]. The agent-based modeling may be used to test theories of the mechanics underlying the interactions among components within the systems and their resulting dynamics. It can also facilitate the mapping of micro-to-macro systems by relaxing assumptions or altering the mechanics of interactions at the individual agent level, to investigate emerging behavior at the system level. ABMs become an instrument of scientific research and therefore a principled approach to their use is required, which is similar to the scientific method, to ensure these fit their purposes. ABMs represent a potent tool to increase our understanding of the mechanical behaviors of complex dynamic systems, and it is suggested that a warning should be announced when the results of the simulations are used with a predictive capability [260].

Agent-based modeling focuses on the rules and interactions among the independent components (agents) of a system, generating populations of those components and simulating their interactions in a virtual world to create an *in-silico* experimental

model. These models are spatial, use parallelism, incorporate stochasticity, reproduce emergent properties, and because of their object-oriented nature, it facilitates the mapping of the expression of current biomedical knowledge. Consequently, researchers may instantiate ideas of experiments in an *in-silico* environment, to test the veracity and validity of their conceptual models by comparing the simulated experiments against the results obtained *in-vivo* and *in-vitro* [261]. ABMs are also stochastic models that reveal unique dynamics from extremely specific spatial configurations or from rare localized events, which could be omitted with other approaches. Particularly, this characteristic turns out to be valuable in the results observed from infectious processes on different individuals, in various point of time [262].

An ABM consists of a number of agents that interact with each other over time and space, usually exchanging messages. An agent is an encapsulated computer system located in some medium, which continuously collects information and reacts to the changes that occur in its environment. An agent is equally capable of executing autonomous actions on behalf of its owner and should be competent to discover for itself what it must do to fulfill its design objectives [263]. Coinciding to Meyer and colleagues (2009), the characteristics of the agents can be summarized as follows: (i) the agents are identifiable as independent individuals who have a set of characteristics and rules that govern their behavior; (ii) they are autonomous and can operate independently of their environment and in their interactions with other agents, at least in some situations of interest; (iii) an agent has the ability to recognize and distinguish the traits of other agents, include protocols to interact with other agents and possess the capability to respond to the environment; (iv) it can be directed to a goal, establishing purposes to be achieved in relation to its behavior; (v) an agent may have the ability to learn and adapt his conduct based on his experiences and may incorporate rules that modify his behavior over time [4].

According to Helbing and colleagues (2012), ABMs are classified into physical, economic and sociological models. One of the objectives proposed in simulations of biological systems modeled by ABM is the generation of population of the components of the simulated systems and their interactions, trying to move from a virtual world to create an *in-silico* experimental model [264]. Among the characteristics presented by An and colleagues (2009), the following properties of ABMs applied to biological systems are highlighted: (i) the spatial nature of ABMs supports modeling of agents with known limitations, introduced by locality rules that determine the near environment. The emphasis of behavior leads to local interactions, which closely coincide with the mechanisms of stimulus and response observed in biology; (ii) each class of agents produces multiple manifestations, like a computational object representing a population of agents that interact in a parallel processing environment. Many local conditions lead to diverse trajectories of the behavior of independent agents; (iii) explicitly, biological systems include behaviors that may be randomized at the observational level, but can be totally deterministic from a mathematical point of view; (iv) ABMs can incorporate a modular structure, where new information can be added through new types of agents or by modifying current rules, without the need to redesign the complete simulation; (v) they reproduce emergent properties because of the parallelism,

intrinsic stochasticity and rules of locally limited agents, which allows them to generate a systemic dynamic that could not be inferred from the examination of the rules of the particular agents; (vi) they can be constructed in the absence of complete knowledge [265].

The modeling of a population of heterogeneous agents with a set of diverse characteristics represents a mark that distinguishes an ABM. The agent perspective is unique among simulation techniques, unlike the processes perspective or the variable-state strategy taken by other approaches [4, 265]. ABM correspond a bottom-up approach to model and investigate complex systems, to explicitly represent the behavior of considerable numbers of agents and the processes by which they interact. These essential characteristics are all necessary to produce in the end rudimentary forms of emerging behavior at the system level [4, 265, 266].

One of the motivations of agent-based modeling is to explore the emerging behavior exhibited by the simulated system. ABM frequently exhibits patterns based on the interaction between system agents [4, 262, 264]. A prominent aspect to use ABMs as an integral modeling structure, which moves towards the goal of representing a dynamic knowledge, is the ease that is achieved by transferring conceptual models supported by biomedical research to executable forms. ABMs have the advantage of mapping well the means by which biomedical knowledge is currently expressed and is generally more intuitive for computer scientists and non-mathematicians [265]. According to North and colleagues (2009), the objective of agent-based modeling is to allow experimentation with simulated complex systems. These models may show how complex adaptive systems can evolve over time in a difficult way to predict based on the knowledge of the independent behavior of agents. The rules of the agent are frequently based on theories of the individual and based on these elementary rules, this type of modeling can be used to examine the emergence of patterns, although they are unobvious before the specific examination of the rules. Modeling and simulation based on agents provide a natural frame of reference in which it is possible to execute artificial life experiments [267]. The ABM highly intuitive approach may be able to reproduce various studied behaviors. Because of the complex nature of interactions among agents, the analysis of ABMs may not provide an understanding of the underlying mechanisms. Nevertheless, an ABM can be disturbed through manipulations of the agent's repertoire of local rules to observe how their results change [263].

### 2.9.2 Artificial Life, Biological Systems and Disease

The use of mathematical and computational models inspired by biological processes allows discovering emergent properties, examining the behavior of the system, and generating new hypotheses. Such models allow performing in-silico experiments that could be extremely expensive or impossible to produce in a laboratory [263]. The work of biologists focuses predominantly on comprehends the system to generate predictions and infer manipulations that lead to the desired changes. Conforming to Pezzulo and colleagues (2016), these objectives have motivated the

use of bottom-up modeling, where the behavior of the molecular components and their local interactions become the focus. Based on this approach, specific process models of probabilistic and inferential calculations can be provided. However, a number of biological systems employ extraordinarily diverse underlying molecular mechanisms to achieve the same high-level target state, suggesting the focus on the global state represent a homeostatic target for cellular and molecular activities. For this reason, the top-down modeling is also proposed, which allows us to focus on the broad states of the system as causal actors in models and in the computational principles that govern the dynamics of global systems. Control models that adopt this approach offer a valuable complement because they provide a mechanical road-map for the explanation and inspection of some complex systems. Under this approach a normative principle describing the collective behavior of the system is provided, and its function is governed by optimization. This approach offers an intuitive point of view for the control of extremely complex results to implement directly. Additionally, there are models that uses both approaches, among which are models of dynamic systems in which trajectories emerge seeking to reach an attractor, stable state or limit of the cycle. While the attractor represents a concept of the bottom-up approach, the use of the concept of attractor as causal factor in networks is currently being explored in cancer and synthetic biology applications [268]. In principle all diseases are complex, and under this concept Loscalzo and colleagues (2011) emphasize the importance of reconsidering and redefining the determinants of human disease. They define the disease as the result of the exit of a complex modular network of nodes, mechanically linked to certain underlying pathophenotypes. This study presents a contemporary approach to the disease, pointing out that it needs to be seen from the perspective of a system. The knowledge of two inter-related categories within a cell or organism, are essential for the understanding of the determinants in the expression of a disease, these are: molecular and phenotypic networks. The first category includes interaction networks at the protein level, metabolic and regulatory networks, which incorporate the networks of transcription factors and non-canonical RNA networks. The second category includes networks of co-expression, in which genes are linked when similar expression patterns are exhibited in various diseases; and genetic networks, in which genes are and define a phenotype that differs from any isolated gene [269].

### 2.9.3 Immune System Simulations

Various efforts have been made in the course of time to interpret the immune system. Many of these works are based on the construction of models that use ABMs and CAs (cellular automata). Some of the most outstanding works in this field are described below.

Seiden and colleagues (1992) present an immune system model, based on a generalized CA that focuses on the phenotype of a clone of cells, establishes some interactions with other cells as well as with antigens and antibodies. This model is known as IMMSIM (*immune system simulator*) [270]. Celada and colleagues (1996)

intend reproducing the affinity maturation of the antibody response in a CA model based on binary chains [271]. Meier-Schellersheim and colleague (1999) present a model named SIMMUNE, with which they try to investigate how the adaptive behavioral context of the immune system can emerge from cell-cell and cell-molecule interactions. SIMMUNE model is based on rules that are defined by the user and focuses on performing molecular interactions and cellular responses to certain stimuli [272]. Bernaschi and colleagues (2001) develop a parallel version of the IMMSIM automaton, and they titled it PARIMM, which belong to the class of bit string models. In this case, the bit strings represent the binding sites of cells and molecules. PARIMM model considers the representation of entities, repertoire, functions of affinity and hypermutation, and interactions among entities [273]. Puzone and colleagues (2002) present a modified CA based on the initial IMMSIM model to simulate the changes found and the discrete effects of cell-cell and cell-molecule interactions on the lymphoid system. The authors consider as a problem, the abundant number of parameters that must be configured to achieve a satisfactory fit, which may cause an incorrect adjustment of the CA [274]. Pappalardo and colleagues (2005) present an *in-silico* model that simulates the immune system responses produced from tumor cells in unvaccinated and vaccinated mice, named SimTriplex [275]. The Simtriplex model is based on a modification to the reference framework used by Seiden and colleagues [270], and also on the generalized versions of CAs proposed by Castiglione and colleagues (1997) [276] and Bernaschi and colleagues (2002) [277]. Bandini and colleagues (2006) describe and analyze several mechanisms that support field diffusion (that is, fields that can propagate through space, influencing several cellular agents), with reference to computational cost in terms of memory occupancy of required structures and time complexity of the related algorithms. This model provides the possibility of defining heterogeneous entities in regular and irregular spatial infrastructures, having as antecedent that the remote action performs a tremendously important role in the management of this action at a distance that has been shown previously as a weakness of the discrete models based on CAs [278]. Baldazzi and colleagues (2006) present a class of discrete models based on stochastic CAs, named C-IMMSIM, which tries to simulate the immune system's response to highly active antiretroviral therapy in patients infected with HIV-1 [279]. Castiglione and colleagues (2007) present an adaptation of the C-IMMSIM model of an immune response to a generic pathogen, with the aim of investigating, and validation the proposed dynamics of the persistent infection caused by Epstein-Barr Virus (EBV). Castiglione and colleagues (2007) present an adaptation of the C-IMMSIM model of an immune response to a generic pathogen, with the goal of investigating and evaluating the proposed dynamic corresponding to the persistent infection caused by Epstein-Barr Virus (EBV) [280]. Mata and colleagues (2007) describe the functionality of an immune system simulator based on CAs, named SIS (*synthetic immune system*). SIS is founded on cellular and descriptive rules that define the transition between these states. SIS proposes a model to observe how the immune system responds to its self-antigens and foreign-antigens based on rules defined by the user [281]. Maeda and colleagues (2007) presents method for identifying CA rules. The collected evidence is then classified

using a decision tree constructed based on genetic programming. The generated tree is used to construct the CA transition rules [282]. Folcik and colleagues (2007) design an agents-based model to explore the interactions between the cells of the innate and adaptive immune system, named BIS (*basic immune simulator*). This model simulates type of primary cells, mediators and antibodies that are incorporated in the three virtual spaces that represent parenchymal tissue, secondary lymphatic tissue and lymphatic/humoral circulation [283]. Dreau and colleagues (2009) present an ABM to simulate the development and progression of solid tumors, including the influences of the tumor, the immune response of the host and the level of vascularization of the tumor [284]. This model corresponds to a modified version of the work previously presented by De Pillis and colleagues (2006), a hybrid model (PDE-CA) that integrates a cellular automaton (CA) and partial differential equations (PDE) [285]. Sneddon and colleagues (2011) present the NFsim (Network-Free stochastic simulator) model, a general-purpose platform that overcomes the combinatorial nature of molecular interactions. This model generalizes a kinetic Monte Carlo method based on agents, which represents each molecule as a different software object or agent and uses rules to describe the reactions, specified in terms of general patterns. The model is intended to follow only the molecular configurations that exist at a given time and not to trace all possible chemical species [286]. Wendelsdorf and colleagues (2012) construct a high-performance computing-oriented agent-based model to investigate mucosal immune responses in the gastrointestinal tract, termed ENISI (*Enteric Immunity Simulator*) [287]. This model is created as an extension of EpiSimdemics (2008), a parallel algorithm to simulate the spread of infectious diseases in extensive realistic social networks, whose main objective is to explore the effects of complex pharmaceutical and non-pharmaceutical interventions on the processes within the real populations affected [288]. Kim and colleagues (2012) implement a hybrid model consisting of an ABM that typifies the site of the tumor and a system of differential equations of delay (DDE) representing the lymph node. This model ABM-DDE considers the initial stimulation of the immune response and the resulting immune attack on the tumor mass [289]. This model is based on the previous work of Mallet and colleagues (2006), who propose a hybrid CA-PDE model that describes the interactions between a growing tumor that follows a source of nutrition and the immune system of the host organism [290]. Pappalardo and colleagues (2011) designed an ABM called SimB16, which summarizes the effects produced by both the strategy of specific immunotherapy against the melanoma B16 in mice and the tumor progression in a section of the genetic tissue. This model tries predicting the failures or successes of the treatment [291]. Von Eichborn and colleagues (2013) present an ABM that simulates immune reactions against cancer, using amino acid sequences and potential knowledge-based interactions to predict cellular interactions. This model, termed as VaccIMM, allows the simulation of the effects of a peptide vaccine applied in cancer therapy. This model represents another extension of C-IMMSIM, although with an independent approach [292]. Santos and colleagues (2015) implement a CA model to represent the behavior of cells when the main markers of cancer cells are present in the avascular phase. This model is designed to simulate the development of multicellular

spheroids of tumor cells and not in a particular type of cancer [293]. Shahmoradi and colleagues (2018) investigate the effects of immunotherapy in treatments of avascular tumors, for which they use a CA model to which they add IL-2 as immune therapy injected at the tumor site [294]. This model is based on a previous work presented by Boondirek and colleagues (2006), a CA model that represents the growth of an avascular tumor whose pattern of formation and population expansion are founded on a Monte Carlo simulation [295].

Although all the previous projects have represented outstanding contributions, the models that are related below have been the source of inspiration for many of the works that have been published since 1997. The work of Celada and colleagues (1996) represents one of the first in the area. This tried to define the immune mechanisms in a model, a study which aimed to capture the dynamics of the immune system and carry out *in-silico* experiments [271]. From that moment, several studies on simulators have been presented, which through the definition of rules and interactions have allowed observing some immune reactions. Such studies include the following pioneer models. (i) C-IMMSIM model [279]: this model describes a simulator of the immune response, which shows some results related to antiretroviral therapy in patients infected with HIV. (ii) IMMSIM model [273, 279, 296–299]: this model has been used in applications to represent the affinity maturation and hypermutation of the humoral immune system [271], to evaluate some approaches in the design of vaccines [299], and investigate mechanisms of tolerance by rheumatoid pathological factors [300]. (iii) SIMMUNE model [272]: this model presents the way in which the adaptive behavior of the immunity system may emerge from cell-cell and cell-molecule interactions. This work has been a notable reference in the National Institute of Allergy and Infectious Diseases (NIAID-USA) [301]. (iv) SIS model [281, 302]: based on the description of cell states and transitions among states, this model allows observing some responses of the immune system in relation to the self-antigens and foreign-antigens.

The works described below have in some way contributed to the development of another project on a grander scale. Emerson and colleagues (2007) present IMMUNOGRID, a project founded by the European Union that aspires to establish an infrastructure to simulate the processes of the immune system at the molecular and cellular level, as well as at the organ level. The central simulator of this project was originally derived from the work of Seiden and colleagues (1992) [303], whose model is based on a generalized cellular automaton that focuses on clonotypic cell types and their interactions with other cells, antigens and antibodies, and it known as ImmSim [270], a version that was later modified to create the C-ImmSim model of Baldazzi and colleagues (2006). On this last model, two adaptations implemented by Castiglione and colleagues are then constructed, a modification presented in 1997 that involves distributed parallel processing [276] and another one in 2007 that studies the infection caused by EBV [280]. The definite goal of ImmunoGrid project is to develop a model of the human immune system on a normal scale, that is, a model that captures both the diversity of the immunological repertoire and the legitimate populations of its cellular compo-

nents. About this project, since 2010 has two central simulators; one complements to a revised version of the C-ImmSim model, and the other corresponds to a modified version of the SimTriplex model [304].

Some of these simulators are built to investigate the evolution of particular diseases. (i) HIV infection: Strain and colleagues (2002) modeling the spread of HIV based on known biophysical properties. The constructed CA attempts to make explicit the spatial effects related to the kinetics of viral propagation [305]. (ii) Tuberculosis: Segovia-Juarez and colleagues (2004) develop a model of granuloma formation in the lung. This ABM combines continuous representations of chemokines with discrete agent descriptions of macrophages and T-cells [306]. (iii) Influenza [307]. (iv) Cancer: breast carcinoma [275, 291, 308]; growth and tumor invasion [284, 289, 293, 294, 309, 310]; and cancer therapies [292].

These techniques have equally been used in the simulation of specific infectious diseases, reproducing a variety of host-pathogen interactions. (i) CYCELLS model: Warrender and colleagues (2006) describe the early infection caused by Mycobacterium, using for this purpose Cycells, a hybrid simulator in which the models can include discrete and continuous components as well as deterministic and stochastic dynamics. The cells (for example: T, B, Macrophages, TNF, IL-10 and IFN- $\gamma$ ) in this model are explicitly incorporated, but instead the molecules are represented by their concentrations [311]. (ii) PATHSIM model: Shapiro and colleagues (2008) construct a model based on broadly accepted characteristics of the infection caused by EBV in relation to infection sites and immune system response (including explicitly the virtual populations of B and T cells). Described by its authors as an extreme simplification of reality, this model provides a geometric approximation of Waldeyer's ring together with the abstract compartment representing lymph and peripheral circulation [312]. (iii) MASyV model: Beauchemin and colleagues (2006–2012) design the dynamics of influenza on an *in-vitro* epithelial cell layer, represented in a two-dimensional CA. Instead of treating each virion explicitly, the model considers the concentration of virions by associating variables with values reported from the real world. Local concentrations may adapt according to a discrete version of the diffusion equation with a term of production [307, 313].

ABMs and CAs both considered artificial life methodologies, either in their individual or combined versions, have equally been applied in: advanced computing, parallel computing applications and evolutionary computations, approaches that have equally been applied in simulation of biological processes. Ebeling and colleagues (2001) use the ABMs to identify the extensions that can serve as a bridge to a known dynamic (physical) and then lead it to a more complex dynamic (biological) [314]. Macal and colleagues (2006) provide insights on the appropriate context for using ABMs with respect to more conventional modeling techniques [315]. Baird and colleagues (2012) show that CAs used as models of physical systems may exhibit conserved functions of relevance to the system under study. They present the first tests to constitute the basis of all conserved functions of trivial energies in the general case, and employ this to derive a number of optimizations to reduce time and memory for the discovery non-

trivial functions. In this way, conserved functions can be used to classify CAs and identify connections among seemingly unrelated systems [316]. In the area of biology, these methodologies have been used in the simulation of multicellular biological processes to test scientific hypotheses, to plan experiments and to deduce relationships among properties of complex systems with the aim to recognizing patterns mediated by simple rules of behavior. Such types of models have been used to describe numerous processes of the immune system, expanding knowledge in the understanding of immunology and pathology of diseases. Such type of models have been used to describe several processes of the immune system, expanding in this way the knowledge about the understanding of immunology and pathology of diseases.

### **3 Artificial Life and Complex Systems to Test Therapeutic Vaccines Against Cancers that Originated in Viruses**

Several techniques have been proposed around the creation of complex system models. One of these approaches is presented by Sayama (2015), who considers that this creation process involves the following activities to be carried out. Several techniques have been proposed around the creation of complex systems models. One of these approaches is presented by Sayama H (2015), who considers that this creation process involves the following activities to be carried out: (i) establish the questions that you wish to address; (ii) define components of the system at the microscopic level and establish the dynamic rules of their behavior; (iii) define the structure of the system at the level of interactions between components; (iv) establish the possible states of the system, that is, determine the class of dynamic states that each component can take; (v) establish by means of dynamic rules the way in which the states of the components change over time as a consequence of their mutual interaction [6]. In the process of developing and running simulations to explore complex systems, using agent-based modeling (ABMs), William RA (2018) defines three fundamental stages: discovery, development and exploration phases. In the discovery phase the biological bases of the project are established, the domain of interest is identified and modeled, and the scientific questions are formulated. In this initial phase the domain of the real world is investigated and a domain model is developed, that is, a conceptual model is established. Additionally, this phase includes definitions related to scope, level of abstraction, assumptions and restrictions presented within the conceptual model. These last two aspects allow establishing the context of how to validate, interpret and evaluate the results. In the development phase, the simulator is constructed as such. In this second phase, the ABM is developed and tested. In the exploration phase, the simulator is used to carry out the experimentation stage, with the purpose of obtaining answers to real-world questions about the complex system [260].

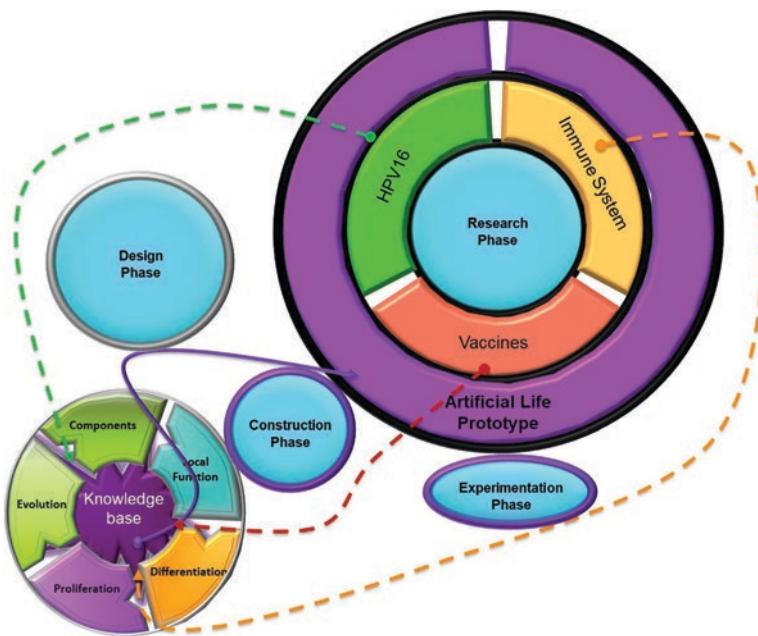
Establishing a conceptual model of the relationship between the immune system and the effects of a vaccine that aims to contain an infectious process caused by a pathogen requires a thorough prior exploration of each of its domains in the real-world, that is: human immune system, virus and therapeutic vaccine under study. It must be also be ensured that whoever defines the conceptual model possesses enough knowledge to understand each component in its own subsystem. The solid knowledge of domains in the real-world allow us to understand and replicate the dynamics of each of its components, establishing how the interactions among components are presented and being clear that some of them can produce to emerging behaviors.

**Creation of an artificial life model** We previously reviewed several of the complexities inherent at the human immune system and life cycle of the DNA viruses that are integrated into the host. Additionally, we consider some challenges involved in the development of therapies against cancers that originate in this class of viruses. Based on this work, and from this point, our interest is focused on producing an artificial life model that allows us to test therapeutic vaccines that try to contain these types of cancers. Comprehending the complexity of the issues involved, and with the aim of presenting a more didactic process, we decided to develop a prototype to evaluate the results of therapeutic vaccines against one of the viruses and their associated diseases studied in this chapter. This is how we decided to work specifically with: innate and adaptive immunity on human beings, human papillomavirus type 16 (HPV16) and therapeutic vaccines. These therapeutic vaccines are based on autologous dendritic cells (DCs) loaded with antigens, with which we attempt to observe their behavior when they additionally use certain checkpoints blockers and/or some adjuvants, within a persistent and cancerous viral infectious environment.

Coinciding with the approaches, both of Sayama H (2015) and William RA (2018), the methodology we use in the construction of our artificial life model involves the following four phases: research, design, construction and experimentation all which we explain below (see Fig. 5).

### 3.1 Research Phase

In the research phase we declare three principal domains that allow us to differentiate the specific study on the following topics: human papillomavirus type-16 life cycle (HPV16), human immune system (HIS), and development of therapeutic vaccines for the control of cancers that originate in this type of virus (TVC). Likewise, we included the study of interactions between HPV16 and HIS domains. Part of the review of these topics is referred to in the previous sections ([2.1](#), [2.2](#), [2.3](#), [2.4](#), [2.5](#), [2.6](#), and [2.7](#)).



**Fig. 5** Methodological approach for the construction of an artificial life model. This figure represents the three domains (HPV16, immune system and vaccines), which are thoroughly researched to build a solid knowledge base that will facilitate the design of a conceptual model and the construction of an artificial life functional prototype. The prototype, in turn, will represent the means to develop own activities of an *in-silico* laboratory

### 3.1.1 HPV16 Domain

During the research phase of the HPV16 domain, our goal is to obtain sufficient knowledge that allows us to recognize this pathogen from several perspectives. Initially, we focused on examining general aspects including issues related to the structure and classification of the virus, life cycle and evolution of Papillomavirus (PVs), risk factors, process of progression to cervical cancer and other associated diseases.

Regarding the HPV configuration, we delve into its genome and molecular structure, reviewing in detail the early (E1, E2, E4, E5, E6 and E7) and late (L1 and L2) proteins, together with the genes p53, p21, Tert and Rb, all of them acting as key components of its life cycle. In addition, we review the classifications currently known of this virus including subtypes, taxonomy and phylogenetic variants of HPV16.

Concerning the evolution of PVs, we included the study of themes related to the regulation of their genetic expression, replication and intracellular growth processes. With particular emphasis, we investigate the behavior of the virus when an immune response is evidenced by the host, in the presence an infectious condition caused by HPV16, including the processes of persistence, regression, progression to

neoplastic intraepithelial lesions and pre-cancer. In addition, we studied the risk factors and deepened in the process of progression to cervical cancer and other related cancers that consider their origin in the infection caused by HPV16.

### 3.1.2 HIS Domain

In the research phase of the HIS domain, we intend to know and understand the components, cycles and patterns that govern it. Based on this, we exhaustively explore several theories and related postulates, levels of understanding of immunology, innate and adaptive immunity, production of cytokines and chemokines, growth factors and transcription factors.

In relation to cellular immunology, we investigate its antecedents, and processes linked to antibodies including: somatic hypermutation, class-switch recombination, affinity maturation, production of antibodies, and properties of human immunoglobulins. In addition, we review the processes linked to antigens, major histocompatibility complex (MHC/HLA), adhesion molecules, and especially the procedures related to the initiation of the human immune response under both normal and defense conditions. Under this same context, we examine the processes of differentiation, proliferation and cellular apoptosis linked to the two common progenitor groups (myeloid and lymphoid) that are part of the immune system. Regarding the cellular groups linked to myeloid progenitors, we studied the groups of macrophages and dendritic cells. Concerning the lymphoid progenitors, we reviewed the populations of B-cells and T-cells.

Regarding B-cells, we study the distinct stages and locations in which their development takes place; that is, B-cells that evolve in bone marrow, the spleen and the lymph nodes. The B-cells in bone marrow include: stem cells, pro-B and pre-B cells, and immature B-cells. The B-cells in spleen include: transitional B-cells (T1 and T2), marginal zone B-cells (MZ), follicular B-cells and germinal centre B-cells (GC). The B-cells in lymph nodes include: memory B-cells and plasma cells (PC). In addition, we examined the role of B-cells in adaptive immunity, the population of cytokines that act on B-cells and the TLR molecules and receptors associated with this cell population. We equally considered the processes linked to the behavior of the B-cell receptor (BCR) signaling pathways and associated with the presentation of antigens. Additionally, we investigate the mechanisms that demonstrate behaviors in the division processes, both stochastic and asymmetric within these cell populations.

By evaluating the T-cell population, we review the associated development and differentiation processes, through which various cell populations are derived. These include; NK-cells, CD4 T-helper cells and CD8 T-cells. In turn, the CD4 T-helper cells can be differentiated into the following cellular populations: Th1 (T helper type-1), Th2, Th9, Th17, Th22, Tfh (follicular helper T), and Treg (regulatory T) cells. Likewise, the CD8 T-cells can be differentiated into memory CD8 T-cells and cytotoxic T-lymphocytes (CTLs). We exhaustively investigate the stages of develop-

ment of these populations during the infectious processes caused by DNA viruses, including the following topics: the role of T-cells in adaptive immunity, population of cytokines acting on T-cells, molecules and TLRs associated with T-cells and behavior of signaling pathways of the T-cell receptor (TCR). With particular attention at this point, we study the processes of plasticity and flexibility. Additionally, we studied the interaction processes that arise between populations of B-cells and T-cells.

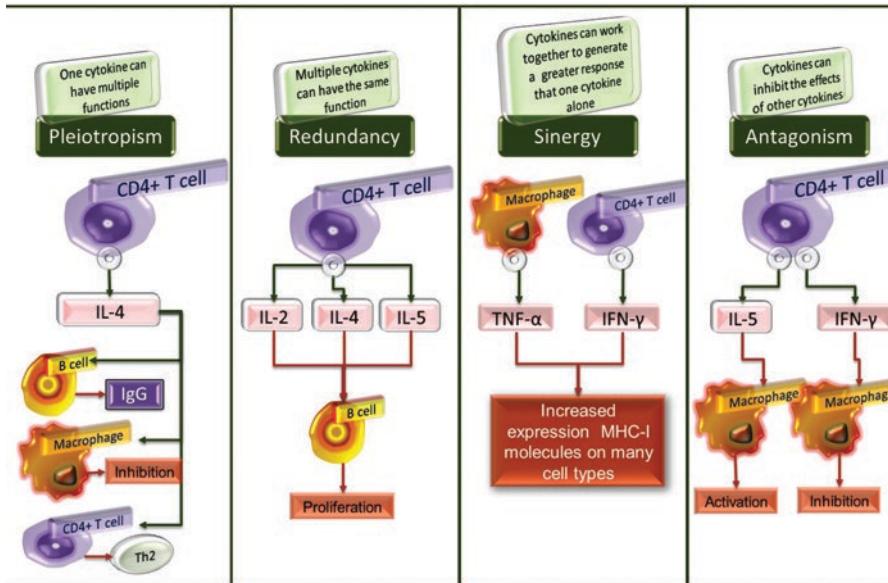
As part of the research phase of the HIS domain, we equally treat with considerable interest the biological behavior of the Toll-like receptors, where we analyze their antecedents, structure, classification and signaling pathways. We examined the behavior of some of its components, considering 13 Toll-like receptors different, included from TLR1 to TLR13; and six families of adaptors, including: MyD88, TIRAP (MAL), TRIF, TRAM, SARM and BCAP (PIK3API). Regarding the TLR signaling pathways, we concentrated on investigating the following components: IRAK complex, TRAF family, TK1 and TAB, MAPK complex, IKKs, NF-KB, CREB and AP-1, IRFs, JNKs and p38; and also, other involved molecules, such as: TOLLIP, PELLINO, PI3K, ECSIT, SRC family and IAPS. Moreover, we analyze the TLR life cycle, and the negative regulation processes linked to the TLR signaling.

Another essential component in the research phase of the HIS domain corresponds to the cytokine populations. Within this component we evaluate the biological behavior of several of its members. These include: TNF family, with its alpha and beta members; TGF family, with its alpha and beta members; Interferon family, with its members: alpha, beta, kappa, gamma and lambda; MIF; Interleukin family, with its members: from IL-1 to IL-27, IL-28A, IL-28B, IL-29, and from IL-30 to IL-39. Likewise, we review the interleukin classification patterns which allow us to identify some cytokines that involve pro-cancer, anti-cancer behavior or both; as well as, the components of the immune system that possess the ability to activate cytokine inhibitory functions (see Fig. 6).

### 3.1.3 Interactions Between HPV16 and HIS Domains

Once fulfilled the objectives we proposed in relation to the essential understanding of both the virus under study and the immune system of its human host, we proceed to examine the processes that arise around its integration with the aim of identifying mechanisms of interaction, response cycles and feedback patterns. In this context, we concentrate our work in two fundamental objectives: to recognize in detail the interactions between the immune system and HPV16 (HIS-HPV16), and the actions that connect TLRs, Cytokines and HPV16.

To deepen the HIS-HPV16 interactions, we investigated the following topics: innate immunity, adaptive immunity, immune evasion, response of antibodies induced by HPV, MHC/HLA molecules and cytokines affected by the infectious process caused by HPV. Under the context of HIS-HPV16 interaction and viral infection processes we evaluated the exchange mechanisms that arise between HPV and the immune system, cancer and immune response, and behavior of immune



**Fig. 6** Properties of cytokines. This figure represents the processes associated with four key properties of cytokines (pleiotropism, redundancy, synergy and antagonism), which can modify the behavior of certain cell populations

cells against HPV. Within this last component, we included the analysis of the major histocompatibility complex and the following cell populations: keratinocytes, macrophages, dendritic cells, T-cells, cytotoxic T-cells, NK cells, NKT cells, B-cells and Treg cells.

To reinforce our understanding of the interactions that arise between TLRs, cytokines and HPV16, we focus on the following topics: Toll-like receptors (TLRs), signaling pathways that are affected by the infectious process HPV16, cytokines, negative regulation of TLR signaling, exchange relationship between HPV16-TLRs-Cytokines, exchange correlation between HPV16-TLRs and cancer. In relation to TLRs and their signaling pathways, we examined the behavior of NF- $\kappa$ B, IRFs and IAPs with particular detail when these are affected by the infectious process caused by HPV16. Regarding cytokines, we analyzed in detail the behavior of TNF (TNF- $\alpha$ , TNF- $\beta$ ) and TGF (TGF- $\alpha$ , TGF- $\beta$ ) against the HPV16 infectious process. Within this component we considered the relationship between interferons, HPV16 and cancer, focusing on: IFN- $\alpha$ , IFN- $\beta$ , IFN- $\kappa$ - IFN- $\gamma$  and IFN- $\lambda$ ; and also the interactions that arise between MIF, HPV16 and cancer. Additionally, under this environment we review in detail the behavior of following interleukins: IL-1, IL-2, IL-3, IL-4, IL-5, IL-6, IL-7, IL-8, IL-9, IL-10, IL-11, IL-12, IL-13, IL-14, IL-15, IL-16, IL-17, IL-18, IL-19, IL-20, IL-21, IL-22, IL-23, IL-24, IL-25, IL-26, IL-27,

IL-28A, IL-28B, IL-29 (IFN- $\lambda$ ), IL-30, IL-31, IL-32, IL-33, IL-34, IL-35, IL-36, IL-37, IL-38 and IL-39.

Within the cytokine component, we also delve into the changes that arise in cytokines due to the HPV16 infectious process, for which we include the study of the following integration processes: relations between cytokines and HPV16 where we examine their interactions with the early proteins of the virus (E2, E5, E6 and E7), the activity between TNFs and HPV16, and the activity between IFNs and HPV16. We also studied the relationships between: interleukins and HPV16; cytokines, HPV16 and CIN stages; biomarkers and HPV; cytokines in a tumor environment; and, cytokines and vaccines. Faced with the interactions between HPV16, TLRs and cytokines, we delve into the following topics: immunity of the HPV host; the functions and expressions of TLRs (TLR1-TLR9); TLRs and cytokines in HPV infections; cervical carcinogenesis; and, HPV evasion strategies.

### 3.1.4 TVC Domain

In the research phase of the therapeutic vaccines domain (TVC), we focus on two contexts. A first context, which allows us to recognize the relationships that exist between cytokines, HPV16, cancer and vaccines; and the second context, under which we try to identify the relationships between TLRs, HPV16, cancer and vaccines.

Under the first context, we focus on six key interaction processes within the TVC domain: (i) signaling pathways: we analyze the behavior of the interactions between HPV16 and some key components that are part of the signaling cascade, such as: NF- $\kappa$ B, IRFs and IAPs; (ii) interactions between cytokines and HPV16; (iii) interactions between interleukins, HPV16 and cancer; (iv) correlation between cytokines and cancer prognosis; (v) behavior of cytokines in a tumor microenvironment; (vi) relationship between cytokines and vaccines.

Belonging to the second context, we analyze four processes that involve the recognition of their behaviors and interactions. These include: (i) negative regulation of TLR signaling; (ii) correlation between TLRs, HPV16 and cytokines: under this component we analyze the HPV host immunity processes, functions and expressions of each of the TLRs, behavior of the TLRs and cytokines during the infectious processes caused by HPV, cervical carcinogenesis and evasion strategies of HPV; (iii) correlation between TLRs, HPV16 and vaccines: under this component we analyze regulation of adaptive responses, clinical studies of TLRs, potential therapeutic targets in TLR signaling, TLR agonists in viral infections, TLR agonists in cancer, TLR agonists in diseases associated with HPV, TLR antagonists, adjuvant activity, and immunotherapy directed at HPV infection and cervical lesions where we examine the development of prophylactic vaccines and therapeutic vaccines; (iv) correlation between TLRs, HPV16 and cancer: focused on these interactions, we reviewed innate immunity, adaptive immunity, T-cell immunity, antibody-dependent

cell-mediated cytotoxicity, TLRs (TLR3, TLR4, TLR7, TLR8 and TLR9) as cancer-positive regulators, autoimmunity, double role of TLRs in cancer (pro- and anti-tumorigenic TLRs), immunosuppression and cancer, immunotherapy and cancer, and then immune markers and cervical cancer.

### **3.1.5 Contributions Obtained in the Research Phase**

The research phase of the HPV16 domain allowed us to know the dynamics of the virus, the injuries they cause in the human and the impact that its infectious process has on the world society. In our research, this component allowed us to define the elements necessary to develop the conceptual model of the HPV16 life cycle, as a fundamental basis to construct its prototype and later integrate it with the other two domains.

Based on the research phase of the HIS domain we obtained the necessary elements to clearly define the different cell populations, the control points, the states, transitions and interactions, which we must incorporate into the conceptual model. Some of the issues that this scope involves consider the processes of differentiation, proliferation and cellular apoptosis as well as activities associated with Toll-like receptors, production of cytokines and patterns linked to the downstream and upstream signaling pathways. This conceptual model becomes the basis to develop an HIS prototype that can interact with the HPV16 prototype, with which it will be possible to produce a virtual environment to simulate the dynamics of the human immune system. In succession, the combination of these two HPV16-HIS domains will provide the basis for integration of the third domain.

The research phase of the TVC domain allowed us to recognize the challenges experiencing the development of therapeutic vaccines that attempt to control cancers that originate in this virus and to get clarity in the elements that we need to define their conceptual model. Once TVC is integrated into the other two domains, the virtual environment that will produce will allow us to test therapeutic vaccines that are based on autologous dendritic cells loaded with E6 and E7 antigens. It will additionally include the possibility of incorporating different checkpoint blockers (PD-1, PD-L1, CTLA-4), and support the alternative of boosting the vaccine with diverse adjuvants on the group of cytokines and interleukins that will be integrated into the model.

## **3.2 Design Phase**

As soon as the conceptual model for the HPV16 domain was established, we proceeded to design and develop a functional prototype of the life cycle of this virus. Its antecedents, design and technical details are published in our article [317]. This is the reason why, from now on, we will concentrate on explaining aspects related to the HIS and TVC domains.

Based on the results obtained in the research phase of the HIS, HPV16 and TVC domains and in addition to the recognition of the interactions between HPV16-HIS domains, at this point we possess the necessary information to establish the general design of the proposed model. This general design includes the specifications related to: techniques used in the development of the proposed model, details of the conceptual model and preliminary considerations. Likewise, aspects related to the activation procedures of checkpoints, states and transitions are incorporated. In the final instance of this phase, we conclude with an exhaustive description of the logical model.

### **3.2.1 Techniques Used in the Development of the Proposed Model**

The immune system represents a complex system that is comprised by different groups of cells and molecules that involve a considerable number of tasks. The cells of the physical world represent a complex universe, with thousands of biochemical reactions that are part of a network in which they are connected and regulated by signals produced in the environment. With the intention of predicting the subsequent actions in these networks it is necessary to identify all the cellular and molecular components, understand where they are located, with whom they interact and how their activity is regulated. To facilitate their understanding, we employ dynamic models based on the definition of the parts (research phase) and the knowledge of the biological processes that govern them, until completing the context of the unified system (proposed conceptual model). The understanding of the biological processes and the interactions that arise among the three proposed domains, represent the raw material for the construction of our model, in addition to the knowledge contributed by the complex systems and the techniques inherent in the design of artificial life.

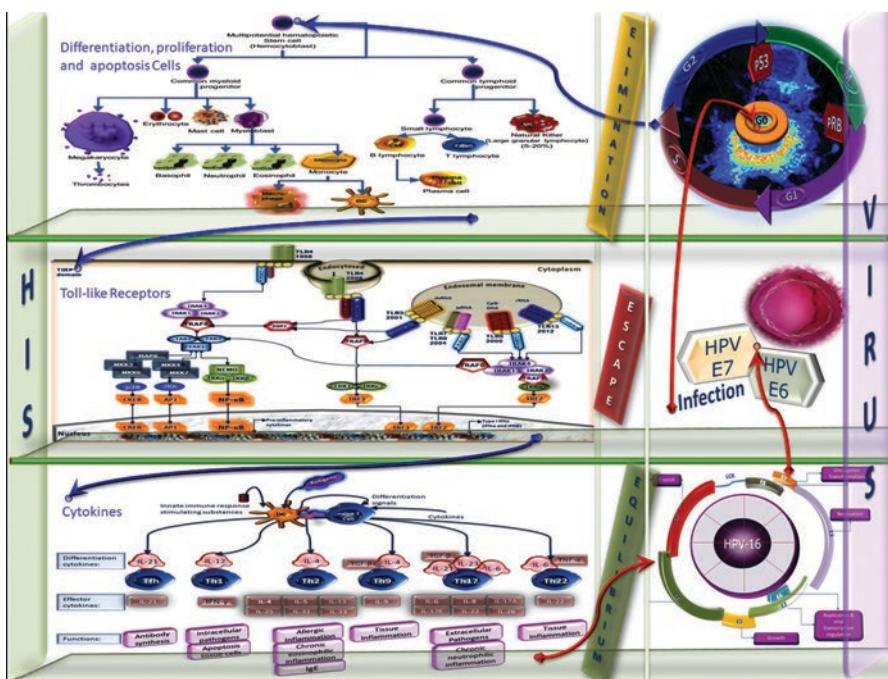
Artificial life is associated with the emergence of order in nature, and modeling based on agents is related to the exploration and understanding of the processes that lead to the emergence of patterns and behaviors through computational means. Regarding the design that involves these concepts, our work considers use of both approaches: bottom-up and top-down. The modeling of agents in the proposed model is built with a “bottom-up” approach, through the explicit representation of the behavior of the agents that exemplify components of the immune system and HPV, as well as the processes by which these domains interact. In biological systems, the emergency represents a central event, which arises from the interactions among its components. In general, agent-based models often exhibit patterns and relationships that emerge from the interactions among their agents, so we use “top-down” modeling to focus on the feedback processes. This dynamic allows us to establish the interactions between the internal environment (HIS-HPV16 domains) and the external environment (TVD domain), with which we attempt to simulate the feedback processes and the collective behavior of the system, in that way trying to find the optimal vaccination cycles.

### 3.2.2 Conceptual Model

To develop the conceptual model of the proposed complex dynamic systems, we support us with the design of diagrams, schemes and figures that allow us to visualize the interactions that arise between the various components associated with each of the previously established domains.

Initially an entire scheme is defined where the interactions that arise between the host and pathogen are represented, that is, between the HIS and HPV16 domains. A second scheme incorporates the components and processes of each of these domains. A third scheme focuses on the relationships among the interacting components between both domains, and that allows activating sensors that guarantee the feedback of the model at any point of time. Subsequently these three schemes are merged into a diagram, on which the processes that feed the model are demarcated in two directions, from the bottom to the top and from the top to the bottom (see Fig. 7).

From this analysis, three stages of work are established in which is sought to specify the details of key processes for each domain and the interactions among both domains. In the first stage, in relation to HIS domain, details are established to simulate differentiation, proliferation and cellular apoptosis processes which include: the behavior of the master regulator and the interactions that arise around



**Fig. 7** General design of an artificial life model. This figure corresponds to the schematic representation of the interactions that arise between the HIS and HPV16 domains, in each of the three established microenvironments (cells, TLRs and cytokines)

cell populations. Regarding the HPV16 domain, details are established to simulate cell cycle phases (G0, G1, S, G2 and M) as well as the procedures associated with p53, pRB and Tert. In the second stage, in relation to HIS domain, the details are defined to simulate the environment corresponding to TLRs, including their downstream signaling pathways. Regarding to the HPV16 domain, the details are incorporated to simulate the evolution of the virus under infection, persistence and lesion development environments. In the third stage, in relation to the HIS domain, the details corresponding to the cytokine environment and its downstream signaling pathways are defined. Regarding the HPV16 domain, the details are defined that allow the incorporation of the signal detector whose function will be to demonstrate the distinct stages of the HPV16 life cycle (growth, replication, regulation, transcription, capsid conformation and oncogenic transformation). Depending on the cell populations that differentiate, proliferate or die, different activation signals are generated. Some of these signals allow activating the behavior of cytokines that each cell population is able to secrete, and the surface molecules that they can express, events that in turn can induce activity in populations of TLRs.

Subsequently, we work on the schematic representation of the activities linked to the processes reported in each of the previous stages, for which we use additional graphics that allow us to visualize key procedures and establish the bases to develop the corresponding microenvironment. Based on this, the three previous stages are complemented describing in their order the activities related to the cellular-microenvironment, TLR-microenvironment and Cytokines-microenvironment.

### **Cellular Microenvironment**

We construct the schematic representation of the cell differentiation process according to the evolution of common myeloid and lymphoid progenitors. This allows us to identify and describe each of the cell populations that can be differentiated, as well as their master regulators, surface markers and cytokines linked to each cell group. The process of cellular differentiation from myeloid precursors involves the populations of monocytes, macrophages and dendritic cells (DCs) which are affected by transcription factors, epigenetic regulators and transcriptional mechanisms. The process of differentiation from lymphoid progenitors involves the lineages of B-cells, T-cells and NK cells. Based on this exploration, four themes are established on which it is deepened to be incorporated into the cellular microenvironment: (i) modeling myeloid precursors, (ii) modeling lymphoid precursors, (iii) establishing initial population, (iv) modeling of cell zones.

**Modeling myeloid precursors** Monocytes circulate in blood, bone marrow and spleen, and these do not proliferate in a stable state. Monocytes occur predominantly in the blood while macrophages are found in lymphoid organs. During inflammation, monocytes can differentiate into macrophages or DCs. The processes of differentiation and migration are conditioned by the inflammatory medium and the pattern recognition receptors (PRRs) associated with HPV16. Monocytes, macrophages and DCs express MHC class-II and interact with helper T-cells during the immune response.

The activation of macrophages involves a set of stimuli and expression programs that depend on the microenvironment and produce polarized phenotypes that vary between classical activation (M1) and alternative activation (M2). Classical activation or M1 includes the expression of iNOS, high levels of IL-12, and reduced levels of IL-10 production. The M1 activation corresponds to the stimulation of macrophages mediated by lipopolysaccharides (LPS), IFN- $\gamma$ , TNF- $\alpha$ , GM-CSF, and TLR ligands like TLR2 and TLR4. This stimulus leads to a proinflammatory phenotype. The M1 macrophages develop a response to the activation of PRRs and allow them to enable their ability to secrete vast amounts of proinflammatory cytokines. They also express elevated levels of MHC, co-stimulatory molecules, and Fc $\gamma$ R receptors. The alternative activation or M2 corresponds to a stimulation of macrophages dependent on IL-4 and IL-13. This stimulus leads to an anti-inflammatory phenotype. The M2 macrophages are involved in responses associated with tissue remodeling, angiogenesis and tumor progression. Antigen-specific T-cells factors can bind to the surface of macrophages.

Dendritic cells (DCs) reside in locations where viruses begin their multiplication process before spreading to secondary sites in the organism. For this reason, the DCs are able to detect viruses through TLR receptors, process and present viral antigens by the major histocompatibility complex (MHC) class-II to T-cells which in turn causes the start of an adaptive immune response. The transcription factors IRF2, IRF4 and IRF8 are involved in the diversification of DCs groups. There are four main groups of DCs: conventional DCs (cDCs), Langerhans cells (LCs), myeloid derived DCs (mDCs) and plasmacytoid DCs (pDCs). The DCs express diverse PRRs, such as: TLRs, NLRs, RLRs and CLRs, which are capable of binding cellular motifs characteristic of HPV16 or associated with cellular damage. The DCs activated by PRRs lead to the activation of antigen-specific T-cells. In general, the DCs express CD11c and MHC class-II molecules, but once activated they increase their levels of expression of peptide-MHC complexes and co-stimulatory molecules which allows them to activate the T-cells effectively. Activated DCs (or mature DCs) loaded with the antigens initiate the process of differentiation of antigen-specific T-cells into effector T-cells, activating functions and unique cytokines profiles.

The cDCs, mDCs and LCs perform functions linked to the maintenance of self-tolerance and to the induction of specific immune responses against invading pathogens. The pDCs posses the function of secreting considerable amount of IFN- $\alpha$  in response to viral infections, and priming T-cells against viral antigens, although they can also act as antigen-presenting cells (APCs) and control the T-cell responses. The expression of MHC class-II and co-stimulatory molecules CD40, CD80 and CD86, allow pDCs to present antigens to CD4+ T-cells. The master regulatory factor for the development of pDCs is E2-2. The migration of pDCs involves CD62L, PSGL1,  $\beta$ 1 and  $\beta$ 2 and multiple chemokine receptors such as: CXCR3, CXCR4, CCR2, CCR5, CCR6, CCR7, CCR8 and CCR10. During inflammation, CXCR3 and CCR5 lead to the migration of pDCs towards inflamed tissues. The mDCs cells signal through TLR3, TLR4 and TLR8. The pDCs express mainly TLR7 and TLR9 receptors, and are unique in activating IRF7, which allows them to rapidly produce

elevated levels of Interferon type I in responses to viral infections. The pDCs can also secrete cytokines such as: IL-6, IL-12, CXCL8, CXCL10, CCL3 and CCL4.

**Modeling lymphoid precursors** Considering the inherent complexity of the process of differentiation B-cells, we construct a graph that allows us to outline their development and differentiated subgroups, following the trace from a hematopoietic precursor cell that initially takes place in the bone marrow and subsequently it is transformed into spleen and lymph nodes. In our model, we work on the development of B-cells based on two approaches. Regarding the first approach, we illustrate the stages of development of antigen-independent B-cells; and with the second approach, we describe the stages of development dependent on the antigen. In the first case the groups of stem cells, pro-B, pre-B and immature cells are considered. In the second case the thymus-independent B-cells, marginal zone (MZ) B-cells, follicular (FO) B-cells and germinal-center (GC) B-cells are included. Likewise, this case considers the B-cells that differentiate between memory B-cells and plasma cells (PC).

Taking into account the complexity linked to the T-cell population, we also work on an additional graph that allows us to outline the processes of differentiation at the levels of this cell group. In our model, the T-cell development and differentiation processes are worked based on three approaches. The first approach involves some transcription factors. The second approach involves cytokine profiles. The third approach involves the profile of cytokines, transcription factors, receptors and the triggering of some of their functions associated with the processes of differentiation, activation, secretion and negative regulation. For the first case, we consider two transcription factors: *Ikaros* and *Gata3*, which are involved in the primitive stages that compromise the T-cell lineage. The *Ikaros* factor by itself is necessary for the early development of B-cells and for a variety of mature T-cell functions. The expression of *Gata3* in the hematopoietic lineage is restricted to T-cells and NK cells. This case additionally includes the transcription factors: T-bet, Eomes, BCL6, Blimp1, ID2, ID3, STAT3 – STAT5, which regulate the development of effector CD8+ T-cells and memory T-cells. For the second case, the population of NKT cells is included. For the third case, we include: the population on helper T-cells (Th1, Th2, Th9, Th17 and Th22), follicular T-cells (Tfh), and regulatory T-cells (Treg).

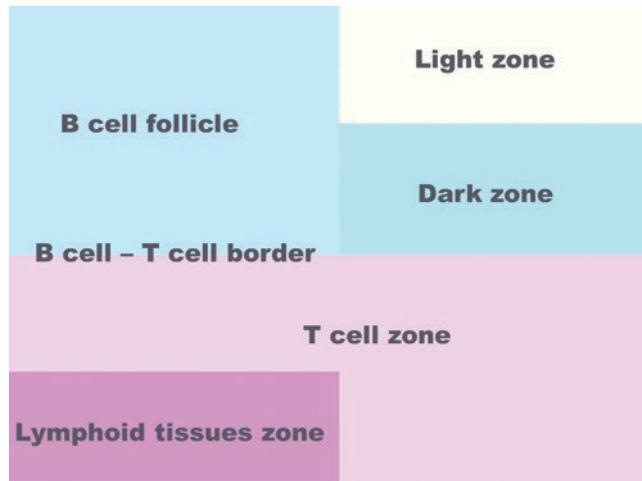
In our model the keratinocytes (KCs) population receives special treatment, since their activation and inactivation processes are coordinated mainly by growth factors and cytokines, which are produced by the KCs themselves and by other cell types at their around. For this reason in our model, this population is incorporated into the microenvironment of cytokines. The KCs function as antigen presenting cells and are capable of inducing the expression of Th1/Th2 type cytokines and cytotoxic responses in CD4+ and CD8+ memory T-cells.

**Establishing initial population** In relation to this topic, we also define the protocols that support the establishment of the initial population of the related cell groups. To achieve this, we consulted key information provided by expert laboratories in human blood processing, and proceeded to elaborate the documentation that explain in detail the estimation of the initial population by each cell group. Based on the

previous work, we ultimately defined the set of cell populations that will be incorporated into the model. These include: macrophages, dendritic cells (DCs), stem cells, B-cells in bone marrow (pro-B, pre-B, immature), B-cells in spleen (T1/T2, MZ, FO, short-lived PC, GC), B-cells in lymph nodes (long-lived PC, memory B-cells), naïve T-cells, Plasma cells (MZ-dependent short-lived PC and GC-dependent long-lived PC), memory B-cells (serological memory and GC-dependent humoral memory), Natural Killer (NK) cells, NKT cells, Memory T-cells, cytotoxic T lymphocytes (CTLs), helper T-cells (Th1, Th2, Th9, Th17, Th22), T follicular helper cells (Tfh), Regulatory T-cells (Tregs). Additionally, populations of the Follicular dendritic cells (FDCs), Keratinocytes (KCs) and Antibodies (Ags) are incorporated into the model.

**Modeling of cell zones** Based on this analysis, we equally establish the zones that the model must differentiate to recreate the simulation of the dynamics among cell groups that activate innate and adaptive immune responses. Our model defines six zones that include: (i) Lymphoid tissues zone; (ii) T-cells zone; (iii) Follicular B-cells zone; (iv) Interaction zone between B-cells and T-cells; (v) Follicular zone; and (vi) Germinal centre zone (see Fig. 8).

**Lymphoid tissues zone** On this zone the mature DCs expressing CD80 (B7.1) and CD86 (B7.2) present the ligand formed by the peptide-MHC class-II complex to the TCR receptor of naïve CD4+ T-cells. On this environment, the binding between molecules and receptors of the antigens presenting cells and CD4+ T-cells is simulated. That is, the binding of pMHCII with TCR and B7 (CD80/CD86) with CD28, which leads to the activation of naïve CD4+ T-cells. T-cells zone: Once the CD4+



**Fig. 8** Schematic representation of cell interaction zones. This figure represents each of the zones delimited in the artificial life model on which the interactions among different cell populations are simulated. Within the model there is a correlation between the zones, and the processes simulated associated with the immune response dynamics that are described in the (Fig. 2)

T-cells are active, they secrete IL-21 and induce the expression of CD28 and ICOS.

Follicular B-cells zone: After the naïve B-cells bind to antigen proteins, they up-regulate the expression of CCR7 and they are attracted by CCL21 towards the exit of the T-cell zone in secondary lymphoid tissues, where the help of the T cells is triggered.

Interaction zone between B-cells and T-cells: Sustained signaling of CD4+ T-cells activated through TCR, CD28 and IL-21R, both in the T-cells zone and in the interaction zone B-T cells leads to the modulation of some receptors. These include the expression of chemokine receptors, such as: CXCR5 (CD185, produced by B-cells and DCs) and CCR7 (CD197, produced by NK, Th2, Tfh, Th17 cells); and co-stimulatory receptors, such as: ICOS (CD278, expressed by Th1, Th2, Tfh, Th17 cells), CD40L (CD154, expressed by Tfh cells), and OX40 (CD134, expressed by Tfh and Treg cells). Populations of B-cells and T-cells form long-lived interactions, resulting in the complete activation of B-cells. The T-cells acquire the characteristics of the T follicular helper cells (Tfh) phenotype. At that point the Tfh cells migrate from the interfollicular region to the follicle. Some of the B-cells activated by the antigen that proliferate, differentiate among cells that secrete antibodies or early plasmablast (precursors of short-lived plasma cells), which then move towards the region adjacent to the subcapsular sinus. Subsequently, B-cells migrate from the interfollicular region towards the center of the follicle.

Follicular zone: This zone is characterized by a network of follicular dendritic cells (FDCs). On this area the B-cells that enter begin to proliferate, and as a result push the resident B-follicular cells to one side, to construct the early germinal centre. This early GC consists of B-cell blasts surrounded by the mantle zone (this structure is also known as a secondary follicle.). In this location, the differentiation process can occur through three distinct ways: (i) a follicular path that leads to the germinal centre; (ii) an extracellular pathway that induces the production of short-lived plasma cells which produce low affinity antibodies; and (iv) a pathway that leads to the generation of a B-cell population of short-lived memory. These cells are adoptively transferred, antigen-independent and non-proliferative. In addition, they recirculate through the lymphatic system and react only when they encounter the antigen, at which time they proliferate and secrete Ig.

Germinal centre zone: Within the group of antigen-specific B cells, only a limited group can access the germinal centre reaction and it corresponds to those cells that possess the relatively higher affinity inside the group, as a result of interclonal competition. The B-cells activated by the antigen, precursors of germinal center (GC), form the early GC where they differentiate between blasts. During the following days they will undergo clonal expansion until they manage to assemble the mature GC. The mature GC is characterized by containing light zones and dark zones [40, 318–320]. To delimit the creation of this GC zone in our model, we implemented the simulation of the following processes: (i) GC formation; (ii) establishment of the dark zone and light zone. The processes that take place in the GC zone depend on the interactions between B-cells, T follicular helper cells (Tfh) and follicular dendritic cells (FDCs). These interactions involve the simulation of the following processes: (iii) clonal expansion of B-cells; (iv) somatic hypermutation (SHM); (v) selection of high-affinity B-cells; (vi) class-switch recombination (CSR); (vii) negative selection;

(viii) GC reaction dynamics; (ix) movement among GC zones; (x) molecular control of the GC reaction; (xi) maintenance of the GC reaction.

## TLR Microenvironment

In the schematic representation of the activities linked to the microenvironment in which the Toll-like receptors (TLRs) are developed, three main work fronts are defined in our model: (i) modeling TLR processes; (ii) modeling negative regulation; and (iii) modeling TLR zones.

**Modeling TLR processes** Considering HPV16 is a double-stranded DNA virus integrated into the host, our model simulates the behavior of Toll-like receptors associated with the viral dsDNA responses. For this reason we incorporate the details of the signaling pathways of TLR3, TLR4, TLR7, TLR8 and TLR9. Having cleared the scope regarding the TLRs of interest, we designed maps that allow us to outline the behavior of each of its signaling pathways, based on three approaches. The first approach corresponds to the behavior of the signaling pathways against stimuli initiated from endogenous TLR ligands. The second approach involves the detail of the behavior of TLR signaling pathways in conditions of infection, tissue damage, and progression to cancer. The third approach corresponds to TLR signals that cause particular effects on tumors. Under the first approach, we analyze and detail the early warning signals that the host's immune system detects from the endogenous ligands. Under the second approach, we analyze and detail the signals that are mediated by the degradation of macromolecules that activate the TLRs. In addition, the stimuli that the TLRs induce to initiate inflammatory responses that attempt to protect and repair damaged tissue are included, as well as the deregulation of their signaling processes either due to inherited deficiencies or chronic activation of the infection. Under the third approach, we consider the changes observed on TLR signaling pathways when these interact with ligands linked to tumor processes. After examining the maps resulting from each approach, we worked on the construction of a consolidated graph that allow us to ultimately establish the components and details of the signaling cascades that we must incorporate in our model.

Individual TLRs expressed on different cell types recognize their specific ligands and activate signaling pathways. The process of activating a certain signaling cascade depends on the fulfillment of three key events. The main activation event is presented from the interaction between a TLR receptor and a TLR ligand, which results in the coupling of the TIR domain and the respective TLR. The second activation event considers the type of receptor-ligand coupling prior, to induce the recruitment of adapter molecules related to it (MyD88, TIRAP, TRIF, TRAM), which equally contain a TIR domain. The third activation event considers the type of receptor-ligand-adapter coupling that is active to induce the signal that allows triggering the related signaling pathway.

Once an active signal that induces a specific signaling pathway is detected, our model follows a downstream path that involves the simulation of the recruitment and possible activation of several complexes and families, such as: IRAK complex (IRAK1, IRAK2, IRAK4); TRAF family (TRAF3, TRAF6); transformation factors

(TAB2, TAB3, TAK1); MAPKs complexes (MAPK, MKK3, MKK4, MKK6, MKK7); IKKs complexes (IKK $\alpha$ , IKK $\beta$ , IKK $\gamma$ ). This process subsequently continues with the coupling of other pathways that include: CREB, AP1, NF- $\kappa$ B, IRF3 and IRF7. Depending on the circumstances of the microenvironment, CREB, AP1 and NF- $\kappa$ B, have the ability to induce the expression of genes involved in inflammation and initiate the activation of adaptive immunity. Some of these genes involve: IL-1 $\beta$ , IL-6, IL-18 and TNF. In this same context, IRF3 and IRF7 are similarly essential in the induction processes of type I Interferons (IFN- $\alpha$ , IFN- $\beta$ ). Additionally, we simulate the behavior of other factors involved in TLR signaling, such as: (i) TOLLIP: it involved in IL-1 receptor trafficking and in the rotation of the associated IL-1R kinase; (ii) PELLINO: it includes Pellino1 - Pellino3. These are able to interact with IRAK1, IRAK4, TRAF6, MAPK3 and MAPK7; (iii) PI3K: as a TLR activation component, it can affect distinct pathways depending on the type of cell involved. In cancer conditions its activity increases; (iv) ECSIT: this adapter protein interacts with TRAF6 and some members of the MAP and MEKK1 kinases families, with which it can phosphorylate and activate the IKK complex; (v) Family of SRC kinases: this family is involved in the regulation of PI3K/PTEN/Akt, through mechanisms like phosphorylation of PI3K and PTEN, which then end in the inhibition of PTEN. Increases in the activity of SRC kinases have been correlated with progression to malignancy, and it is identified as a key molecule in tumor process that can provide oncogenic signals [321]; (vi) IAPs: this component includes: XIAP, cIAP1, cIAP2, Survivin, BRUCE, ML-IAP, ILP2. These proteins serve as inhibitors of programmed cell death of anti-apoptotic. IAPs act as physiological inhibitors of caspases, a group of cysteine proteases that are considered central executors of apoptosis. Within this group, the cIAPs act as positive regulators of canonic signaling NF- $\kappa$ B downstream from TNF-R1 receptors, and also function as negative regulators of the non-canonical NF- $\kappa$ B pathways.

Regarding the activities linked to each of the TLRs, and in addition to the signaling pathways, our model also considers the processes of positive regulation, negative regulation, deficiencies, behavior on cancer conditions, and blocking factors.

**Modeling negative regulation in TLR signaling** When the inflammatory cytokines that are produced as a result of TLR signaling, are released in excess, they induce disorders that are associated with an elevated rate of cell death. From this event, evolved mechanisms are activated that try to modulate the responses mediated by TLRs. The negative regulation of TLR-induced responses is relevant in suppressing inflammatory immune responses, and inflammation is generated through its interruption or mutation. Some of the molecules that negatively regulate TLR signaling include: IRAK-M, SOCS1, MyD88s, SIGIRR, ST2, TANK, Atg16l1, and A20. These molecules, specifically, are equally considered in our model.

**Modeling TLR zones** Based on this analysis, we equally designate the zones that the model must differentiate to recreate the simulation of the TLR dynamics. These include: (i) endogenous ligands zone; (ii) extracellular receptors zone; (iii) TIR domain zone; (iv) cytoplasm zone; (v) endocytosis zone (linked to TLR4); (vi)

endosomal membrane zone (linked to TLR3, TLR7, TLR8, TLR9); (vii) core zone (linked to proinflammatory cytokines and type I interferons).

### Microenvironment Cytokines

As in the previous microenvironments, we elaborate graphs that allow us to visualize the key processes associated with cytokines. We work on the schematic representation of the signaling pathways linked to five distinct families of cytokines, including: TNF, TGF, IFNs, MIF and ILs. Each family involves its respective members, including: TNF- $\alpha$ , TNF- $\beta$ , TGF- $\alpha$ , TGF- $\beta$ , IFN- $\alpha$ , IFN- $\beta$ , IFN- $\gamma$ , IFN- $\lambda$ , MIF, IL-1 - IL-13, IL-15, IL-17 - IL-28, IL-30 - IL-39. Additionally, related cytokine receptors are included, such as: IL1R1, IL1RA, IL1R2, IL2Ra, IL2Rb, IL3Ra, IL3Rb, IL3Ra, IL-3Rb, IL4Ra, IL5R, IL6Ra, IL7Ra, IL8Ra, IL8Rb, IL9Ra, IL10R1, IL10R2, IL11Ra, IL12R, IL12Rb1, IL12Rb2, IL13Ra1, IL13Ra2, IL15R, IL17Rb, IL18R, IL20R1, IL20R2, IL21R, IL22R1, IL23R, IL24R, IL27R, IL31Ra, IL33Ra, IL35R, IL36R, CD132 ( $\gamma$ c chain), OSMR, IFNAR1M IFNAR2, IFNgR1, IFNgR2, IFNIR1, TNFR1, TNR2, TGFbR1, TGFbR2, CXCR1, CXCR2.

From this exercise, we specify the behavior of several components that interact with HPV16 life cycle, among which we consider: adhesion molecules, growth factors, enzymes, proteins that correspond to the group of signaling units in cytokine receptors, metalloproteinases, kinases ligands, and other transcription factor proteins. The plan of induction of cytokines is conditioned by the class of TLRs, the nature of the ligand and the type of cell that are activated. All cells that release cytokines can influence the immune regulation of the cytokine network. The immunostimulatory cytokines (Th1 type) are produced typically by APCs and NK cells, and the immunoinhibitory cytokines (Th2 type) are secreted principally by lymphocytes and monocytes. Cytokine signaling pathways are triggered from the junction between the cytokine and its functional receptor. Once the binding is activated, the cytokines can act on their target cells. However, HPV16 has the ability to alter the synthesis of host cytokines, implementing as mechanisms the degradation of proinflammatory cytokines or using cytokine receptors as entry portals for cellular evasion.

This microenvironment in our model also incorporates the activation processes of the keratinocytes population (KC). The most common initiator of keratinocytes is IL-1, whose forms (alpha and beta) are present in the cytoplasm of KCs. When the injury occurs, the keratinocytes process and release IL-1 s. These interleukins act as chemoattractants of lymphocytes, which in turn help the KCs in their migration to the sites of the lesion. Additionally, IL-1 also acts as an autocrine signal that activates KCs, causing them to proliferate, migrate and express a set of specific activation genes, among which are: GM-CSF, TNF- $\alpha$ , TGF- $\alpha$  and AREG (amphiregulin). Once activated, the KCs also produce adhesion molecules like ICAM-1 and integrins like fibronectin, a component that promotes the migration of KCs. The TNF- $\alpha$  has the ability to maintain KCs in activated state, inducing the production of other signaling molecules, cytokines, growth factors and their receptors [29].

Along with the processes linked to each of the cytokines and their signaling pathways, our model also incorporates the processes of positive regulation, negative regulation, deficiency, behavior in cancer conditions, and related blocking factors.

### 3.2.3 Preliminary Considerations

In these stages we try presenting in a very general way the key processes that will generate the global dynamics of the model, that is, those actions that will contribute to the emergence of possible interactions between the host and pathogen. This dynamic provides a broader understanding of the proposed conceptual model and allows proposing the basic scheme for the subsequent stage in which the prototype implementation is planned. The execution of this stage implies the elaboration of figures and graphs that summarize the dynamics of the particular processes and the explanation of the corresponding detail.

In our particular case, we present in a general way the interactions generated between the life cycle of HPV16 during the development of its infectious process and the mechanisms of immune response by the host, both innate and adaptive. During the development of this stage, we work on the schematic representation and the detail of five key components. These are: (i) Immune response to HPV16 infection: in this component we illustrate how the process of activating the immune response depends upon three signals. The first signal is presented during the maturation process of the DCs. The second signal occurs during the interaction process between T-cells and B-cells. The third signal is produced from the generation of cytokines. Processes associated with cases of regression in infectious HPV environment and evasion mechanisms are also detailed. Regarding to the persistent and high risk infection process, both the behavior of the expression of the HPV16 viral oncoproteins and the interactions between oncoproteins and immune signaling components are illustrated, which cause the activation of the immune response to be retarded and therefore persistent infection is facilitated. (ii) Cell differentiation caused by the detection of a natural infection: in addition to defining the process that allow the generation of several cell groups, we explain the management of the times associated with division and death in each cellular population. (iii) Detection of a viral infection: in this component we explain the process in detail around the stages of growth, division and cell death, controls that govern cell fate, type of division, frequency and average times. Within this component, we additionally include the interaction processes between B-cells and T-cells. We define the procedures related to antigen presentation, B-cell differentiation, somatic hypermutation (SHM), class switch recombination (CSR) including isotype exchange and affinity maturation. (iv) Detection HPV16 infection: within this component we explain the sequence of steps that this process follows, which involves TLR signaling, expression of viral oncoproteins, behavior of cell populations, trafficking of viral antigens, target cell infected by this pathogen and secretion of cytokines. (v) Cancer and immunity: in this component we detail the essential phases of interactions between tumor cells and the host's immune system.

### 3.2.4 Checkpoints Activation

The complexity associated with physical systems frequently involves the interaction among critical points that produce chaos or not, and where self-organization is more than likely occurs [3]. In this stage, it is sought to introduce the components and key interaction processes that guarantee the general dynamic of the model.

In our particular model, we present in detail the checkpoints that start and dynamize the following five processes: (i) recognition of viral pathogens through the participation of PAMPs; (ii) cytokines production that promote the activation of T-cells and their relationship with TLR receptors; (iii) activation of T-cells dependent on TLRs; (iv) dynamics among cell groups that activate innate and adaptive immune responses; (v) dynamics on the germinal centre reaction.

### 3.2.5 States and Transitions

In this stage the key concepts that define the checkpoints, set of rules, states, transitions and interactions are established.

In our model, the principal checkpoints are given by the activation of TLR receptors that recognize viral PAMPs. Once these are activated, they stimulate the transcription of inflammatory genes and trigger signaling pathways, which leads to the stimulation of transcription factors that allows the secretion of some cytokines that contribute to the manifestation or not of an immune response.

Regarding the rules, in general our model is governed by five fundamental rules: co-selection, network connectivity, rotation, deletion, and change of states. In front of the HPV16 domain, the rules are given by the behavior of viral proteins. In relation to the HIS and TVC domains, the specific rules that affect all the processes incorporated into the model are due to biological conditions.

When addressing states, our model considers four distinct approaches. The first approach is based on the perspective of the cell as a member of a cellular population. The second approach considers a general perspective of the host's immune system. The third approach is based on the characterization of differentiated degrees of injuries caused by viral infection. The fourth approach focuses on the environment linked to the application of a therapeutic vaccine that counteracts the conditions of a disease generated by viral infection. Then, for each one of the previously declared approaches, the possible states that represent an associated characteristic at some point of time are established. The change among states will represent the result of the conditions generated in the microenvironment en which the component that receives the action develops and evolves, accordingly modifying its current situation.

A transition corresponds to a key criterion that allows the development of its own dynamics associated with the components of the environment. In our model, we work with three key criteria that correspond to: elimination, evasion (or escape) and equilibrium.

In the task of defining the interactions that will be incorporated into the model, and due to the complexity that it represents, we need to propose nine key scenarios. These are: (i) cell populations and HPV16; (ii) HPV16 and cancer; (iii) HPV16 and TLRs; (iv) TLRs and cancer; (v) cytokines and HPV16; (vi) cytokines and cancer; (vii) HIS and HPV16; (viii) TLRs, cytokines, HPV16 and TVC; (ix) cell population and TVC. From the review of these interactions, we define key specific components on which we examine the behavior of their main markers. From the review of the interactions within these scenarios, we define key specific components on which we examine the behavior of their main markers. Based on previously exposed, in our model we consider four types of fundamental interactions, defined as: stimulation, inhibition, suppression and death. In stimulation, one component of the model forces the other with which it interacts, to generate a state of activation and consequently to be able to induce an action akin to its essential function. In inhibition, the opposite happens. One component of the model causes the other with which it interacts, to be temporarily blocked and be unable to perform its essential function, as long as this interaction continues being active. In suppression, a component prevents the action of its counterpart. In death condition, the result of the interaction among two components causes one of them to disappear from the model.

### **3.2.6 Description of the Logical Model**

This stage tries to provide better clarity and precision on the logic to be adopted in the implementation phase, especially addressing the most critical processes. Additionally, the sequence of steps that must be followed in the construction of the corresponding prototype is elaborated.

In our model for example, in relation to critical processes, we work to define the specific logic around the cellular microenvironment where the conditions are generated that allow predicting the level of affinity maturation and also the appropriate environment for some cell populations to differentiate. At the level of the extracellular microenvironment, we define the necessary logic to be able to observe the modifications and results that occur when incorporating the control of therapeutic vaccines. To describe the sequence of steps, we addressed this task for each of the established environments, initially focusing on the intracellular and extracellular microenvironments, and ultimately, considering the set of interactions that relate the three domains HIS, HPV16 and TVC.

## **3.3 Construction Phase**

This phase aims to define and prepare all the components required in the construction process of a functional prototype that can be used initially as a means of verification of the conceptual and logical models, and later as an experimental laboratory. As reported by An and colleagues (2009), a model aims to formalize the process of

knowledge representation particularly in terms of dynamic instantiation of knowledge, as a means of viewing and evaluating hypotheses [265]. Embodying a conceptual model in a functional prototype offers the researcher the ability to visualize, analyze and evaluate the consequences associated with each hypothesis being tested, and also a frame of reference in the planning and design of novel experiments.

During the construction phase of our model, we focus on detailing three key stages that furnish us the basis to generate the structure and develop the code that will allow us to produce our functional prototype. These three stages involve the description of the following topics: characteristics of the development tool, description of the prototype and discussion of the global parameters of the artificial life model.

### 3.3.1 Development Tools

The first step in the construction phase is to define the programming tool that will allow the functional prototype of the proposed model to be developed. Given that our problem is of biological origin and that we focus on a technique of agent-based models, these two criteria must be taken into account when choosing the software tool that will allow producing the code to build the simulator. Additionally, it is essential to consider complementary reasons that support the selection of the development tool, such as: the usability, the compatibility of exchange with various platforms, the capabilities to connect to databases, the tools for the design of the graphical user interfaces, among many other aspects [267].

Currently, several software products are available that have the ability to reproduce fundamental characteristics of biological systems and at the same time, allow to focus on the use of methods derived from artificial life. The software oriented to ABMs, as an artificial life technique, allows the creation of new laboratory environment that can contribute to traditional methods (*in-vitro* and *in-vivo*) in research processes. Currently, several software tools are available can be considered.

To advance the study of options that allow us to select the software tools appropriate to our particular case, in which we investigate complex biological dynamic systems, we propose two approaches. A first approach corresponds to platforms used for modeling and simulation supported in agents; and a second approach, corresponds to high-level programming languages for the creation of agent-based models together with the tools that allow an integrated development environment (IDEs). Regarding the first approach, Railsback and colleagues (2006) compare five platforms oriented to scientific models based on agents. They evaluate aspects such as: the modeling structure, dynamic programming, the generation of aleatory numbers, the development experience, the speed of execution, and some factors keys [322]. Abar and colleagues (2017) present a characterization of 85 groups of agents-base tools that can assist designers and system developers in the construction of their models, with the possibility of displaying the results of the simulation in real time using tabulation formats, graphs and visual records. In their analysis process, they consider several comparison criteria and outstanding characteristics of each platform. Some of these criteria include: source code specifications, online accessibil-

ity, licensing category, software availability, integrated development environment (IDE), programming language to develop the model, application programming interface (API) or native libraries, and nature or type of agent's evolution. Additionally, they compare information related to: compiler, operating system, hardware requirements, modeling power, usability, scalability and domains covered by each platform [323]. Regarding the second approach, there are also products, such as: *Java*, *C++*, *Phyton*, *Microsoft.net*, *Smalltalk*, among others. In addition, there are other agent-oriented programming languages (AOP). Products like *Oragent* [324] and *Ulam* [325], correspond to this category. Along with the previously mentioned platforms, many other options are accessible and in constant evolution. So the most significant event in this process of choosing the appropriate software tool is that the chosen product represents the result of a well-informed decision.

### 3.3.2 Description of the Functional Prototype

During this stage, which form part of the construction phase, various activities are carried out to prepare the key components that help to obtain better clarity and order at the moment of initiating the writing of the code on the previously selected software tool. In our particular case, this set of activities includes: (i) a general description of the graphical user interface; (ii) details of the logic conceived for the construction of the prototype; (iii) description of means for displaying the output data in real-time; and (iv) discussion of the parameters of the artificial life model.

In the process of describing the graphical user interface (GUI) in general, we prepared an interface design segmented by blocks of information, which allows us to clearly indicate the location of: the input parameters, the principal panel where the simulation is exhibited, and the monitors where the trend graphs, statistical details and status data are displayed. In front of each these components we explain in detail what it represents in the model, what its functionality is, and how the information it produces should be interpreted. Regarding the description of the logic implemented, we build a figure that allow us to outline the reasoning we follow in the implementation process and to propound the information flows that arise from the interactions among the three planned domains. Our logic of implementation considers two perspectives, one external and one internal. Based on the first, we define the properties of the global dynamics and the particularities of the external microenvironment, which allows us to establish the global parameters of the prototype. Based on the internal perspective, we define the properties of the local dynamics and the particularities of the internal microenvironment, which allows us to determine the control parameters of the prototype. The logic that we propose within the dynamics of the model involves the specification of the downstream feedback processes, where the global parameters allow the control parameters to be fed again. Subsequently, the upstream feedback processes start their activity from the interaction procedures between the external and internal microenvironments thereby influencing the control parameters, which in turn feed the global parameters again. From this point, we specify the unit of time and describe the com-

putational logic linked to configuration issues, processes and the interactions between the two microenvironments and the three domains proposed. After that, we develop the detail of each perspective, as a result of the stimulus-response relationship. In addition, we specify each of the properties linked to the global and local dynamics, and the detail of the state attributes at the level of the external microenvironment together with the peculiarities of the sensors associated with the agents.

The description of means of visualization of output data involves the comprehensive explanation of several components. Among these, the explanation of one of the icons incorporated into the graphical user interface is included. In addition, the association established between the source data and the visualization tools is detailed, considering the characteristics that are to be displayed in the presentation of each of the monitors (principal, state and signaling pathways) together with the scatter plots and line graphs (trends and statistics). The design defined for the integration of these components must allow the user to observe the behavior of a simulation in real-time and extract the data required in any period of time.

In the discussion of the parameters of the artificial life model, we work on the description of both the global and control parameters. Under this context, we explain how these should be incorporated into the GUI, and for each one of them, we include the details of its functionality, specifications, restrictions, state variables that must be controlled and the way in which it ought to interact with end-user, both in execution time and stop time.

Once this stage is finished, the code writing process begins, the result of which will be a functional prototype ready to be tested and documented.

### **3.4 *Experimentation Phase***

The experimentation phase focuses on the construction of the virtual laboratory based on simulation, for which the functional prototype that has been developed is used. In this phase we also try to examine the behavior of the artificial life model and establish if the finite nature of the simulation affects the results. To achieve this, we prepare a set of strategies that demand the planning of a coordinated group of experiments that must be run in the prototype.

The preparation of the set of experiments involves developing the following activities: (i) describe the conditions under which the experiment runs in the prototype; (ii) define an experimental design diagram that makes it possible to clearly explain the various scenarios that are to be tested; (iii) establish the set of experiments linked to each of the scenarios envisaged in the test, explaining their meaning, relevance and the objectives that are intended to be achieved; (iv) define the combination of parameters used in each experiment; (v) establish the set of variables of the model to be followed and the behaviors expected to be observed; (vi) define the rules necessary for an appropriate interpretation of the results produced; and (vii) establish the registration patterns of the results for further analysis.

It is clear all the data obtained from the experiments run in the functional prototype are documented as the results of the proposed artificial life model. Considering the enormous volume of data that can be generated during a simulation, it is beneficial to delimit the results that will constitute part of the documentation for each one of the planned experiments. This facilitates the subsequent consolidation and analysis of the results.

### **3.5 Results, Analysis and Conclusions Phase**

During this phase the relevant data obtained from each scenario that has been tested in the simulator is explained and documented, corroborating the information with the tables, monitors and graphs that interpret the results within the functional prototype. It is pertinent to explain how the records are processed, what types of tables are presented and how the resulting data are interpreted according to the objectives that each scenario has proposed within the study. Based on these results, we proceed to explain the data consolidation strategy.

Having access to the consolidated data, an analysis of the trends observed in the various experiments performed is carried out, considering each of the monitoring variables previously determined. Generally, in this stage of analysis, statistical tools are used to facilitate the process of analyzing and interpreting the data set. Regarding the analysis of the data, several strategies are equally available. Some of them allow investigating if there are statistical significance and scientific relevance among the average distribution of the simulation data, based on experimental control conditions and dynamics. Consequently, causal relationships can be inferred between the simulation data and the configuration changes within the functional prototype.

Considering the consolidated data and the trend analysis of the monitoring variables, we try comparing them, although it is not always possible, against reports of comparable clinical cases to recognize if the artificial life model generates behaviors similar to those observed in the real-world.

Depending on the results, the data analysis and the comparisons, the set of conclusions is prepared, with which this phase is finalized. As a complement, it is equally valuable to highlight the benefits and restrictions of the model, as well as describe the potential it has as a basis for future developments.

## **4 Conclusions**

Inspired by the exploration of novel approaches that try to improve the care of patients suffering from some type of cancer, we work in the construction of innovative technological options based on the modeling of artificial life. Under this context, we create software tools that aim to represent a support in the design of novel

therapeutic treatments when cancer is already present. Specifically, we focus on constructing options that try to contribute to the development of therapeutic vaccines to treat human cancers, which consider their origin in infectious processes caused by DNA viruses that are integrated into the host.

Our purpose in this chapter is to introduce the methods we use to understand and abstract some of the biological information we need. This information concedes us the possibility of building an artificial life model to simulate the behaviors of the human immune system when it is exposed to a persistent viral infection process, which can eventually develop cancer conditions. This immune system is capable to react when a therapeutic treatment based on vaccines is tried. For this purpose, we advance research in three biological components: human immune system, DNA viruses and therapeutic vaccines. Fully interpreting the dynamics that arise among these three components implies facing several challenges related to acquiring a more precise understanding of their biological processes. Part of these challenges involves understanding the heterogeneity that exists among the distinct types of cells present in the organism, recognizing intercellular signaling processes and identifying the controls exerted by the transduction pathways when they communicate signals with the extracellular environment to induce intracellular effects; in general, be able to recognize the changes that occur in the human immune system. Some of these changes arise when the immune system faces environments of persistent viral infection could later evolve and develop cancer. Other changes are evident when patients are given some therapies that try to mitigate the effects caused by these diseases.

Another of the fundamental objectives that we propose consists to reproduce the processes with the greatest possible biological rigor. For this reason, we conducted an exhaustive study of each of the three domains proposed: immune system, pathogen and therapeutic vaccine. With the aim of expanding our knowledge by evaluating each of the components that are part of these domains, we propose four key approaches: (i) behavior under natural conditions; (ii) behavior under both natural and persistent viral infection conditions; (iii) behavior in condition of progression to cancer; and (iv) behavior in therapeutic treatment condition. By acquiring a more precise understanding of the dynamics that arise among the three proposed domains and achieving translate it into an artificial life model, we have the opportunity to test potential objectives for therapeutic intervention and to show possible adverse effects that could affect some of the components that are simulate.

Both the complex systems approach and the artificial life techniques, we employ them to evaluate signaling among diverse types of cells submitted to processes of persistent infection and progression to cancer. This class of models helps to elucidate the key cellular and molecular processes that govern the procedures of differentiation, proliferation and cell apoptosis when these respond to specific therapies, in our particular case treatments based on therapeutic vaccines against cancer of viral origin.

In dynamics systems like those involved in the response of the host to a viral infection, progression to cancer or treatment based on therapeutic vaccines, our analysis of the absolute system and the dynamics that arise from the interactions among the proposed domains represent a way of to interpret the biological results

(emergent properties). However, its behavior cannot be predicted only by knowing the independent components of each of the subsystems that comprise it, it is completely necessary to comprehend the flow of the dynamics in its entirety. Based on this, artificial life models like ours can become a remarkable approach for the establishment of cooperative and synergistic relationships in the development of novel therapeutic and personalized treatments.

The tools based on artificial life are tremendously potent in the creation of models capable of predicting ideal therapies for the treatment of cancer, as long as there is an explicit and comprehensive understanding of the biological behaviors associated with the complex systems discussed here. However, there are yet gaps in the comprehensive knowledge of their behavior, and the specific biological mechanics linked to several of their processes is not fully understood either. Nonetheless, artificial life allows us to simulate behaviors that present us with an extraordinarily valuable approximation. Artificial life models like ours offer us the possibility of abstracting patterns, obtaining new knowledge, predicting some behaviors and discerning how viruses cause disease. In this way, we identify a valuable opportunity to improve the care of cancer patients, but we will have to keep in mind – for now – that we are working in a virtual world and not in the real-world.

This work describes the path we traveled and some the tools we use in the construction of our model and in the development of our functional prototype. However, the complete operability of the functional prototype as such and the documentation that contains the detail of its processes (some mentioned in this chapter), constitute part of another article.

## References

1. Globocan. Estimated cancer incidence, mortality and prevalence worldwide in 2012. France. 2012. <http://globocan.iarc.fr/Default.aspx>. Accessed 24 Mar 2018.
2. International Agency for Research on Cancer (IARC). Cancers attributable to infections. France. 2018. <https://gco.iarc.fr/infections/help>. Accessed 24 Mar 2018.
3. Katz JS. What is a complex innovation system? In SPRU Working paper Series. Ciarli T, Rotolo D, editors. University of Sussex. Montreal, Quebec, Canada. 2015, Jul. ISSN: 2057-6668.
4. Meyers RA, editor. Encyclopedia of complexity and systems science. SpringerScience+BusinessMedia, LLC., New York; 2009; p. 92–271. ISBN: 978-0-387-30440-3.
5. Merelli E, Rucco M, Sloot P, Tesei L. Topological characterization of complex systems: using persistent entropy. Entropy. 2015;17(10):6872–92. <https://doi.org/10.3390/e17106872>.
6. Sayama H, editor. Introduction to the modeling and analysis of complex systems. Binghamton University, SUNY. 2015. ISBN 978-1-942341-08-6.
7. Mitleton-Kelly E, editor. Complex systems and evolutionary perspectives on organisations: the applications of complexity theory to organisations. Advances series in management. Oxford, UK: Elsevier Science Ltd; 2003. ISBN 9780080439570.

8. Mitchell M, Newman M. Complex systems theory and evolution. In: Pagel M, editor. Encyclopedia of evolution. New York: Oxford University Press; 2002. ISBN: 978-0-195-12200-8.
9. Martínez-García M, Hernández-Lemus E. Health systems as complex systems. *Am J Oper Res.* 2013;3(1A):113–26. <https://doi.org/10.4236/ajor.2013.31A011>.
10. Ellis B, Herbert SI. Complex adaptive systems (CAS): an overview of key elements, characteristics and application to management theory. *Inform Prim Care.* 2011;19(1):33–7. <https://doi.org/10.14236/jhi.v19i1.791>.
11. Pathak SD, Day JM, Nair A, Sawaya WJ, Kristal MM. Complexity and adaptivity in supply networks: building supply network theory using a complex adaptive systems perspective. *Decis Sci.* 2007;38(4):547–80. <https://doi.org/10.1111/j.1540-5915.2007.00170.x>.
12. Alcocer-Cuarón C, Rivera AL, Castaño VM. Hierarchical structure of biological systems. A bioengineering approach. *Bioengineered.* 2014;5(2):73–9. <https://doi.org/10.4161/bioe.26570>.
13. Qian H. Stochastic physics, complex systems and biology. *Quantitative Biol.* 2013;1(1):50–3. <https://doi.org/10.1007/s40484-013-0002-6>.
14. Voit EO, editor. A first course in systems biology. Chapter 1: biological systems. New York: Garland Science; 2012; 496 p. ISBN: 9789-0815344674.
15. Gunawardena J. Biological systems theory. *Science.* 2010;328(5978):581–2. <https://doi.org/10.1126/science.1188974>.
16. Harvard Medical School. What is immunology? Boston, MA. 2015. <https://immunology.hms.harvard.edu/about-us/what-is-immunology>. Accessed 20 Nov 2015.
17. Affymetrix-eBioscience. Cytokines-Atlas. Headquarters San Diego, CA 92121.USA. 2015. <http://www.ebioscience.com/knowledge-center/antigen/cytokines.htm>. Accessed 20 Nov 2015.
18. American Cancer Society (ACS). Cancer prevention and early detection facts and figures 2015–2016. Atlanta: American Cancer Society. p. 2015.
19. Parkin J, Cohen B. An overview of the immune system. *Lancet.* 2001;357(9270):1777–89. [https://doi.org/10.1016/S0140-6736\(00\)04904-7](https://doi.org/10.1016/S0140-6736(00)04904-7).
20. Medzhitov R. Recognition of microorganisms and activation of the immune response. *Nature.* 2007;449(7164):819–26. <https://doi.org/10.1038/nature06246>.
21. Arazi A, Pendergraft WF III, Ribeiro RM, Perelson AS, Hacohen N. Human systems immunology: hypothesis-based modeling and unbiased data-driven approaches. *Semin Immunol.* 2013;25(3):193–200. <https://doi.org/10.1016/j.smim.2012.11.003>.
22. Stern PL, Einstein MH. Chapter 3 the immunobiology of human papillomavirus associated oncogenesis. In: Borruto F, De Ridder M, editors. HPV and cervical cancer: Springer Sience+Business Media, LLC; 2012a. p. 45–61. [https://doi.org/10.1007/978-1-4614-1988-4\\_3](https://doi.org/10.1007/978-1-4614-1988-4_3).
23. Stern PL, van der Burg SH, Hampson IN, Broker TR, Fiander A, Lacey CJ, et al. Therapy of human papillomavirus-related disease. *Vaccine.* 2012;30S(5):F71–82. <https://doi.org/10.1016/j.vaccine.2012.05.091>.
24. Timmis J, Knight T, de Castro LN, Hart E. An overview of artificial immune systems. In: Paton R, Boulouri H, Holcombe M, Tateson R, editors. Computation in cells and tissues, Natural Computing Series. Berlin, Heidelberg: Springer; 2004. ISBN: 978-3-642-05569-0.
25. Chu LH, Gangopadhyay A, Dorfleutner A, Stehlik C. An updated view on the structure and function of PYRIN domains. *Apoptosis.* 2015;20(2):157–73. <https://doi.org/10.1007/s10495-014-1065-1>.
26. Illumina Technology. Immunology research review. An overview of recent immunology research. Publications Featuring Illumina®Technology. 2014. <https://www.illumina.com/science/publication-reviews.html>. Accessed 30 Dec 2014.
27. Bourke CD, Prendergast CT, Sanin DE, Oulton TE, Hall RJ, Mountford AP. Epidermal keratinocytes initiate wound healing and pro-inflammatory immune responses following percu-

- taneous schistosome infection. *Int J Parasitol.* 2015;45(4):215–24. <https://doi.org/10.1016/j.ijpara.2014.11.002>.
28. Bernard FX, Morel F, Camus M, Pedretti N, Barrault C, Garnier J, et al. Keratinocytes under fire of proinflammatory cytokines: bona fide innate immune cells involved in the physiopathology of chronic atopic dermatitis and psoriasis. *J Allergy.* 2012;718725:1–10. <https://doi.org/10.1155/2012/718725>.
29. Freedberg IM, Tomic-Canic M, Komine M, Blumenberg M. Keratins and the keratinocyte activation cycle. *J Invest Dermatol.* 2001;116(5):633–40. <https://doi.org/10.1046/j.0022-202x.2001.doc.x>.
30. Hoffmann GW. Immune network theory. 2nd ed. Burnaby, Canada: Printed by Still Creek Press; 2011. ISBN 978-0-9812196-0-8.
31. Markham JF, Wellard CJ, Hawkins ED, Duffy KR, Hodking PD. A minimum of two distinct heritable factors are required to explain correlation structures in proliferating lymphocytes. *J R Soc Interface.* 2010;7(48):1049–59. <https://doi.org/10.1098/rsif.2009.0488>.
32. Tarlinton D. B-cell differentiation: instructive one day, stochastic the next. *Curr Biol.* 2012;22(7):R235–7. <https://doi.org/10.1016/j.cub.2012.02.045>.
33. Kondo M. Lymphoid and myeloid lineage commitment in multipotent hematopoietic progenitors. *Immunol Rev.* 2010;238(1):37–46. <https://doi.org/10.1111/j.1600-065X.2010.00963.x>.
34. Tobón GJ, Izquierdo JH, Cañas CA. B lymphocytes: development, tolerance, and their role in autoimmunity-focus on systemic lupus erythematosus. *Autoimmune Dis.* 2013;2013(827254):1–17. <https://doi.org/10.1155/2013/827254>.
35. Ginhoux F, Jung S. Monocytes and macrophages developmental pathways and tissue homeostasis. *Nat Rev Immunol.* 2014;14(6):392–404. <https://doi.org/10.1038/nri3671>.
36. Álvarez-Errico D, Vento-Torm R, Sieweke M, Ballestar E. Epigenetic control of myeloid cell differentiation, identity and function. *Nat Rev Immunol.* 2015;15(1):7–17. <https://doi.org/10.1038/nri3777>.
37. Geissmann F, Manz MG, Jung S, Sieweke M, Merad M, Ley K. Development of monocytes, macrophages and dendritic cells. *Science.* 2010;327(5966):656–61. <https://doi.org/10.1126/science.1178331>.
38. Bortnick A, Allman D. What is what should always have been: long-lived plasma cells induced T-cell independent antigens. *J Immunol.* 2013;190(12):5913–8. <https://doi.org/10.4049/jimmunol.1300161>.
39. Kalia V, Sarkar S, Gourley TS, Rouse BT, Ahmed R. Differentiation of memory B and T cells. *Curr Opin Immunol.* 2006;18(3):255–64. <https://doi.org/10.1016/j.co.2006.03.020>.
40. Nutt SL, Hodking PD, Tarlinton DM, Corcoran LM. The generation of antibody-secreting plasma cells. *Nat Rev Immunol.* 2015;15(3):160–71. <https://doi.org/10.1038/nri3795>.
41. Wells A, Gudmundsdottir H, Turka LA. Following the fate of individual T cells throughout activation and clonal expansion. Signals from T cell receptor and CD28 differentially regulate the induction and duration of a proliferative response. *J Clin Investig.* 1997;100(12):3173–83. <https://doi.org/10.1172/JCI119873>.
42. Gudmundsdottir H, Wells AD, Turka LA. Dynamics and requirements of T cell clonal expansion in vivo at the single-cell level: effector function is linked to proliferativa capacity. *J Immunol.* 1999;162(9):5212–23.
43. Sebzda E, Mariathasan S, Ohteki T, Jones R, Bachmann MF, Ohashi PS. Selection of the T cell repertoire. *Annu Rev Immunol.* 1999;17:829–74. <https://doi.org/10.1146/annurev.immunol.17.1.829>.
44. Efroni S, Harel D, Cohen IR. Emergent dynamics of thymocyte development and lineage determination. *PLoS Comput Biol.* 2007;3(1):e13. <https://doi.org/10.1371/journal.pcbi.0030013>.
45. Fiúza UM, Arias AM. Cell and molecular biology of Notch. *J Endocrinol.* 2007;194(3):459–74. <https://doi.org/10.1677/JOE-07-0242>.
46. Zabriskie J, editor. Essential clinical immunology. New York: The Rockefeller University, Cambridge University Press; 2009. ISBN-13 978-0-521-51681-5.

47. Henderson A, Calame K. Transcriptional regulation during B cell development. *Annu Rev Immunol.* 1998;16:163–200. <https://doi.org/10.1146/annurev.immunol.16.1.163>.
48. De Wit J, Jorritsma T, Makuch M, Remmerswaal EBM, Bos HK, Souwer Y, et al. Human B cells promote T-cell plasticity to optimize antibody response by inducing coexpression of TH1/TFH signatures. *J Allergy Clin Immunol.* 2015;135(4):1053–60. <https://doi.org/10.1016/j.jaci.2014.08.012>.
49. Bachmann MF, Zinkernagel RM. Neutralizing antiviral B cell responses. *Annu Rev Immunol.* 1997;15:235–70. <https://doi.org/10.1146/annurev.immunol.15.1.235>.
50. Mempel TR, Henrickson SE, von-Andrian UH. T-cell priming by dendritic cells in lymph nodes occurs in three distinct phases. *Nature.* 2004;427(2970):154–9. <https://doi.org/10.1038/nature02238>.
51. Zhu J, Mohan C. Toll-Like receptor signaling pathways-therapeutic opportunities. *Mediat Inflamm.* 2010;2010(781235):1–7. <https://doi.org/10.1155/2010/781235>.
52. DeCarlo CA, Rosa B, Jackson R, Niccoli S, Escott NG, Zehbe I. Toll-like receptor transcriptome in the HPV-positive cervical cancer microenvironment. *Clin Dev Immunol.* 2012;2012(785825):1–9. <https://doi.org/10.1155/2012/785825>.
53. Kanneganti T-D. Central roles of NLRs and inflammasomes in viral infection. *Nat Rev Immunol.* 2010;10(10):688–98. <https://doi.org/10.1038/nri2851>.
54. So EY, Ouchi T. The application of Toll like receptors for cancer therapy. *Int J Biol Sci.* 2010;6(7):675–81. <https://doi.org/10.7150/ijbs.6.675>.
55. Frazao JB, Errante PR, Condino-Neto A. Toll-like receptors' pathway disturbances are associated with increased susceptibility to infections in humans. *Arch Immunol Ther Exp.* 2013;61(6):427–43. <https://doi.org/10.1007/s00005-013-0243-0>.
56. Zhou Q, Zhu K, Cheng H. Toll-like receptors in human papillomavirus infection. *Arch Immunol Ther Exp.* 2013;61(3):203–15. <https://doi.org/10.1007/s00005-013-0220-7>.
57. Jensen S, Thomsen AR. Sensing of RNA viruses: a review of innate immune receptors involved in recognizing RNA virus invasion. *J Virol.* 2012;86(6):2900–10. <https://doi.org/10.1128/JVI.05738-11>.
58. Hayashi F, Smith KD, Ozinsky A, Hawn TR, Yi EC, Goodlett DR, et al. The innate immune response to bacterial flagellin is mediated by toll-like receptor 5. *Nature.* 2001;410(6832):1099–103. <https://doi.org/10.1038/35074106>.
59. Basith S, Manavalan B, Yoo TH, Kim SG, Choi S. Roles of Toll-like receptors in cancer: a double-edged sword for defense and offense. *Arch Pharm Res.* 2012;35(8):1297–316. <https://doi.org/10.1007/s12272-012-0802-7>.
60. Ellerman JE, Brown CK, de Vera M, Zeh HJ, Billiar T, Rubartelli A, et al. Masquerader: high mobility group Box-1 and cancer. *Clin Cancer Res.* 2007;13(10):2836–48. <https://doi.org/10.1158/1078-0432.CCR-06-1953>.
61. Goutagny N, Estornes Y, Hasan U, Lebecque S, Caux C. Targeting pattern recognition receptors in cancer immunotherapy. *Target Oncol.* 2012;7(1):29–54. <https://doi.org/10.1007/s11523-012-0213-1>.
62. Kim YK, Shin J-S, Nahm MH. Nod-like receptors in infection, immunity, and diseases. *Yonsei Med J.* 2016;57(1):5–14. <https://doi.org/10.3349/ymj.2016.57.1.5>.
63. Eisenbarth SC, Williams A, Colegio OR, Meng H, Strowig T, Rongvaux A, et al. NLRP10 is a NOD-like receptor essential to initiate adaptive immunity by dendritic cells. *Nature.* 2012;484:510–3. <https://doi.org/10.1038/nature11012>.
64. Motta V, Soares F, Sun T, Philpott DJ. Nod-like receptors: versatile cytosolic sentinels. *Physiol Rev.* 2015;95(1):149–78. <https://doi.org/10.1152/physrev.00009.2014>.
65. Man SM, Kanneganti T-D. Regulation of inflammasome activation. *Immunol Rev.* 2015;265(1):6–21. <https://doi.org/10.1111/imr.12296>.
66. Sun Q, Fan J, Billiar TR, Scott MJ. Inflammasome and autophagy regulation: a two-way street. *Mol Med.* 2017;23:188–95. <https://doi.org/10.2119/molmed.2017.00077>.
67. Anand PK, Malireddi RKS, Lukens JR, Vogel P, Bertin J, Lamkanfi M, et al. NLRP6 negatively regulates innate immunity and host defence against bacterial pathogens. *Nature.* 2012 Aug;488:389–93. <https://doi.org/10.1038/nature11250>.

68. Yan A, Farmer E, Wu TC, Hung CF. Perspectives for therapeutic HPV vaccine development. *J Biomed Sci.* 2016;23(1):75. <https://doi.org/10.1186/s12929-016-0293-9>.
69. Morrone SR, Matyszewski M, Yu X, Delannoy M, Egelman EH, Son J. Assembly-driven activation of the AIM2 foreign-dsDNA sensor provides a polymerization template for downstream ASC. *Nat Commun.* 2015;6(7827):1–13. <https://doi.org/10.1038/ncomms8827>.
70. Khare S, Ratsimandresy RA, de Almeida L, Cuda CM, Rellick SL, Misharin AB, et al. The pyrin domain-only protein POP3 inhibits ALR inflammasomes and regulates responses to infection with DNA viruses. *Nat Immunol.* 2014;15(4):343–53. <https://doi.org/10.1038/ni.2829>.
71. Reinholtz M, Kawakami Y, Salzer S, Kreuter A, Dombrowski Y, Koglin S, et al. HPV16 activates the AIM2 inflammasome in keratinocytes. *Arch Dermatol Res.* 2013;305(8):723–32. <https://doi.org/10.1007/s00403-013-1375-0>.
72. Uniprot.org. UniProt: a hub for protein information. *Nucleic Acids Res.* 43: D204–D212. 2015. <http://www.uniprot.org/uniprot/?query=tlr+&sort=score>. Accessed 5 Jan 2015.
73. Akira S, Takeda K. Toll-like receptor signalling. *Nat Rev Immunol.* 2004;4(7):499–511. <https://doi.org/10.1038/nri1391>.
74. O'Neill LAJ, Golenbock D, Bowie AG. The history of Toll-like receptors—redefining innate immunity. *Nat Rev Immunol.* 2013;13(6):453–60. <https://doi.org/10.1038/nri3446>.
75. Lim KH, Staudt LM. Toll-like receptor signaling. *Cold Spring Harb Perspect Biol.* 2013;5(1):a011247. <https://doi.org/10.1101/cshperspect.a011247>.
76. Amador-Molina A, Hernández-Valencia JF, Lamoyi E, Contreras-Paredes A, Lizano M. Role of innate immunity against human papillomavirus (HPV) infections and effect of adjuvants in promoting specific immune response. *Viruses.* 2013;5(11):2624–42. <https://doi.org/10.3390/v5112624>.
77. Daud II, Scott ME, Ma Y, Shibuski S, Farhat S, Moscicki AB. Association between toll-like receptor expression and human papillomavirus type 16 persistence. *Int J Cancer.* 2011;128(4):879–86. <https://doi.org/10.1002/ijc.25400>.
78. Hasimu A, Ge L, Li QZ, Zhang RP, Guo X. Expressions of toll-like receptors 3, 4, 7, and 9 in cervical lesions and their correlation with HPV16 infection in Uighur women. *Chin J Cancer.* 2011;30(5):344–50. <https://doi.org/10.5732/cjc.010.10456>.
79. Thompson MR, Kaminski JJ, Kirt-Jones EA, Fitzgerald KA. Pattern recognition receptors and the innate immune response to viral infection. *Viruses.* 2011 Jun;3(6):920–40. <https://doi.org/10.3390/v3060920>.
80. Kanzler H, Barrat FJ, Hessel M, Coffman RL. Therapeutic targeting of innate immunity with Toll-like receptor agonist and antagonist. *Nat Med.* 2007;13(5):552–9. <https://doi.org/10.1038/nm1589>.
81. Kawai T, Akira S. The role of pattern-recognition receptors in innate immunity: update on Toll-like receptors. *Nat Immunol.* 2010;11(5):373–84. <https://doi.org/10.1038/ni.1863>.
82. Bonjardim CA. Interferons (IFNs) are key cytokines in both innate and adaptive antiviral immune responses and viruses counteract IFN action. *Microbes Infect.* 2005;7(3):569–78. <https://doi.org/10.1016/j.micinf.2005.02.001>.
83. Hennessy E, Parker AE, O'Neill LAJ. Targeting Toll-like receptors: emerging therapeutics? *Nat Rev Drug Discov.* 2010;9(4):293–307. <https://doi.org/10.1038/nrd3203>.
84. Yarovinsky F. Innate immunity to *Toxoplasma gondii* infection. *Nat Rev Immunol.* 2014;14(2):109–21. <https://doi.org/10.1038/nri3598>.
85. Barton GM, Kagan JC. A cell biological view of Toll-like receptor function: regulation through compartmentalization. *Nat Rev Immunol.* 2009;9(8):535–42. <https://doi.org/10.1038/nri2587>.
86. Morrison DK. MAP kinase pathways. *Cold Spring Harb Perspect Biol.* 2012;4:a011254. <https://doi.org/10.1101/cshperspect.a011254>.
87. Häcker H, Tseng PH, Karin M. Expanding TRAF function: TRAF3 as a tri-faced immune regulator. *Nat Rev Immunol.* 2011;11(7):457–68. <https://doi.org/10.1038/nri2998>.

88. Hirsch I, Caux C, Hasan U, Bendriss-Vermare N, Olive D. Impaired Toll-like receptor 7 and 9 signaling: from chronic viral infections to cancer. *Trends Immunol.* 2010;31(10):391–7. <https://doi.org/10.1016/j.it.2010.07.004>.
89. Lam LT, Wright G, Davis E, Lenz G, Farinha P, Dang L, et al. Cooperative signaling through the signal transducer and activator of transcription 3 and nuclear factor- $\kappa$ B pathways in subtypes of diffuse large B-cell lymphoma. *Blood.* 2008;111(7):3701–13. <https://doi.org/10.1182/blood-2007-09-111948>.
90. Ngo V, Young RM, Schmitz R, Jhavar S, Xiao W, Lim KH, et al. Oncogenically active MYD88 mutations in human lymphoma. *Nature.* 2011;470(7332):115–21. <https://doi.org/10.1038/nature09671>.
91. American Cancer Society (ACS). Cancer immunotherapy. 2018. <https://www.cancer.gov/>. Accessed 28 May 2018.
92. Goazigo AR, Steenwinckel JV, Rostène W. Current status of chemokines in the adult CNS. *Prog Neurobiol.* 2013;104:67–92. <https://doi.org/10.1016/j.pneurobio.2013.02.001>.
93. Hinck AP. Structural studies of the TGF- $\beta$ s and their receptors – insights into evolution of the TGF- $\beta$  superfamily. *FEBS Lett.* 2012;586(14):1860–70. <https://doi.org/10.1016/j.febslet.2012.05.028>.
94. Lata S, Raghava GP. Prediction and classification of chemokines and their receptors. *Protein Eng Des Sel.* 2009;22(7):441–4. <https://doi.org/10.1093/protein/gzp016>.
95. Turner MD, Medjai B, Hurst T, Pennington DJ. Cytokines and chemokines: at the crossroads of cells signalling and inflammatory disease. *Biochim Biophys Acta.* 2014;1843(11):2563–82. <https://doi.org/10.1016/j.bbamcr.2014.05.014>.
96. Shaikh PZ. Cytokines & their physiologic and pharmacologic functions in inflammation: a review. *Int J Pharm Life Sci.* 2011;2(11):1247–63.
97. Akdis M, Burgler S, Crameri R, Eiwegger T, Fujita H, Gomez E, et al. Interleukins, from 1 to 37, and interferon- $\gamma$ : receptors, functions, and roles in diseases. *J Allergy Clin Immunol.* 2011;127(3):701–721.e1-70. <https://doi.org/10.1016/j.jaci.2010.11.050>.
98. Rosa MI, Morales MV, Vuolo F, Petronilho F, Bozzetti MC, Medeiros LR, et al. Association of interleukin-6 in women with persistence of DNA-HPV: a nested case-control study. *Arch Gynecol Obstet.* 2012;285(1):143–8. <https://doi.org/10.1007/s00404-011-1925-7>.
99. Fernandes APM, Goncalves MAG, Duarte G, Cunha FQ, Simoes RT, Donadi EA. HPV16, HPV18, and HIV infection may be influence cervical cytokine intraleisional levels. *Virology.* 2005;334(2):294–8. <https://doi.org/10.1016/j.virol.2005.01.029>.
100. Bohnhorst J, Rasmussen T, Moen SH, Flottum M, Knudsen L, Borset M, et al. Toll-like receptors mediate proliferation and survival of multiple myeloma cells. *Leukemia.* 2006;20(6):1138–44. <https://doi.org/10.1038/sj.leu.2404225>.
101. Jego G, Bataille R, Geffroy-Luseau A, Descamps G, Pellat-Deceunynck C. Pathogen-associated molecular patterns are growth and survival factors for human myeloma cells through Toll-like receptors. *Leukemia.* 2006;20(6):1130–7. <https://doi.org/10.1038/sj.leu.2404226>.
102. Lippitz B. Cytokine patterns in patients with cancer: a systematic review. *Lancet Oncol.* 2013;14(6):e218–28. [https://doi.org/10.1016/S1470-2045\(12\)70582-X](https://doi.org/10.1016/S1470-2045(12)70582-X).
103. Padhan K, Varma R. Immunological synapse: a multi-protein signalling cellular apparatus for controlling gene expression. *Immunology.* 2010;129(3):322–8. <https://doi.org/10.1111/j.1365-2567.2009.03241.x>.
104. Li X, Jiang S, Tapping RI. Toll-like receptor in cell proliferation and survival. *Cytokine.* 2010;49(1):1–9. <https://doi.org/10.1016/j.cyto.2009.08.010>.
105. Macian F. NFAT proteins: key regulators of T-cell development and function. *Nat Rev Immunol.* 2005;5(6):472–84. <https://doi.org/10.1038/nri1632>.
106. Hu F, Meng Y, Gou L, Zhang X. Analysis of promoters and CREB/AP-1 binding sites of the human TMEM174 gene. *Exp Ther Med.* 2013;6(5):1290–4. <https://doi.org/10.3892/etm.2013.1275>.

107. Wen AY, Sakamoto KM, Miller LS. The role of the transcription factor CREB in immune function. *J Immunol.* 2010;185(11):6413–9. <https://doi.org/10.4049/jimmunol.1001829>.
108. Oeckinghaus A, Hayden MS, Ghosh S. Crosstalk in NF- $\kappa$ B signalling pathways. *Nat Immunol.* 2011;12(8):695–708. <https://doi.org/10.1038/ni.2065>.
109. Hoesel B, Schmid JA. The complexity of NF- $\kappa$ B signaling in inflammation and cancer. *Mol Cancer.* 2013;12:86. <https://doi.org/10.1186/1476-4598-12-86>.
110. Huang TT, Wuerzberger-Davis SM, Wu ZH, Miyamoto S. Sequential modification of NEMO/IKKgamma by SUMO-1 and ubiquitin mediates NF- $\kappa$ pA activation by genotoxic stress. *Cell.* 2003;115(5):565–76. [https://doi.org/10.1016/S0092-8674\(03\)00895-X](https://doi.org/10.1016/S0092-8674(03)00895-X).
111. Yuan H, Fu F, Zhuo J, Wang W, Nishitani J, An DS, et al. Human papillomavirus type 16 E6 and E7 oncoproteins upregulate c-IAP2 gene expression and confer resistance to apoptosis. *Oncogene.* 2005;24:5069–78. <https://doi.org/10.1038/sj.onc.1208691>.
112. Snow WM, Stoesz BM, Kelly DM, Albensi BC. Roles for NF- $\kappa$ B and gene targets of NF- $\kappa$ B in synaptic plasticity, memory, and navigation. *Mol Neurobiol.* 2014;49(2):757–70. <https://doi.org/10.1007/s12035-013-8555-y>.
113. Micheau O, Tschoch J. Induction of TNF receptor I-mediated apoptosis via two sequential signaling complexes. *Cell.* 2003;114(2):181–90. [https://doi.org/10.1016/S0092-8674\(03\)00521-X](https://doi.org/10.1016/S0092-8674(03)00521-X).
114. Sen R. The origins of NF- $\kappa$ B. *Nat Immunol.* 2011;12:686–8. <https://doi.org/10.1038/ni.2071>.
115. Bassères DS, Baldwin AS. Nuclear factor- $\kappa$ pA and inhibitor of  $\kappa$ pA kinase pathways in oncogenic initiation and progression. *Oncogene.* 2006;25(51):6817–30. <https://doi.org/10.1038/sj.onc.1209942>.
116. Spitzkovsky D, Hehner SP, Hofmann TG, Möller A, Schmitz ML. The human papillomavirus oncoprotein E7 attenuates NF- $\kappa$ B activation by targeting the I $\kappa$ B kinase complex. *J Biol Chem.* 2002;277(28):25576–82. <https://doi.org/10.1074/jbc.M201884200>.
117. Pradeu T, Kostyrka G, Dupré J. Understanding viruses: philosophical investigation. *Studies History Philo Biolog Biomed Sci.* 2016;59:57–63. <https://doi.org/10.1016/j.shpsc.2016.02.008>.
118. Gibbs AJ, Gibbs MJ. A broader definition of ‘the virus species’ brief report. *Arch Virol.* 2006;151(7):1419–22. <https://doi.org/10.1007/s00705-006-0775-2>.
119. International Committee on Taxonomy of Viruses (ICTV). ICTV Taxonomy. 2018. <https://talk.ictvonline.org/taxonomy/w/ictv-taxonomy>. Accessed 29 Mar 2018.
120. Morgan GJ. What is a virus species? Radical pluralism in viral taxonomy. *Stud Hist Phil Biol Biomed Sci.* 2016;59:64–70. <https://doi.org/10.1016/j.shpsc.2016.02.009>.
121. Calisher CH. The taxonomy of viruses should include viruses. *Arch Virol.* 2016;161(5):1419–22. <https://doi.org/10.1007/s00705-016-2779-x>.
122. McLaughlin-Drubin ME, Munger K. Viruses associated with human cancer. *Biochim Biophys Acta.* 2008;1782(3):127–50. <https://doi.org/10.1016/j.bbadi.2007.12.005>.
123. Ahuja R, Jamal A, Nosrati N, Pandley V, Rajput P, Saxena N, et al. Human oncogenic viruses and cancer. *Curr Sci.* 2014;107(5):768–85.
124. Santiago DN, Heidbuechel JPW, Kandell WM, Walker R, Djeu J, Engeland CE, et al. Fighting cancer with mathematics and viruses. *Viruses.* 2017;9(9):239. <https://doi.org/10.3390/v9090239>.
125. Flippot R, Malouf GG, Su X, Khayat D, Spano J-P. Oncogenic viruses: lessons learned using next-generation sequencing technologies. *Eur J Cancer.* 2016;61:61–8. <https://doi.org/10.1016/j.ejca.2016.03.086>.
126. Moore PS, Chang Y. Why do viruses cause cancer? Highlights of the first century of human tumour virology. *Nat Rev Cancer.* 2010;10(12):878–89. <https://doi.org/10.1038/nrc2961>.
127. Nomaguchi M, Fujita M, Miyazaki Y, Adachi A. Viral Tropism. *Front Microbiol.* 2012;3:281. <https://doi.org/10.3389/fmicb.2012.00281>.
128. Crow MS, Javitt A, Cristea LLM. A proteomics perspective on viral DNA sensors in host defense and viral immune evasion mechanisms. *J Mol Biol.* 2015;427(11):1995–2012. <https://doi.org/10.1016/j.jmb.2015.02.016>.

129. Law GL, Korth MJ, Benecke AG, Katze MG. Systems virology: host-directed approaches to viral pathogenesis and drug targeting. *Nat Rev Microbiol.* 2013;11(7):455–66. <https://doi.org/10.1038/nrmicro3036>.
130. Schäfer G, Blumenthal MJ, Katz AA. Interaction of human tumor viruses with host cell surface receptors and cell entry. *Viruses.* 2015;7(5):2592–617. <https://doi.org/10.3390/v7052592>.
131. Nowak MA, May RM. *Viral dynamics: mathematical principles of immunology and virology.* Oxford, UK: Oxford University Press; 2001. p. 2–3. ISBN 0-19-850417-9.
132. Van Regenmortel MHV. The metaphor that viruses are living is alive and well, but it is no more than a metaphor. *Stud Hist Phil Biol Biomed Sci.* 2016;59:117–24. <https://doi.org/10.1016/j.shpsc.2016.02.017>.
133. Lin CZ, Xiang GL, Zhu XH, Xiu LL, Sun JX, Zhang XY. Advances in the mechanisms of action of cancer-targeting oncolytic viruses. *Oncol Lett.* 2018;15(4):4053–60. <https://doi.org/10.3892/ol.2018.7829>.
134. Kuss-Duerkop SK, Westrich JA, Pyeon D. DNA tumor virus replication of host DNA methylation and its implications for immune evasion and oncogenesis. *Viruses.* 2018;10(2):E82. <https://doi.org/10.3390/v10020082>.
135. Jiang M, Imperiale MJ. Design starts: how small DNA viruses remodel the host nucleus. *Futur Virol.* 2012;7(5):445–59. <https://doi.org/10.2217/FVL.12.38>.
136. Turnell AS, Grand RJ. DNA viruses and the cellular DNA-damage response. *J Gen Virol.* 2012;93(Pt 10):2076–97. <https://doi.org/10.1099/vir.0.044412-0>.
137. Saldivar JC, Cortez D, Cimprich KA. The essential kinase ATR: ensuring faithful duplication of a challenging genome. *Nat Rev Mol Cell Biol.* 2017;18(10):622–36. <https://doi.org/10.1038/nrm.2017.67>.
138. Zhao J, Dang X, Zhang P, Nguyen LN, Cao D, Wang L, et al. Insufficiency of DNA repair enzyme ATM promotes naïve CD4 T-cell loss in chronic hepatitis C virus infection. *Cell Discovery.* 2018;4:16. <https://doi.org/10.1038/s41421-018-0015-4>.
139. American Cancer Society (ACS). *Cancer prevention and early detection facts and figures 2017–2017.* Atlanta: American Cancer Society; 2017.
140. Seeger C, Mason WS. Molecular biology of hepatitis B virus infection. *Virology.* 2015;479–480:672–86. <https://doi.org/10.1016/j.virol.2015.02.031>.
141. Lamontagne RJ, Bagga S, Bouchard MJ. Hepatitis B virus molecular biology and pathogenesis. *Hepatoma Res.* 2016;2:163–86. <https://doi.org/10.20517/2394-5079.2016.05>.
142. Xu W, Yu J, Wong VW-S. Mechanism and predictions of HCC development in HBV infection. *Best Pract Res Clin Gastroenterol.* 2017;31(3):291–8. <https://doi.org/10.1016/j.bpg.2017.04.011>.
143. Levrero M, Zucman-Rossi J. Mechanism of HBV-induced hepatocellular carcinoma. *J Hepatol.* 2016;64(1 Suppl):S84–S101. <https://doi.org/10.1016/j.jhep.2016.02.021>.
144. De Villiers EM, Fauquet C, Broker TR, Bernard HU, zur Hausen H. Classification of papillomaviruses. *Virology.* 2004;324(1):17–27. <https://doi.org/10.1016/j.virol.2004.03.033>.
145. Picconi MA, Alonio LV, García-Carrancá A, Lizano M, Cervantes-Vazquez G, Distefano AL, et al. Molecular variants of human papillomavirus (HPV) types 16 and 18 in adenocarcinomas of the cervix. *Medicina (Buenos Aires).* 2000;60(6):889–94.
146. Travasso CM, Anand M, Samarth M, Deshpande A, Kumar-Sinha C. Human papillomavirus genotyping by multiplex pyrosequencing in cervical cancer patients from India. *J Biosci.* 2008;33(1):73–80.
147. Cal CM. El virus del papiloma humano. *Cadernos de Atención Primaria.* 2008;15(1):72–4.
148. Goering RV. Molecular epidemiology of nosocomial infection: analysis of chromosomal restriction fragment patterns by pulsed-field gel electrophoresis. *Infect Control Hosp Epidemiol.* 1993;14(10):595–600.
149. Chan SY, Bernard HU, Ratterree M, Birkebak TA, Faras AJ, Ostrow RS. Genomic diversity and evolution of papillomaviruses in rhesus monkeys. *J Virol.* 1997;71(7):4938–43.

150. Vanegas VA, Rubio AI, Bedoya AM, Sánchez GI. Estructura molecular y antigenética de la vacuna contra el virus de papiloma humano 16 (VPH 16). *Acta Biológica Colombiana.* 2008;13(3):37–48.
151. Yamada T, Wheeler CM, Halpern AL, Stewart AC, Hildesheim A, Jenison SA. Human papillomavirus type 16 variant lineages in United States populations characterized by nucleotide sequence analysis of the E6, L2, and L1 coding segments. *J Virol.* 1995;69(12):7743–53.
152. Yamada T, Manos MM, Peto J, Greer CE, Munoz N, Bosch FX, Wheeler CM. Human papillomavirus type 16 sequence variation in cervical cancer: a worldwide perspective. *J Virol.* 1997;71(3):2463–72.
153. Cornet I, Gheit T, Franceschi S, Vignat J, Burk RD, Sylla BS, et al. Human papillomavirus type 16 genetic variants: phylogeny and classification based on E6 and LCR. *J Virol.* 2012;86(12):6855–61. <https://doi.org/10.1128/JVI.00483-12>.
154. Taylor ER, Morgan IM. A novel technique with enhanced detection and quantitation of HPV-16 E1- and E2-mediated DNA replication. *Virology.* 2003;315(1):103–9. [https://doi.org/10.1016/S0042-6822\(03\)00588-9](https://doi.org/10.1016/S0042-6822(03)00588-9).
155. Okoye A, Cordano P, Taylor ER, Morgan IM, Everett R, Campo MS. Human papillomavirus 16 L2 inhibits the transcriptional activation function, but not the DNA replication function, of HPV-16 E2. *Virus Res.* 2005;108(1–2):1–14. <https://doi.org/10.1016/j.virusres.2004.07.004>.
156. International Agency for Research on Cancer (IARC). IARC monographs on the evaluation of carcinogenesis risks to humans. vol. 100B. IARC Press 2012. ISBN 978 92 832 1319 2. ISSN 1017-1606. Lyon, France.
157. Flores ER, Allen-Hoffman BL, Lee D, Sattler CA, Lambert PF. Establishment of the human papillomavirus type 16 (HPV-16) life cycle in an immortalized human foreskin keratinocyte cell line. *Virology.* 1999;262(2):344–54. <https://doi.org/10.1006/viro.1999.9868>.
158. Frazer IH. Prevention of cervical cancer through papillomavirus vaccination. *Nat Rev Immunol.* 2004;4(1):46–55. <https://doi.org/10.1038/nri1260>.
159. Polyomaviridae Study Group of the International on Taxonomy of viruses, Calvignac-Spencer S, Feltkamp MCW, Dauherty MD, Moens U, Ramqvist T, et al. A taxonomy update for the family polyomaviridae. *Arch Virol.* 2016;161(6):1739–50. <https://doi.org/10.1007/s00705-016-2794-y>.
160. Liu W, MacDonald M, You J. Merkel cell polyomavirus infection and Merkel cell carcinoma. *Curr Opin Virol.* 2016;20:20–7. <https://doi.org/10.1016/j.coviro.2016.07.011>.
161. Bhart H, Solis M, Kack-Kack W, Soulier E, Velay A, Fafi-Kremer S. In vitro and in vivo models for the study of human polyomavirus infection. *Viruses.* 2016;8(10):292. <https://doi.org/10.3390/v8100292>.
162. Wendzicki JA, Moore PS, Chang Y. Large T and small T antigens of Merkel cell polyomavirus. *Curr Opin Virol.* 2015;11:38–43. <https://doi.org/10.1016/j.coviro.2015.01.009>.
163. Spurgeon ME, Lambert PF. Merkel cell polyomavirus: a newly discovered human virus. *Virology.* 2013;435(1):118–30. <https://doi.org/10.1016/j.virol.2012.09.029>.
164. Cook DL, Frieling GW. Merkel cell carcinoma: a review and update on current concepts. *Diagn Histopathol.* 2016;22(4):127–33. <https://doi.org/10.1016/j.mpdhp.2016.04.002>.
165. Van der Meijden E, Kazem S, Dargel CA, Vuren NV, Hensbergen PJ, MCW F. Characterization of T antigens, including middle T and alternative T, expressed by the human polyomavirus associated with trichodysplasia spinulosa. *J Virol.* 2015;89(18):9427–39. <https://doi.org/10.1128/JVI.00911-15>.
166. Shuda M, Kwun HJ, Feng H, Chang Y, Moore P. Human Merkel cell polyomavirus small T antigen is an oncoprotein targeting the 4E-BP1 translation regulator. *J Clin Invest.* 2011;121(9):3623–34. <https://doi.org/10.1172/JCI46323>.
167. Esau D. Viral causes of lymphoma: the history of Epstein-Barr virus and human T-lymphotropic virus 1. *Virology.* 2017;8:1–5. <https://doi.org/10.1177/1178122X17731772>.
168. International Agency for Research on Cancer (IARC). Section of infections – infections and cancer biology group. 2018a. <http://www.iarc.fr/en/research-groups/ICB/index.php>. Accessed 27 Mar 2018.

169. Feng H, Shuda M, Chang Y, Moore PS. Clonal integration of a polyomavirus in human Merkel cell carcinoma. *Science.* 2008;319(5866):1096–100. <https://doi.org/10.1126/science.1152586>.
170. Kassem A, Schöpflin A, Diaz C, Weyers W, Stickeler E, Werner M, et al. Frequent detection of Merkel cell polyomavirus in human Merkel cell carcinomas and identification of a unique deletion in the VP1 gene. *Cancer Res.* 2008;68(13):5009–13. <https://doi.org/10.1158/0008-5472.CAN-08-0949>.
171. Shuda M, Feng H, Kwun HJ, Rosen ST, Ghoerup O, Moore PS, et al. T antigen mutations are a human tumor-specific signature for Merkel cell polyomavirus. *Proc Natl Acad Sci U S A.* 2008;105(42):16272–7. <https://doi.org/10.1073/pnas.0806526105>.
172. Moore PS, Chang Y. Common commensal cancer viruses. *PLoS Pathog.* 2017;13(1):e1006078. <https://doi.org/10.1371/journal.ppat.1006078>.
173. Mizuguchi Y, Takizawa T, Yoshida H, Uchida E. Dysregulated miRNA in progression of hepatocellular carcinoma; a systematic review. *Hepatol Res.* 2016;46(5):391–406. <https://doi.org/10.1111/hepr.12606>.
174. Yue D, Zhang Y, Cheng L, Ma J, Xi Y, Yang L, et al. Hepatitis B virus X protein (HBx) induced abnormalities of nucleic acid metabolism revealed by H-NMR-based metabolomics. *Sci Rep.* 2016;6(24430):1–13. <https://doi.org/10.1038/srep24430>.
175. World Health Organization (WHO). Weekly epidemiological record. 2014;43(89):465–92. ISSN 0049-8114.
176. Kim R, Emi M, Tanabe K. Cancer immunoediting from immune surveillance to immune escape. *Immunology.* 2007;121(1):1–14. <https://doi.org/10.1111/j.1365-2567.2007.02587.x>.
177. Yuzhalin A, Kutikhin A. Interleukins in cancer biology: their heterogeneous role. Chapter 1–10, edited by Arseniy E. Yuzhalin Anton G. Kutikhin, Academic Press, Amsterdam, 2015. ISBN 9780128011218.
178. Read SA, Douglas MW. Virus induced inflammation and cancer development. *Cancer Lett.* 2014;345(2):174–81. <https://doi.org/10.1016/j.canlet.2013.07.030>.
179. Torres-Poveda K, Bahena-Román M, Madrid-González C, Burguete-García AI, Bermúdez-Morales VH, Peralta-Zaragoza O, et al. Role of IL-10 and TGF- $\beta$ 1 in local immunosuppression in HPV-associated cervical neoplasia. *World J Clin Oncol.* 2014;5(4):753–63. <https://doi.org/10.5306/wjco.v5.i4.753>.
180. Jeon S, Allen-Hoffmann BL, Lambert PF. Integration of human papillomavirus type 16 into the human genome correlates with a selective growth advantage of cells. *J Virol.* 1995;69(5):2989–97.
181. Zhou F, Leggatt GR, Frazer IH. Human papillomavirus 16 E7 protein inhibits interferon- $\gamma$ -mediated enhancement of keratinocytes antigen processing and T-cell lysis. *FEBS J.* 2011;278(6):955–63. <https://doi.org/10.1111/j.1742-4658.2011.08011.x>.
182. Ren C, Cheng X, Lu B, Yang G. Activation of interleukin-6/signal transducer and activator of transcription 3 by human papillomavirus early proteins 6 induces fibroblast senescence to promote cervical tumourigenesis through autocrine and paracrine pathways in tumour microenvironment. *Eur J Cancer.* 2013;49(18):3889–99. <https://doi.org/10.1016/j.ejca.2013.07.140>.
183. Song D, Li H, Li H, Dai J. Effect of human papillomavirus infection on the immune system and its role in the course of cervical cancer. *Oncol Lett.* 2015;10(2):600–6. <https://doi.org/10.3892/ol.2015.3295>.
184. Aggarwal R, Misra S, Guleria C, Suri V, Mangat N, Sharma M, et al. Characterization of toll-like receptor transcriptome in squamous cell carcinoma of cervix: a case-control study. *Gynecol Oncol.* 2015;138(2):358–62. <https://doi.org/10.1016/j.ygyno.2015.05.029>.
185. Bhat P, Mattarollo SR, Gosmann C, Frazer IH, Leggatt GR. Regulation of immune responses to HPV infection and during HPV-directed immunotherapy. *Immunol Rev.* 2011;239(1):85–98. <https://doi.org/10.1111/j.1600-065X.2010.00966.x>.

186. Bermúdez-Morales VH, Peralta-Zaragoza O, Alcocer-González JM, Moreno J, Madrid-Marina V. IL-10 expression is regulated by HPV E2 protein in cervical cancer cells. *Mol Med Rep.* 2011;4(2):369–75. <https://doi.org/10.3892/mmr.2011.429>.
187. Conesa-Zamora P. Immune responses against virus and tumor in cervical carcinogenesis: treatment strategies for avoiding the HPV-induced immune escape. *Gynecol Oncol.* 2013;131(2):480–8. <https://doi.org/10.1016/j.ygyno.2013.08.025>.
188. Song SH, Lee JK, Seok OS, Saw HS. The relationship between cytokines and HPV-16, HPV-16 E6, E7, and high-risk HPV viral load in the uterine cervix. *Gynecol Oncol.* 2007;104(3):732–8. <https://doi.org/10.1016/j.ygyno.2006.10.054>.
189. Vandermark ER, Deluca KA, Gardner CR, Marker DF, Schreiner CN, Strickland DA, et al. Human papillomavirus type 16 E6 and E7 proteins alter NF- $\kappa$ B in cultured cervical epithelial cells and inhibition of NF- $\kappa$ B promotes cell growth and immortalization. *Virology.* 2012;425(1):53–60. <https://doi.org/10.1016/j.virol.2011.12.023>.
190. Houben R, Angermeyer S, Haferkamp S, Aue A, Goebeler M, Schrama D, et al. Characterization of functional domains in the Merkel cell polyoma virus large T antigen. *Int J Cancer.* 2015;136(5):E290–300. <https://doi.org/10.1002/ijc.29200>.
191. Sauer CM, Haugg AM, Chteinberg E, Rennspies D, Winneperenninckx V, Speel E-J, et al. Reviewing the current evidence supporting early B-cells as the cellular origin of Merkel cell carcinoma. *Crit Rev Oncol Hematol.* 2017;116:99–105. <https://doi.org/10.1016/j.critrevonc>.
192. Van der Meijden E, Feltkamp M. The human polyomavirus middle and alternative T-antigens: thoughts on roles and relevance to cancer. *Front Microbiol.* 2018;9:398. <https://doi.org/10.3389/fmicb.2018.00398>.
193. Wang RF, Wang H. Immune targets and neoantigens for cancer immunotherapy and precision medicine. *Cell Res.* 2017;27(1):11–37. <https://doi.org/10.1038/cr.2016.155>.
194. Obeid JM, Hu Y, Slingluff CL Jr. Vaccines, adjuvants and dendritic cell activators – current status and futures challenges. *Semin Oncol.* 2015;42(4):549–61. <https://doi.org/10.1053/j.seminoncol.2015.05.006>.
195. Ophir E, Bobisse S, Coukos G, Harari A, Kandalaft LE. Personalized approaches to active immunotherapy in cancer. *Biochim Biophys Acta.* 2016;1865(1):72–82. <https://doi.org/10.1016/j.bbcan.2015.07.004>.
196. Song Q, Zhang C-D, Wu X-H. Therapeutic cancer vaccines: from initial findings to prospects. *Immunol Lett.* 2018;196:11–21. <https://doi.org/10.1016/j.imlet.2018.01.011>.
197. Zamarin D, Postow MA. Immune checkpoint modulation: rationale design of combination strategies. *Pharmacol Ther.* 2015;150:23–32. <https://doi.org/10.1016/j.pharmthera.2015.01.003>.
198. Whiteside TL. Inhibiting the inhibitors: evaluating agents targeting cancer immunosuppression. *Expert Opin Biol Ther.* 2010;10(7):1019–35. <https://doi.org/10.1517/1471259.8.2010.48220>.
199. Handisurya A, Lázár S, Papay P, Primas C, Haitel A, Horvat R, et al. Anogenital human papillomavirus prevalence is unaffected by therapeutic tumour necrosis factor-alpha inhibition. *Acta Derm Venereol.* 2016;96(4):494–8. <https://doi.org/10.2340/00015555-2298>.
200. Werberich GM, Strava T, Vizioli C, Fernandes GDS. Human papillomavirus-induced cancer: late relapse in a patient treated with tumor necrosis factor-alpha inhibitor. *J Global Oncol.* 2016;3(3):275–7. <https://doi.org/10.1200/JGO.2016.005835>.
201. Neuzillet C, Tijeras-Raballand A, Cohen R, Cros J, Faivre S, Raymond E, et al. Targeting the TGF $\beta$  pathway for cancer therapy. *Pharmacol Ther.* 2015;147:22–31. <https://doi.org/10.1016/j.pharmthera.2014.11.001>.
202. Takaoka A, Hayakawa S, Yanai H, Stoiber D, Negishi H, Kikuchi H, et al. Integration of interferon-alpha/beta signalling to p53 responses in tumour suppression and antiviral defence. *Nature.* 2003;424(6948):516–23. <https://doi.org/10.1038/nature01850>.
203. DeCarlo CA, Severini A, Edler L, Escott NG, Lambert PF, Ulanova M, et al. IFN- $\kappa$ , a novel type I IFN, is undetectable in HPV-positive human cervical keratinocytes. *Lab Investig.* 2010;90(10):1482–91. <https://doi.org/10.1038/labinvest.2010.95>.

204. Cancer Research UK (CRUK). Other treatments. 2018. <http://www.cancerresearchuk.org/about-cancer/cancer-in-general/treatment/other>. Accessed 30 May 2018.
205. Abdo J, Cornell DL, Mittal SK, Agrawal DK. Immunotherapy plus cryotherapy: potential augmented abscopal effect for advanced cancers. *Front Oncol.* 2018;8(85):1–16. <https://doi.org/10.3389/fonc.2018.00085>.
206. Russell SJ, Peng K-W, Bell JC. Oncolytic virotherapy. *Nat Biotechnol.* 2012;30(7):658–70. <https://doi.org/10.1038/nbt.2287>.
207. Chaurasiya S, Chen NG, Warner SG. Oncolytic virotherapy versus cancer stem cells: a review of approaches and mechanisms. *Cancers.* 2018;10(4):E124. <https://doi.org/10.3390/cancers10040124>.
208. Liu Y, Sethi MS, Hinoue T, Schneider BG, Cherniack AD, Sanchez-Vega F, et al. Comparative molecular analysis of gastrointestinal adenocarcinomas. *Cancer Cell.* 2018;33(4):721–35. <https://doi.org/10.1016/j.ccr.2018.03.010>.
209. Saleh T, Shojaosadati SA. Multifunctional nanoparticles for cancer immunotherapy. *Hum Vaccin Immunother.* 2016;12(7):1863–75. <https://doi.org/10.1080/21645515.2016.1147635>.
210. Hu Z, Ott PA, Wu CJ. Towards personalized, tumour-specific, therapeutic vaccines for cancer. *Nat Rev Immunol.* 2018;18(3):168–82. <https://doi.org/10.1038/nri.2017.131>.
211. Melief CJM, van Hall T, Arens R, Ossendorp F, van der Burg SH. Therapeutic cancer vaccines. *J Clin Invest.* 2015;125(9):3401–12. <https://doi.org/10.1172/JCI80009>.
212. Palucka K, Banchereau J. Dendritic cell-based cancer therapeutic vaccines. *Immunity.* 2013;39(1):38–48. <https://doi.org/10.1016/j.immuni.2013.07.004>.
213. Sabado RL, Balan S, Bhardwaj N. Dendritic cell-based immunotherapy. *Cell Res.* 2017;27:74–95. <https://doi.org/10.1038/cr.2016.157>.
214. Shang N, Figini M, Shangguan J, Wang B, Sun C, Pan L, et al. Dendritic cells based immunotherapy. *Am J Cancer Res.* 2017;7(10):2091–102.
215. Melief CJM. Cancer immunotherapy by dendritic cells. *Immunity.* 2008;29(3):372–83. <https://doi.org/10.1016/j.immuni.2008.08.004>.
216. Kumai T, Kobayashi H, Harabuchi Y, Celis E. Peptide vaccines in cancer – old concept revisited. *Curr Opin Immunol.* 2017;45:1–7. <https://doi.org/10.1016/j.co.2016.11.001>.
217. Kuai R, Ochyl LJ, Bahjat KS, Schwendeman A, Moon JJ. Designer vaccine nanodisc for personalized cancer immunotherapy. *Nat Mater.* 2017;16(4):489–96. <https://doi.org/10.1038/nmat4822>.
218. Guo C, Manjili MH, Subjeck JR, Sarkar D, Fisher PB, Wang XY. Therapeutic cancer vaccines: past, present and future. *Adv Cancer Res.* 2013;119:421–75. <https://doi.org/10.1016/B978-0-12-407190-2.00007-1>.
219. Li L, Pretrovsky N. Molecular mechanisms for enhanced DNA vaccine immunogenicity. *Expert Rev Vaccines.* 2016;15(3):313–29. <https://doi.org/10.1586/14760584.2016.1124762>.
220. Yang B, Jeang J, Yang A, Wu TC, Hung CF. DNA vaccine for cancer immunotherapy. *Hum Vaccin Immunother.* 2014;10(11):3153–64. <https://doi.org/10.4161/21645515.2014.980686>.
221. Guo P, Wang J, Liu J, Xia M, Li W. Macrophage immigration inhibitory factor promotes cell proliferation and inhibits apoptosis of cervical adenocarcinoma. *Tumour Biol.* 2015;36(7):5095–102. <https://doi.org/10.1007/s13277-015-3161-4>.
222. Sim GC, Radvanyi L. The IL-2 cytokine family in cancer immunotherapy. *Cytokine Growth Factor Rev.* 2014;25(4):377–90. <https://doi.org/10.1016/j.cytofr.2014.07.018>.
223. Soares KC, Rucki AA, Wu AA, Olino K, Xiao Q, Chai Y, et al. PD-1/PD-L1 blockade together with vaccine therapy facilitates effector T cell infiltration into pancreatic tumors. *J Immunother.* 2015;38(1):1–11. <https://doi.org/10.1097/CJI.0000000000000062>.
224. Linch SN, Kasiewicz MJ, McNamara MJ, Hilgart-Martiszus IF, Farhad M, Redmond WL. Combination OX40 agonism/CTLA-4 blockade with HER2 vaccination reverses T-cell anergy and promotes survival in tumor-bearing mice. *Proc Natl Acad Sci U S A.* 2016;113(3):E319–27. <https://doi.org/10.1073/pnas.1510518113>.

225. Schwartzentruber DJ, Lawson DH, Richards JM, Conry RM, Miller DM, Treisman J, et al. gp100 peptide vaccine and interleukin-2 in patients with advanced melanoma. *N Engl J Med.* 2011;364(22):2119–27. <https://doi.org/10.1056/NEJMoa1012863>.
226. Dutta R, Mahato RI. Recent advances in hepatocellular carcinoma therapy. *Pharmacol Ther.* 2017;173:106–17. <https://doi.org/10.1016/j.pharmthera.2017.02.010>.
227. Hochnadel I, Kossatz-Boehlert U, Jedicke N, Lenzen H, Manns MP, Yevsa T. Cancer vaccines and immunotherapeutic approaches in hepatobiliary and pancreatic. *Hum Vaccin Immunother.* 2017;13(12):2931–52. <https://doi.org/10.1080/21645515.2017.1359362>.
228. Bann DV, Deschler DG, Goyal N. Novel immunotherapeutic approaches for head and neck squamous cell carcinoma. *Cancers.* 2016;8(10):E87. <https://doi.org/10.3390/cancers8100087>.
229. Tello TL, Coggshall K, Yom SS, Yu SS. Merkel cell carcinoma: an update and review: current and future therapy. *J Am Acad Dermatol.* 2018;78(3):445–54. <https://doi.org/10.1016/j.jaad.2017.12.004>.
230. Harms PW. Update on Merkel cell carcinoma. *Clin Lab Med.* 2017;37(3):485–501. <https://doi.org/10.1016/j.cll.2017.05.004>.
231. Delhalle S, Bode SFN, Balling E, Ollert M, He FQ. A roadmap towards personalized immunology. *NPJ Syst Biol Appl.* 2018;4(9):1–14. <https://doi.org/10.1038/s41540-017-0045-9>.
232. Silva JM, Videira M, Gaspar R, Préat V, Florindo HF. Immune system targeting by biodegradable nanoparticles for cancer vaccines. *J Control Release.* 2013;168(2):179–99. <https://doi.org/10.1016/j.jconrel.2013.03.010>.
233. Sahin U, Türeci O. Personalized vaccines for cancer immunotherapy. *Science.* 2018;359(6382):1355–60. <https://doi.org/10.1126/science.aar7112>.
234. Zhang X, Sharma PK, Goedeggembu P, Gillanders WE. Personalized cancer vaccines: targeting the cancer mutanome. *Vaccine.* 2017;35(7):1094–100. <https://doi.org/10.1016/j.vaccine.2016.05.073>.
235. Yarchoan M, Johnson BA 3rd, Lutz ER, Laheru DA, Jaffee EM. Targeting neoantigens to augment antitumour immunity. *Nat Rev Cancer.* 2017;17(4):209–22. <https://doi.org/10.1038/nrc.2016.154>.
236. Gruijl TD, van den Eertwegh AJM, Pinedo HM, Schepé RJ. Whole-cell cancer vaccination: from autologous to allogeneic tumor- and dendritic cell-based vaccines. *Cancer Immunol Immunother.* 2008;57(10):1569–77. <https://doi.org/10.1007/s00262-008-0536-z>.
237. Bencherif SA, Sands RW, Ali OA, Li WA, Lewin SA, Braschler TM, et al. Injectable cryogel-based whole cell cancer vaccines. *Nat Commun.* 2015;6:7556. <https://doi.org/10.1038/ncomms8556>.
238. Kandalaft LE, Chiang CL, Tanyi J, Motz G, Balint K, Mick R. A phase I vaccine trial using dendritic cells pulsed with autologous oxidized lysate for recurrent ovarian cancer. *J Transl Med.* 2013;11:149. <https://doi.org/10.1186/1479-5876-11-149>.
239. Yang YW, Luo WH. Cellular biodistribution of polymeric nanoparticles in the immune system. *J Control Release.* 2016;227:82–93. <https://doi.org/10.1016/j.jconrel.2016.02.011>.
240. Bolhassani A, Javanzad S, Saleh T, Hashemi M, Aghasadeghi MR, Sadat SM. Polymeric nanoparticles: potent vector s for vaccine delivery targeting cancer and infectious diseases. *Hum Vaccin Immunother.* 2014;10(2):321–32. <https://doi.org/10.4161/hv.26796>.
241. Le Gall CM, Weiden J, Eggermont LJ, Figdor CG. Dendritic cells in cancer immunotherapy. *Nat Mater.* 2018;17:472–7. <https://doi.org/10.1038/s41563-018-0093-6>.
242. Zhu M, Wang R, Nie G. Applications of nanomaterials as vaccine adjuvants. *Hum Vaccin Immunother.* 2014;10(9):2761–74. <https://doi.org/10.4161/hv.29589>.
243. Zhu G, Lynn GM, Jacobson O, Chen K, Liu Y, Zhang H, et al. Albumin/vaccine nano-complexes that assemble in vivo for combination cancer immunotherapy. *Nat Commun.* 2017;8(1954):1–15. <https://doi.org/10.1038/s41467-017-02191-y>.
244. Fecek RJ, Storkus WJ. Combination strategies to enhance the potency of monocyte-derived dendritic cell-based cancer vaccines. *Immunotherapy.* 2016;8(10):1205–18. <https://doi.org/10.2217/imt-2016-0071>.

245. Koido S. Dendritic-tumor fusion cell-based cancer vaccines. *Int J Mol Sci.* 2016;17(6):E828. <https://doi.org/10.3390/ijms17060828>.
246. Castiglione F, Bernaschi M. Epitope screening and cell cooperation in the immune response. Intelligent Systems, Modelling and Simulations (ISMS). Proceedings – 2011 2nd International Conference on Intelligence Systems, Modeling and Simulations. ISMS. 2011 Feb;127–132. <https://doi.org/10.1109/ISMS.2011.30>.
247. Daudi J. An overview of application of artificial immune system in swarm robotic systems. *Adv Robotics Automat.* 2015;4:127. <https://doi.org/10.4172/2168-9695.1000127>.
248. Zeeshan M, Javed H, Haider A, Khan A. An immunology inspired flow control attack detection using negative selection with r-contiguous bit matching for wireless sensor networks. *Int J Distrib Sensor Netw* 2015;11(11):1–7. doi:<https://doi.org/10.1155/2015/169654>.
249. Khan MT, de Silva, CW. Autonomous fault tolerance multi-robot cooperation using artificial immune system. *Automation and Logistics. ICAL 2008. IEEE International Conference on* 2008. 2008 Sep;623–8. <https://doi.org/10.1109/ICAL.2008.4636225>.
250. Nigam D, Kumar V. Artificial immune system: a potential tool to handle bioinformatics issues. *Int J Artif Intell Knowl Discov.* 2012;2(1):1–5.
251. Saybani MR, Shamshirband S, Hormozi SG, Wah TY, Aghabozorgi S, Pourhoseingholi MA, et al. Diagnosing tuberculosis with a novel support vector machine-based artificial immune system recognition system. *Iran Red Crescent Med J.* 2015;17(4):e24557. [https://doi.org/10.5812/ircmj.17\(4\)2015.24557](https://doi.org/10.5812/ircmj.17(4)2015.24557).
252. Onomza WV, Alhassan J, Alelere M, Tunde A. Development of secure plus antivirus with the artificial immune system model. *Int J Innov Technol Res.* 2015;3(2):1882–96.
253. Rai N, Singh A. Improved clonal selection algorithm (ICLONALG). *Int J Current Eng Technol.* 2015;5(4):2459–64.
254. Ali NIM, Malek MA, Ismail AR. Immune network algorithm in monthly streamflow prediction at Johor river. *ARPN J Eng Appl Sci.* 2015;10(3):1352–6.
255. Zeng J. Computer malicious executables detection based on real-valued negative selection algorithm. *Appl Math Inform Sci.* 2015;9(2):1089–94. <https://doi.org/10.12785/amis/090260>.
256. Liò P, Migliolo O, Nicosia G, Nolfi S, Pavone M. Advances in artificial life: synthesis and simulation of living systems: editorial. *Artif Life.* 2015;21(4):395–7. [https://doi.org/10.1162/ARTL\\_e\\_00189](https://doi.org/10.1162/ARTL_e_00189).
257. Langton CG. Artificial Life: proceedings of an interdisciplinary workshop on the synthesis and simulation of living systems, held September, 1987, in Los Alamos, New Mexico, Santa Fe Institute Studies in the Sciences of Complexity. 1989; vol VI. Addison-Wesley.
258. Ray TS. An evolutionary approach to synthetic biology: zen and the art of creating life. *Artif Life.* 1994;1(1/2):195–226.. MIT Press. <https://doi.org/10.1162/artl.1993.1.179>.
259. Aguilar W, Santamaría-Bonfil G, Froese T, Gershenson C. The past, present, and future for artificial life. *Front Robotics AI.* 2014;1:8. <https://doi.org/10.3389/frobt.2014.00008>.
260. Williams RA. Lesson learned on development and application of agent-based models of complex dynamical systems. *Simul Model Pract Theory.* 2018;83:201–12. <https://doi.org/10.1016/j.simpat.2017.11.001>.
261. Komosinski M, Adamatzky A, editors. Artificial life models in software. Second ed: Springer; 2009. <https://doi.org/10.1007/978-1-84882-285-6>.
262. Bauer AL, Beauchemin CAA, Perelson AS. Agent-based modeling of host-pathogen systems: the successes and challenges. *Inf Sci.* 2009;179(10):1379–89. <https://doi.org/10.1016/j.ins.2008.11.012>.
263. Elkalaawy N, Wassal A. Methodologies for the modeling and simulation of biochemical networks, illustrated for signal transduction pathways: a primer. *Biosystems.* 2015;129:1–18. <https://doi.org/10.1016/j.biosystems.2015.01.008>.
264. Helbing D, Baiti S. How to do agent-based simulations in the future: from modeling social mechanism to emergent phenomena and interactive systems design. Chapter 2: agent-based modeling of the book Social Self-Organization. Springer, Berlin. 2012 Feb;25–70. [https://doi.org/10.1007/978-3-642-24004-1\\_2](https://doi.org/10.1007/978-3-642-24004-1_2).

265. An G, Mi Q, Dutta-Moscato J, Vodovotz Y. Agent-based models in translational systems biology. *WIREs Syst Biol Med.* 2009;1(2):159–71. <https://doi.org/10.1002/wsbm.45>.
266. Hwang M, Garbey M, Berceli SA, Tran-Son-Tay R. Rule-based simulation of multi-cellular biological systems – a review of modeling techniques. *Cell Mol Bioeng.* 2009;2(3):285–94. <https://doi.org/10.1007/s12195-009-0078-2>.
267. North MJ, Macal CM. Foundations of and recent advances in artificial life modeling with repast 3 and repast symphony. In: Komosinski M, Adamatzky A, editors. *Artificial life models in software*. London: Springer; 2009;. Chapter 2. p. 37–60.
268. Pezzullo G, Levin M. Top-down models in biology: explanation and control of complex living systems above the molecular level. *J R Soc Interface.* 2016;13(24):1–16. <https://doi.org/10.1098/rsif.2016.0555>.
269. Loscalzo J, Barabasi AL. Systems biology and future of medicine. *Wiley Interdiscip Rev Syst Biol Med.* 2011;3(6):619–27. <https://doi.org/10.1002/wsbm.144>.
270. Seiden PE, Celada F. A model for simulating cognate recognition and response in the immune system. *J Theor Biol.* 1992;158(3):329–57. [https://doi.org/10.1016/S0022-5193\(05\)80737-4](https://doi.org/10.1016/S0022-5193(05)80737-4).
271. Celada F, Seiden PE. Affinity maturation and hypermutation in a simulation of the humoral immune response. *Eur J Immunol.* 1996;26(6):1350–8. <https://doi.org/10.1002/eji.1830260626>.
272. Meier-Schellersheim M, Mack G. SIMMUNE, a tool for simulating and analyzing immune system behavior. Cornell University Library. 1999 Mar; [arXiv:cs/9903017v1](https://arxiv.org/abs/cs/9903017v1).
273. Bernaschi M, Castiglione F. Design and implementation of an immune system simulator. *Comput Biol Med.* 2001;31(5):303–31. [https://doi.org/10.1016/S0010-4825\(01\)00011-7](https://doi.org/10.1016/S0010-4825(01)00011-7).
274. Puzone R, Kohler B, Seiden P, Celada F. IMMSIM, a flexible model for in machina experiments on immune system responses. *Futur Gener Comput Syst.* 2002;18(7):961–72. [https://doi.org/10.1016/S0167-739X\(02\)00075-4](https://doi.org/10.1016/S0167-739X(02)00075-4).
275. Pappalardo F, Lollini PL, Castiglione F, Motta S. Modeling and simulation of cancer immunoprevention vaccine. *Bioinformatics.* 2005;21(12):2891–7. <https://doi.org/10.1093/bioinformatics/bti426>.
276. Castiglione F, Bernaschi M, Succi S. Simulating the immune response on a distributed parallel computer. *Int J Modern Phys C.* 1997;8(3):527–45. <https://doi.org/10.1142/S0129183197000424>.
277. Bernaschi M, Castiglione F. Selection of escape mutants from immune recognition during HIV infection. *Immunol Cell Biol.* 2002 Jun;80(3):307–313. <https://doi.org/10.1046/j.1440-1711.2002.01082>.
278. Bandini S, Mauri G, Vizzari G. Supporting action-at-a-distance in situated cellular agents. *Fundamenta Informaticae.* 2006;69(3):251–71.
279. Baldazzi V, Castiglione F, Bernaschi M. An enhanced agent based model of the immune system response. *Cell Immunol.* 2006;244(2):77–9. <https://doi.org/10.1016/j.cellimm.2006.12.006>.
280. Castiglione F, Duca K, Jarrah A, Laubenbacher R, Hochberg D, Thorley-Lawson D. Simulating epstein–barr virus infection with C-ImmSim. *Bioinformatics.* 2007;23(11):1371–7. <https://doi.org/10.1093/bioinformatics/btm044>.
281. Mata J, Cohn M. Cellular automata-based modeling program: synthetic immune system. *Immunol Rev.* 2007;216(1):198–212. <https://doi.org/10.1111/j.1600-065X.2007.00511.x>.
282. Maeda K, Sakama C. Identifying cellular automata rules. *J Cell Autom.* 2007;2(1):1–20.
283. Folcik VA, An GC, Orosz CG. The basic immune simulator: an agent-based model to study the interactions between innate and adaptive immunity. *Theor Biol Med Model.* 2007;4(39):1–18. <https://doi.org/10.1186/1742-4682-4-39>.
284. Dréau D, Dimitre S, Ted C, Mirsad H. An agent-based model of solid tumor progression. In: Rajasekaran S, editor. *Bioinformatics and Computational Biology*. BiCoB 2009. Lecture Notes in Computer Science. 2009;5462. Springer, Berlin, Heidelberg. doi:[https://doi.org/10.1007/978-3-642-00727-9\\_19](https://doi.org/10.1007/978-3-642-00727-9_19).

285. De Pillis LG, Mallet DG, Radunskaya AE. Spatial tumor-immune modeling. *Comput Math Methods Med.* 2006;7(2–3):159–76. <https://doi.org/10.1080/10273660600968978>.
286. Sneddon M, Faeder JR, Emonet T. Efficient modeling, simulation and coarse-graining of biological complexity with NFsim. *Nat Methods.* 2011;8(2):177–83. <https://doi.org/10.1038/nmeth.1546>.
287. Wendelsdorf KV, Alam M, Bassaganya-Riera J, Bisset K, Eubank S, Hontecillas R, et al. Enteric immunity simulator: a tool for *in silico* study of gastroenteric infections. *IEEE Trans Nanobioscience.* 2012;11(3):273–88. <https://doi.org/10.1109/TNB.2012.2211891>.
288. Barret CL, Bisset KR, Eubank SG, Feng X, Marathe MV. Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In: SC’2008: Proceedings of the 2008 ACM/IEEE Conference on Supercomputing. 2008 Nov;1–12. <https://doi.org/10.1109/SC.2008.5214892>.
289. Kim PS, Lee PP. Modeling protective anti-tumor immunity via preventative cancer vaccines using a hybrid agent-based and delay differential equation approach. *PLoS Comput Biol.* 2012;8(10):e1002742. <https://doi.org/10.1371/journal.pcbi.1002742>.
290. Mallet DG, De Pillis LG. A cellular automata model of tumor-immune system interactions. *J Theor Biol.* 2006;239(3):334–50. <https://doi.org/10.1016/j.jtbi.2005.08.002>.
291. Pappalardo F, Forero IM, Pennisi M, Palazon A, Melero I, Motta S. SimB16: modeling induced immune system response against B16-melanoma. *PLoS One.* 2011;6(10):e26523. <https://doi.org/10.1371/journal.pone.0026523>.
292. Von Eichborn J, Woelke AL, Castiglione F, Preisnner R. VaccImm: simulating peptide vaccination in cancer therapy. *BioMed Central Bioinform.* 2013;14(127):1–8. <https://doi.org/10.1186/1471-2105-14-127>.
293. Santos J, Monteagudo A. Analysis of behavior transitions in tumour growth using a cellular automaton simulation. *IET Syst Biol.* 2015;9(3):75–87. <https://doi.org/10.1049/iet-syb.2014.0015>.
294. Shahmoradi S, Rahatabad FN, Maghooli K. A stochastic cellular automata model of growth of avascular tumor with immune response and immunotherapy. *Inform Med Unlocked.* 2018; <https://doi.org/10.1016/j.imu.2018.06.008>.
295. Boondireck A, Lenbury Y, Wong-Ekkabut J, Triampo W, Tang IM, Picha P. A stochastic model of cancer growth with immune response. *J Korean Phys Soc.* 2006;49(4):1652–66.
296. Bezzi M, Celada F, Ruffo S, Seiden PE. The transition between immune and disease states in a cellular automaton model of clonal immune response. *Phys A Stat Mech Its Appl.* 1997;245(1–2):145–63. [https://doi.org/10.1016/S0378-4371\(97\)00290-2](https://doi.org/10.1016/S0378-4371(97)00290-2).
297. Celada F, Seiden P. Modeling immune cognition. *IEEE International Conference on Systems, Man, and Cybernetics.* San Diego, CA, USA. 1998 Oct; vol. 4, p. 3787–3792. <https://doi.org/10.1109/ICSMC.1998.726677>.
298. Kleinstein SH, Seiden PE. Simulating the immune system. *Comput Sci Eng.* 2000;2(4):69–77. <https://doi.org/10.1109/5992.852392>.
299. Kohler B, Puzone R, Seiden PE, Celada F. A systematic approach to vaccine complexity using an automaton model of the cellular and humoral immune system. I. Viral characteristics and polarized responses. *Vaccine.* 2000;19(7–8):862–76. [https://doi.org/10.1016/S0264-410X\(00\)00225-5](https://doi.org/10.1016/S0264-410X(00)00225-5).
300. Stewart JJ, Agosto H, Litwin S, Welsh JD, Shlomchik M, Weigert M, Seiden PE. A solution to the rheumatoid factor paradox: pathologic rheumatoid factors can be tolerized by competition with natural rheumatoid factors. *J Immunol.* 1997;159(4):1728–38.
301. Bardi JS. New NIAID program aims to model immune responses and key infectious diseases. NIH/National Institute of Allergy and Infectious Diseases 2012 Jul. <http://www.nih.gov/news/pr/jul2006/niaind-12.htm>. Accessed 8 Sep 2012.
302. Langman RE, Mata J, Cohn M. A computerized model for the self-non-self discrimination at the level of the Th (Th genesis). II. The behavior of the system upon encounter with non-self antigens. *Int Immunol.* 2003;15(5):593–609. <https://doi.org/10.1093/intimm/dxg059>.

303. Emerson A, Rossi E. ImmunoGrid – the virtual human immune system project. *Stud Health Technol Inform.* 2007;126:87–92.
304. Halling-Brown M, Pappalardo F, Rapin N, Zhang P, Alemani D, Emerson A, et al. ImmunoGrid: towards agent-based simulations of the human immune system at a natural scale. *Phil Trans R Soc A.* 2010;368:2799–815. <https://doi.org/10.1098/rsta.2010.0067>.
305. Strain MC, Richman DD, Wong JK, Levine H. Spatiotemporal dynamics of HIV propagation. *J Theor Biol.* 2002;218(1):85–96. <https://doi.org/10.1006/jtbi.2002.3055>.
306. Segovia-Juarez JL, Ganguli S, Kirschner D. Identifying control mechanisms of granuloma formation during *M. tuberculosis* infection using an agent-based model. *J Theor Biol.* 2004;231(3):357–76. <https://doi.org/10.1016/j.jtbi.2004.06.031>.
307. Beauchemin C. MASyV: A Multi-Agent System Visualization program. Free open-source GNU GPL software available online on SourceForge.net. <http://masyv.sourceforge.net/>. Accessed 9 Sep 2012.
308. Motta S, Castiglione F, Lollini P, Pappalardo F. Modelling vaccination schedules for a cancer immunoprevention vaccine. *Immunome Res.* 2005;1(5):1–18. <https://doi.org/10.1186/1745-7580-1-5>.
309. Alarcon T, Byrne HM, Maini PK. A multiple scale model for tumor growth. *Society for Industrial and Applied Mathematics. Multiscale Model Simul.* 2005;3(2):440–75. <https://doi.org/10.1137/040603760>.
310. Zhang Y, Wallace DL, de Lara CM, Ghattas H, Asquith B, Worth A, et al. In vivo kinetics of human natural killer cells: the effects of ageing and acute and chronic viral infection. *Immunology.* 2007;121(2):258–65. <https://doi.org/10.1111/j.1365-2567.2007.02573.x>.
311. Warrender C, Forrest S, Koster F. Modeling intercellular interactions in early Mycobacterium infection. *Bull Math Biol.* 2006;68(8):2233–61. <https://doi.org/10.1007/s11538-006-9103-y>.
312. Shapiro M, Duca KA, Lee K, Delgado-Eckert E, Hawkins J, Jarrah AS, et al. A virtual look at Epstein–Barr virus infection: simulation mechanism. *J Theor Biol.* 2008;252(4):633–48. <https://doi.org/10.1016/j.jtbi.2008.01.032>.
313. Beauchemin C, Forrest S, Koster FT. Modeling influenza viral dynamics in tissue. In: Bersini H, Carneiro J, editors. *Artificial immune systems. ICARIS 2006. Lecture notes in computer science*, vol. 4163. Berlin: Springer; 2006. p. 23–36. [https://doi.org/10.1007/11823940\\_3](https://doi.org/10.1007/11823940_3).
314. Ebeling W, Schweitzer F. Swarms of particle agents with harmonic interactions. *Theory Biosci.* 2001;120(3–4):207–24. <https://doi.org/10.1007/s12064-001-0019-7>.
315. Macal CM, North MJ. Tutorial on agent-based modeling and simulation part 2: how to model with agents. *Simulation Conference 2006, WSC 06. Proceedings of the 38th conference on winter simulation.* IEEE. 2006 Dec; pp.73–83. ISBN:1-4244-0501-7.
316. Baird L, Fagin B. Conserved energy functions for cellular automata: finding nontrivials faster through a complete theory of the trivials. *J Cell Autom.* 2012;7(2):115–50.
317. Escobar-Ospina ME, Gómez-Perdomo J. A growth model of human papillomavirus type 16 designed from cellular automata and agent-based models. *Artif Intell Med.* 2013;57(1):31–47. <https://doi.org/10.1016/j.artmed.2012.11.001>.
318. De Silva N, Klein U. Dynamics of B cells in germinal centres. *Nat Rev Immunol.* 2015;15(3):137–48. <https://doi.org/10.1038/nri3804>.
319. Hwang JK, Alt FW, Yeap LS. Related mechanisms of antibody somatic hypermutation and class switch recombination. *Microbiol Spectr.* 2015;3(1):MDNA3-0037-2014. <https://doi.org/10.1128/microbiolspec.MDNA3-0037-2014>.
320. Kurosaki T, Kometani K, Ise W. Memory B cells. *Nat Rev Immunol.* 2015;15(3):149–59. <https://doi.org/10.1038/nri3802>.
321. Sen B, Johnson FM. Regulation of Src family kinases in human cancers. *J Signal Transduc.* 2011;865819:1–14. <https://doi.org/10.1155/2011/865819>.
322. Railsback SF, Lytinen SL, Jackson SK. Agent-based simulation platforms: review and development recommendations. *SIMULATION.* 2006;82(9):609–23. <https://doi.org/10.1177/0037549706073695>.

323. Abar S, Theodoropoulos GK, Lemarinier P, O'Hare GMP. Agent based modelling and simulation tools: a review of the state-of-art software. *Comput Sci Rev.* 2017;24:13–33. <https://doi.org/10.1016/j.cosrev.2017.03.001>.
324. Hu C, Mao X, Li M, Zhu Z. Organization-based agent-oriented programming: model, mechanisms, and language. *Front Comp Sci.* 2014;8(1):33–51. <https://doi.org/10.1007/s11704-013-2345-6>.
325. Ackley DH, Ackley ES. The ulam programming language for artificial life. *Artif Life.* 2016;22(4):431–50. [https://doi.org/10.1162/ARTL\\_a\\_00212](https://doi.org/10.1162/ARTL_a_00212).
326. Yan Q, Li M, Liu Q, Li F, Zhu B, Wang J, et al. Molecular characterization of woodchuck IFI16 and AIM2 and their expression in woodchucks infected with woodchuck hepatitis virus (WHV). *Sci Rep.* 2016;6(28776):1–11. <https://doi.org/10.1038/srep28776>.
327. Yi Z, Lin WW, Stunz LL, Bishop GA. Roles for TNF-receptor associated factor 3 (TRAF3) in lymphocyte functions. *Cytokine Growth Factor Rev.* 2014;25(2):147–56. <https://doi.org/10.1016/j.cytofr.2013.12.002>.

# Mystery of HIV Drug Resistance: A Machine Learning Perspective



**Mohanapriya Arumugam, Nirmaladevi Ponnusamy,  
Sajitha Lulu Sudhakaran, Vino Sundararajan,  
and Pandjassarame Kanguane**

**Abstract** Human immunodeficiency virus (HIV) is one of the fastest developing pathogens known. HIV/AIDS is an incurable disease which causes severe damage to the immune system. The recommended treatment for HIV/AIDS is a combination of three antiretroviral (ARV) drugs from two or more different drug groups and is known as highly active antiretroviral therapy (HAART). Drug resistance is a major impediment experienced by therapist in treating HIV infected patients. Theoretically, drug resistance can be predicted from the presence of specific mutations in the viral genome. With the current disease burden and lack of resources in developing countries, phenotypic tests are not viable. Developing a computational prediction of drug resistance phenotype will enable efficient and timely selection of the best treatment regimens. Nevertheless, the very large range of possible drug combinations and of viral mutational patterns leads to an extremely complex scenario, making prediction of in vivo treatment response extremely challenging. To deal with such complexity, machine learning methods are being increasingly explored. Clinical and technological developments has generated and stored large volumes of data in public databases which facilitates the use of machine learning methods for predicting drug resistance. Quite a lot of machine learning approaches such as neural networks, support vector machine, Bayesian networks, decision trees and linear regression have been proposed for the prediction of phenotypic drug resistance. Therefore, conducting resistance testing is certainly significant in order to administer appropriate antiviral drugs to HIV infected patients.

---

M. Arumugam (✉) · N. Ponnusamy · S. L. Sudhakaran  
Department of Biotechnology, School of Biosciences and Technology,  
Vellore Institute of Technology, Vellore, India  
e-mail: [mohanapriya@vit.ac.in](mailto:mohanapriya@vit.ac.in)

V. Sundararajan  
Department of Biosciences, School of Bio Sciences and Technology,  
Vellore Institute of Technology, Vellore, Tamilnadu, India

P. Kanguane  
Biomedical Informatics, Irulan Sandy Annexe, Puducherry, India

## Key Concepts

Key Phrases:

1. The rapid selection of drug resistant viral mutations creates a massive challenge for therapy.
2. There is an urgent need to develop rapid and cost effective drug resistance prediction methods
3. Global and country specific drug resistance databases are important sources HIV genetic data
4. Drug resistance can be measured using phenotyping and genotyping tests
5. Machine learning methods enhances the speed and accuracy of resistance prediction and therapy plan decision making
6. The input for machine learning methods is the viral genome; the output is the resistance values to a particular drug
7. Supervised learning methodolog relies on labeled data
8. Unsupervised model relies on unlabeled data
9. Semi-supervised learning takes an intermediate ground. It relies on both labeled and unlabeled data
10. Rule based methods characterize the drugs by the mutational patterns which produce resistance
11. Both rule based and machine learning based prediction servers are available online

**Keywords** Drug resistance · Prediction · Machine learning · Supervised · Unsupervised · Rule-based methods · Databases

## 1 Introduction

### 1.1 *HIV/AIDS: A Global Disease*

More than 70 million people have been infected with the HIV virus and 35 million people have died of HIV so far. Globally, 36.9 million [31.1–43.9 million] people were living with HIV at the end of 2017. An average of 0.8% adults of age range between 15 and 49 are estimated to live with HIV infection worldwide, however the number of infective cases considerably vary between countries [1]. The African population remains most adversely affected, which accounts for 4.1% (1 in every 25 adults) living with HIV and which is nearly two-thirds of the people living with HIV worldwide and is followed by south and south-east Asia. Since the start of the epidemic the number of deaths from HIV AIDS related infections is 35.4 million.

These enormous numbers of mortalities and infections related to HIV reveal the threat posed by HIV since the past few years. The life expectancy of individuals in the countries majorly affected by HIV has reduced [2].

Joint United Nations Program on HIV/AIDS (UNAIDS) and the World Health Organization (WHO), have devoted to the goals of ending the AIDS pandemic as a public health threat by 2030 and guaranteeing that by 2020, 90% of people with HIV infection know that they have it, 90% of infected people are receiving antiretroviral therapy, and persistent viral suppression is attained in 90% of those getting treatment [3].

## 2 The Drug Resistance Challenge

It is a huge challenge for researchers to develop an effective drug against HIV. The rapid selection of drug resistant viral mutations creates a whopping challenge for treatment. These resistance mutations in the genome of infecting virus is a significant contraindication for an effective virological response to HAART [4, 5].

There are several ARVs developed to inhibit HIV by different targets. Nonetheless, some approved ARVs are presently available for the treatment of HIV infection. Most of them focus on two of the most important viral enzymes, namely Protease and Reverse transcriptase. The viral enzymes, HIV-1 protease (PR) and reverse transcriptase (RT), are main and well characterized drug targets. The activity of these two proteins is inhibited by the antiviral PR inhibitors (PIs) and the active site (NRTIs) and non-active site RT inhibitors (NNRTIs) [6]. Despite the amount of drugs there is not a single drug that can completely inhibit the viral multiplication. The standard treatment for patients infected with human HIV, referred to as highly active antiretroviral therapy (HAART), consists of three or more HIV drugs, most commonly two nucleoside reverse transcriptase inhibitors (NRTIs) in combination with either a nonnucleoside reverse transcriptase inhibitor (NNRTI), a protease inhibitor (PI), or more recently, an integrase inhibitor (INI). Table 1 illustrates the most used drugs, grouped by the target.

## 3 Drug Resistance Databases

International and country specific drug resistance databases are important repositories of HIV-1 genetic data [7]. The UK HIV Drug Resistance Database contains over 10,000 non-B subtype isolates. Proposals needs to be submitted for the use of data. The Stanford HIV Drug Resistance Database (HIVDB) curates all published data and contains nearly 150,000 sequences. This is presented in many statistical and graphical formats. A number of treatment-experienced mutation differences have been highlighted in the literature and are stored and curated in

**Table 1** Anti-retroviral therapy drug chart grouped based on the viral targets

Categories	NRTIs	NNRTIs	PIs	Entry inhibitors (CCR5 and fusion inhibitors)	Integrase inhibitors
Treatment naïve patients	Zidovudine (AZT) Stavudine (d4T) Lamivudine (3TC) Didanosine (ddI) Zalcitabine (ddC) Abacavir (ABC)	Nevirapine (NVP) Delavirdine (DLV) Efavirenz (EFV) Rilpivirine (RPV)	Ritonavir (RTV) Indinavir (IDV) Saquevir (SQV) Nelfinavir (NFV) Amprenavir (APV) Fosamprenavir (FPV) Atazanavir (ATV)	–	Dolutegravir (DTG)
Treatment experienced patients prevents	Tenofovir (TDF) Emtricitabine (FTC)	Etravirine (ETV)	Tipranavir (TPV) Darunavir (DRV)	Maraviroc (MVC) Enfuvirtide (T20)	Raltegravir (RAL)
Prevention of vertical transmission (WHO guidelines)	Zidovudine (AZT) Lamivudine (3TC) Tenofovir (TDF) Emtricitabine (FTC)	Nevirapine (NVP) Efavirenz (EFV)	–	–	–
Children	Zidovudine (AZT) Stavudine (d4T) Lamivudine (3TC) Didanosine (ddI) Emtricitabine (FTC) Abacavir (ABC)	Nevirapine (NVP)	Ritonavir (RTV) Nelfinavir (NFV) Darunavir* (DRV)	Enfuvirtide (T20)	–
For post exposure prophylaxis	Zidovudine (AZT) Stavudine (d4T) Lamivudine (3TC) Didanosine (ddI) Emtricitabine (FTC) Tenofovir (TDF)	Efavirenz (EFV)	Indinavir (IDV) Nelfinavir (NFV) Ritonavir (RTV) Saquevir (SQV) Fosamprenavir (FPV) Atazanavir (ATV) Ritonavir (RTV)/Lopinavir (LPV)	Enfuvirtide (T20)	–
For pre exposure prophylaxis	Tenofovir (TDF) Emtricitabine (FTC) in MSM	–	–	–	–

HIVDB. Using these databases, several models based on machine learning algorithms have been developed for the prediction of phenotypic drug resistance from genotypes [8–10].

## 4 Assessing the Drug Resistance

Drug resistance can be measured using two biological tests: phenotyping and genotyping. The most widely used tool is the genotypic test where the sequence of the viral genome is analysed for the presence of known drug resistance mutations [11]. In the phenotypic test, the susceptibility to drugs is measured with cells infected with the viral strain *in vitro* [12]. It determines whether a mutation of the virus might be resistant to a given drug or not. Hence the first one quantifies drug susceptibility while the second one determines the mutational pattern. Phenotypic testing is slower, very expensive and is practically impossible to examine the mutational resistance which is constantly emerging. Currently, genotyping is recommended for new HIV infections or for individuals failing therapy in order to identify the presence of resistant mutations and guide the choice of drugs. Thus, the development of computational algorithms to predict the resistance from a given genotype is the only choice [13].

## 5 Basis of Machine Learning (ML) Methods

Over the past two decades, there has been a consistent increase in the medical research involving ML, progressing intensively from laboratory curiosity to a practical clinical application. This is largely because of growing dimensions of clinical, social, epidemiological, genetic and other types of data that are overflowing for humans to deduce from. Machine learning systems are expected to enhance the speed and accuracy of prediction and decision making among physicians thereby reducing costs, time and boosting patients' health [14].

Commonly, the input for machine learning methods is the viral genome sequence; while the output of the algorithms is the resistance values of the virus to certain drug. The general procedure followed by ML algorithms is: first, the algorithm studies the training data set of both the input and their corresponding output; then by using statistical, classification, or other algorithms, a computational model is learned and constructed for these data; finally, a new viral genotype is given to the model, and the predicted resistance value is generated by the model. From this predicted resistance value, the given genotype could be predicted as drug resistant or not, or somewhere in between, to certain drug. The different algorithms used to construct the computational model could further classify as the statistical learning methods, classification methods together with the molecular structure based methods [15].

## 6 Classification of ML Methods

ML methods can be broadly classified into supervised, unsupervised and semi-supervised learning methods. In a supervised learning methodology, the algorithmic rule learns on a labeled dataset, providing a solution key that the algorithmic rule will use to evaluate its accuracy on training data. An unsupervised model, in contrast, relies on unlabeled data that the algorithm tries to make sense of by mining features and patterns on its own. Semi-supervised learning takes a middle ground. It uses a small amount of labeled data supporting a larger set of unlabeled data, hence it relies on both labelled and unlabelled data.

## 7 Supervised Learning Methods

A number of supervised learning methods have been developed. These include multiple linear regression, decision trees and forests [6], logistic regression, support vector machines [16–19], artificial neural networks [18–20], Bayesian classifier, classification and regression trees [21, 22], K-nearest neighbours among others [23]. Supervised learning methods search for a function  $f(x)$  that predicts a target/output variable ( $y$ ) given a set of predictor/input variables( $x$ ). The training data is called labeled data because it consists of  $(x,y)$  pairs of variables. The inputs  $x$  may be DNA sequences, molecular structures, images, graphs or videos. Outputs (or labels) may include continuous outcomes or the more common binary yes or no outcomes. Ensemble methods combine outputs of multiple independently trained weaker models to make an overall prediction. The selection of the combination of weaker learning methods is made in such a way as to maximize the prediction power of the ensemble algorithm. Ensemble methods include boosting, bootstrap aggregation, stacking/blending, random forests [19, 24] and their modifications [25].

While predicting HIV drug resistance using supervised learning method from genotypic data, it is important to consider the following points (a) analysis and pre-processing of data(b)features selection or extraction, if it is necessary (c) choosing of classification or regression method [26].

## 8 Unsupervised Learning Methods

Unsupervised learning involves the analysis of unlabeled (no distinction between input and output) data under assumptions about the structural properties of the data (e.g., algebraic, combinatorial, or probabilistic). Since there are no training examples used in this process, the learning algorithm aims to identify patterns and correlations in the given data. The main applications of these algorithms include clustering and dimensionality reduction. Dimensionality reduction algorithms,

including principal components analysis, manifold learning, factor analysis, random projections, and autoencoders, identify and eliminate redundancies in the data so as to remain with only the variables that account for the most variability in the data. Clustering algorithms partition data into coherent clusters and determine the partitioning rule for predicting clusters in future data. The K-means clustering algorithm is the most commonly used method [23]. Computational complexity is a major concern in both clustering and dimension reduction since the datasets to exploit are large and unlabelled [27].

## 9 Semi-Supervised Learning Method

A third major ML concept is semi-supervised learning. Here, the data is a combination of small quantities of labeled and large quantities of unlabelled training data. The algorithm learns the structures of the data from the labeled examples and makes assumptions about the unlabeled data in order to make predictions. Semi-supervised learning is useful when the cost associated with labeling is too high to allow for a fully labelled training process [28]. Semi-supervised learning is subclassified into inductive learning whereby the goal is to learn from both the labeled and unlabeled dataset to predict labels for future datasets and transductive learning whereby the goal is to predict labels for the unlabeled portion of the data [29].

The choice of ML method to use is guided by the objectives of the analysis and the data available. Important data considerations include the number of predictor variables/features available in the data and quality of data. In general, a small but informative feature space results in higher generalizability of the model and avoids overfitting [30] while improving data quality and greatly improves the analysis. Several approaches using machine learning, such as linear regression [31], decision trees [32], neural networks [33], support vector regression [34], and Bayesian networks [35], and rule-based methods, such as Stanford HIVdb [36], HIV-GRADE [37], and ANRS [38], have been proposed for the interpretation of genotypic tests [39].

## 10 Classification and Regression

Using *in-vitro* and/or *in-vivo* experiments, databases based on genotype-phenotype pairs have been constructed. Based on this knowledge, it is possible to perceive the problem as a classification or regression. Regression learning methods predict outcomes in a continuous spectrum while classification learning methods predict outcomes of a categorical (binary) nature [40]. Either of these methods helps to decide which drugs or combination of drugs should be used for therapy. Genotype-phenotype pairs can be used directly as the training data to regression methods. In regression method, the objective is to predict the resistance of the mutational pattern

to a specific drug which is given as a phenotypic value. When considering classification problems, the objective variable (phenotype value) needs to be transformed into class labels before being utilized. The method commonly used for the interpretation of phenotypic data is based on biological cut-offs. Cut-offs are measured experimentally for each drug in order to classify a mutational pattern either as “drug susceptible” or as “drug resistant” [41, 42]. The problem is divided into three classes: sensitive, intermediate and resistant. Profuse research has been done on problems such as multiclass classification, multi-label classification, ranking problems, and general structured prediction problems [43].

## 11 Support Vector Machine (SVM)

The SVM is a method developed by Vapnik in 1996 [44] from statistical learning theory. SVMs have become an important machine learning technique for many pattern recognition problems like face recognition and text recognition, especially in computational biology. It can be divided into linear and nonlinear SVM. A simple linear SVM could be effectively trained for both classification and regression. This shows that the encoding is highly linear and can effectively represent the features in the sequence. It has also been used as a regression model to predict resistance [34]. For training and testing the SVM LIBSVM software library developed by Chang can be used and is available at <http://www.csie.ntu.edu.tw/cjlin/libsvm>.

Araya et al. [45] used SVMs in predicting the drug resistance of an HIV strain extracted from a patient. They used cross-validation to measure the unbiased estimate on 2045 data sets to evaluate the performance of SVM. The classification accuracy and the area under the receiver operation characteristics (ROC) curve was used as a performance measure. They also developed models based on neural networks, decision trees and logistic regression. The results show that SVMs are a highly successful classifier and perform better than the other techniques with performance ranging between 94.1% and 96.3% accuracy and 81.3% and 97.5% area under the ROC curve.

Yu et al. [31] performed five fold cross validation test by implementing in MATLAB SVM toolbox [46, 47] along with the linear kernel. Their classification showed high accuracy, sensitivity and specificity for all inhibitors. For PIs the accuracy values ranges from a low of 0.93 to a high of 0.96, while sensitivity and specificity range from 0.92–0.96 and 0.94–0.98, respectively. Resistance to NRTIs is classified with even higher accuracies of 0.97–0.99, sensitivities of greater than 0.98 and specificities of 0.95–0.99, while for NNRTIs the classification performance was superior with all values of over 0.97 for accuracy, sensitivity and specificity. The excellent performance with the linear SVM kernel demonstrates conclusively that the novel encoding using Delaunay triangulation separates the resistant and non-resistant data into two distinct categories.

## 12 Bayesian Network (BN)

A BN is a probabilistic model that describes statistical independencies between multiple variables [48]. Dependencies are represented in a directed acyclic graph and form the qualitative component of a BN. A BN may be learned from data by searching for the network that explains a maximum of the observed correlations in the data using a minimum number of arcs [49]. BN learning has previously successfully been used to discover relationships between amino acids for predicting protein secondary structure [50], they are well suited to learn interactions between different protein residues. The same way the BN can be postulated with respect to drug resistance for the presence of arcs in the network between amino acids and for the network structure around the drug node.

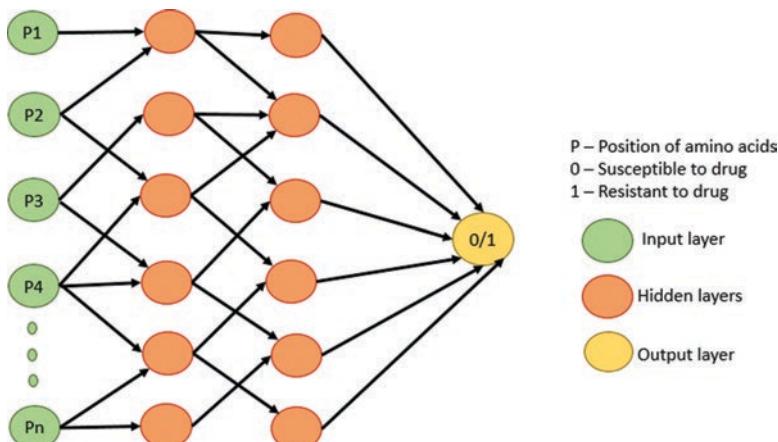
Deforche et al. [35] used BN to analyse the direct influences between protein residues and exposure to treatment in clinical HIV-1 proteases sequences from diverse subtypes. Three hundred forty sequences from patients on nelfinavir treatment of various subtypes, and 967 sequences from PI naive patients were considered as a dataset for BN learning. Network semantics were based on the correspondence between an unconditional dependency implied by an arc between two mutations and a possible epistatic fitness interaction between these two mutations [52]. A mutation that confers phenotypic resistance on its own and plays a key role in drug resistance is considered a major mutation [51]. A minor mutation further increases resistance mostly only in presence of a major mutation, or compensates for a possible fitness impact of other mutations, and is therefore selected only in the presence of these other mutations. Since a minor mutation interacts epistatically with a corresponding major mutation, the BN indicates this relationship by an arc between these mutations. The presence of a minor mutation is predicted mostly by the presence of the corresponding major mutation, and thus expected to be unconnected to treatment in the network. Beerewinkel et al. [52], used a mixture of Bayesian tree models, which are a constrained version of BNs. This mixture of Bayesian tree models were used to model HIV drug resistance evolution from similar cross-sectional data. Their model represents ordered accumulation of mutations following a number of possible trees. This model captures antagonistic epistatic fitness effects between resistance mutations.

## 13 Artificial Neural Networks (ANN)

ANN is represented as a diagram of an interrelated group of artificial neurons, with information on the weights of the arcs that connect the neurons. It is characterized for the neuron model, the topology and the training algorithm. ANNs can be divided into two categories based on the topology: feed-forward and recurrent neural network. Feed-forward neural network are characterised by directed acyclic

graphs and recurrent neural networks are characterised by cycle graphs. Multilayer Perceptron is a generic example of feed-forward ANN [53]. The grouping of mutations in the genome sequences and its influence in phenotypic fold resistance values are non-linear in nature. While a mutation may increase the fold resistance towards a particular ARV drug, the presence of that mutation with an extra or added mutation in a genomic sequence may decrease the fold resistance towards the same ARV drug. ANN models help in discovering these non-linear attributes of the relationship between mutations and phenotypic fold resistance. A simple three layer structural design of a neural network consists of an input layer, a hidden layer and an output layer. A neural network architecture may have one or many hidden layers. Classification of genomic sequences as resistant or susceptible to a specific ARV drug involves processing a large number of inputs disseminating through the multiple hidden layers [54] to reach the preferred output while training. In this context the inputs are the mutations at each codon and the output is the phenotypic binary variable which indicates the sequence as resistant if 1 and susceptible if 0 (Fig. 1).

Xiaxia et al. [31] used ANN to classify genotype-phenotype data for resistance. Precisely, the three-layer feedforward network was used in Matlab [55, 56]. The network had one hidden layer of 20 nodes and was trained with backpropagation with a maximum of 50 training epochs. The values calculated for accuracy, sensitivity and specificity for resistance to PIs have a low of 0.91 and reach 0.97. Improved performance was achieved for classifying resistance to RT inhibitors compared with PIs. Results for NRTIs gave accuracy, sensitivity and specificity values of 0.96–0.99, while for NNRTIs all values were greater than 0.98.



**Fig. 1** The structural architecture of a neural network

## 14 K Nearest Neighbour (KNN) Algorithm

The KNN algorithm is a non-parametric method that uses the full training data set. It finds the K nearest neighbours to a query point and reports either their class by majority vote or the average of their resistance value. The advantage of applying KNN is that the training step is faster, unlike SVM, KNN uses the whole training data in the prediction stage because the result is reported based on the training data. Updating a KNN predictor with new experimental resistance data is especially straightforward and simply requires performing the feature extraction step on the new data [57].

## 15 The Random Forest (RF) Algorithm

The RF algorithm is an ensemble based classifier/regression that works with multiple decision trees to enhance the accuracy. The drawback of individual decision trees is that they are sensitive to small changes of selected features in the training space. Therefore an individual decision tree is a weak learner with a poor ability to generalize and a strong tendency to be unstable. RF uses the ensemble votes from multiple decision trees to improve the stability of trained machine as well as the prediction accuracy. In practice, the RF algorithm calculates the averaged value voted from different sub-trees that randomly built from the training dataset [57].

Shen et al. [6] used random forest and K-nearest neighbour,to handle genotype-phenotype datasets of HIV protease (PR) and reverse transcriptase (RT).The prediction accuracies were examined by five-fold cross-validation on the genotype-phenotype datasets. It predicts the drug resistance phenotype and its relative severity from a query sequence. The accuracy of the classification was higher than 0.973 for eight PR inhibitors and 0.986 for ten RT inhibitors, respectively. The overall cross-validated regression R<sup>2</sup>-values for the severity of drug resistance were 0.772–0.953 for 8 PR inhibitors and 0.773–0.995 for 10 RT inhibitors.

## 16 Rule Based Prediction

Simple rule based methods characterize the drugs by the mutational patterns which produce resistance. After mutational patterns are detected it has to be analyzed and interpreted in order to choose the suitable drug combination. Genotypic Interpretation Systems (GIS), have been developed [58] by an expert panel to predict the resistance to ARV from mutational pattern. These rules have information about the mutation or combination of mutations studied or seen before in experiments. List of mutations from the International Aids Society (IAS) is used to develop these systems [59].

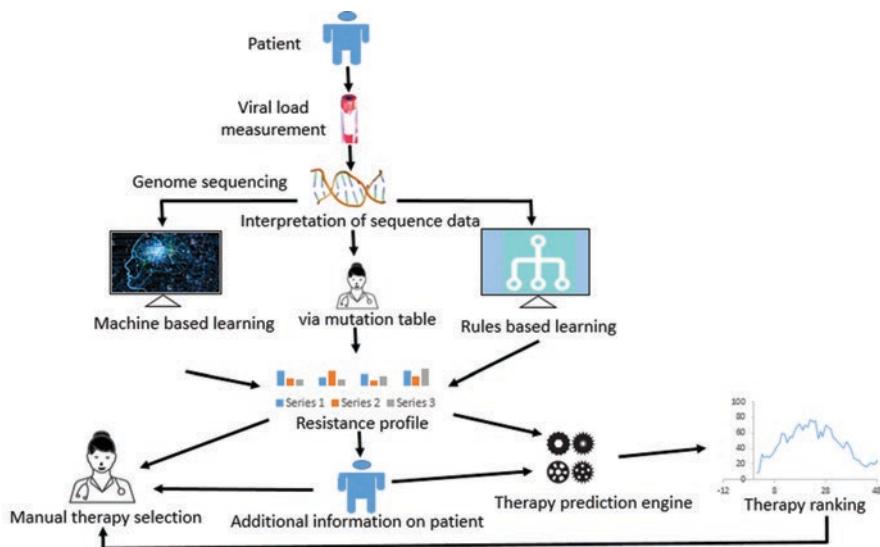
Rule-based algorithms have been shown to have many advantages over laboratory tests. One widely used rule based interpretation algorithm was built by the French ANRS (Agence Nationale de Recherches sur le SIDA). ANRS is observed as a gold standard in interpreting HIV drug resistance using mutations in genomes. ANRS classifies ARV resistance according to three levels: susceptible, intermediate, and resistant. ‘Susceptible’ designates that a particular ARV drug will be effective against HIV; ‘intermediate’ designates that the ARV drug is partially effective; and if the ARV is not effective at all, it is classified ‘resistant’. ANRS algorithm is based on a linear combination of mutations. If a particular mutation or a group of mutations are present in the genome, the algorithm returns a resistance profile applicable to that particular sequence. Each rule consists of a Boolean expression. For example, an ANRS rule for abacavir states: “If there are five or more of the following RT mutations (M41L,D67N, L74V, M184V/I, L210W, T215Y/F), report resistance to abacavir” [36, 60].

Alike ANRS [58, 61], the algorithms Rega-5.5 [62] and Visible Genetics v. 6 (VGI-6) [63] report three levels of the resistance: susceptible, resistant, and intermediate. The rules used in these algorithms are sets of the Boolean expressions. These are designed to provide reasonable interpretations for the large number of remaining possible mutation combinations. The Stanford University HIV Drug Resistance Database (Stanford HIVdb) algorithm [64] and mutation rate based score [65] outputs a total of five levels of the resistance: susceptible, potential low-level resistance, low-level resistance, intermediate resistance, and high-level resistance. The penalty score used in the algorithm is described as follows: for each mutation, a drug penalty score is given by the algorithm. The total scores are added and the sum is used to infer the final result from which the drug resistance category is obtained. Furthermore, a combined rule-based and penalty-based method called AntiRetroScan (ANS) is proposed and applied to both PI and RTI [66]. This system is developed and maintained by University of Sienna and Italian Antiretroviral Resistance Cohort Analysis Website respectively.

## 17 Molecular Structure Based Prediction Methods

Basically, the HIV drug resistance is caused by the alteration of the structure and the enzymes’ drug target sites. The molecular structure can also be used to predict the drug resistance value to the mutations to certain drug. This approach includes the molecular docking methods, the homology-based modeling methods [67], as well as the molecular dynamics simulations [68].

The computational structure-based methods are often used for structure optimization and scoring docked protein-ligand complex. Such procedures are similar to the drug resistance prediction, and could be used to solve this problem [69–71]. Combined sequence-structure approaches are also used to solve this task: a Delaunay tessellation derived four-body statistical potential mutagenesis method together with SVM and random forest classification methods is applied to predict the drug resistance for HIV-1 reverse transcriptase inhibitor, Nevirapine (NVP) and more



**Fig. 2** Drug resistance prediction and therapy plan pipeline

inhibitors [72]. Ekachai et al. [70] used protein–inhibitor docking with a molecular dynamics protocol that takes protein–inhibitor flexibility into account to determine the correlation between the experimentally determined inhibitory concentration (phenotype) and the computer calculated protease inhibitor binding affinities based on the HIV-1 protease gene mutations (genotype). Genotypic and phenotypic evaluations were performed by ViroLogic, Inc., [73] using population sequencing and PhenoSense HIV assays [74], respectively. They further compared the predictions with an established HIV-1 genotypic interpretation systems: a rule-based method Stanford HIV drug resistance database and an SVM-based method geno2pheno [75]. The overall HIV drug resistance prediction methods and therapy planning schema is depicted in Fig. 2. Various prediction systems both machine learning and rule based methods are listed in Table 2.

## 18 Conclusion

HIV infections and drug resistance are likely to continue as a problem in the absence of an effective vaccine, due to the high genetic variation and existence of poorly accessible reservoirs of virus. This highlights the critical need for the development of targeted treatment based on genotype data and new antiretroviral drugs for both therapy and pre-exposure prophylaxis. Both rule based and machine learning based prediction service providers are available online. Any scientist or physician can use them as a valuable resource for therapy decision making. Based on the prediction there is a scope for future research to develop effective, cheap and robust treatment regimen for HIV in lower and middle income countries.

**Table 2** Rule based and machine learning based web services for drug resistance prediction and combinatorial ARV regimen selection

S. No	System	Model	References
1	IAS/USA	Rule based learning	[59]
2	HIVdb	Rule based learning	[9]
3	Rega	Rule based learning	[62]
4	ANRS	Rule based learning	[58]
5	HIV-GRADE	Rule based learning	[37]
6	ResRIS	Rule based learning	[76]
7	AntiRetroScan	Rule based learning	[77]
8	Geno2pheno	Support vector machine	[75]
10	WebPSSM	Position Specific Scoring Matrices	[78]
11	Trophix	Logistic regression	[79]
13	HIV-TRePS	Random forest	[80]
14	EuResist	Logistic regression, Bayesian networks, Alternating decision trees	[81]

IAS International AIDS society, HIVdb HIV Drug Resistance Database, ANRS Agence Nationale de Recherches sur le SIDA, PSSM Position Specific Scoring Matrices

## References

1. World Health Organization.. HIV drug resistance report 2017. [www.who.int/hiv/pub/drugresistance/hivdr-report-2017/en/](http://www.who.int/hiv/pub/drugresistance/hivdr-report-2017/en/).
2. Kallings LO. The first postmodern pandemic: 25 years of HIV/AIDS. J Intern Med. 2008;263(3):218–43.
3. Beyer C, Pozniak A. HIV drug resistance – an emerging threat to epidemic control. N Engl J Med. 2017;377(17):1605–7.
4. Greene WC. A history of AIDS: looking back to see ahead. Eur J Immunol Nov:2007. <http://www.ncbi.nlm.nih.gov/pubmed/17972351/>.
5. Popovic M, Sarnagdharan MG, Read E, Gallo RC. Detection, isolation, and continuous production of cytopathic retroviruses (HTLV-III) from patients with AIDS and pre-AIDS. Science. 1984;224(4648):497–500.
6. Shen C, Yu X, Harrison RW, Weber IT. Automated prediction of HIV drug resistance from genotype data. BMC Bioinf. 2016;17(Suppl 8):278.
7. de Oliveira T, Shafer RW, Seebregts C. Public database for HIV drug resistance in southern Africa. Nature. 2010;464(7289):673.
8. UK HIV Drug Resistance Database, UCL Institute for Global Health in London. Available at: <http://www.hivrd.org.uk>.
9. Stanford HIV drug resistance database. Available at: <https://hivrd.stanford.edu>.
10. Cordes F, Kaiser R, Selbig J. Bioinformatics approach to predicting HIV drug resistance. Expert Rev Mol Diagn. 2006;6(2):207–15.
11. Carvajal-Rodríguez A. The importance of bio-computational tools for predicting HIV drug resistance. Recent Pat DNA Gene Seq. 2007;1(1):63–8.
12. Prosperi MCF, De Luca A. Computational models for prediction of response to antiretroviral therapies. AIDS Rev. 2012;14(2):145–53.
13. Durant J, et al. Drug-resistance genotyping in HIV-1 therapy: the VIRADAPT randomised controlled trial. Lancet. 1999;353(9171):2195–9.
14. Obermeyer Z, Emanuel EJ. Predicting the future – big data, machine learning, and clinical medicine. N Engl J Med. 2016;375(13):1216.

15. Yu X. "HIV drug resistant prediction and featured mutants selection using machine learning approaches". Dissertation, Georgia State University. 2014.
16. Singh Y, Mars M. Support vector machines to forecast changes in CD4 count of HIV-1 positive patients. *Sci Res Essays*. 2010;5(17):2384–90.
17. Goldbaum MH, Falkenstein I, Kozak I, Hao J, Bartsch DU, Sejnowski T, Freeman WR. Analysis with support vector machine shows HIV-positive subjects without infectious retinitis have MFERG deficiencies compared to normal eyes. *Trans Am Ophthalmol Soc*. 2008;106:196–204; discussion 204–5.
18. Singh Y, Narsai N, Mars M. Applying machine learning to predict patient-specific current CD 4 cell count in order to determine the progression of human immunodeficiency virus (HIV) infection. *Afr J Biotechnol*. 2013;12(23):3724–30.
19. Wang D, et al. A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy. *Artif Intell Med*. 2009;47(1):63–74.
20. Larder B, Wang D, Revell A. Application of artificial neural networks for decision support in medicine. *Methods Mol Biol*. 2008;458:123–36.
21. Li Y, Rapkin B. Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS. *J Clin Epidemiol*. 2009;62(11):1138–47.
22. Munoz-Moreno JA, et al. Classification models for neurocognitive impairment in HIV infection based on demographic and clinical variables. *PLoS One*. 2014;9(9):e107625.
23. Choi I, et al. Machine learning methods enable predictive modeling of antibody feature: function relationships in RV144 vaccines. *PLoS Comput Biol*. 2015;11(4):e1004185.
24. Revell AD, et al. The use of computational models to predict response to HIV therapy for clinical cases in Romania. *Germs*. 2012;2(1):6.
25. Dietterich TG. Ensemble methods in machine learning. [cited 2016 30th/Aug/]; Available from: <http://www.cs.orst.edu/~tgd>.
26. Bonet I, Rodríguez A, Grau Ábalo R, García MM, Saeys Y, Nowé A. In: Gelbukh A, Morales EF, editors. *MICAI 2008, LNAI 5317: Comparing distance measures with visual methods*. Berlin/Heidelberg: Springer; 2008. p. 90–9.
27. Jordan M, Mitchell T. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349(6245):255–60.
28. White AD. Complexity of human immunodeficiency virus management in developing countries. *Epidemiology*. 1998;9(6):593–5.
29. Zhu X, Goldberg AB. Introduction to semi-supervised learning. *Synth Lect Artif Intell Mach Learn*. 2009;3(1):1–130.
30. Tan P-n, Steinbach M, Kumar V. *Introduction to data mining*. New York: Pearson Education, Limited; 2014.
31. Yu X, Weber IT, Harrison RW. Prediction of HIV drug resistance from genotype with encoded three-dimensional protein structure. *BMC Genomics*. 2014;15(Suppl 5):S1.
32. Beerenwinkel N, et al. Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. *Proc Natl Acad Sci U S A*. 2002;99(12):8271–6.
33. Wang D, Larder B. Enhanced prediction of Lopinavir resistance from genotype by use of artificial neural networks. *J Infect Dis*. 2003;188(5):653–60.
34. Beerenwinkel N, et al. Geno2pheno: estimating phenotypic drug resistance from HIV-1 genotypes. *Nucleic Acids Res*. 2003;31(13):3850–5.
35. Deforche K, et al. Analysis of HIV-1 pol sequences using Bayesian networks: implications for drug resistance. *Bioinformatics*. 2006;22(24):2975–9.
36. Liu TF, Shafer RW. Web resources for HIV type 1 genotypic-resistance test interpretation. *Clin Infect Dis*. 2006;42(11):1608–18.
37. Obermeier M, et al. HIV-GRADE: a publicly available, rules-based drug resistance interpretation algorithm integrating bioinformatic knowledge. *Intervirology*. 2012;55(2):102–7.
38. Brun-Vezinet F, et al. Clinically relevant interpretation of genotype for resistance to abacavir. *AIDS*. 2003;17(12):1795–802.

39. Humphris-Narayanan E, Akiva E, Varela R, Ó Conchúir S, Kortemme T. Prediction of mutational tolerance in HIV-1 protease and reverse transcriptase using flexible backbone protein design. *PLoS Comput Biol*. 2012;8(8):e1002639.
40. Hastie T, Tibshirani R, Friedman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer; 2009.
41. Perno CF, Bertoli A. Clinical cut-offs in the interpretation of phenotypic resistance. In: Geretti AM, editor. Antiretroviral Resistance in Clinical Practice. London: Mediscript; 2006.
42. Winters B, et al. Determination of clinically relevant cutoffs for HIV-1 phenotypic resistance estimates through a combined analysis of clinical trial and cohort data. *J Acquir Immune Defic Syndr*. 2008;48(1):26–34.
43. Bonet I. Machine learning for prediction of HIV drug resistance: a review. *Curr Bioinforma*. 2015;10(5):579–85.
44. Vapnik VN. The nature of statistical learning theory. New York: Springer; 2000.
45. Araya ST, Hazelhurst S. Support vector machine prediction of HIV1 drug resistance using the viral nucleotide patterns. *Trans Roy Soc S Afr*. 2009;64(1):62–72.
46. Canu S, Grandvalet Y, Guigue V, Rakotomamonjy A. SVM and kernel methods MATLAB toolbox. Perception Systemes et Information. Rouen: INSA de Rouen; 2005.
47. The MathWorks Inc. <http://www.mathworks.com>.
48. Pearl J, Gabbay DM, Smets P. Graphical models for probabilistic and causal reasoning, Handbook of defeasible reasoning and uncertainty management systems, Volume 1: quantified representation of uncertainty and imprecision. 1998;1:367–389.
49. Heckerman D. A tutorial on learning with Bayesian networks. Learning in graphical models. Cambridge: MIT Press; 1999. p. 301–54.
50. Klingler TM, Brutlag DL. Discovering structural correlations in  $\alpha$ -helices. *Protein Sci*. 1994;3(10):1847–57.
51. Shafer RW. Genotypic testing for human immunodeficiency virus type 1 drug resistance. *Clin Microbiol Rev*. 2002;15(2):247–77.
52. Beerenwinkel N, et al. Learning multiple evolutionary pathways from cross-sectional data. *J Comput Biol*. 2005;12:584–98.
53. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, editors. Parallel distributed processing: explorations in the microstructure of cognition. Cambridge: MIT Press; 1986. p. 318–62.
54. Tamura S, Tateishi M. Capabilities of a four-layered feedforward neural network: four layers versus three. *IEEE Trans Neural Netw*. 1997;8(2):251–5.
55. Hornik K, Stinchcombe M, White H. Multilayer feedforward networks are universal approximators. *Neural Netw*. 1989;2(5):359–66.
56. Howard D, Beale M. Neural network toolbox, for use with MATLAB, User's guide, version 4. Natick: The MathWorks Inc; 2000. p. 133–05.
57. Adeniyi DA, Wei Z, Yongquan Y. Automated web usage data mining and recommendation system using K-Nearest Neighbor (KNN) classification method. *Appl Comput Inform*. 2016;12(1):90–108.
58. Agence Nationale de recherches sur le SIDA. <http://www.hivfrenshresistance.org>. Accessed 2 Feb 2013.
59. Johnson VA, Calvez V, Gunthard HF, et al. Update of the drug resistance mutations in HIV-1. *Top Antivir Med*. 2013;21(1):6–14.
60. Yashik S, Maurice M. Predicting a single HIV drug resistance measure from three international interpretation gold standards. *Asian Pac J Trop Med*. 2012;5(7):566–72.
61. Meynard JL, et al. Phenotypic or genotypic resistance testing for choosing antiretroviral therapy after treatment failure: a randomized trial. *AIDS*. 2002;16(5):727–36.
62. Van Laethem K, et al. A genotypic drug resistance interpretation algorithm that significantly predicts therapy response in HIV-1-infected patients. *Antivir Ther*. 2002;7(2):123–9.
63. Ravela J, et al. HIV-1 protease and reverse transcriptase mutation patterns responsible for discordances between genotypic drug resistance interpretation algorithms. *J Acquir Immune Defic Syndr*. 2003;33(1):8.

64. Rhee SY, et al. Human immunodeficiency virus reverse transcriptase and protease sequence database. *Nucleic Acids Res.* 2003;31(1):298–303.
65. Schmidt B, et al. Simple algorithm derived from a geno-/phenotypic database to predict HIV-1 protease inhibitor resistance. *AIDS.* 2000;14(12):1731–8.
66. Zazzi M, et al. Comparative evaluation of three computerized algorithms for prediction of antiretroviral susceptibility from HIV type 1 genotype. *J Antimicrob Chemother.* 2004;53(2):356–60.
67. Shenderovich MD, et al. Structure-based phenotyping predicts HIV-1 protease inhibitor resistance. *Protein Sci.* 2003;12(8):1706–18.
68. Jenwitheesuk E, Samudrala R. Improved prediction of HIV-1 protease-inhibitor binding energies by molecular dynamics simulations. *BMC Struct Biol.* 2003;3(1):2.
69. Cao ZW, et al. Computer prediction of drug resistance mutations in proteins. *Drug Discov Today.* 2005;10(7):521–9.
70. Jenwitheesuk E, Samudrala R. Prediction of HIV-1 protease inhibitor resistance using a protein-inhibitor flexible docking approach. *Antivir Ther.* 2005;10(1):157–66.
71. Ravich VL, Masso M, Vaisman II. A combined sequence–structure approach for predicting resistance to the non-nucleoside HIV-1 reverse transcriptase inhibitor Nevirapine. *Biophys Chem.* 2011;153(2):168–72.
72. Masso M, Vaisman II. Sequence and structure based models of HIV-1 protease and reverse transcriptase drug resistance. *BMC Genomics.* 2013;14(Suppl 4):S3.
73. Kellam P, Larder BA. Recombinant virus assay: a rapid, phenotypic assay for assessment of drug susceptibility of human immunodeficiency virus type 1 isolates. *Antimicrob Agents Chemother.* 1994;38(1):23–30.
74. Hertogs K, et al. A rapid method for simultaneous detection of phenotypic resistance to inhibitors of protease and reverse transcriptase in recombinant human immunodeficiency virus type 1 isolates from patients treated with antiretroviral drugs. *Antimicrob Agents Chemother.* 1998;42(2):269–76.
75. de Mendoza C, et al. HIV-1 genotypic drug resistance interpretation rules – 2009 Spanish guidelines. *AIDS Rev.* 2009;11(1):39–51.
76. Anta L, et al. Resistance to the most recent protease and non-nucleoside reverse transcriptase inhibitors across HIV-1 non-B subtypes. *J Antimicrob Chemother.* 2013;68(9):1994–2002.
77. ARCA AntiRetroScan. Available at: [http://www.hivarca.net/hiv\\_resistance.asp](http://www.hivarca.net/hiv_resistance.asp).
78. WebPSSM. Available at: <http://indra.mullins.microbiol.washington.edu/webpssm/>.
79. Trophix (prediction of HIV-1 tropism). Available at: <http://sourceforge.net/projects/trophix/>.
80. RDI HIV-TRePS. Available at: <http://www.hivrdi.org/treps/login.php>.
81. The EuResist engine. Available at: <http://engine.euresist.org>.

# Swarm Intelligence in Cell Entry Exclusion Phenomena in Viruses and Plasmids: How to Exploit Intelligent Gene Vector Self-Scattering in Therapeutic Gene Delivery



Oleg E. Tolmachov

**Abstract** Many naturally occurring viral and bacterial gene transfer systems present ‘superinfection interference’ phenomena, where expression of additional viral genomes or bacterial plasmids in cells previously infected with a homologous virus or plasmid is limited or altogether absent. The lack of expression of superinfecting homologous genetic material could be due to either denial of its cell entry where a failed cell entry attempt is followed by the return of the superinfecting genomes back into the infecting pool, or, alternatively, to a block of transfer, expression or replication of the superinfecting genomes, where these genomes are being deactivated and eventually destroyed. The distinction between these two scenarios is important because the return of the superinfecting genomes back into the pool of circulating infectious genomes allows productive examination of such superinfection interference phenomena within the abstract framework of swarm intelligence, where viral or plasmid genomes are being viewed as individual ‘agents’. Signaling between mobile agents is a crucial element of swarm intelligence. Thus, in superinfection interference, the denial of cell entry to the superinfecting viral/plasmid genomes (‘circulating’ agents) by the cell-resident viral/plasmid genomes (‘settler’ agents) can be regarded as a signaling event. In virology and plasmid biology, occurrences of denial of cell entry to superinfecting genomes can be met in various settings; the most commonly used terms for these phenomena are ‘cell entry exclusion’ and ‘cell surface exclusion’.

So, being concerned with the intelligent swarm-level behavior, this chapter is focused only on the ‘strict non-admittance’ subset of the cell entry exclusion phenomena where the superinfecting agents are recycled back into the ‘circulating’ infectious pool and are not retained and destroyed by the recipient cells. Considered within the framework of swarm intelligence, these genome agents will exhibit intelligent swarm-based self-scattering behavior. In general, such signaling from the ‘settler’ agents to the homologous ‘circulating’ agents can be accomplished through the knockout of a viral/plasmid receptor or another positive gene transfer factor (passive denial of entry) or through the expression of a gene transfer rejection factor

---

O. E. Tolmachov (✉)  
Imperial College London, London, UK  
e-mail: [epsilon@tinyworld.co.uk](mailto:epsilon@tinyworld.co.uk); [oleg.tolmachev@ieee.org](mailto:oleg.tolmachev@ieee.org)

(active denial of entry). This chapter illustrates swarm-based intelligent self-scattering behavior with appropriate examples from known superinfection interference phenomena and conjectures how such behavior of viral/plasmid genomes can be exploited for gene vector scattering needs in therapeutic gene delivery with viral and non-viral gene vectors.

It is anticipated that that intelligent self-scattering of gene delivery agents in gene vector swarms can be employed in artificial gene transfer systems delivering therapeutic genes to human cells in order to avoid undesired multicopy gene conveyance and, thus, to achieve uniform transgene copy-number distribution and transgene expression at an unvaried curatively-effective level in target cells. Thus, it is expected that the exploitation of the intelligent self-scattering capability of swarms of therapeutic gene vectors will find a wide range of applications in gene therapy, particularly where, on the one hand, therapeutic gene delivery is required to be highly concentrated and efficient and, on the other hand, uniform transgene dosage is critical for the optimal curative effect.

**Keywords** Self-scattering · Self-focusing · Superinfection · Swarm · Intelligent behavior · Naturally occurring viruses · Naturally occurring plasmids

### Key Phrases

Gene therapy; Gene vectors; Swarm intelligence in viruses; Swarm intelligence in plasmids; Superinfection interference; Entry exclusion; Surface exclusion; Intra-swarm signaling; Inter-agent signaling; Extracellular receptor knockdown.

## 1 Introduction

While viruses and bacterial plasmids belong to diverse biological systems, they share an intra-cellular parasitic lifestyle and can commonly be viewed as gene vectors, capable of delivering their genome cargo to recipient cells through viral infection, transfection and transduction or plasmid conjugation, mobilization, and transformation. The similarities of lifestyle between various viruses and plasmids, in particular, the similarities in their cell-to-cell transmission modes, result in the similarity of their adaptations. One example of strikingly similar behavior in many natural viral and bacterial gene transfer systems is the phenomenon of ‘superinfection interference’ [1], where the expression of surplus viral or plasmid genomes in cells, which have been previously infected with a homologous or related genome, is limited or altogether absent.

The universal occurrence of superinfection interference in widely diverse viral and plasmid systems suggests that it is adaptive. So, why could this behavior of viruses and plasmids be useful to them? Firstly, the ‘egoistic genome’ of the original infection suppresses any possible competitor genomes, which can be similar but

still genetically distinct, thereby ensuring its own preferential propagation. Secondly, reduction of initial parasitic gene load and cytotoxicity impact could be beneficial for the host cell and hence for the orderly propagation of the original infection agent strategically exploiting cell resources. Thirdly, rejection of secondary infection agents, after failed entry attempts, back into the circulating pool could support the parasitic goal of infecting a greater number of recipient cells using the available pool of infectious agents, with the additional advantage of generating potentially greater total infectious progeny from recipient cells that were not weakened by a prior infection. Fourthly, minimization of the initial parasitic gene load could be useful to avoid overexpression of viral genes in order to camouflage a viral infection event to avoid detection by the immune system. Fifthly, superinfection by virions with defective genomes, known as ‘defective interfering particles’, or defective plasmids could waste cell energy and material resources and, thereby interfere with the production of infectious propagation-competent progeny. Sixthly, the production of infectious propagation-competent progeny could also be undermined through genetic recombination between different versions of infectious agents (e.g., reassortment in the segmented RNA viruses) generating attenuated offspring. Seventhly, rapid establishment of the superinfection interference epigenetic cell regime, after initial infection, provides a clear selective pressure for non-attenuated viral/plasmid versions that are capable of prompt cell entry and fast transgene expression.

In different contexts and settings within virology and plasmid biology, the particular occurrences of superinfection interference are variously known as ‘immunity to superinfection’ or ‘superinfection immunity’ [2, 3], ‘plasmid incompatibility’ [4], ‘superinfection exclusion’ [5], ‘superinfection resistance’ [6], ‘homologous interference’ [7], ‘cell entry exclusion’ and ‘cell surface exclusion’ [8].

The specific molecular mechanisms of superinfection interference for various viruses and plasmids are extremely diverse. In general, superinfection is blocked because superinfecting genomes are not expressed. It is possible to identify two alternative scenarios. Firstly, in some naturally occurring systems, the lack of expression of superinfecting genomes is due to blocks of transfer, expression or replication of the superinfecting genomes, with these genomes being deactivated by the infected recipient cell and eventually destroyed. In this scenario a superinfecting agent succeeds in entering the infected cell and the expression blocking mechanism is predominantly intracellular. Secondly, in some other naturally occurring systems, the lack of expression of superinfecting genomes is due to the outright denial of their cell entry where failed cell entry attempts are followed by the return of the superinfecting genomes back into the infecting pool. These ‘strict non-admittance’ mechanisms are predominantly extracellular and are typically met among phenomena referred to as ‘cell surface exclusion’ or ‘cell entry exclusion’. Obviously, extracellular mechanisms, where superinfecting viral particles or plasmids are rejected and free to infect susceptible non-infected cells, result in the ‘economy’ of infectious particles for further infections. In contrast, for intracellular mechanisms of superinfection interference such ‘economy’ is lacking as the superinfecting agents lose their infecting potential after penetration of the cell and instead are reutilized by the cell.

While intracellular homologous interference mechanisms do not offer the benefit of maximizing the infection potential of the pool of infectious agents, they are still extremely wide-spread in nature. This indicates the critical importance of other advantages of superinfection interference, including, most prominently, safeguarding the ‘egoistic gene’ interests of the original infecting agent through blocking the reproduction of related but potentially competing genetic material. In addition, there is a clear disproportion between a multitude of intracellular molecular opportunities to preclude the expression of the superinfecting genome and the smaller number of molecular opportunities to prevent cell entry of a superinfecting agent, which follows exactly the same entry route as the resident infecting agent. Indeed, the molecular possibilities of the cell modifying its surface according to the instructions of the resident virus or plasmid are limited by the fact that for their cell entry, viruses and plasmids exploit molecular features of the cell surface, which are often linked to essential cellular functions such as nutrient trans-membrane transport.

For technological innovation it is important to study closely how naturally-occurring systems perform, but not necessarily to trace them literally. For example, modern aircraft do copy some features of flying birds but only a few, as the purpose and technological possibilities differ in natural and artificial systems. Similarly, while ‘strict non-admittance’ cell entry exclusion constitutes only a small subset of the general superinfection interference phenomena, this chapter argues that it is this form of superinfection interference that has the substantial potential to be developed to optimize artificial gene transfer systems, most prominently in gene therapy. With this in mind, the distinction between two superinfection interference scenarios above is important, because the return of the superinfecting genomes back into the pool of circulating genomes allows the productive examination of such superinfection interference phenomena within the framework of swarm intelligence.

Swarm intelligence is the ability of a set of simple agents to behave in a highly optimized way relying on information transfer between the individual simple agents, which are typically mobile. So, complex behavior of ant colonies, bird flocks, fish schools, and growing bacterial colonies was analyzed using the swarm intelligence framework, as reviewed in [9]. Swarm-level (‘global’) optimization was also described for various technological systems of agents like cars moving in traffic [10], small satellites [11], and small drones [12]. Within a biological context, the concept of swarm intelligence was further expanded to include the globally optimized behavior of such diverse systems as immune networks [13] and growing plant root apices [14].

Thus, in the ‘strict non-admittance’ cell entry exclusion scenario, viral or plasmid genomes are being viewed as individual ‘agents’ and the denial of cell entry to the superinfecting viral/plasmid genomes (‘circulating agents’) by the cell-resident viral/plasmid genomes (‘settler agents’) is being viewed as a signaling event. Clearly, ‘circulating’ agents and ‘settler’ agents could be structurally different, as ‘circulating’ agents are true physical particles with intact infectivity while ‘settler’ swarm members are non-infectious and gradually lose their form of individual compact particles within the recipient cells, e.g., by integrating into the host genome and retaining just their genetic content. This situation can be accommodated within the

swarm intelligence approach by thinking of the swarm as a swarm of heterogeneous agents [15].

The advantage of the ‘swarm intelligence’ view of the entry exclusion phenomena is that it provides guidance for the creation of artificial gene transfer systems with desired qualities, by borrowing some select features of the naturally occurring gene transfer systems. So, this chapter focuses only on the ‘strict non-admittance’ subset of the cell entry exclusion phenomena where the superinfecting agents are recycled back into the circulating pool without loss of their infectivity and are not retained and destroyed by the primarily infected recipient cells.

It is useful to employ some effects of swarm intelligence for therapeutic gene delivery, where there are currently multiple challenges on the way to successful curative applications [16–18]. Thus, artificial gene vectors exhibiting swarm intelligence were proposed to improve cell targeting [19]. In these gene vectors, inter-agent (peer-to-peer or intra-swarm) signaling, a pivotal element of swarm intelligence, is accomplished through the expression of extra-cellular receptors on cells, which were initially infected with epigenome-probing vector particles and assessed by them as suitable targets for circulating vector particles with a therapeutic gene load. So, displayed extra-cellular receptors allow cell entry of circulating therapeutic vector particles, thereby creating the effect of ‘self-focusing’ of gene vectors on suitable target cells. Thus, gene vector particles that are capable of peer-to-peer (inter-agent) signaling could form swarms with useful intelligent behavior, such as cell-specific self-focusing on target cell populations, relying on ‘scout’ vector particles interrogating various cellular biomarkers and reporting cell-targeting information to ‘therapeutic’ vector particles.

With their sophisticated cell-to-cell transmission machinery, plasmids and viral genomes are essentially efficient gene vectors. So, in the naturally occurring gene transfer systems displaying ‘strict non-admittance’ cell entry exclusion phenomena, a cell-resident ‘settler’ agent, which has already infected the cell, signals to ‘circulating’ agents to prohibit their entry into the particular occupied host cell. Thus, the globally optimized behavior of the swarm of suitable viral or plasmid genomes is to avoid previously infected cells and to spread. This behavior can be described as intelligent self-scattering of the swarm of viral or plasmid genomes.

Effective and side-effect-free gene therapy relies on targeted delivery of a therapeutic gene load to a specific cell population where transgenes are then uniformly expressed at a defined optimal level [17]. These requirements are often self-contradictory as an efficient focused gene delivery results in high multiplicity of infection (MOI) with a varied number of gene copies being introduced into the individual cells by the vector, producing a non-uniform distribution of the level of therapeutic transgene expression among the treated cells. Still, notwithstanding the disadvantages of high MOI, the risk of undesired side-effects dictates the overriding need for the focusing of gene transfer on a specific curatively-important cell population. The focusing is usually accomplished using: (1) ‘non-intelligent’ cell-specific gene delivery relying on the attachment of cell-targeted gene vector particles and pre-existing extracellular cell-specific ligands [20]; (2) physical focusing using targeted vector administration *in vivo* [21] or isolation of the target cell population for

gene transfer *ex vivo* followed by implantation of the genetically modified cells back into the body [22]. These methods of gene vector focusing could be supplemented with combinatorial targeting, which attains increased targeting resolution through *in situ* intra-body assembly of functional gene transfer systems from several elements capable of independent spread [23] and through intelligent swarm-based gene vector self-focusing [19].

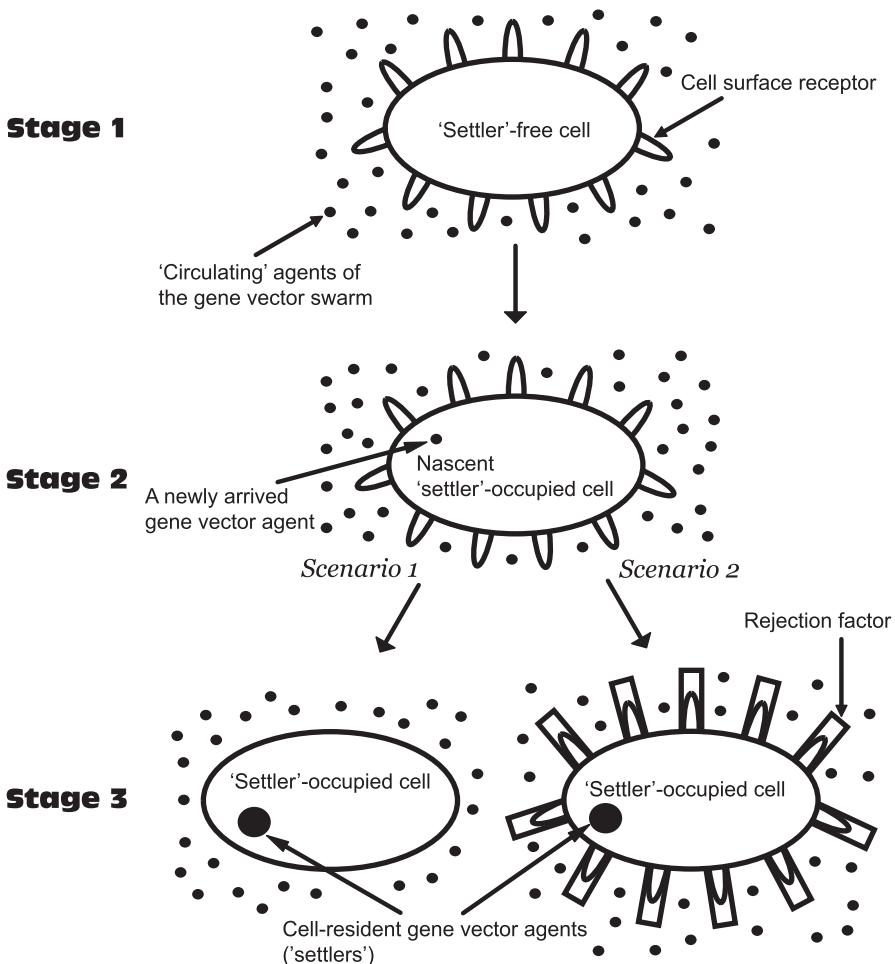
So, while gene delivery should be focused on a specific tissue requiring genetic treatment, the spread of therapeutic vector particles within the tissue is necessary to ensure the maximal number of target cells are reached, through the uniform distribution of the vector particles among the target cells where therapeutic transgenes are expressed at a uniform optimal level. Therefore, it is promising to create therapeutic gene vectors with inbuilt self-scattering mechanisms, mimicking the intelligent self-scattering molecular machinery of naturally occurring viruses and plasmids.

This chapter illustrates swarm-based intelligent self-scattering behavior in naturally occurring viruses and plasmids with appropriate examples from known super-infection interference phenomena and conjectures how such behavior of viral/plasmid genomes can be exploited for gene vector scattering needs in therapeutic gene delivery with viral and non-viral gene vectors.

## 2 Mechanisms of the Superinfection Interference Phenomena Viewed as a Particular Case of Swarm Intelligence

When cell entry exclusion phenomena are viewed within the framework of swarm intelligence, viral and plasmid genomes can be seen to exhibit intelligent swarm-based self-scattering behavior. As swarm agents are typically simple and carry only basic communication machinery, it is not unusual for the intra-swarm signaling between the agents to be indirect, with substantial reliance on the elements of the swarm's environment. Such indirect communication of agents that is mediated by modification of the environment is known as 'stigmergy'. A classic example of stigmergy is inter-agent interaction through pheromone-scented trails in ant colonies, which plays a key role in generating optimized foraging behavior [24]. Similarly, information transfer between gene vector particles with swarm-based global behavior can be modeled to be indirect, being achieved through manipulation of the host cells by the gene vector agents to express or, alternatively, to knock out a non-diffusible extracellular ear-marking receptor [19].

In general, within the context of the 'strict non-admittance' subset of the cell entry exclusion phenomena, such signaling from the 'settler' agents to the homologous 'circulating' agents can be accomplished: (1) through the knockout or knock-down of an extracellular viral/plasmid receptor (passive denial of entry); (2) through the expression of a rejection factor (active denial of entry). Thus, signaling for



**Fig. 1** Two scenarios for the self-scattering behavior of viral or plasmid agents with signaling via a non-diffusible cell surface receptor or a rejection factor

self-scattering using non-diffusible extracellular markers can be achieved through: (1) the knockout or knockdown of a cell surface receptor for vector particles in a 'settler'-occupied cell; or (2) the flag-up of a rejection factor, which blocks gene transfer to a 'settler'-occupied cell (Fig. 1).

So, in Fig. 1, viral or plasmid particles/genomes are abstractly viewed as 'gene vector agents'. Agents can be 'circulating', if they are outside the cells and infectious, and 'settlers', if they are cell-resident and non-infectious. The diagram shows a primary cell infection with a gene vector agent resulting in the 'strict non-admittance' cell entry exclusion of any superinfecting agents. Denial of the cell entry constitutes a signaling event within the swarm of gene vector agents. There are two alternative scenarios, which are detailed below, for the cell entry exclusion to arise.

In Stage 1, a naïve (non-infected, that is, ‘settler’-free) recipient cell is shown surrounded by ‘circulating’ agents. In Stage 2, a transient state of the primary infection is shown where a ‘circulating’ agent has successfully entered the recipient cell and is establishing itself within the cell through expression of its genes, turning itself into the ‘settler’ agent. In Stage 3, the ‘settler’ agent is shown to have enacted the cell entry exclusion regime in the recipient cell. This could have happened through Scenario 1, where expression of the ‘settler’-agent genes caused knockout of the cell surface receptors preventing secondary infections. Alternatively, this could have happened through Scenario 2, where expression of the ‘settler’-agent genes resulted in the emergence (‘flag-up’) of a ‘rejection factor’ blocking secondary infections.

‘Circulating’ agents, which are denied entry into the ‘settler’-occupied cells, are free to infect new ‘settler’-free cells, thereby creating the effect of the swarm-based self-scattering behavior. The signaling within the swarm is mediated by the cell surface receptors or rejection factors, which are cell-bound and, therefore, non-diffusible.

Indeed, one such signaling mechanism, which exists in ecotropic Murine Leukemia Virus (MLV), is based on the downregulation of the extracellular virus receptor CAT-1 during initial MLV infection, which results in a block of any subsequent MLV infections. The CAT-1 receptor is removed from the cell surface through its intracellular sequestration by the viral envelope anti-receptor protein expressed by a viral particle, which has already successfully infected the cell [6]. The fact that extracellular ecotropic MLV receptors, which are responsible for virus-cell attachment and membrane fusion, decay without being replenished in a primarily infected cell means that virions’ secondary collisions with the cell leave their infectivity undiminished. So, in the naturally occurring self-scattering setting, an intelligent swarm of ecotropic MLV retroviral particles delivering their cargo of genomic RNA uses knockout of the receptor protein CAT-1 to self-scatter and, thus, to achieve a more even distribution of delivered viral genomes throughout the recipient cell population. Thus, a particular type of superinfection interference in ecotropic MLV can be proposed as a prototype of swarm intelligence in viruses.

Often in naturally occurring gene transfer systems there are restrictions on the receptor-knockout method of communication between agents. Indeed, it is common for extracellular receptors to perform several functions. Therefore, the receptor knockout strategy for inter-agent signaling can be limited by essential cellular functions of the receptors to be knocked out. In the ecotropic MLV infection system, CAT-1 also performs an essential function of the cationic amino acid transporter, so MLV possesses a molecular mechanism for the receptor knockout to be non-lethal for the cell. The mechanism relies on one of the glycosylated versions of CAT-1 lacking affinity to its viral anti-receptor but still, fortuitously, retaining its essential cationic-amino-acid transporter activity [25].

Influenza A virus from the Orthomyxoviridae family follows a superinfection resistance strategy, which is somewhat similar to the above strategy of ecotropic MLV. One of the major Influenza A viral proteins, enzyme neuraminidase (NA), was shown to block sialic acid modification of intracellularly maturing glycoproteins,

thereby downregulating the extracellular sialic acids' residues, which are well-established Influenza A viral receptors [26]. This ensures superinfection exclusion for the infected cells expressing NA and, in addition, clears the way for the safe release of the progeny virions, which could otherwise be trapped by the sialic acids' residues. Sendai virus from the Paramyxoviridae family, which also uses sialic acids' residues as cell receptors, was also shown to employ NA both for superinfection interference and for efficient release of the viral progeny [27]. As with ectropic MLV, a block of superinfection through the denial of virus-to-cell attachment, leaves superinfecting Influenza A and Sendai virions intact and infectious, yielding to the clear-cut self-scattering swarm interpretation.

It is common for several superinfection interference mechanisms to co-exist in a single viral or plasmid system. For example, in Vaccinia virus from the Poxviridae family, viral proteins A56 and K2 in the cell membrane are known to mediate superinfection resistance by interacting with the entry-fusion complex of re-infecting virions subsequent to membrane fusion [28]. In addition, a second superinfection exclusion mechanism in Vaccinia virus was discovered, which acted at the post-attachment stage prior to the fusion of viral and cell membranes [28]. Poxviridae are the largest viruses infecting mammalian cells and have a substantial genetic capacity (Vaccinia virus genome size is about 190 kbp), so it is not surprising that in these viruses a single function is achieved through several molecular pathways. However, indicating the importance of superinfection interference for the viral lifestyle, dual mechanisms of superinfection exclusion were also reported for a much smaller Bovine Viral Diarrhea Virus (BVDV) from the Flaviviridae family with the genome size of about 12.5 kb [7]. Secondary BVDV infections were blocked both at the level of viral entry and at the level of viral RNA replication [7]. Notwithstanding their attachment to a primarily infected cell, the bulk of rejected superinfecting Vaccinia virions retained their infectivity for susceptible cells [28], while the infectivity of rejected BVDV virions was not reported [7]. Provided the rejected cell-attached superinfecting Vaccinia virions are given a chance to re-infect susceptible naïve cells, they could be assumed returned back to the infecting pool, and the observed superinfection interference can be straightforwardly interpreted as self-scattering behavior of an intelligent swarm of viral agents.

In transmission systems of conjugative plasmids, molecular machinery for superinfection interference often overlaps with machinery for blocking conjugative transfer from one donor cell to another donor cell. Thus, in *Escherichia coli* F-factor plasmid system, entry exclusion involves two constitutively-expressed independently-operating rejection factors, TraT and TraS, which are active both in the denial of secondary conjugation events with exconjugant cells and in donor-to-donor DNA transfer [8]. In addition, F-factor plasmids interfere with each other through a plasmid incompatibility mechanism, which involves gradual plasmid loss in dividing cells [4]. Some naturally occurring non-conjugative plasmids, such as mobilized by F-factor plasmid ColE1, code for their own rejection factors, so for these plasmids with helper-dependent DNA transmission, superinfection interference occurs through their own systems of replicon incompatibility and cell entry exclusion [29]. Although in the F-factor and ColE1 entry exclusion systems the actual copies of the abortively transferred plasmids might not be used for productive

re-infection, the plasmid transfer scenario involves disruption of secondary mating events and ensuing return of the unsuccessful donor cells back to the pool of the available donors, consequently yielding to the interpretation within the framework of the self-scattering gene vector swarms.

### 3 A Unified Swarm Intelligence Framework for the Self-Scattering and Self-Focusing Behaviors of Gene Vectors

Thus, in the viral and plasmid systems with intelligent self-scattering there are two functional types of agents: (1) ‘settler’ agents, which have successfully reached their final destination within susceptible cells; (2) ‘circulating’ agents, which continue to move in the vicinity of the susceptible cells. The self-scattering of gene vector agents depends on the successfully intracellularly established ‘settler’ agents signaling to ‘circulating’ agents to prohibit their cell entry. In a broader perspective, intelligent self-scattering of viruses and plasmids is a form of spatial re-distribution of gene delivery. Thus, self-scattering is due to passive non-admittance or, theoretically, even active repulsion of circulating agents from the settler-populated cells. In contrast, if the settler agents would offer passive admittance or, theoretically, active attraction of circulating agents to the settler-populated cells, this would result in another mode of swarm-based spatial redistribution, intelligent self-focusing. Self-focusing on the best nectar source is described for foraging honeybees [30] but so far was not recognized in naturally occurring gene transmission systems of viral and plasmid agents. Undoubtedly, this is because powerful selective pressures drive the evolution of viruses and plasmids in the opposite direction, towards superinfection interference, and, therefore, in many cases to intelligent self-scattering. However, it was proposed that spatial distribution of artificial therapeutic gene vectors in the human body could be optimized through self-focusing swarm-level behavior [19].

In swarms of therapeutic gene vectors with intelligent cell-specific self-focusing, there are two functional types of agents: (1) ‘scout’ vector particles, which interrogate cellular biomarkers and generate appropriate cell ear-marking signals; and (2) ‘therapeutic’ vector particles, which deliver therapeutic genetic cargo to the cells ear-marked by the ‘scout’ agents [19].

It is common for swarm members (agents) to play different roles in intelligent global behavior. In general, the distinction between swarm agents can be purely functional, with all functional types of agents being structurally identical (a homogeneous swarm). Alternatively, swarm agents playing different functional roles within a swarm can be structurally distinct (a heterogeneous swarm). For example, self-focusing swarms of gene vectors can be homogeneous, with structurally identical ‘scout’ and ‘therapeutic’ agents, or heterogeneous, with structurally distinct ‘scout’ and ‘therapeutic’ agents [19].

As ‘settler’ agents in self-scattering gene vector swarms are cell-bound and cell-modified, they cannot be structurally identical to ‘circulating’ agents. Therefore, all

'settler'-dependent self-scattering vector swarms are necessarily heterogeneous. However, clearly, the membership in 'settler' and 'circulating' pools of agents is not permanent due to the continuous unidirectional progression of some 'circulating' agents to the 'settler' status within the available susceptible 'settler'-free cells. Consequently, self-scattering gene vector swarms are heterogeneous swarms with a dynamic unidirectional switch of an agent type. In this model, another unidirectional feature of self-scattering swarms is signaling, which happens exclusively from the 'settler' agents to the 'circulating' agents. Self-focusing gene vector swarms share this unidirectional signaling feature, as in these swarms signaling occurs exclusively from the settled 'scout' agents to the 'circulating' agents.

## 4 Pheromone-Based Inter-Agent Signaling Channels for Self-Scattering and Self-Focusing of Gene Vector Swarms

While efficient inter-agent (intra-swarm) signaling *per se* is an absolute requirement for intelligent swarm-level behavior, communication in the naturally occurring swarms of genetic agents is not universally mediated by non-diffusible surface markers. Indeed, many conjugative plasmids of Gram-positive bacteria employ diffusible chemicals for inter-plasmid messaging. Thus, diffusible peptide 'sex pheromones' of plasmidless Gram-positive bacteria, such as *Enterococcus faecalis*, induce male conjugation machinery in cells bearing conjugative plasmids, leading to aggregation ('clumping') of recipient and donor cells [31]. Once the susceptible cell has acquired a particular plasmid, the production of the respective pheromone is shut down, while the cell continues to excrete unrelated pheromones specific to other conjugative plasmids [31]. Considering the plasmids as independent agents within a plasmid swarm, knockout of the pheromone production in the recipient cell by the newly arrived plasmid issues a signal to cognate plasmids in other cells to shut down expression of their conjugation-enabling genes. As cell 'clumping' is eventually reversed, this signaling results in swarm-based self-scattering of the plasmid-bearing cells. Similarly to F-factor-determined surface exclusion mechanism in *E. coli*, the pheromone-mediated signaling in *E. faecalis* is a dual purpose system; it serves both to prevent gene transfer between induced donor cells (established donors) and to prevent gene transfer between established donors and nascent donors, which have just received their 'settler' plasmid. In *E. faecalis*, the convoluted intertwining between mechanisms of obstruction of donor-to-donor gene transfer and superinfection interference was studied in pheromone-inducible conjugative plasmids pCD10 and pAD1. In another layer of complexity, in addition to the pheromone-mediated signaling described above, plasmids pCD10 and pAD1 were demonstrated to possess a surface exclusion mechanism mediated by a plasmid-encoded cell surface protein which blocks conjugation between donor cells harboring induced plasmids [31]. If the pheromone concentration is sufficiently high, then newly arrived plasmid in a recipient cell acquires the induced epigenetic

state, the plasmid-encoded rejection factor becomes expressed and prevents further conjugation events through surface exclusion. If the pheromone concentration is not sufficiently high, then newly arrived plasmid in a recipient cell is in the non-induced epigenetic state, so the plasmid-encoded rejection factor is not expressed and surface exclusion does not happen.

Signaling via a diffusible chemical was suggested for artificial self-focusing vector swarms [32]. Toxicity of artificially synthesized diffusible chemicals might present an obstacle for the application of vector swarms *in vivo*. Thus, peptide-based pheromone signaling in *E. faecalis* can be of interest for the development of *in vivo* drug delivery systems exploiting optimal global behaviors of vector swarms.

Another lesson to learn from the pheromone-mediated signaling in conjugative gene transmission in *E. faecalis* is that any change in the molecular state, both positive and negative, can constitute a signal. Thus, for swarm-based self-focusing and self-scattering, the signals can be either of a ‘flag-up’ type, when a non-diffusible marker is established on the cell surface or when a diffusible chemical is released or, alternatively, of a ‘knockout’ type, when a non-diffusible marker is knocked out from the cell surface or when the release of a diffusible chemical is stopped.

The pheromone-based signaling of conjugative plasmids in *E. faecalis* does not rely on the self-propelled movement (motility) of agents relative to the pheromone source, that is, chemotropism, and depends exclusively on the passive cell movement and random collisions [31]. However, many swarms with signaling via diffusible chemicals do involve chemotropism [33], which could affect the autonomous spatial movement behavior of circulating agents either attractively (positive chemotaxis) or repulsively (negative chemotaxis). Accordingly, the released diffusible signaling chemicals are referred to as chemoattractants, which exert a pulling effect on the circulating agents in positive chemotaxis, or chemorepellents, which exercise a pushing effect on the circulating agents in negative chemotaxis.

Within the context of gene delivery to individual cells, the differences in the effects of signaling through a non-diffusible extracellular marker and through a diffusible chemical are substantial. Indeed, cell-surface-marker-based signaling has single-cell precision, since the marker unequivocally identifies individual cells, not a group of neighboring cells within the expanding zone of the diffusible chemical’s release.

## 5 Potential Use of Intelligent Self-Scattering in Therapeutic Gene Delivery

### 5.1 Problems in Therapeutic Gene Delivery

Gene vectors are particles, which are designed to deliver genes into the living cells. Gene vectors are used, in particular, in gene therapy, where they deliver therapeutic genetic cargo to human cells requiring curative genetic intervention [34]. Depending on the prevailing origin of their backbone, gene vectors can be split into two catego-

ries: (1) viral vectors, which are derivatives of naturally occurring viruses; (2) non-viral vectors, which include naked nucleic acids and nucleic acids artificially packaged by macromolecules. Whatever the treatment scenario, focusing of gene delivery on specific target cells requiring curative treatment is important in gene therapy because it minimizes the risk of side-effects of gene delivery to non-targeted cells and creates a leeway for the reduction of vector dose to minimize undesirable side-effects of gene delivery both in non-targeted and targeted cells. The typical unwanted side-effects of gene delivery are cytotoxicity, immunogenicity, accumulation of ‘genetic garbage’ [35], and insertional mutagenesis leading to malignancies and unethical changes in germline cells [23].

However, highly focused gene delivery is associated with a number of its own problems. Indeed, a single infectious particle is normally sufficient for the effective infection of the cell. This is typically true both for naturally occurring viruses/plasmids and their modified versions, therapeutic gene vectors. Thus, in general, assuming no surface exclusion takes place, after viral preparation is incubated with recipient cells, the MOI for individual cells follows Poisson distribution [36]. As a result, in intensely focused gene delivery, a considerable portion of the recipient cells receive multiple vector particles. With many types of therapeutic gene vectors, high MOI of individual cells results in cytotoxic effects, which undermine the therapeutic effects of the vectors [37]. In addition, the random character of the Poisson distribution means that even at the relatively high average MOI not all the susceptible target cells could be infected. Thus, while many vector particles from the infecting pool are wasted on the recipient cells infected at the high individual MOI, there is concomitant waste of uninfected susceptible target cells even at the high average MOI. As a result, the infecting potential of therapeutic vector preparation, which is often expensive to produce, is squandered because of the non-uniform distribution of MOI for individual cells.

There is another serious problem associated with highly focused gene delivery. Therapeutic transgenes should be expressed at a curatively effective level and uniformly within the targeted cell population in order to achieve maximal therapeutic effect and to avoid undesired side-effects [17]. Transgene copy number has a significant bearing on the level of transgene expression and its variability. Typically, the level of mammalian transgene expression is proportional to small transgene copy numbers and unpredictable for high transgene copy numbers [18, 38, 39]. In general, multicopy transgene status can result from: (1) numerous entries of gene vectors to a cell; (2) numerous gene copies present within an infecting therapeutic particle; and (3) faster-than-chromosomal episomal replication of transgenes within the target cell. In the most common gene transfer scenarios, the variation in the number of vector entries into a single cell is the most difficult to control factor of the cell-to-cell variability in transgene copy numbers.

Therefore, as intensely focused gene transfer can result in wide distribution of transgene abundance in the targeted cells, targeting of gene vectors can lead to unfavorable highly-varied distribution of the therapeutic transgene expression level at the delivery site. This unwanted variability of transgene expression can be due to both the uneven distribution of the transgene copy number among targeted cells and

the irregular level of transgene expression in cells with a high copy-number of transgenes. In different disease settings, different ranges of the level of transgene expression are required to produce therapeutic action. Thus, highly concentrated gene transfer presents a risk of irregular expression of transgenes with stoichiometrically-dependent therapeutic effect.

The problems mentioned above could be resolved through focused gene delivery with a more uniform, non-Poisson, distribution of MOI by gene vectors among the targeted cells. Relying on the model of the ‘strict non-admittance’ entry exclusion behavior in gene transmission systems of naturally occurring viruses and plasmids, it can be proposed that controlled spread of gene transfer among the targeted cells can be achieved through swarm-based self-scattering of gene vectors.

## **5.2 *How to Employ Swarm Intelligence of Gene Vectors to Address Problems in Therapeutic Gene Delivery?***

Self-focusing of swarms of genetic vectors to enhance targeting of specific cell populations was proposed [19] and is likely to contribute to the minimization of undesired side-effects through additional targeting of gene delivery. However, the distribution of MOI among individual cells using self-focusing vector swarms is still expected to be uneven, and, in fact, even more skewed than the Poisson distribution of non-swarm gene vectors because of the multiple infections of the cells in which resident ‘scout’ agents ‘open the gates’ for secondary infections. Therefore, targeted delivery with self-focusing gene vector swarms would not avoid the cytotoxicity associated with high MOI, would not utilize the full gene delivery potential of the vector preparation and would be vulnerable to irregular transgene expression due to the varied number of therapeutic gene copies transferred into the target cells.

Taking a cue from naturally occurring intelligent self-scattering gene transfer agents, it is attractive to attain uniform distribution of MOI for individual cells in highly focused therapeutic gene delivery through intelligent self-scattering of therapeutic gene vectors away from the cells which have already been effectively occupied by a single gene vector agent. Thus, self-scattering due to ‘strict non-admittance’ cell entry exclusion could restrict the MOI of individual cells, resolving the problem of excessive copies of therapeutic transgenes and, consequently, could limit vector cytotoxicity and give valuable economy of gene vector particles for extra infections of susceptible target cells. In addition, the self-scattering property of the gene vector swarms could offer a uniform level of therapeutic transgene expression within the population of target cells.

Similarly to the self-scattering in gene transmission systems of naturally occurring viruses and plasmids, the intelligent self-scattering could be achieved through

signaling by the cell-resident ‘settler’ agents, prohibiting ‘circulating’ agents from cell entry. To avoid ‘signal degradation’ and the ensuing loss of precision associated with the signal transfer via a diffusible chemical, signaling could be accomplished through knockout of a non-diffusible extracellular gene transfer receptor or through the expression of an extracellular factor blocking gene transfer with any additional therapeutic vector particles.

To find a suitable receptor to knock out, the gene therapist can use an approach, which is similar to the ‘strict non-admittance’ cell entry exclusion by the cell-resident MLV virus, exploiting structural heterogeneity in native receptors with essential cellular functions. Thus, selective knockout should be directed at the gene-transfer-facilitating version of the receptor, leaving intact the gene-transfer-inactive versions of the receptor that are still capable of performing a particular essential cellular function. This strategy would allow the engineering of a receptor-knockout-based self-scattering effect in gene vector swarms without disruption of any indispensable function of the target cell.

Alternatively, the gene therapist could employ a knockout of an artificial receptor, which has been designed to perform the only function of allowing therapeutic particles to enter and which is, therefore, non-essential for cell survival and specialized tissue functions. One option, which seems convenient, is to pre-position an artificial receptor on the surface of target cells using gene vectors which are targeted using intelligent ‘scout’-dependent self-focusing [19]. Indeed, intelligent self-focusing and self-scattering of gene vector swarms are two compatible strategies, which can be combined through employment of heterogeneous self-scattering swarms of gene vectors containing specialized ‘scout’ agents and specialized ‘therapeutic’ agents. Such scout agents probe the epigenetic state of recipient cells and then, depending on the specific features of the epigenetic state, issue an ‘enter’ or ‘entry denied’ signal to circulating therapeutic particles. After a required number of the therapeutic particles (e.g., one or two particles) have arrived, the installed genetic machinery issues a definitive ‘entry denied’ signal to any further approaching vector particles. In this scenario, the same artificial extracellular receptor is used to ‘open the gate’ for therapeutic vector particles to enter specific target cells and then to ‘close the gate’ for any extra unwanted vector particles. That is, the ‘enter’ signal is issued through the expression of an artificial ‘flag-up’ extracellular receptor by scout vector particles, while the ‘entry denied’ signal is generated by scout and/or therapeutic ‘settler’ particles knocking out the extracellular artificial ‘flag-up’ receptor, which is no longer required after the cell has received an appropriate dose of curative genetic cargo.

It is expected that the self-scattering capability of intelligent ‘settler’-dependent swarms of therapeutic gene vectors could be a useful upgrade for gene therapy vectors, particularly in curative scenarios where, on the one hand, highly focused vector administration creates a substantial variance in the individual MOI, and, on the other hand, uniform transgene expression dosage in target cells is critical for the effectiveness of genetic treatment [17].

### ***5.3 What Could Be the Challenges and Drawbacks of Using Swarm Intelligence of Self-Scattering Gene Vectors in Therapeutic Gene Delivery?***

Self-scattering relying on native cellular receptors looks particularly challenging as native extracellular proteins are typically essential either for cells' survival or for their tissue function. Therefore, any manipulation of native extracellular factors which successfully separates their gene transfer function from their essential function is likely to be dependent on a particular cellular gene transfer setting and difficult to expand to other settings, where the chosen extracellular factor might be absent or where its gene transfer function and essential cellular function are more difficult to split.

Alternative self-scattering approaches relying on an artificial cell surface receptor or an artificial extracellular gene-transfer-blocking factor could have their own challenges and drawbacks: (1) such foreign-sourced markers could still interfere with the cell's functionality; (2) cellular molecular surveillance machinery (e.g., proteases) could attack and inactivate foreign proteins; (3) the immune system could attack cells with foreign antigens exposed at their surface.

### ***5.4 How to Proceed with Application of Swarm Intelligence of Self-Scattering Gene Vectors in Gene Therapy to Avoid its Drawbacks?***

So, cell surface markers used for signaling between gene delivery agents need to be evaluated in terms of their: (1) non-interference with essential cellular functions; (2) invisibility or resistance to intracellular molecular surveillance systems; (3) invisibility to extracellular (immune) molecular surveillance systems. In general, it is vital to evaluate all possible adverse side-effects due to total genetic machinery being installed in the target cells, including cytotoxicity, induction of oncogenesis and undesired immune reactions. Clinical trials with therapeutic self-scattering gene vectors are expected to be preceded with the testing of the vectors in animal models.

Clearly, single cell MOI and single cell therapeutic transgene expression level after gene delivery with therapeutic vector particles need to be determined to evaluate the efficiency of gene delivery to the target cell population, the uniformity of individual MOI, the uniformity of therapeutic transgene copy number and the uniformity of the level of transgene expression within the target cell population. Gene vectors carrying easily-detectable marker genes could be good candidates to obtain the experimental evidence to assess the self-scattering phenomenon for swarms of artificial gene vectors.

## 5.5 *What Further Developments in Application of Intelligent Self-Scattering of Gene Vectors in Gene Therapy Could Be Expected?*

It is hoped that the amalgamation of intelligent ‘scout’-dependent self-focusing and intelligent ‘settler’-dependent self-scattering can become a starting point in the accumulation of further useful intelligent features by swarms of therapeutic gene vectors. Future developments might include the acquisition of additional intelligent global behaviors by therapeutic gene vector swarms, which could offer self-optimized content of therapeutic messages, self-optimized timing of gene transfer or further improvements in self-optimized spatial distribution of gene transfer. Clearly, the implementation of such extra features of swarm intelligence would require a substantial expansion in the bandwidth of inter-agent signaling via the cell surface.

## References

1. Reinhart TA, Ghosh AK, Hoover EA, Mullins JI. Distinct superinfection interference properties yet similar receptor utilization by cytopathic and noncytopathic feline leukemia viruses. *J Virol.* 1993;67(9):5153–62.
2. Abedon ST. Bacteriophage secondary infection. *Virol Sin.* 2015;30(1):3–10.
3. Dunny GM, Zimmerman DL, Tortorello ML. Induction of surface exclusion (entry exclusion) by *Streptococcus faecalis* sex pheromones: use of monoclonal antibodies to identify an inducible surface antigen involved in the exclusion process. *Proc Natl Acad Sci U S A.* 1985;82(24):8582–6.
4. Novick RP. Plasmid incompatibility. *Microbiol Rev.* 1987;51(4):381–95.
5. Lu MJ, Henning U. Superinfection exclusion by T-even-type coliphages. *Trends Microbiol.* 1994;2(4):137–9.
6. Nethe M, Berkhouit B, van der Kuyl AC. Retroviral superinfection resistance. *Retrovirology.* 2005;2:52.
7. Lee YM, Tscherne DM, Yun SI, Frolov I, Rice CM. Dual mechanisms of pestiviral superinfection exclusion at entry and RNA replication. *J Virol.* 2005;79(6):3231–42.
8. Arutyunov D, Frost LS. F conjugation: back to the beginning. *Plasmid.* 2013;70(1):18–32. Epub 2013/05/02.
9. Krause J, Ruxton GD, Krause S. Swarm intelligence in animals and humans. *Trends Ecol Evol.* 2010;25(1):28–34.
10. Toutouh J, Alba E. A swarm algorithm for collaborative traffic in vehicular networks. *Veh Commun.* 2018;12:127–37.
11. Nag S, Sumner L. Behaviour based, autonomous and distributed scatter manoeuvres for satellite swarms. *Acta Astronaut.* 2013;82(1):95–109.
12. Alfeo AL, Cimino MGCA, De Francesco N, Lega M, Vaglini G. Design and simulation of the emergent behavior of small drones swarming for distributed target localization. *J Comput Sci.* 2018;29:19–33.
13. Weng LG, Liu QS, Xia M, Song YD. Immune network-based swarm intelligence and its application to unmanned aerial vehicle (UAV) swarm coordination. *Neurocomputing.* 2014;125:134–41.

14. Baluska F, Lev-Yadun S, Mancuso S. Swarm intelligence in plant roots. *Trends Ecol Evol.* 2010;25(12):682–3.
15. Inácio FR, Macharet DG, Chaimowicz L. PSO-based strategy for the segregation of heterogeneous robotic swarms. *J Comput Sci.* 2019;31:86–94.
16. Tolmachov O, Harbottle R, Bigger B, Coutelle C. Minimized, CpG-depleted, and methylated DNA vectors: towards perfection in nonviral gene therapy. In: Schleef M, editor. *DNA pharmaceuticals: formulation and delivery in gene therapy, DNA vaccination and immunotherapy.* Weinheim: Wiley-VCH; 2006. p. 43–54.
17. Weber W, Fussenegger M. Pharmacologic transgene control systems for gene therapy. *J Gene Med.* 2006;8(5):535–56.
18. Tolmachov OE, Subkhankulova T, Tolmachova T. Silencing of transgene expression: a gene therapy perspective. In: Martin-Molina F, editor. *Gene therapy – tools and potential applications.* Rijeka, Croatia: InTech; 2013. p. 49–68.
19. Tolmachov OE. Self-focusing therapeutic gene delivery with intelligent gene vector swarms: intra-swarm signalling through receptor transgene expression in targeted cells. *Artif Intell Med.* 2015;63(1):1–6. Epub 2014/12/31.
20. Buchholz CJ, Friedel T, Büning H. Surface-engineered viral vectors for selective and cell type-specific gene delivery. *Trends Biotechnol.* 2015;33(12):777–90.
21. Borroni E, Miola M, Ferraris S, Ricci G, Žužek Rožman K, Kostevšek N, et al. Tumor targeting by lentiviral vectors combined with magnetic nanoparticles in mice. *Acta Biomater.* 2017;59:303–16.
22. Gowing G, Svendsen S, Svendsen CN. Chapter 4 – Ex vivo gene therapy for the treatment of neurological disorders. In: Dunnett SB, Björklund A, editors. *Progress in brain research.* Cambridge San-Diego Oxford London: Elsevier; 2017. p. 99–132.
23. Tolmachov OE. Split vector systems for ultra-targeted gene delivery: a contrivance to achieve ethical assurance of somatic gene therapy in vivo. *Med Hypotheses.* 2014;83(2):211–6.. Epub 2014/05/24
24. Dorigo M, Bonabeau E, Theraulaz G. Ant algorithms and stigmergy. *Futur Gener Comput Syst.* 2000;16(8):851–71.
25. Wang H, Klamo E, Kuhmann SE, Kozak SL, Kavanaugh MP, Kabat D. Modulation of ectopic murine retroviruses by N-linked glycosylation of the cell surface receptor/amino acid transporter. *J Virol.* 1996;70(10):6884–91. Epub 1996/10/01.
26. Huang IC, Li W, Sui J, Marasco W, Choe H, Farzan M. Influenza A virus neuraminidase limits viral superinfection. *J Virol.* 2008;82(10):4834–43.
27. Goto H, Ohta K, Matsumoto Y, Yumine N, Nishio M. Evidence that receptor destruction by the Sendai virus hemagglutinin-neuraminidase protein is responsible for homologous interference. *J Virol.* 2016;90(17):7640–6.
28. Laliberte JP, Moss B. A novel mode of poxvirus superinfection exclusion that prevents fusion of the lipid bilayers of viral and cellular membranes. *J Virol.* 2014;88(17):9751–68.
29. Yamada Y, Yamada M, Nakazawa A. A ColE1-encoded gene directs entry exclusion of the plasmid. *J Bacteriol.* 1995;177(21):6064–8.
30. Schmickl T, Thenius R, Crailsheim K. Swarm-intelligent foraging in honeybees: benefits and costs of task-partitioning and environmental fluctuations. *Neural Comput Applic.* 2012;21(2):251–68.
31. Clewell DB. Bacterial sex pheromone-induced plasmid transfer. *Cell.* 1993;73(1):9–12.
32. Grancic P, Stepanek F. Active targeting in a random porous medium by chemical swarm robots with secondary chemical signaling. *Phys Rev E Stat Nonlinear Soft Matter Phys.* 2011;84(2–1):021925. Epub 2011/09/21.
33. Shkлярш А, Фінкельштейн А, Ариль Г, Каліман О, Ингем C, Бен-Яаков Е. Collective navigation of cargo-carrying swarms. *Interface Focus.* 2012;2(6):786–98.
34. Misra S. Human gene therapy: a brief overview of the genetic revolution. *J Assoc Physicians India.* 2013;61(2):127–33.. Epub 2014/01/30

35. Tolmachov OE. Transgenic DNA modules with pre-programmed self-destruction: universal molecular devices to escape 'genetic litter' in gene and cell therapy. *Med Hypotheses*. 2015;85(5):686–9.
36. Ellis EL, Delbrück M. The growth of bacteriophage. *J Gen Physiol*. 1939;22(3):365–84.
37. Castellani S, Di Gioia S, Trotta T, Maffione AB, Conese M. Impact of lentiviral vector-mediated transduction on the tightness of a polarized model of airway epithelium and effect of cationic polymer polyethylenimine. *J Biomed Biotechnol*. 2010;2010:11.
38. Zielske SP, Lingas KT, Li Y, Gerson SL. Limited lentiviral transgene expression with increasing copy number in an MGMT selection model: lack of copy number selection by drug treatment. *Mol Ther*. 2004;9(6):923–31.
39. Kong Q, Wu M, Huan Y, Zhang L, Liu H, Bou G, et al. Transgene expression is associated with copy number and cytomegalovirus promoter methylation in transgenic pigs. *PLoS One*. 2009;4(8):e6679.

# A Combinatorial Computational Approach for Drug Discovery Against AIDS: Machine Learning and Proteochemometrics



Sofia D'souza, Prema K. V., and Seetharaman Balaji

**Abstract** Computational methods have been widely used in drug discovery including identification of novel targets, studying drug target interactions, and in virtual screening of compounds against known targets. Machine learning techniques have been used in predictions of novel targets and drugs with greater accuracy compared to other methods. Machine learning algorithms have also been widely used in predicting the progression of disease, resistance of a drug to a virus, treatment efficacy prediction, and also in predicting the effectiveness of combinational therapy with respect to HIV-1. In this article, we have focused on some of the machine learning techniques in the context of viral disease. In brief, machine learning methods have great potential in drug discovery, drug repurposing, and in precision medicine.

**Keywords** Human immunodeficiency virus-1 (HIV-1) · Machine learning (ML) · Support vector machines (SVM) · Decision tree (DT) · Random Forest (RF) · Artificial neural network (ANN) · Proteochemometric modeling (PCM)

## Key Phrases

AIDS is one of the world's biggest epidemic. There are approximately 20 million people who are receiving treatment every year and yet there are millions of deaths due to HIV-1 and ineffective treatments. The management of HIV-1 is challenging because of wide variability in the genetic makeup of the virus and their circulating recombinant forms. This necessitate personalized treatment. Computational methods have greatly improved the disease management by incorporating machine learning to predict novel targets, drug target-interaction, compound activity, drug resistance, treatment efficacy,

S. D'souza · Prema K. V.

Department of Computer Science & Engineering, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India

S. Balaji (✉)

Department of Biotechnology, Manipal Institute of Technology, Manipal Academy of Higher Education, Manipal, Karnataka, India

effectiveness of treatment regimen, effective treatment of comorbid conditions in HIV-1 infected patients, etc. In a futuristic view, the machine learning methods along with proteochemometrics may be useful for personalized treatment of HIV infected patients.

## 1 Introduction

Over the past few decades a tremendous increase in data is witnessed due to the ‘digital revolution’ across many walks of life. The overwhelming data, for instance, medical, genomic, personal, and social data is impossible to analyse and interpret, without proper tools and techniques. Among the technologies of the twenty-first Century, artificial intelligence (AI) and machine learning (ML) drive the development of new learning theory and algorithms to automate the processing of the data explosion with low-cost computation. These techniques have created an enormous impact in every discipline such as health care, manufacturing, education, financial modeling, marketing, and management [1]. It is estimated that about 40% of the potential value today that was created by analytics comes from AI techniques, which falls under the “umbrella of deep learning”. “Deep learning” utilizes multiple layers of artificial networks of neurons resembling the functions of the human brain. It is estimated that deep learning, globally, generates an annual value between \$3.5 and \$5.8 trillion [2].

Many new AI technologies such as neural networks, fuzzy systems, Bayesian networks, support vector machines, decision trees, evolutionary, and swarm intelligence algorithms are not just theoretical inventions that are built into user friendly software packages [3]. The experimentation and implementation of AI in virology and medicine creates new disciplines such as ‘systems medicine’ in addition to medical informatics and bioinformatics. These disciplines can implement AI approaches effectively to contravene viral diseases especially acquired immune deficiency syndrome (AIDS), caused by human immunodeficiency virus 1 (HIV-1). These approaches can help predict HIV-1 disease progression due to a wide variability in the genetic structure of the virus, multiple infections, recombination, demographic and socio-cultural differences, and greatly help in understanding drug resistance and treatment efficacy.

Machine learning tools have enabled processing and analysing complex relationships of data in making effective decisions. The ML algorithms make use of each dataset and predict the outcomes of interest. ML mimics human pattern recognition and learning processes, through a series of complex computations for large datasets [4]. ML algorithms may be broadly classified as follows (i) supervised, (ii) unsupervised, and (iii) semi-supervised learning methods. The supervised learning methods predict the label or outcome of a sample input given a set of labels for the input samples in the training data. They make use of known labels to predict the output label of a sample. A number of supervised learning methods like support vector

machines, decision trees, random forests and classification and regression trees are used in predicting the outcomes of interest. Unsupervised learning methods try to learn the relationship between the input variables and the output variable by learning patterns and correlations in the data, when no training samples are available. These algorithms are mainly used in clustering and dimensionality reduction. K-means clustering is the most commonly used method [5]. The semi-supervised method makes use of small quantities of labeled data and large aggregates of unlabeled data for training. The algorithm learns from the labeled data and makes predictions on unlabeled data [6]. Supervised machine learning techniques have been used in virus studies to predict drug resistance, virological response to anti-retroviral therapy, and also in inhibitor efficacy prediction.

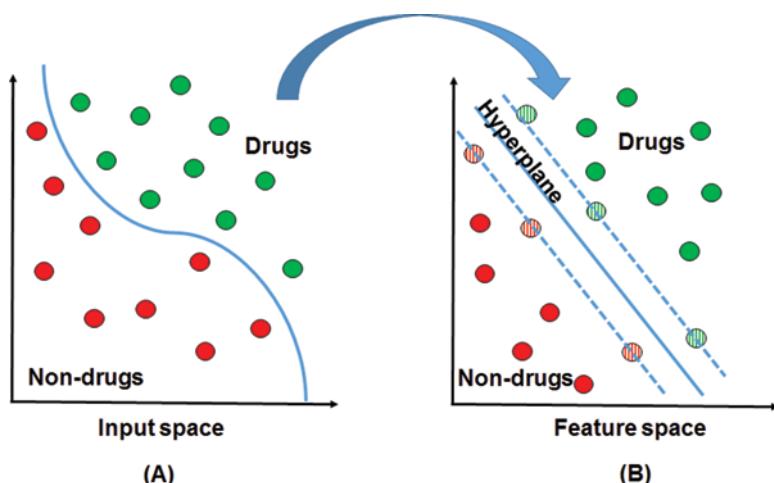
## 2 Machine Learning Methods

Several ML approaches use non-linear methods such as support vector machines (SVM), decision tree (DT), random forest (RF) and neural network (NN), respectively. These methods have been implemented to analyse viral proteins and drugs. In the following sections, the most widely used machine learning models are reviewed, highlighting the method, benefits, and limitations of each.

### 2.1 *Support Vector Machines*

Support vector machines (SVM) are supervised machine learning algorithms used in classification of dataset regression and prediction [7, 8]. SVMs are generally used for classification of input datasets such as drugs and non-drugs, drugs binding to protein targets can be classified as better binding, and poor binding based on their affinity. Firstly, the compounds are projected onto a high dimensional space such that the features become highly separable. This can be achieved by using kernel functions. Some of the kernel functions used can be linear, polynomial, sigmoid and radial basis functions (RBFs). RBFs are local kernel, while others are global in nature. RBF outperforms the other three kernels and thus, it is widely used [9]. The linearly separable data are classified into two regions by a hyperplane. Out of the many hyperplanes, SVM selects the hyperplane in such a way, as to maximize the margin between the two classes in order to keep the error margin minimal, while classifying unknown data (Fig. 1). The ‘support hyperplanes’ define the margins and the data points on these hyperplanes and are termed ‘support vectors’. The soft-margin hyperplane is used in case of non-separable classes, so as to reduce misclassified samples.

In ligand-based virtual screening [10], SVM ranks compounds based on their activity values. Many optimization methods are available to minimize ranking errors, and also a variety of kernel functions for SVMs are available to predict



**Fig. 1** The projection of drugs (green balls) and non-drugs (red balls) that are not linearly separable in low-dimensional input space (**a**) are transformed into high-dimensional feature space (**b**) that are separated by the maximum margin hyperplane. The balls on the intercept (dotted lines) are called ‘support vectors’

target-ligand binding based on binding site similarity of targets and structure similarity of ligands [7–9]. SVM is also used to predict changes in CD4 count of HIV-1 patients [11]. The prediction of CD4 cell count helps in understanding disease progression and also in disease treatment. In this study, the degree of CD4 cell count change was predicted by SVM. The prediction is based on the dataset obtained from Stanford drug resistance database [12], protease sequences, CD4 counts, virus load, and number of weeks from baseline measure of CD4 counts for each patient. The data were refined to remove redundant entries. The ML algorithm was provided with three different inputs to form three models. The input groups are (a) sequence data, (b) sequence data and current viral load, (c) sequence data, current viral load, and number of weeks from current CD4 count to baseline CD4 count.

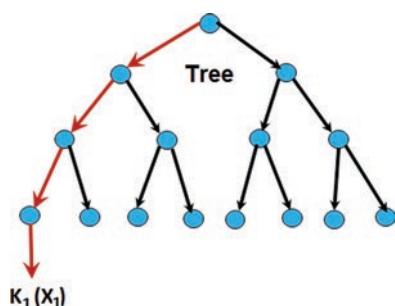
The SVM classifier model outputs the change in CD4 count values. The classifier used RBF, linear and quadratic functions in the models to predict the output results. It was found that the input (c) model is more accurate than other models. The RBF model outperformed linear and quadratic function model for input (c) due to its localized and finite responses across the entire range of predictors. The method has certain limitations. Firstly, the dataset used had fewer samples for prediction using ML algorithms. This could be improved by incorporating larger data set. Secondly, the method needs sequencing and viral load data which is not affordable especially in developing countries. Hence, a dataset containing more common treatment information could be included. Lastly, this method predicts CD4 count in a range instead of actual CD4 count, which is required for disease effective treatment.

## 2.2 Decision Trees

The Decision Tree (DT) method consists of a set of rules that associate molecular descriptors of compounds with their activities. DT is used to predict biological activity and drug-likeness for compound identification. It is not only used for sub-structure identification within a class of compounds; but, also to distinguish between active and inactive compounds as well as to classify compounds into drugs and non-drugs. DT is commonly depicted as a tree in a top-down manner. The root splits into a number of branches at each level. Further, each branch splits repeatedly into two or more branches until the last level of single leaf nodes is reached (Fig. 2). The roots and the leaves are referred as the “nodes”. Each node is allocated with a ‘molecular descriptor’ and each leaf is allocated with a ‘protein target’. Nodes are tested for the condition of molecular descriptor which further splits till the leaf is reached. The leaf node is used for classifying an unknown compound. The quality of the test determines the ‘purity’ of the split. To achieve the best classification, metrics such as entropy, information-gain ratio or Gini-index are used [13]. The info gain metric prefers test conditions such that the tree is balanced where purity is increased.

Decision trees are simpler to understand, easy to interpret and validate the results. Nonetheless, their prediction suffers from high degree of variance. A small training set has a significant effect on the learning process. However, overfitting is a problem with large training sets. The performance of the DT depends on the quality of decision criteria as well as on the sequence of decision attributes. The decision attributes have to be arranged in decreasing order of importance.

Classification and regression trees [14] were built to predict the quality of life (QOL) response shift in HIV/AIDS patients [15]. Content analysis of over 6700 responses were collected at baseline and after 6 months of follow-up was carried out. The primary and subsequent goal content dimensions were coded, and decision tree was built to make prediction of QOL response shift score. It was observed that



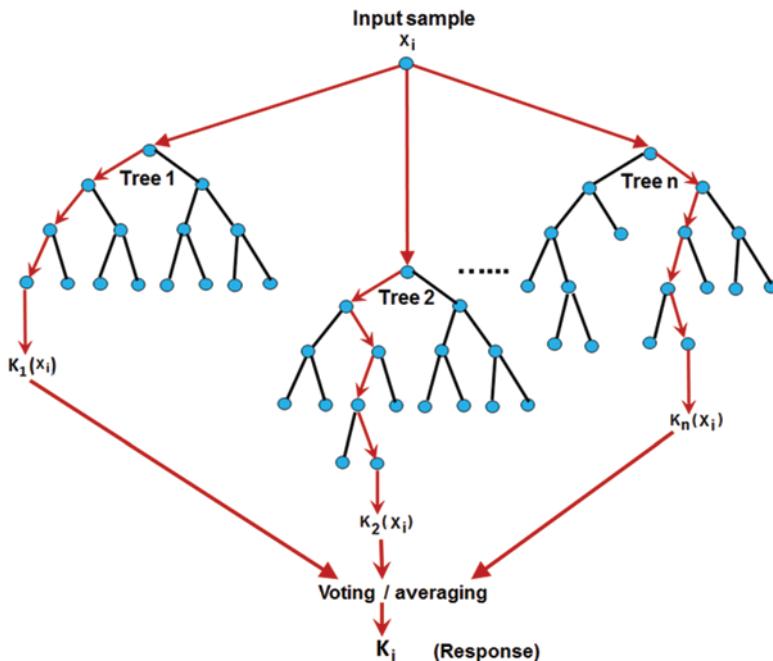
**Fig. 2** Decision tree: A tree with a single root node, followed by a set of yes/no decisions (binary splits) that finally results in a set of leaf nodes. For classification, a test event is passed from the root node down the tree and will emerge in a specific leaf node depending on how it responds to the various split criteria

cognitive variables accounted for substantial health related QOL response shift. The Classification and Regression Tree (CART) models were built to study the neurocognitive impairment in HIV-1 patients by using demographic and clinical data [16]. Neurocognitive Impairment (NCI) affects approximately one-third to two-third of the HIV-1 infected patients. This complication leads to impaired daily functioning, poor quality of life and higher death rates. Therefore, rapid and reliable detection of NCI is crucial for the treatment of HIV-1 infected patients. The study was based on 331 HIV-1 infected patients in Spain. The demographic variables included age, gender, employment status, years of education and route of transmission. The clinical variables included current CD4 cell count, nadir CD4 cell count (i.e., the lowest point to which the CD4 count has dropped), plasma viral load and other parameters. The CART model was used to predict the absence or presence of NCI disorder. The study identified CD4 cell count and age as the most significant variables for prediction of impairment in antiretroviral-naïve group patients. In patients on antiretroviral therapy, nadir CD4 cell count, and CNS penetration effectiveness (CPE) score. The limitation of this study was that two small groups were studied with less samples. Additional clinical information such as cardiovascular risk factors, hormonal markers which are known to have associations with cognitive impairment with infected as well as non-infected individuals could have been incorporated in the study. The outcome of this study was NCI which is a part of HIV-1 associated neuro-cognitive disorders [17].

### **2.3 Random Forest**

A commonly used method to restrict high variance in a DT is to prune the tree, using cross validation or complexity parameters. The Random Forest (RF) method is an ensemble of decision trees that works to improve accuracy of classification/regression (Fig. 3). An ensemble of DTs was created from the subgroup of the total description set to make predictions [18]. This is called ‘random decision forest’ and is a high performance model compared to a single DT. Individual decision trees are sensitive to small changes in the features of the training space. Therefore, individual decision trees are weak learners and are not capable of generalization. The RF algorithm calculates value from multiple sub-trees by voting and is able to give accurate prediction (Fig. 3). The RF algorithm calculates values from multiple sub-trees by voting and is able to give accurate prediction. Bagging, boosting, and stacking are the ensemble techniques that are better predictors compared to individual learning algorithms and have minimum variance.

Random Forest [19] methods makes use of bagging and subset selection at each node of DT to further improve prediction. A training set consists of samples from the original dataset chosen by random sampling. The test set is constructed by the left out samples, the size that is approximately one-third of the samples used for tree building. The test set is called “out-of-bag” cases. The performance evaluation of



**Fig. 3** Random forest (RF): The tree structures indicate yes/no rules at each branching, with the associated subspace partitioning a hypothetical space. The individual predictions from each tree are collected and combined for further prediction by voting (for classification) or averaging (for regression). The descriptor  $X_i$  is the split depending upon the target property at each node

the classification of the test set depends on “out-of-bag” error rates. The elimination of the features is decided by the confirmation of many trees. Another advantage of RF is that it is relevant to data with high dimensions and with large amount of noise, also associate data with highly correlated variables and with a smaller number of observations. It is less vulnerable to overfitting, which is a problem due to training of large samples of data. Moreover, the class imbalance problem is also tackled by RF. Compared to other ML algorithms, RF has been tested to improve the accuracy of the prediction of quantitative structure activity relationship (QSAR) data [20]. It also has built-in descriptor selection and a method to assess the importance of each descriptor. It has also been used to predict protein-ligand binding affinity and in scoring functions of molecular docking. It is a robust method that is used to predict datasets with noise.

RF methods have been used to predict the drug resistance in HIV-1 PR and RT using genotype data [21]. The virological response to combination HIV-1 therapy was studied by Wang et al. [22]. The authors used unified encoding of protein sequences and molecular structures using Delaunay triangulation. A comparison of methods for combination therapy was studied. Genotypic drug resistance testing is about 60–65% predictive of responses to combination antiretroviral therapy

(cART). However, HIV genotyping costs more and is unavailable in many resource-limited settings. Nonetheless, cART helps reduce mortality and morbidity in these resource-limited settings. With more than 25 antiretroviral drugs used in cART and more than 100 mutations involved in drug resistance, finding an optimal treatment is challenging. In another study, the HIV resistance Response Database Initiative (RDI) database was used to predict the virological response using ANN, RF and SVM. The virological response to cART was predicted using RF without a genotype in resource limited settings [23]. Drug resistance prediction without using genomic data was done by using HIV Resistance RDI database [24] which consists data of 84,000 patients from 30 countries. The TCE from the database was used to construct RF models to predict the output variable, with follow-up viral load. Remarkably, the models were able to predict regimens that were effective.

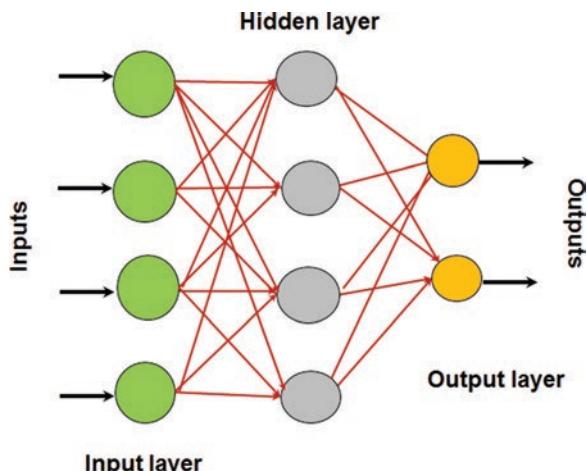
A computational model to study response to HIV-1 therapy was conducted by Tarasova et al. [25]. The high variability of HIV-1 is an important means by which for antiviral resistance occurs for RT and PR inhibitors. This method aims to predict viral resistance to antiretroviral drugs using RF approach. The amino acid and nucleotide sequences were represented as fragments. They were correlated with the Stanford HIV drug resistance database. Nucleotide descriptors were preferred over peptide descriptors for drug resistance prediction as they do not require processing or pre-alignment, as well as the number of descriptors was large (500 nucleotide descriptors) compared to 130 peptide descriptors. A phenotypic webserver to predict effects of drugs against various viral infections was also developed by Beerenwinkel et al. [26]. The prediction is not limited to classify the virus resistance/non-resistance against a particular drug but also to predict the strength of resistance in order to select the most effective drug. The strength of resistance may be categorized as low, medium, or high. The query amino acid sequence is mapped onto a structure to generate a feature vector of 210 dimensions. The machine learning model takes this input and predicts the resistance. The model is trained by using known samples using KNN and RF algorithms. The publicly available drug resistance database which includes drug susceptibility tests for PR and RT was used for this study and eight PR inhibitors and 10 RT inhibitors were used. There is a specific value called the ‘cut-off’ indicating the virus that has become drug resistant. The cut-off values for all inhibitors were obtained from prior publications [27–29]. If the resistance were less than the cut-off value, then the mutant was considered as ‘non-resistant’ or ‘susceptible’ to the drug and the value assigned was ‘0’, otherwise, it was considered as ‘resistant’ and assigned a value ‘1’. Supervised ML algorithm was used. The K-Nearest Neighbor (KNN) method is a non-parametric method that uses full training data set [30]. It finds the K-nearest points to the query point and classifies them according to the majority or by the average values of resistance. KNN is faster compared to SVM but it uses the entire dataset for prediction because the result is based on the training data.

## 2.4 Neural Network

Drug resistance using neural network methods was studied by Dragichi et al. [31]. The general architecture of neural network is depicted in Fig. 4. Drug resistance for protease inhibitors such as Indinavir and Saquinavir was predicted using the structural features of protease-inhibitor complex. Self-organizing maps (SOM) are a type of artificial neural network (ANN) used to extract important features and cluster patterns in an unsupervised manner. The effects of structural features on drug resistance was studied. It was observed that different point mutations cause similar structural changes in the active site. The inhibitors block the activity of the PR thereby halting or minimizing virus replication. PR and RT, which have roles in virus replication, are attacked by a combination of PR and RT inhibitors. However, treatment failures may occur. Once treatment failure occurs, the treatment is changed, and the virus is usually attacked by a combination of drugs. However, there are two problems: (i) FDA approved drugs are limited and therefore effective drugs combinations are also limited. (ii) Cross-resistance may occur, which reduces the number of effective combination therapies. In order to predict drug resistance, 3D-structures of mutant genotypes were generated and analysed. SOMs were used to extract such structural features.

The reason behind HIV treatment failure is mainly due to drug resistance. To re-establish viral suppression by selection of a new treatment regimen, genotypic resistance testing is recommended [32–34]. A model was developed [35] to predict the response to viral therapy by using HIV genotype and clinical information. Three models ANN, RF and SVM were developed. Data from treatment change episodes were identified from HIV resistance response database initiative. The output variable, follow-up viral load was predicted using 76 input variables. The goal was to predict treatment response accurately from genotype and clinical information which will help in treatment selection. The input variables included baseline genotypic

**Fig. 4** General architecture of a multilayer feed-forward artificial neural network (ANN) representing input, hidden and output layers. Some of the feed forward networks are multilayer perceptrons (MLP), radial basis function (RBF) networks, and self-organising maps (SOM)



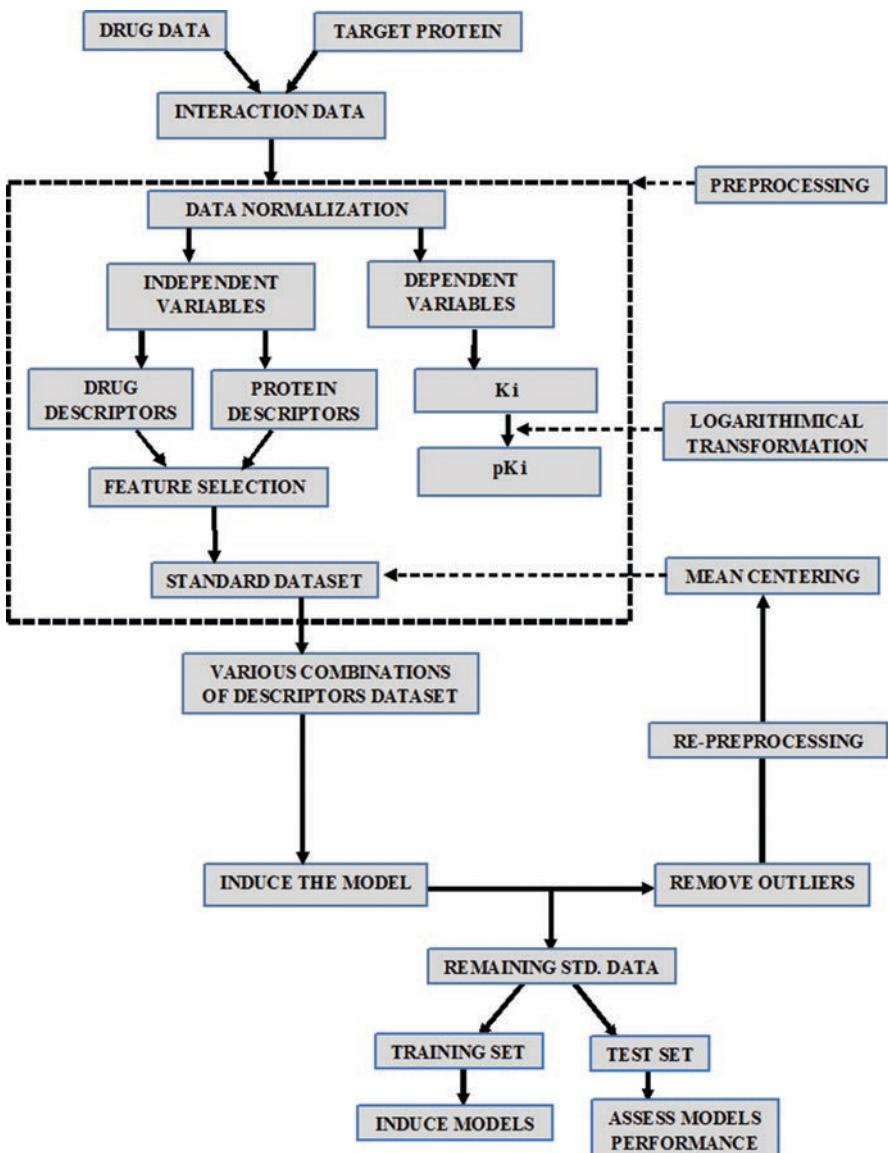
mutations, drugs in new combination regimen, treatment history variables, baseline CD4 count, baseline viral load and time to follow-up viral load. The predictions of the three models were compared. The combination of ANN and RF model achieved the highest correlation ( $r^2 = 0.747$ ) with actual responses. The only limitation of this study was that the dataset used for the study was small.

### 3 Proteochemometric Approach for Finding New Drugs

The Proteochemometric (PCM) model can be applied to predict activity profile of ligands against viral protein variants, to predict the sensitivity of viral mutants to antiretroviral drugs, to learn features relevant for activity, to predict drug susceptibility of HIV-1 PR and RT. This is of interest to pharmaceutical companies so that the information from similar targets can be used to guide lead/drug discovery for a target of interest. An explosion of data related to drugs [36], proteins [37], and diseases [38] are available. These data can be intelligently mined, and meaningful insights can be explored, related to the interaction of ligands and proteins, drug-drug interactions, protein-protein interactions, prediction of drug resistance/susceptibility, efficacy and toxicity. The study of drugs interacting with the targets can be used to gain knowledge pertaining to side effects, drug repurposing, and finding alternative drugs for treatment of a disease as well as for personalized medicine. Drug target interaction is mainly studied using conventional methods such as QSAR [39] and molecular docking [40]. The QSAR method maps a particular ligand to a particular target based on the structure-activity relationship between the two. The improved QSAR method screens multiple ligands against a single target. Molecular docking involves a complex optimisation task of finding the most favourable conformation (pose) of a ligand binding to a specific target. However, this is a computationally expensive method for a large dataset. Moreover, the 3D-structures of most of the proteins are unavailable and docking method is not suitable for a large scale virtual screening of databases for ligands. Hence, computationally fewer intensive methods such as PCM may be considered to infer protein-ligand interactions [41].

PCM is an effective statistical approach to predict drug-protein interactions even for the targets whose structures are unknown. A PCM model can be constructed by incorporating similarities among ligand and target spaces, simultaneously. A primary advantage of a PCM model is that the multiple interactions of a group of ligands with a number of targets can be described, apart from describing particular interactions among individual ligands and targets in the dataset. Therefore, activity data of compounds on multiple targets will be helpful to construct an effective PCM model. PCM can predict the output variable e.g. affinity, by creating a model that accepts the descriptors of the compound as well as the target. The activity results of the known compounds can be extrapolated to new targets and also the activity results of known targets can be extrapolated to new set of compounds. The PCM approach can be integrated with ML approaches by considering ligand descriptors, target descriptors and known interaction data to predict the biological activity

(Fig. 5). The ligand and target features are combined in the PCM approach, in order to predict the ligands binding affinity. As the number of features of ligands and targets increases, the prediction accuracy of the machine learning model decreases, due to the addition of variables that have noise. Feature selection and feature extrac-



**Fig. 5** The workflow of proteochemometric (PCM) modeling using machine learning for predicting protein-ligand interactions adapted from Huang et al. [44]. The PCM model makes use of protein and ligand descriptors along with known binding affinity values, for training and testing

tion techniques can be employed to increase the performance level of the model. Feature selection is used when most relevant subsets of features are selected from the existing feature set to construct the model. Filter methods or wrapper methods may be applied [42]. Some of the examples of filter methods used are chi square test, information gain, and correlation coefficient scores. Wrapper methods include recursive feature elimination algorithms, where features are chosen to improve the performance of the model. In order to reduce the dimension of feature space, feature extraction can be employed by combining correlated features thereby reducing the feature space. This improves the performance of the model. Principal component analysis and factor analysis are some of the common dimensionality reduction techniques.

PCM has also been employed to model HIV-1 protease susceptibility. Lapins and Wikberg used 240 z-scale descriptors representing physico-chemical properties of mutated HIV-1 PR and six orthogonal descriptors for 3D-structural property descriptions of protease inhibitors [43]. The descriptors were correlated to the susceptibility data of 828 unique HIV-1 PR variants for 7 protease inhibitors. Three PCM models were developed. The model with intra-protease cross terms and protease-inhibitor cross terms provided better results with  $r^2 = 0.9$  [43]. In another study, PCM model was constructed on bio-activity spectra of HIV-1 protease inhibitors using protein ligand interaction fingerprint (PLIF) as cross terms. The cross terms describe interaction of proteins residues with its ligands. It was observed that the use of cross terms improved the performance of the PCM model [44]. A study using PCM models were also used to predict susceptibility of mutated variants of HIV-1 to PR and RT [45, 46]. Cross terms along with target and inhibitor descriptors were used to construct PCM models for PR and RT. Also, HIV inhibitor efficacy prediction was implemented using PCM models. The prediction of phenotypic resistance using genotypic data for novel PR and RT mutants was also carried out. The authors used clinical database of genotypic and phenotypic information to build the PCM model [47].

## 4 Summary and Perspectives

Since the beginning of this century, computational methods have been proposed to study various aspects of viral diseases such as drug resistance, viral mutations, effectiveness of viral treatments, prediction of disease progression, study of quality of life responses of patients, and combination therapy. Due to the availability of drugs, genomic, transcriptomic and proteomic data along with computational power to analyse them, much progress have been made to predict HIV-1 drug resistance and efficacy. Machine learning algorithms have contributed significantly to improve the management of the disease. Prospectively, these approaches may help to control the burden of viral diseases globally. Despite the success of these methods, there are still some limitations. The requisite amount of data to make accurate predictions is not available, as it requires sequence and viral load information of the patients,

which are expensive to obtain. The experimental data of known interactions between drugs and target proteins, as well as the unknown interactions between them are not balanced leading to class imbalance problem, where the classifier tends to be biased towards the class with more number of samples. Moreover, as the dimensionality of dataset increases, the relevant feature subset selection to train the ML model is difficult. To overcome these limitations, a combinatorial computational approach is recommended that can automatically extract the relevant features.

**Acknowledgements** The corresponding author acknowledges the grant (No. VGST/GRD-533/2016-17/241) received from Karnataka Science and Technology Promotion Society (KSTePS), India, for supporting the ‘Centre for Interactive Biomolecular 3D-literacy (C-in-3D)’ under the VGST scheme – Centres of Innovative Science, Engineering and Education (CISEE) for the year 2016-17.

## References

1. Jordan MI, Mitchell TM. Machine learning: trends, perspectives, and prospects. *Science*. 2015;349(6245):255–60. <https://doi.org/10.1126/science.aaa8415>.
2. Chui M, Henke N, Miremadi M. Most of AI's business uses will be in two areas. *Harv Bus Rev*. 2018. <https://hbr.org/2018/07/most-of-ais-business-uses-will-be-in-two-areas>.
3. Singh Y. Machine learning to improve the effectiveness of ANRS in predicting HIV drug resistance. *Healthc Inform Res*. 2017;23(4):271–6. <https://doi.org/10.4258/hir.2017.23.4.271>.
4. Evans D, Pottier C, Fletcher R, Hensley S, Tapley I, Milne A, Barbetti M. A comprehensive archaeological map of the world's largest preindustrial settlement complex at Angkor, Cambodia. *Proc Natl Acad Sci*. 2007;104(36):14277–82. <https://doi.org/10.1073/pnas.0702525104>.
5. Montgomery EB Jr, Huang H, Assadi A. Unsupervised clustering algorithm for N-dimensional data. *J Neurosci Methods*. 2005;144(1):19–24. <https://doi.org/10.1016/j.jneumeth.2004.10.015>.
6. Garla V, Taylor C, Brandt C. Semi-supervised clinical text classification with Laplacian SVMs: an application to cancer case management. *J Biomed Inform*. 2013;46(5):869–75. <https://doi.org/10.1016/j.jbi.2013.06.014>.
7. Vapnik VN. The nature of statistical learning theory. New York: Springer; 2000. p. 314.
8. Vapnik VN. Statistical learning theory. New York: John Wiley & Sons, Inc; 1998.
9. Camps-Valls G, Bruzzone L. Kernel-based methods for hyperspectral image classification. *IEEE Trans Geosci Remote Sens*. 2005;43(6):1351–62. <https://doi.org/10.1109/TGRS.2005.846154>.
10. Jacob L, Hoffmann B, Stoven V, Vert JP. Virtual screening of GPCRs: an in silico chemogenomics approach. *BMC Bioinformatics*. 2008;9(1):363. <https://doi.org/10.1186/1471-2105-9-363>.
11. Singh Y, Mars M. Support vector machines to forecast changes in CD4 count of HIV-1 positive patients. *Sci Res Essays*. 2010;5(17):2384–90.
12. Shafer RW. Rationale and uses of a public HIV drug-resistance database. *J Infect Dis*. 2006;194(Supplement\_1):S51–8. <https://doi.org/10.1086/505356>.
13. Raileanu LE, Stoffel K. Theoretical comparison between the gini index and information gain criteria. *Ann Math Artif Intell*. 2004;41(1):77–93.
14. Breiman L. Classification and regression trees. Taylor & Francis Group, LLC 1984, Boca raton, FL, pp368
15. Li Y, Rapkin B. Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS. *J Clin Epidemiol*. 2009;62(11):1138–47. <https://doi.org/10.1016/j.jclinepi.2009.03.021>.

16. Muñoz-Moreno JA, Pérez-Álvarez N, Muñoz-Murillo A, Prats A, Garolera M, Jurado MÀ, Fumaz CR, Negredo E, Ferrer MJ, Clotet B. Classification models for neurocognitive impairment in HIV infection based on demographic and clinical variables. PLoS One. 2014;9(9):e107625. <https://doi.org/10.1371/journal.pone.0107625>.
17. Schouten J, Cinque P, Gisslen M, Reiss P, Portegies P. HIV-1 infection and cognitive impairment in the cART era: a review. AIDS. 2011;25(5):561–75. <https://doi.org/10.1097/QAD.0b013e3283437f9a>.
18. Ho TK. The random subspace method for constructing decision forests. ITPAM. 1998;20:832–44.
19. Breiman L. Random forests. Mach Learn. 2001;45:5–32.
20. Svetnik V, Liaw A, Tong C, Culberson JC, Sheridan RP, Feuston BP. Random forest: a classification and regression tool for compound classification and QSAR modeling. J Chem Inf Comput Sci. 2003;43(6):1947–58. <https://doi.org/10.1021/ci034160g>.
21. Shen C, Yu X, Harrison RW, Weber IT. Automated prediction of HIV drug resistance from genotype data. BMC Bioinformatics. 2016;17(8):278. <https://doi.org/10.1186/s12859-016-1114-6>.
22. Wang D, Larder B, Revell A, Montaner J, Harrigan R, De Wolf F, Lange J, Wegner S, Ruiz L, Pérez-Elías MJ, Emery S. A comparison of three computational modelling methods for the prediction of virological response to combination HIV therapy. Artif Intell Med. 2009;47(1):63–74. <https://doi.org/10.1016/j.artmed.2009.05.002>.
23. Revell AD, Wang D, Wood R, Morrow C, Tempelman H, Hamers RL, Alvarez-Uria G, Streinu-Cercel A, Ene L, Wensing AM, DeWolf F. Computational models can predict response to HIV therapy without a genotype and may reduce treatment failure in different resource-limited settings. J Antimicrob Chemother. 2013;68(6):1406–14. <https://doi.org/10.1093/jac/dkt041>.
24. Larder BA, DeGruttola V, Hammer S, Harrigan R, Wegner S, Winslow D, Zazzi M. The international HIV resistance response database initiative: a new global collaborative approach to relating viral genotype and treatment to clinical outcome. In: Antiviral therapy, vol. 7. London: International Medical Press Ltd; 2002. p. S111.
25. Tarasova O, Biziukova N, Filimonov D, Poroikov V. A computational approach for the prediction of HIV resistance based on amino acid and nucleotide descriptors. Molecules. 2018;23(11):2751. <https://doi.org/10.3390/molecules23112751>.
26. Beerenswinkel N, Schmidt B, Walter H, Kaiser R, Lengauer T, Hoffmann D, Korn K, Selbig J. Diversity and complexity of HIV-1 drug resistance: a bioinformatics approach to predicting phenotype from genotype. Proc Natl Acad Sci. 2002;99(12):8271–6. <https://doi.org/10.1073/pnas.112177799>.
27. Deeks SG, Hellmann NS, Grant RM, Parkin NT, Petropoulos CJ, Becker M, Symonds W, Chesney M, Volberding PA. Novel four-drug salvage treatment regimens after failure of a human immunodeficiency virus type 1 protease inhibitor-containing regimen: antiviral activity and correlation of baseline phenotypic drug susceptibility with virologic outcome. J Infect Dis. 1999;179(6):1375–81. <https://doi.org/10.1086/314775>.
28. Harrigan PR, Hertogs K, Verbiest W, Pauwels R, Larder B, Kemp S, Bloor S, Yip B, Hogg R, Alexander C, Montaner JS. Baseline HIV drug resistance profile predicts response to ritonavir-saquinavir protease inhibitor therapy in a community setting. AIDS. 1999;13(14):1863–71.
29. Walter H, Schmidt B, Rascu A, Helm M, Moschik B, Paatz C, Kurowski M, Korn K, Überla K, Harrer T. Phenotypic HIV-1 resistance correlates with treatment outcome of nelfinavir salvage therapy. Antivir Ther. 2000;5(4):249–56.
30. Weinberger KQ, Saul LK. Distance metric learning for large margin nearest neighbor classification. J Mach Learn Res. 2009;10(Feb):207–44.
31. Drăghici S, Potter RB. Predicting HIV drug resistance with neural networks. Bioinformatics. 2003;19(1):98–107. <https://doi.org/10.1093/bioinformatics/19.1.98>.
32. Hirsch MS, Günthard HF, Schapiro JM, Vézinet FB, Clotet B, Hammer SM, Johnson VA, Kuritzkes DR, Mellors JW, Pillay D, Yeni PG. Antiretroviral drug resistance testing in adult HIV-1 infection: 2008 recommendations of an International AIDS Society-USA panel. Clin Infect Dis. 2008;47(2):266–85. <https://doi.org/10.1086/589297>.

33. Department of Health and Human Services Panel on Antiretroviral Guidelines for Adults and Adolescents. Guidelines for the use of antiretroviral agents in HIV-1-infected adults and adolescents. Washington, DC: Department of Health and Human Services; 2006.
34. Vandamme AM, Sönnborg A, Ait-Khaled M, Albert J, Asjo B, Bacheler L, Banhegyi D, Boucher C, Brun-Vezinet F, Camacho R, Clevenbergh P. Updated European recommendations for the clinical use of HIV drug resistance testing. *Antivir Ther.* 2004;9(6):829–48.
35. Schmidt B, Walter H, Moschik B, Paatz C, Van Vaerenbergh K, Vandamme AM, Schmitt M, Harrer T, Überla K, Korn K. Simple algorithm derived from a geno-/phenotypic database to predict HIV-1 protease inhibitor resistance. *AIDS.* 2000;14(12):1731–8.
36. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, Sajed T, Johnson D, Li C, Sayeeda Z, Assempour N. DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* 2017;46(D1):D1074–82. <https://doi.org/10.1093/nar/gkx1037>.
37. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res.* 2000;28(1):235–42. <https://doi.org/10.1093/nar/28.1.235>.
38. Chen X, Ji ZL, Chen YZ. TTD: therapeutic target database. *Nucleic Acids Res.* 2002;30(1):412–5. <https://doi.org/10.1093/nar/30.1.412>.
39. Hansch C. Quantitative approach to biochemical structure-activity relationships. *Acc Chem Res.* 1969;2(8):232–9. <https://doi.org/10.1021/ar50020a002>.
40. Ballester PJ, Mitchell JB. A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics.* 2010;26(9):1169–75. <https://doi.org/10.1093/bioinformatics/btq112>.
41. Shaikh N, Sharma M, Garg P. An improved approach for predicting drug–target interaction: proteochemometrics to molecular docking. *Mol BioSyst.* 2016;12(3):1006–14. <https://doi.org/10.1039/C5MB00650C>.
42. Kohavi R, John GH. Wrappers for feature subset selection. *Artif Intell.* 1997;97(1–2):273–324. [https://doi.org/10.1016/S0004-3702\(97\)00043-X](https://doi.org/10.1016/S0004-3702(97)00043-X).
43. Lapins M, Wikberg JE. Proteochemometric modeling of drug resistance over the mutational space for multiple HIV protease variants and multiple protease inhibitors. *J Chem Inf Model.* 2009;49(5):1202–10. <https://doi.org/10.1021/ci800453k>.
44. Huang Q, Jin H, Liu Q, Wu Q, Kang H, Cao Z, Zhu R. Proteochemometric modeling of the bioactivity spectra of HIV-1 protease inhibitors by introducing protein-ligand interaction fingerprint. *PLoS One.* 2012;7(7):e41698. <https://doi.org/10.1371/journal.pone.0041698>.
45. Lapins M, Eklund M, Spjuth O, Prusis P, Wikberg JE. Proteochemometric modeling of HIV protease susceptibility. *BMC Bioinformatics.* 2008;9(1):181. <https://doi.org/10.1186/1471-2105-9-181>.
46. Junaid M, Lapins M, Eklund M, Spjuth O, Wikberg JE. Proteochemometric modeling of the susceptibility of mutated variants of the HIV-1 virus to reverse transcriptase inhibitors. *PLoS One.* 2010;5(12):e14353. <https://doi.org/10.1371/journal.pone.0014353>.
47. van Westen GJ, Hendriks A, Wegner JK, IJzerman AP, van Vlijmen HW, Bender A. Significantly improved HIV inhibitor efficacy prediction employing proteochemometric models generated from antivirogram data. *PLoS Comput Biol.* 2013;9(2):e1002899. <https://doi.org/10.1371/journal.pcbi.1002899>.

# Application of Support Vector Machines in Viral Biology



Sonal Modak, Swati Mehta, Deepak Sehgal, and Jayaraman Valadi

**Abstract** Novel experimental and sequencing techniques have led to an exponential explosion and spiraling of data in viral genomics. To analyse such data, rapidly gain information, and transform this information to knowledge, interdisciplinary approaches involving several different types of expertise are necessary. Machine learning has been in the forefront of providing models with increasing accuracy due to development of newer paradigms with strong fundamental bases. Support Vector Machines (SVM) is one such robust tool, based rigorously on statistical learning theory. SVM provides very high quality and robust solutions to classification and regression problems. Several studies in virology employ high performance tools including SVM for identification of potentially important gene and protein functions. This is mainly due to the highly beneficial aspects of SVM. In this chapter we briefly provide lucid and easy to understand details of SVM algorithms along with applications in virology.

**Keywords** Support vector machines · Supervised learning · Classification · Regression function identification · Epitope prediction · Quantitative structure activity relationships · Domain attributes · Attribute selection viral biology

---

S. Modak · S. Mehta · D. Sehgal

Life Sciences and Healthcare Unit, Persistent Systems Ltd., Pune, Maharashtra, India

J. Valadi (✉)

Life Sciences and Healthcare Unit, Persistent Systems Ltd., Pune, Maharashtra, India

Center for Development of Advanced Computing, Savitri Bai Phule Pune University, Pune, India

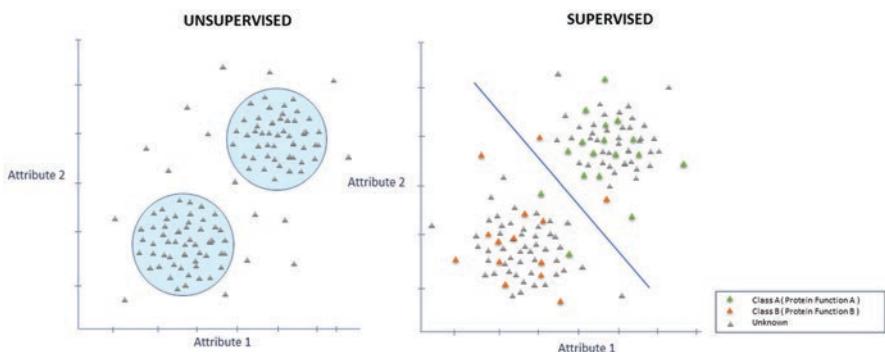
e-mail: [jayaraman@cms.unipune.ac.in](mailto:jayaraman@cms.unipune.ac.in)

## 1 Introduction

Accurate annotation employing domain information extracted from sequence/structure and related attributes immensely enhances our current understanding of viral genomes. A major role is played by data driven modelling in recent advances made in vaccine development, epidemiology studies, pathogenicity determination, and drug design [1]. Introduction of NGS technology coupled with novel experimental techniques have provided very large volumes of data requiring accurate machine learning based modelling techniques. These techniques can be broadly categorized into supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning can be explained with a classic example of function annotation (see Fig. 1). In this task we have knowledge of certain number of sequences belonging to functional ‘class1’ from prior experimental annotation and knowledge of another set of sequences known not to be annotated as ‘class2’. As shown in Fig. 1, a knowledge based model is built which separates data into two classes. This knowledge may be in terms of domain attributes extracted from sequences\structure, etc. The set of domain attributes are known as input data. Experimentally annotated class information is known as output data. The supervised learning model derives a functional relation between input and output. This model can be used to classify a query example to identify the functional class employing this model. This approach can be extended to classification into multiple functional classes.

In Unsupervised learning, we do not have prior knowledge about the classes. Unsupervised Learning is a class of Machine Learning techniques which enables us to discover patterns in the data. The data given to the unsupervised algorithm are not labelled, which means only the input variables (X) representing sequences\structure are presented to the algorithm with no corresponding output variables (Fig. 1). This type of learning is used extensively in viral biology to infer Phylogeny. The unsupervised learning method groups data without any prior knowledge of class labels. After the model is built one can derive knowledge about examples clustered in any



**Fig. 1** Supervised vs. unsupervised learning

specific group. While supervised and unsupervised learning learn from data, the reinforcement learning paradigm learns from experience. In the following sections we provide details of SVM algorithms, a list of domain attributes presented to the algorithm, selection of informative attributes, and finally a discussion on some applications of SVM in viral biology.

## 2 Support Vector Machines for Classification

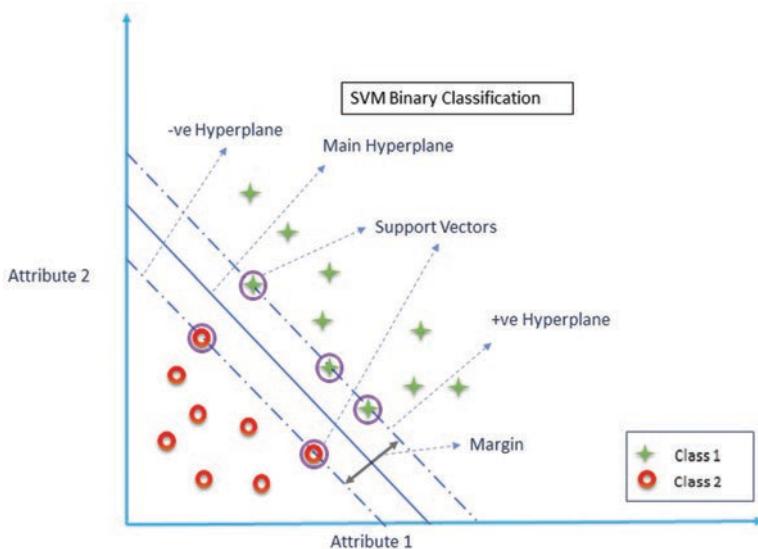
Support Vector Machines can be used both for supervised and unsupervised learning tasks. In viral biology, SVM is used mainly for supervised learning. SVM classifiers are a set of universal feed-forward network-based algorithms that have been rigorously formulated from statistical learning theory by Vapnik [2]. They are very popular machine learning paradigms which are routinely used in different branches of science and engineering.

### 2.1 SVM Binary Classifier for Linearly Separable Data

Let us take a simple case study to explain principle of SVM Linear Classification. The task is to build a model to separate a set of sequences belonging to functional class 1 from another set of sequences belonging to functional class 2. Class 1 examples can be peptides having antiviral activity while class 2 examples are not known to possess any antiviral activity. The input data vector for *i*th example is denoted by  $x_i$  and the corresponding class label is denoted  $y_i$ . The output of any example belonging to class 1 is represented by the subset  $y_i = +1$  and those belonging to class 2 are represented by the subset  $y_i = -1$ . The hyperplane for the linearly separable data can be defined as:

$$\mathbf{w} \cdot \mathbf{x}_i + b = 0$$

This hyperplane (Fig. 2) separates the data into two different classes. ‘ $\mathbf{w}$ ’ refers to the weight vector with elements equal to the number of attributes. The problem here is to find out the best values of the elements of the weight vector, which maximize separation of the two classes with reference to a given performance measure (e.g. accuracy). This amounts to finding a hyperplane which maximizes the margin. This implies that at the training stage the examples belonging to class 1 should be maximally separated from examples belonging to class 2. It can be shown that such a problem can be formulated as a Convex Quadratic Optimization problem [2]. The solution for such a convex optimization problem has only one global optimum as opposed to multiple local optimum solutions (algorithm can get stuck up in any of

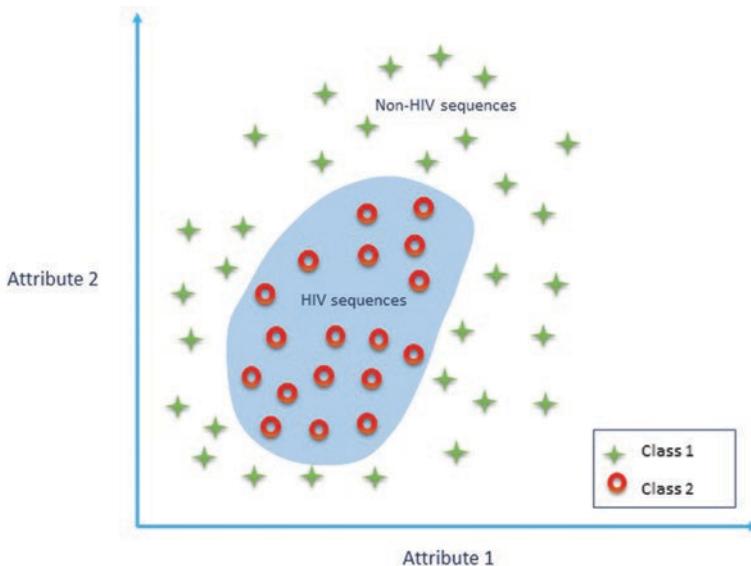


**Fig. 2** Maximum margin-minimum norm classifier

the inferior local optima) like other candidate algorithms like neural network etc. have. It is this highly beneficial aspect coupled with superior performance has attracted researchers and practitioners from different fields to employ Support Vector Machines. After model building, the weight vectors can be obtained from only a subset of training examples. This subset is known as Support Vectors and hence the name Support Vector Machines. It must be noted here that SVM converts the original “N” dimensional problem into a one dimensional problem using dot products between the examples.

## 2.2 Non-linear Support Vector Machines

Biological data are inherently non-linear. A linear hyperplane cannot satisfactorily separate such non-linear data (Fig. 3). To handle these data SVM first transforms the data to a higher dimensional feature space and then employs a linear hyperplane. There are two inherent difficulties in the above approach: (i) It is difficult to find a suitable transformation by trial-and-error. (ii) We may have to employ a transformation to a very high dimensional space for reasonable classification accuracy which becomes computationally intractable. To solve these problems SVM employs appropriate kernel functions. Kernel functions are defined as a function of dot products in the original space and they are equivalent to the dot products in the higher dimensional feature space. SVM separating surface can now be defined as a linear

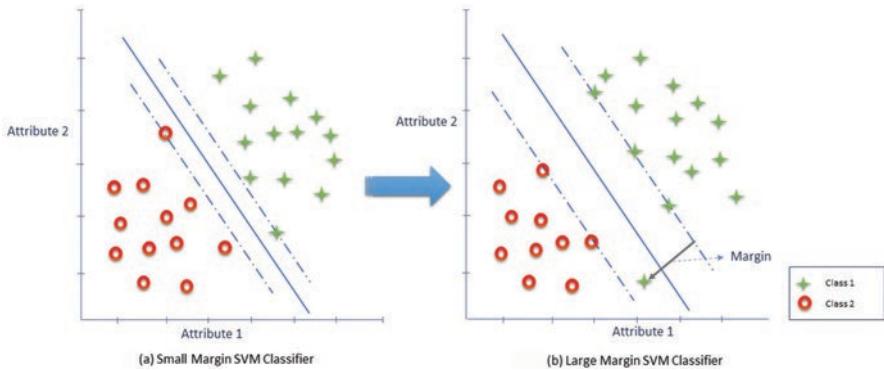


**Fig. 3** Non-linearly separable data

hyperplane in the high dimensional feature space and introduction of appropriate kernel functions make it possible to do all the computations in the original space itself. Kernel functions have to satisfy Mercers Theorem; They have to satisfy the axioms of Hilbert space and have to be positive definite. The most popular kernel functions are Polynomial, Gaussian Radial Basis Function (RBF), and Multi-layer Perceptron kernel functions. Apart from these there are several domain dependent kernel functions. In computational biology, string kernels and Fisher kernels are very popular. Formulation as described above is known as Hard-margin SVM classification.

### 2.3 Soft Margin SVM

If we try to find a hyperplane which yields the maximum possible training accuracy, the margin obtained may become very narrow. Such a hyperplane while classifying the training set very well, over-fits the data and may fail miserably in unseen query test examples. It may be possible to increase the margin with slight loss of training accuracy (Fig. 4). This will generalize better than the one having a narrow margin and has more robust prediction capabilities. This trade-off between margin maximization and misclassification error in soft margin formulations can be obtained by optimizing a new parameter ‘C’.



**Fig. 4** Trade off: increasing margin/reducing misclassification

### 3 Brief Details of Classification of Real-Life Binary Datasets

Given a dataset we must first find the optimal hyperplane in the original dimension. In SVM terminology this is known as a linear kernel and after building the model we must estimate the required performance measure (e.g. accuracy). If it is not satisfactory, we must resort to nonlinear separation and employ conventional kernels like Polynomial, Gaussian Radial Basis Function (RBF), Exponential Radial Basis Function, Multi-layer Perceptron kernel functions etc. For every kernel, there are kernel parameters. With each kernel, apart from finding the best kernel parameters one must also tune the ‘C’ parameter as discussed in earlier section. If these kernels also are not satisfactory then we must resort to domain dependent kernels.

### 4 Support Vector Machines for Regression

In classification examples are grouped into discrete sets. In regression, a functional relationship is found between input data and output having continuous values. Many problems require a nonlinear model to adequately regress the data. The methodology described in the previous sections can be easily extended to employ SVM to handle nonlinear regression (Schölkopf et al. 1999) [2]. The methodology for linear regression is same as that of conventional models for regression; examples which are linearly classifiable can be done in the original dimension itself. What is different in SVM linear regression is that a novel epsilon-insensitive loss function is defined, which is robust against outliers in the data [3, 4]. For data that cannot be regressed linearly, a principle similar to the one implemented in classification problems can be extended easily; for such kind of problems, data needs to be taken to a higher dimensional feature space and subsequently regressed linearly. Appropriate kernel functions can again be defined to simplify computation.

## 5 Attributes Used in Viral Biology Problems

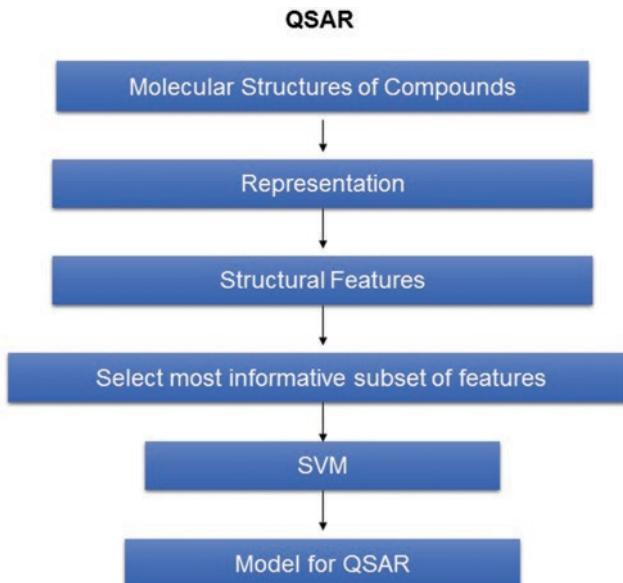
In viral biology we encounter a variety of attribute types, with each type providing huge magnitudes of domain attributes. Broadly, these attributes can be classified as sequence based, structure based, spectrum of light or radiation based (i.e. spectroscopic), microarray gene expression profiles etc. Protein sequence k-mer features range from amino acid (AA) ( $k = 1$ ), dipeptide ( $k = 2$ ), tripeptide ( $k = 3$ ) to tetrapeptide ( $k = 4$ ) and so on. It is possible to extract physiochemical properties like hydrophobicity, charge, hydrophilicity etc., from each of the AA alphabets. The simplest discrete set of features is the AA composition. Conversion of sequence information in terms of AA composition reduces the protein sequence into a 20 letter alphabet. While this is beneficial, we lose all sequence information. Recently Chou defined and introduced different types of pseudo-AA (PseAA) compositional attributes of protein sequences; these are a set of discrete numbers derived from AA sequences possessing some sort of sequence order or pattern information [5]. Ever since the first PseAA composition was formulated, these attributes have been successfully employed in several protein function identification tasks. Two classes of attributes frequently used in viral biology are listed below:

### 5.1 QSAR Descriptors

In quantitative structure activity relationship modelling, domain information about a molecule is provided in terms of different types of descriptors. The initially developed QSAR descriptors comprise hydrophobic, electric, and stearic parameters. Currently, descriptors of different dimensions ranging from 0 to 3 are routinely employed in modern QSAR analysis. Zero-dimensional descriptors comprise of atom counts, bond counts, molecular weight, sum of atomic properties; one dimensional descriptors two-dimensional descriptors deal with topological descriptors and three dimensional descriptors provide geometrical information. Originally QSAR is regression problem in which a functional relationship is obtained between activity of a molecule and the descriptors. This relationship can be linear or non-linear so a regressor like SVM or random forest can be employed for this job. This is illustrated in Fig. 5.

### 5.2 PSSM Descriptors

Evolutionary information, one of the most important types of information in assessing functionality in biological analysis, has been successfully used to encode protein in many applications. PSIBLAST is used to repeatedly search specific databases, using a multiple alignment of high scoring sequences found in each search as input



**Fig. 5** QSAR regression using SVM

in the next round of searching. Normally iterations are continued until user specified number of iterations and at the end, the final Position Specific Scoring Matrix (PSSM) is generated. Such a matrix provides remote homology information and using PSSM attributes as descriptors in SVM would be useful if remotely connected sequences have similar functionalities. In the view of the fact that SVM requires the fixed length feature vectors, a vector of dimension 400 can be recovered from PSSM score matrix for use as input in SVM classifier.

Apart from the attributes described above, many different types of attributes are used, depending on the particular domain problem encountered.

### 5.3 Attribute Selection

Not all attributes are informative in data sets. Features which are non-informative will act as noise, do not have discriminative power & interfere with the classification process. Hence the model will have very little predictive accuracy. In Protein function identification in viral biology, several sequence and structural features can be extracted [6, 7]. For example the AA, dipeptide & tri-peptide compositional features put together amount to 8400 in number & not all of them will be important in a particular function annotation task. To select a subset of informative features by brute force, we need to evaluate huge number of subsets of features which becomes

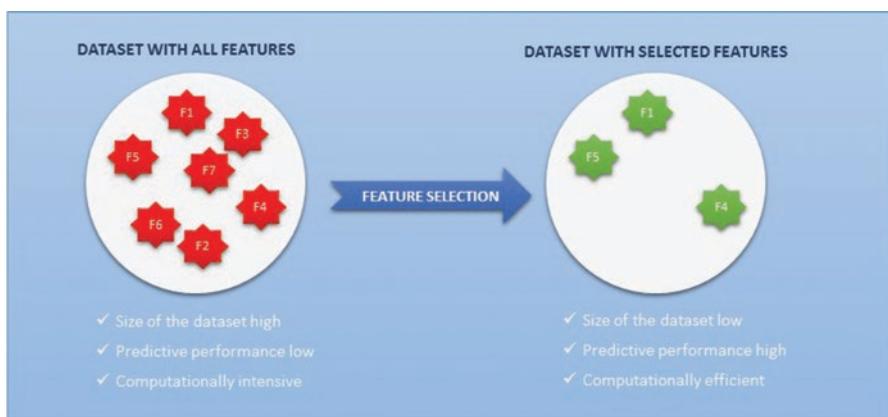
computationally time consuming. Various feature/attribute selection methods are available to simplify the process of subset selection. Feature selection techniques help us to avoid overfitting and improve model performance to provide faster and more cost-effective models; they also provide invaluable domain information. However, feature selection techniques have to employ appropriate search techniques, they bring in an additional level of complexity and computational cost. Feature selection techniques differ from each other in the way they incorporate this search in the added space of feature subsets in the model selection. Figure 6 illustrates the advantages of feature selection. These methods can be broadly classified as filter, wrapper and embedded methods.

### 5.3.1 Filter Ranking Methods

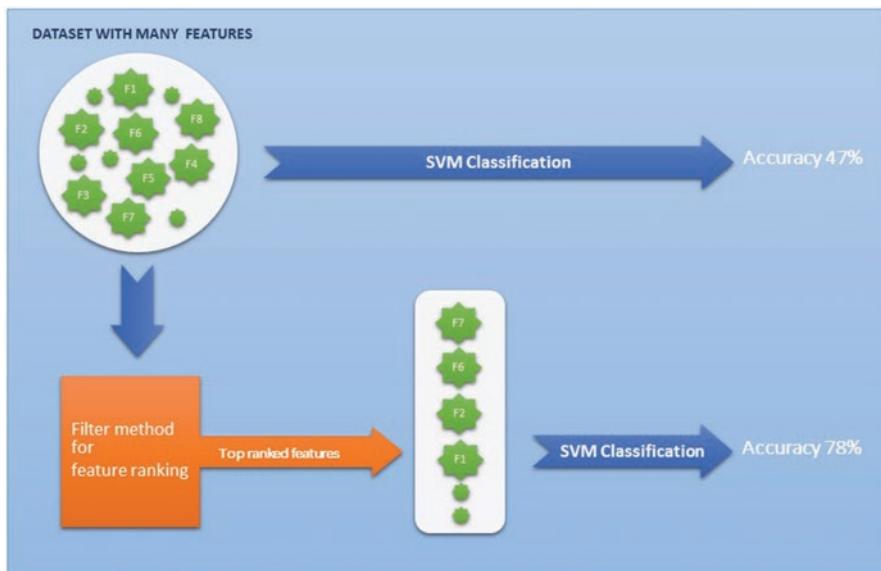
Filter ranking methods use some heuristics to score and rank the features (Fig. 7). In the example given above once the 8400 features are ranked by an appropriate filter method, the most informative subset of features can be selected, and the model can be built on this subset to maximize performance. Most popular filter methods include mutual information, student t-test, correlation-based feature selection (CFS) and several variants of the Markov blanket filter method, Minimum Redundancy-Maximum Relevance (mRmR) and Uncorrelated Shrunken Centroid (USC) algorithm. We give below some of the methods used in viral biology related problems:

#### 5.3.1.1 Information Gain

Information gain score for any given attribute is calculated as the difference between entropy of the entire data set and the conditional entropy of for each possible value of the attribute. This can be done by binning each attribute and counting the fre-



**Fig. 6** Advantages of feature selection



**Fig. 7** Filter ranking & classification accuracy

quency of occurrence of different labels for the range of the attribute in each bin. Based on the score, top ranking attribute subset can be easily identified to build the model.

#### 5.3.1.2 mRmR

The attributes are selected in such a way they are mutually dissimilar, non-redundant and maximally relevant simultaneously.

#### 5.3.1.3 Mutual Information

Mutual information is a measure between random variables, that quantifies the information obtained about one of them, through the other. For the purpose of feature selection, mutual information between the subset of selected features and the target variable should be maximal.

#### 5.3.1.4 Correlation Filter

The Correlation Feature Selection (CFS) selects subset of features that uncorrelated to each other but maximally correlated to the output variable.

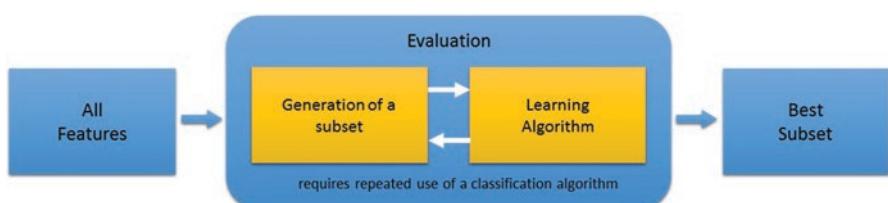
### 5.3.1.5 Chi-Square

The chi-square test is a statistical test computes a score reflecting of independence to determine the dependency of two variables. We need to calculate chi-square statistics between every feature variable and the target variable and observe the existence of a relationship between the variables and the target. If the target variable is independent of the feature variable, we can discard that feature variable. If they are dependent, the feature variable is very important. For continuous variables, chi-square can be applied after “Binning” the variable.

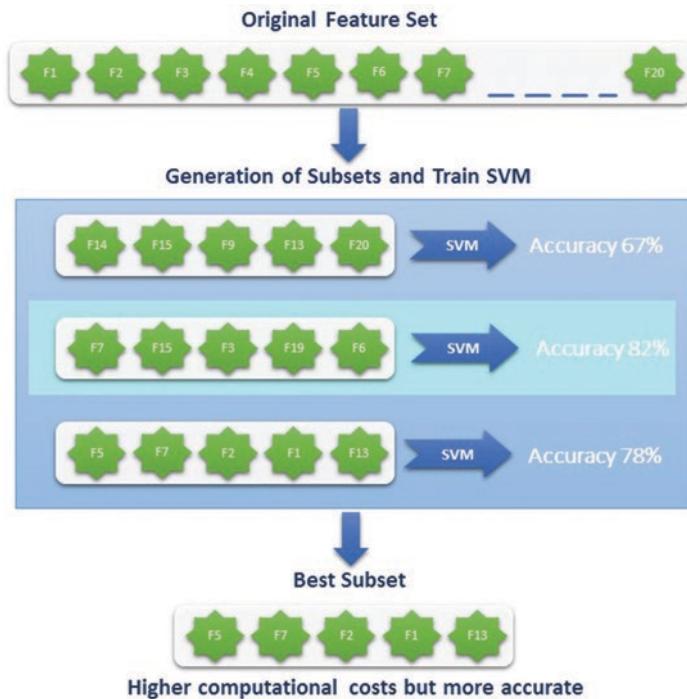
### 5.3.2 Wrapper Methods

While filter methods are fast, they are not very accurate as they do not encode feature correlation. Wrapper methods employs a learning classifier for repeated evaluation of different subsets of features. These methods include forward selection & backward selection algorithms. In forward selection we start with an empty set and add features one by one which maximally improve accuracy until all features are added in the set. A subset can then be chosen which exhibits maximum accuracy. In backward selection we start with all features and remove least significant features one by one.

Recently recursive feature elimination wrappers have become very popular. In SVM recursive feature elimination algorithm, viz., SVM-RFE, the simulations start with all features and the algorithm weights are determined. Then features with least absolute value of weight are recursively removed until no feature is left out. Here again best performing subset can be easily identified which is used in the final model (see Figs. 8 and 9). Several wrapper based methods are population based and use Genetic algorithms, Ant Colony Optimization or other swarm intelligent methods. These methods mimic some nature inspired phenomena and evolve optimal solutions. Fie e.g. ACO is based on co-operative search behaviour of live ants. Biogeography is the study of distribution and dynamics of a large number of species geographically over a period of time. Biogeography based optimization (BBO)



**Fig. 8** Wrappers: schematic representation



**Fig. 9** Wrappers & classification accuracy

involves mimicking the natural processes of migration over a population in iterative generations, simulating discrete time. Atulji Srivatsava et al. employed BBO Simultaneous Feature Selection and MHC Class I Peptide Binding Prediction using Support Vector Machines and Random Forests [8].

### 5.3.3 Embedded Methods

In embedded class of feature selection techniques, optimal subset search is facilitated within the classification model. In random forest there are two inbuilt feature ranking methods, viz., Gini importance and variable importance. In SVM recursive feature elimination algorithm, viz., SVM-RFE, the simulations start with all features and the algorithm weights are determined. Then features with the least absolute value of weight are recursively removed until no feature is left out. Here again best performing subset can be easily identified which is used in the final model.

## 6 Performance Measures

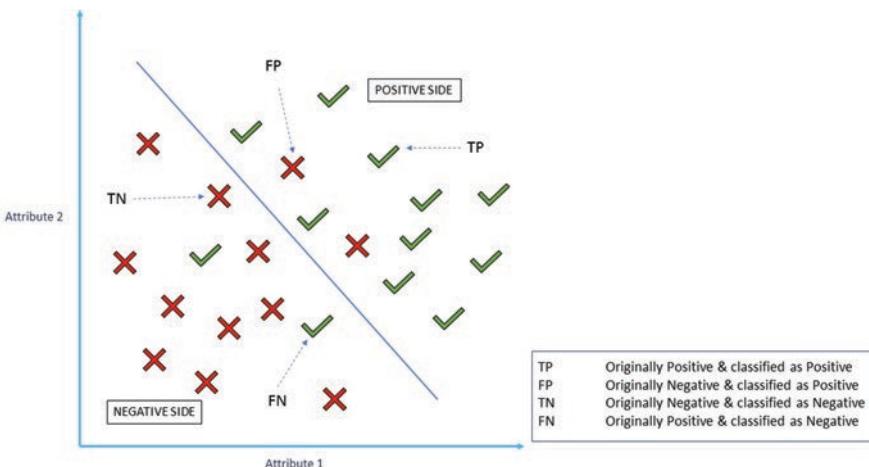
While accuracy is the conventional performance measure, it may not be appropriate in all situations. In some examples we may require maximizing the positive accuracy while in some other situations negative accuracy may be the desired performance measure. Also, in imbalanced datasets, where we have more examples in one class than the other we have to optimize both positive and negative accuracies.

Referring to Fig. 10, true positives are the examples which are originally positive and are predicted positive by SVM. True negatives are the examples which are originally negative and predicted negative. False positives are the examples which are originally negative but predicted positive. False negatives are the examples which are originally positive, but predicted negative. With these definitions, we can define positive and negative accuracies. True positive rates or sensitivities are defined as;

$$TPR = \frac{\text{number of true positive examples}}{\text{total number of positive examples}} = \frac{TP}{TP + FN}$$

True negative rate or specificity can be defined as:

$$TNR = \frac{\text{number of true negative examples}}{\text{total number of negative examples}} = \frac{TN}{TN + FP}$$



**Fig. 10** Distribution of examples classified by the model

Precision or positive predictive value can be defined as:

$$PPV = \frac{TP}{TP + FP}$$

F1 score is a harmonic mean of precision and sensitivity:

$$PPV = 2 \left[ \frac{PPV * TPR}{PPV + TPR} \right] = \frac{2TP}{2(TP + FP + FN)}$$

Apart from these Matthew Correlation Coefficient is used as measure which provides optimal positive and negative accuracy and can be defined as:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

MCC score of  $-1$  indicates very poor classification and  $+1$  indicates highest possible performance. In case of imbalance datasets it is customary to use MCC as the desired performance measure.

## 6.1 Cross Validation Measures

A simple way to test the performance is to split the data with 80% train and 20% test. The model is built on the 80% train data and tested on 20% test data. While this can be done for quickly estimating the performance of the model may not be fully adequate. To remove statistical bias two different cross validation measures are used to gauge the performance and obtain the best algorithm parameters. In K-fold cross validation, the training set is randomly divided into K-folds. To start with, the first fold is used as the test set and the remaining  $k - 1$  folds are used to build SVM hyperplane model. This model is evaluated by using the examples in the first fold. Similarly, each of the other  $k$  folds are used as test sets and the remaining  $k - 1$  folds are employed to build the models respectively. From these  $k$  experiments the cross validation accuracy is estimated as the average of  $k$  test accuracies. In leave one out cross validation procedure, each time one example is left out as a test example and the remaining  $n - 1$  examples are used to build the model. The built model is tested with the left out example. Conventionally fivefold or tenfold measures are used ( $k = 5$  or  $k = 10$ ). In k-fold cross validation, irrespective of the number of examples in the datasets,  $k$  different models are always built, whereas in leave-one-out cross validation, number of models is equal to number of examples in the training set.

## 7 SVM Extension to Solve Multi-class Type of Classification Problems

There are different algorithms, which address multi-class classification problem. Two well-known techniques include one-against-all method (Weston and Watkins 1999) and one-against-one technique [9]. One-against-all method considers the multi-class problem as a collection of binary classification problems. In general,  $k$  classifiers are needed to solve the  $k$  class problem. The  $k$ th classifier constructs a hyper-plane between class  $k$  and the  $k - 1$  other classes. A majority vote across the classifiers is applied to classify the new test point. In one-against-one technique  $k(k + 1)/2$  classifiers are needed. In each classifiers a model is built with examples of one class against examples of another class. Here again for a test example majority vote is needed to decide the class label.

## 8 Other SVM Types

### 8.1 Least Square SVM (LSSVM)

Least Square SVM classifier were proposed by Suykens and Vandewalle [10]. In their version of least square SVM they add a term in the objective function which penalizes square of error between prediction and actual class label. In this version, the problem is now formulated as a set of linear equations, instead of the convex quadratic problem for classical SVMs. Such a formulation makes computation simpler and faster. Several problems in bioinformatics has been solved using LSSVM. LSSVM formulation has also been extended for solving SVM problems.

### 8.2 One Class SVM

Several real-life datasets are highly imbalanced. Function annotation problems in viral biology have a small number of positive examples, while the negative examples can be very large. So such a distribution causes imbalance in the datasets and the minority class prediction accuracy will be very poor. One class SVM has been proposed in the literature to overcome this issue. In One class SVM only the data belonging to the majority class examples is used to build the model. There are two different version proposed in the literature for One class SVM. In the Tax and Duin's version [11], a model for the smallest hyper-sphere including all the majority class examples is formed. A new example is predicted as a majority class example if it falls inside the sphere. Otherwise it is predicted as a minority class example. For

non-linearly separable patterns appropriate kernel functions can be defined as in the case of binary SVMs. In the other version of the One class SVM, a hyperplane model is used instead of a hypersphere model [12]. One class SVM can also be used to detect anomalies and faults.

## 9 Applications of SVM in Virology

In this section, we outline a few important problems in viral biology where SVMs have been successfully applied on many case studies.

### 9.1 *Quantitative Structure Activity Relationship (QSAR) Applications*

Rapid assessment of desired activities of a large number of small-molecule compounds can be achieved by High throughput screening (HTS). QSAR analysis has been playing a key role in screening of compounds by building knowledge-based models [13]. This greatly reduces the experimental screening load. QSAR methodology focuses on finding a model, which allows for correlating the experimentally determined activity of a family of compounds with their molecular structure. Once a high performance model is built, it can be used to identify the activity of any new compound based on appropriate domain attributes extracted from their molecular structure. The set of atoms and covalent bonds between them can define a molecular structure. However, creation of structure-activity relationship models cannot be directly done from the structure of the molecule. Domain information has to be presented to the algorithm in the form of descriptors; molecular descriptors range from physicochemical and quantum-chemical to geometrical and topological features. The methodology of building QSAR models consists of four steps: (a) extracting descriptors from molecular structure (b) choosing most informative descriptors as per activity (c) building a model based on filtered molecular descriptors (d) screening molecule for activity in question. In Table 1, different example of descriptors as listed. These examples are categorised based on structural conformations [13].

Quantitative structure–activity relationship (QSAR) modelling with descriptor selection has become increasingly important because of a large number of descriptors of different types can be extracted in principle. Descriptor selection can improve the accuracy of QSAR classification studies and reduce their computation complexity by removing the irrelevant and redundant descriptors. Descriptor selection is an important pre-processing tool for QSAR studies. The sparse support vector machine (SSVM), one of the embedded methods, is of particular interest because it can perform descriptor selection and classification simultaneously.

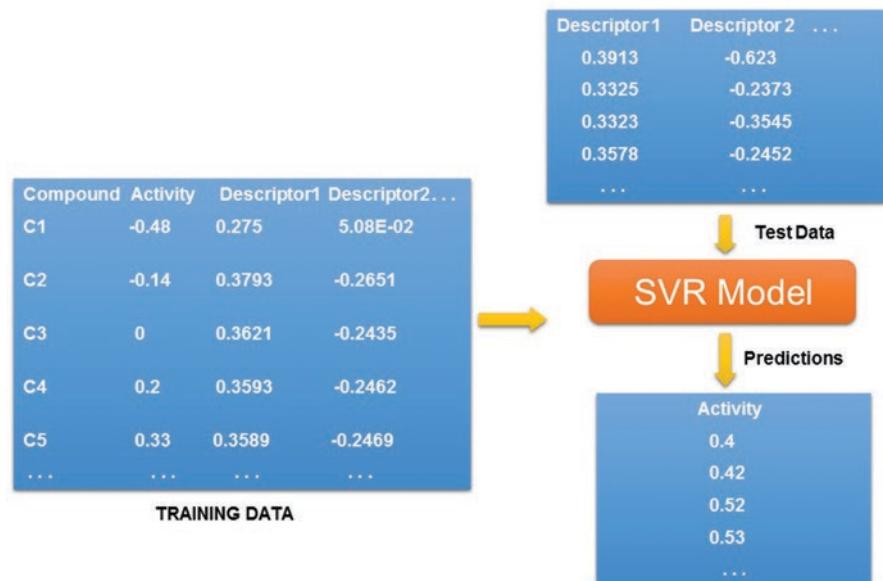
**Table 1** Examples of different descriptors based on structural conformation [15]

Category	Descriptors
2D QSAR descriptors	Constitutional descriptors Electrostatic and quantum-chemical descriptors Topological descriptors Geometrical descriptors Molecular fingerprints & fragment-based descriptors
3D QSAR descriptors	Comparative molecular similarity indices analysis Comparative molecular moment analysis Weighted holistic invariant molecular descriptors VolSurf approach Grid-independent descriptors

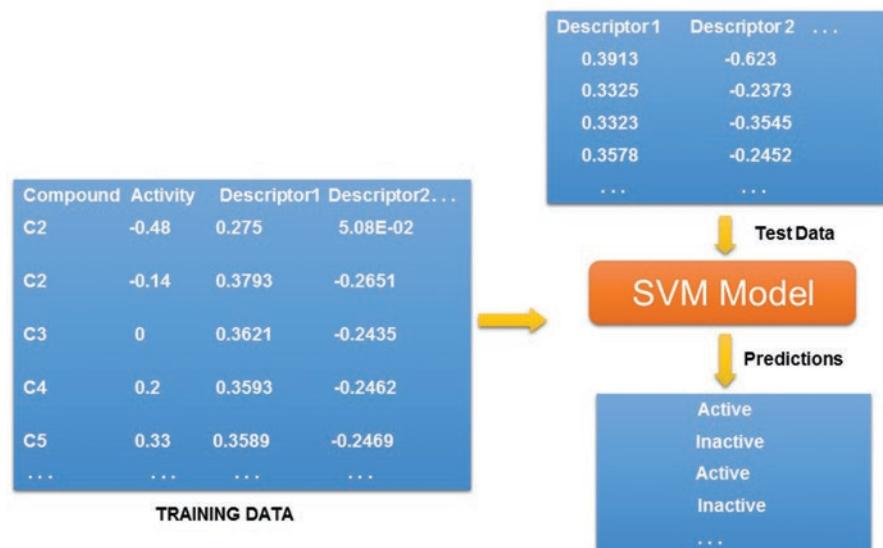
Further explanation is included in Sect. 5.3.

SVMs have been found to provide robust and accurate QSAR models for several problems encountered in viral biology. Two types of QSAR models can be built. First one is a regression problem in which a model is built against descriptors vs. experimentally annotated activities. This is a regression problem, schematically shown in Fig. 11. The second problem can be posed as classification problem. For this a threshold value for the experimental activities has to be defined. Compounds having activities less than these threshold activities are grouped into ‘class1’. The other compounds are grouped into ‘class2’. SVM classification model is built to separate compounds into two groups. A new query compound can then be classified as active or inactive as schematically represented in Fig. 12.

Human immunodeficiency virus (HIV) affects and destroys the immune system and causes acquired immunodeficiency syndrome (AIDS) disease. As per the **UNAIDS report** [14], 77.3 million [59.9 million–100 million] people have become infected with HIV since the start of the epidemic and 35.4 million [25.0 million–49.9 million] people have perished from AIDS-related illnesses since the start of the epidemic. Numerous molecular modelling approaches have been attempted to address the design of new anti-HIV compounds. Most of them are based on QSAR [15]. In an interesting and comprehensive study [15] QSAR based attributes were selected for predicting inhibiting activity of the compound against HIV proteins including protease (PR), reverse transcriptase (RT) and integrase (IN). Around 18,000 molecular descriptors which include geometric, electrostatic, structural, constitutional, path and graph fingerprints etc. were extracted utilizing the open source PaDEL software. To reduce the number of descriptors Attributes selection was carried out using ‘Best-First’ as the search method in Waikato Environment for Knowledge Analysis (Weka) suite. SMO regression algorithm in the Weka suite was



**Fig. 11** SVM regression model



**Fig. 12** SVM classification model

used to classify the data into active and inactive sets. The models were able to achieve excellent values of Pearson correlation coefficient for all the three data sets, *viz*, PR, RT, IN. An integrated web based Platform **HIVprotI** [16] was further developed using this model.

The tetra-hydro-imidazo[4,5,1-jk][1,4]-benzodiazepines (TIBOs), constitute a group of potent system inhibitors of HIV-1 reverse transcriptase. With a view to segregate TIBO compounds into high and low classes of inhibitors of HIV-1 reverse transcriptase, Hdoufane et al. carried out SAR studies on 89 TIBO derivatives using different classifiers, such as support vector machines, artificial neural networks, random forests, and decision trees [17]. They successfully employed seven molecular descriptors characterizing hydrophobic, electronic, and topological aspects of the molecules and obtained excellent training and test accuracies.

The successful identification of HIV proteins may have important significance in treatment since epidemiological and biological characteristics of HIV-1 and HIV-2 are quite different., Juan Mei et al. employed SVM along with other classifiers to predicted HIV-1 and HIV-2 proteins based on pseudo AA compositions and increment of diversity (ID) algorithm [18]. With jack knife tests, SVM models gave the highest prediction accuracy of 0.9909.

Both HBV and HCV are of immense significance as leading causes of liver cancer as well as co-infection with HIV. A potentially important study included 172 positives and 8998 negative cases and built a classification model of the HBV dataset; in the same study HCV dataset included 533 positives and 7287 negatives [19]. The data had obvious imbalance in the number of examples in the positive and negative data sets. Three different imbalance handling methods, *viz.*, (i) Downsize, (ii) Multi downsize, and (iii) SMOTE were used. SMOTE provided the best performance; SVM prediction accuracies of 64% for HBV and 71% for HCV were reported for this model.

Influenza, a respiratory virus, is correlated with high morbidity and mortality rates. Neuraminidase (NA) and haemagglutinin (HA) are two major glycoproteins found on the surface of the influenza virus. Compounds that inhibit neuraminidase can protect host cells from viral infection and retard the spread of the virus among cells. A two staged approach has been used to build a QSAR classification model separating neuraminidase as active and inactive [20]. In the first stage minimum redundancy maximum relevance criteria was employed to select the most informative descriptors. The second stage employs the selected descriptors as input to SSVM L1-norm classifiers. The dataset consisted 479 neuraminidase inhibitors of H1N1 virus whose experimentally measured IC<sub>50</sub> values were available. These set of training compounds were separated by thresholding the activity into two categories: active compounds with IC<sub>50</sub> < 20 μM, while those with IC<sub>50</sub> > 20 μM were considered to be weakly active compounds. The 7 top descriptors selected gave an SVM classifier accuracy of 90.62% which is far higher than the earlier SVM approaches.

The classification of protein quaternary structure complex is of significant interest in computational biology research. Chi-Chou Huang et.al have developed a two-staged architecture for five class classification of grouping protein quaternary structure of a complex; the five classes are monomer, dimer, trimer, tetramer, and other subunit classes [21]. AA frequency, Shannon entropy and accessible surface areas were employed as domain attributes. One against all SVM classifiers were used of in which positive data consisted of examples of given class and negative data consisted of all the remaining classes. Due to this division number of examples in the positive side of the classifier was much less than the negative side. This created imbalance and reduced classification accuracy. To counter this, the author employs a bootstrap method for repeated sampling and generated different subsets of data. The majority class was further subjected to random sub-sampling. Mathews Correlation coefficient was used as performance measure. The bootstrapping method was able to produce an MCC of 0.696 and above. List of examples are given in Table 2.

**Table 2** Illustrative examples for QSAR applications

Sr. no.	Reference	Brief description of work	Attributes and attribute selection
1.	A binary QSAR model for classifying neuraminidase inhibitors of influenza A viruses (H1N1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine [20]	To predict the neuraminidase inhibitors as active and inactive based on QSAR using 479 neuraminidase inhibitors of H1N1 virus with experimentally measured IC50 values.	Molecular structures of the compounds were sketched using Chem3D software. Dragon software (version 6.0) was used to generate 4885 molecular descriptors including all 29 blocks based on the optimized molecular structures, 2881 left after cleaning up
2.	In Silico SAR Studies of HIV-1 Inhibitors [17]	Classify TIBO compounds into two groups: High and low inhibitors of HIV-1reverse transcriptase based on QSAR studies.	500 molecular descriptors from five different classes (geometrical, topological, constitutional, electrostatic, and quantum-chemistry descriptors).
3.	HIVprotI: An integrated web based platform for prediction and design of HIV proteins inhibitors [16]	A web server to predict inhibition activity of a compound against HIV proteins namely protease (PR), reverse transcriptase (RT) and integrase (IN).	18,000 molecular descriptors extracted using PaDEL software which include geometric, electrostatic, structural, constitutional, path and graph fingerprints
4.	PClass: Protein quaternary structure classification by using bootstrapping strategy as model selection [21]	A web server for protein quaternary structure complex classification into 5 categories namely: Monomer, dimer, trimer, tetramer, and other subunit classes	AA freq. Shannon entropy and accessible surface areas

(continued)

**Table 2** (continued)

Sr. no.	Reference	Brief description of work	Attributes and attribute selection
5.	Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers [18]	To predict of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions	(i) 20 AA compositions (A1) (ii) 400 dipeptide compositions (A2) (iii) AA hydropathy compositions (H1) (iv) 36 hydropathy dipeptide compositions (H2)
6.	iPHLoc-ES: Identification of bacteriophage protein locations using evolutionary and structural features [96]	Tool to discriminates between host located and non-host located phage proteins (PH & non-PH) and membrane and cytoplasm located host proteins (PHM & PHC).	PSSM, AA composition and structural features of the sequences Above features are generated using (1) PSSM file generated from PSI-BLAST and (2) SPD file generated from SPIDER2 software.
7.	Enhancement of hepatitis virus immunoassay outcome predictions in routine pathology data by data balancing and feature selection before the application of support vector machines [19]	Prediction of HBV and HCV for negative and positive using Balancing methods to counter negative samples.	25 variables from laboratory
8.	QSAR studies of the bioactivity of hepatitis C virus (HCV) NS3/4A protease inhibitors by multiple linear regression (MLR) and support vector machine (SVM) [97]	Predict bioactivity of hepatitis C virus (HCV) NS3/4A protease inhibitors	MACCS fingerprint, 20 global molecular descriptors and 88 2D property-weighted autocorrelation descriptors calculated using CORINA Symphony
9	A computational model for predicting transmembrane regions of retroviruses [98]	Identify transmembrane regions in envelope glycoproteins of retroviruses (HERV, HIV, HTLV, SIV, MLV)	10 physicochemical and PSSM score features

## 9.2 SVM Applications Based on Next Generation Sequencing (NGS) Data

The term ‘Next-Generation’ Sequencing (NGS)’ refers to the advancement in nucleic acid sequencing technologies. Numbers of sequence reads generated per run has progressively increased with time, due to improved understanding of molecular biology as well as technological advances. Current sequencing platforms are capable of generating enormous numbers of sequence reads in quick turnaround time, allowing researchers to explore all possible aspects of biomedical studies at molecular level and dig deeper in the genetic aspects. NGS has proven to be an efficient,

fast and reliable approach to solve problems in studies of evolution, ecology and genetics, overcoming the limitation of traditional molecular approaches [22]. Another great advantage of NGS approach over traditional molecular studies is that it is also cost efficient. End-to-end human genome can be sequenced in few hours using NGS technology, whereas, it took over a decade to sequence and assemble human genome using Sanger Sequencing. Based upon the chemistry, a number of NGS platforms have been developed since last decade. Bioinformatics knowledge plays an important role in assembling the fragments sequenced in parallel by mapping all the read sequences to the human genome reference. Depth of the sequencing, i.e. number of times the template has been sequenced, assures accuracy of sequencing, making sure that observed variation in sequenced data is the result of mutations, and not of sequencing errors. NGS can be used to sequence targeted regions identified in a genetic study, or entire genome including all coding genes (whole exome sequencing).

The variations in human genome can be a few nucleotide base changes (substitutions), insertions, and deletions of DNA, large genomic deletions of exons or whole genes and rearrangements such as inversions and translocations. All these anomalies are collectively termed ‘mutations’. Traditional methods of sequencing were only able to discover handfuls of mutations including small insertions and deletions. This led to the development of dedicated assays, to discover additional types of variations. Some of the examples includes fluorescence in situ hybridization (FISH) for conventional karyotyping, or comparative genomic hybridization (CGH) microarrays to detect sub-microscopic chromosomal copy number changes such as microdeletions.

With recent advancements in NGS technologies and better understanding of life at genomic level, various questions have been answered using whole genome sequencing. Areas of applications includes genome diversity, metagenomics, epigenetics, discovery of non-coding RNAs and protein-binding sites, and gene-expression profiling by RNA sequencing [22–26]. Apart from high-throughput whole genome sequencing, typical applications of NGS methods in microbiology and virology are discovery of new microorganisms and viruses by using metagenomic approaches, investigation of microbial communities in the environment and in human body for understanding healthy and disease conditions, analysis of viral genome variability within the host, detection of antiviral drug-resistance mutations in patients with human immunodeficiency virus (HIV) infection or viral hepatitis, etc.

In the context of Microbial Analysis, the term metagenomics designates the analysis of all of the nucleic acid present in a given sample. Without isolating and culturing individual microbial species, entire communities of microorganisms can be explored. NGS applications in metagenomic studies include the discovery of novel viruses from clinical samples in human and animal diseases, e.g. the new Ebola virus Bundiubugyo [27], identification of a viral etiology of disease outbreak in honeybees [28], and involvement of a new arenavirus in transplant-associated disease clusters [29]. Scope of applications also include characterization of the viral community in the environment [30, 31], in animals [32], and viral community in

humans [33–36]. Due to high replication capacity and low fidelity of the replication enzyme, high intra-host variability is shown by reverse transcriptase-dependent viruses (e.g. hepatitis B virus, human immunodeficiency virus) and RNA viruses (e.g. hepatitis C virus, influenza virus). Such a set of closely related genomes within a given host allows a viral population to swiftly adapt to dynamic environments and evolve resistance to vaccines and antiviral drugs [37]. Significant work using NGS has been done for the characterization of intra-host variability of influenza virus [38, 39], HCV, HIV and HBV.

Jian'an Jia et al. designed an approach to distinguish between 2 disease groups caused by Hepatitis B Virus – Chronic Hepatitis B (CHB) and Hepatocellular Carcinoma (HCC) [40]. NGS was used to sequence the pre-S region of a large number of CHB and HCC individuals. The attributes used were word pattern frequency vector of various lengths ranging from  $k = 2$  to  $k = 8$ . Maximum CV mean AUC of 0.93  $k = 5$ . The prediction accuracy was found to be much higher than prediction results using KNN classifiers.

To investigate HBV genotypes and predict HCC status, Xin Bai et al. used NGS to sequence the pre-S region of the HBV sequence of 94 HCC patients and 45 chronic HBV (CHB) infected individuals [41]. Word pattern frequencies among the sequence data of all individuals were calculated and compared using the Manhattan distance. The individuals were grouped using principal coordinate analysis (PCoA) and hierarchical clustering. Word pattern frequencies were also used to build prediction models for HCC status using both K-nearest neighbours (KNN) and support vector machine (SVM). In the independent data set of 46 HCC patients and 31 CHB individuals, a good AUC score of 0.77 was obtained using SVM.

Apart from applications viral disease diagnosis, a recent study demonstrates usefulness of a hybrid approach in early assessment of the risk by predicting the host of influenza viruses using the Support Vector Machine (SVM) classifier based on the word vector, representation and feature extraction method for biological sequences [42]. Accuracies for host prediction in avian, human & swine influenza viruses were 99.7%, 96.9% & 90.6%, respectively. Table 3 contains some examples of SVM application using NGS data to address problems in virology studies.

### **9.3 SVM Applications Based on Spectroscopy Data**

From array of several spectroscopic techniques, Raman spectroscopy and Infrared (IR) absorption spectroscopy have led to major breakthroughs in biological, pharmaceutical, and clinical research [43–45]. With use of visible-light laser beams, Raman spectroscopy can be used as a non-invasive characterization technique and achieve resolution same as fluorescence microscopy. The inelastic scattering of light photons by vibrating molecules in the samples is called as Raman scattering. Information about molecular vibrations produced due to change in frequencies of the photons are useful in diagnostic studies. Such change in frequencies are result of interactions of molecular bonds. Initial changes in almost all the types of diseases

**Table 3** Illustrative examples for SVM applications based on NGS approach

Sr. no.	Reference	Brief description of work	Attributes and attribute selection
1.	Next-generation sequencing revealed divergence in deletions of the preS region in the HBV genome between different HBV-related liver diseases [40]	Distinguish between 2 disease groups caused by Hepatitis B Virus – Chronic Hepatitis B (CHB) and Hepatocellular Carcinoma (HCC)	Nucleotide deletion % obtained from sequences. It is defined as $100 \times (\text{counts of reads with deletion in single nucleotide site}) / (\text{total number of reads including such a nucleotide site})$
2.	Deep sequencing of HBV pre-S region reveals high heterogeneity of HBV genotypes and associations of word pattern frequencies with HCC [41]	Investigate HBV genotypes and to predict HCC status using sequences of pre-S region of the HBV sequence of HCC and HBV patients.	Word pattern frequency vector of various lengths ranging from $k = 2$ to $k = 8$
3.	Predicting the host of influenza viruses based on the word vector [42]	Predict the host(human, avian & swine) of influenza viruses based on the word vector	200-dimension vectors of all proteins and DNA sequences generated using “word2vec” To vectorize protein, the sequence is separated into overlapping words of size 2–4. The word vector of all the words are summed up and averaged that results in 200-dimension vector for each protein. Same done for DNA

(including cancer and viral infections) occur at molecular level. Laboratory tests are inadequate in identifying such changes due to some limitations. Raman spectroscopy has the potential to monitor these changes at molecular level at early stage of the disease [46]. Information about abnormalities can be retrieved from the spectral differences between normal and diseased samples, which is used for the purpose of diagnosis. With diverse areas of applications, Spectroscopy is a promising clinical tool for the real-time diagnosis of diseases and assessment of living healthy and cancerous tissue, cells and their subcellular compounds and structures. It can also be used to track the mode of action of drugs on a molecular level.

Due to its high sensitivity and selectivity Raman spectroscopy requires only a small sample volume and minimal preparation efforts. The high resolution, ease of sample preparation, and very short data collection time required make the technology ideal for use in the study of viruses and virally infected cells. As the acquisition can be fast, processes in real time can be studied. In different conditions and environments, informative molecular details can be extracted since water environment can disturb these spectra to a slight extent. Therefore, this technique is ideal for studies like viral protein assembly, dynamics, interactions and structural alterations, compared to other available methods. The stereochemistry and structures of pro-

teins and nucleic acid components of viruses, can be determined using spectroscopy [47, 48]. The conformational changes that leads to viral procapsid and capsid assembly was identified using Raman spectroscopy [49, 50]. Raman spectroscopy is effective also in distinguishing between even the homogenous viruses, thereby increasing its possible role even further in diagnostic medicine.

Dengue fever, Yellow fever, Japanese encephalitis, Murray Valley encephalitis, tick-borne encephalitis and West Nile encephalitis are diseases attributed to flavivirus infection. Early detection is important to prevent these diseases from progressing into the severe or terminal stages. Non-structural protein 1 (NS1) is acknowledged as one of the biomarkers for flavivirus related diseases. Radzol AR et al. defined a model for PCA-SVM with MLP kernel for classification of flavivirus biomarker, NS1 molecule, from Surface Enhanced Raman Spectroscopic (SERS) spectra of saliva [51]. Best PCA-SVM (MLP) model defined in this study yielded accuracy of 96.9%.

Another example of life-threatening viral infection is Hepatitis B, that attacks the liver. In a study analysing hepatitis B virus (HBV) infection in human blood serum using Raman spectroscopy combined with pattern recognition technique, SVM model with two different kernels i.e. polynomial function and Gaussian radial basis function (RBF) were investigated for the classification of normal blood sera from HBV infected sera based on Raman spectral features [52]. Best performance achieved for polynomial kernel of order-2 with accuracy of 98% using fivefold cross-validation.

In case of chronic hepatitis C, liver biopsy has been the reference for staging the degree of fibrosis until the last decade. For obvious reasons, non-invasive tests e.g. blood tests measuring the markers that are either involved in the synthesis or degradation of extracellular matrix, has to be the preferred alternatives for assessment of hepatic fibrosis. However, the performance of these non-invasive methods is limited in differentiating between mild and moderate stages of fibrosis and in evaluating the effect of treatments on liver fibrosis process. Use of Fourier transform infrared (FTIR) spectroscopy applied to the serum in the assessment of hepatic fibrosis, was demonstrated by Scaglia et al. [53]. Infrared spectral characteristics exhibited by serum from patients, were employed in differentiation of chronic hepatitis C patients with extensive hepatic fibrosis from those without fibrosis and thus predicting the degree of hepatic fibrosis. With leave-one-out cross-validation, the accuracy achieved was 97.7%.

A similar study was performed for the classification of dengue suspected in human sera. SVM models built on the basis of three different kernel functions including Gaussian radial basis function (RBF), polynomial function and linear function were employed to classify the human blood sera based on features obtained from Raman Spectra [54]. With the tenfold cross validation method, best results were obtained for the polynomial kernel of order 1 with diagnostic accuracy of about 85%.

The applications are not limited to only medicinal diagnosis. Viruses could infect over hundreds of different species of plants, including crops of tobacco, tomato,

pepper, cucumber, etc. Viruses can survive outside the plant, and remain in a dormant state to infect growing crops. Once the plant is infected, no chemical cure is effective, and usually all the infected crops should be removed. For detecting seeds infestation caused by cucumber green mottle mosaic virus (CGMMV), near-infrared (NIR) hyperspectral imaging system was used to discriminate virus-infected seeds from healthy seeds with partial least square discriminant analysis (PLS-DA) and least square support vector machine (LS-SVM) [55]. The classification accuracy for virus-infected watermelon seeds were 83.3% with the best model.

Whereas Jiyu Peng et al. proposed an approach to discriminate TMV-infected tobacco based on laser-induced breakdown spectroscopy (LIBS) [56]. Two different kinds of tobacco samples (fresh leaves and dried leaf pellets) were collected for spectral acquisition, and partial least squared discrimination analysis (PLS-DA) was used to establish classification models. In prediction set, 94.4% and 94.7% accuracies obtained for observed emission lines of dried & fresh leaves. Compared to PLS-DA, SVM was proved to be efficient to eliminate influences of moisture content. Some other examples are listed in Table 4.

#### ***9.4 SVM Applications for Epitope Prediction***

An epitope is a specific target of a few AA residues on an antigen molecule that is recognized by B-cells or T-cells of the immune system [57, 58]. A B-cell epitope is the antigen portion that binds to B-cell Receptor (BCR) on B-cells, where BCR contains membrane-bound antibody. There are 2 types of B-cell epitopes based on their orientation. One is linear epitope that comprises of a continuous string of AA s. The second one consisting of most B-cell epitopes is conformational epitope which is made up of discontinuous AAs that comes close with protein folding [59, 60]. A T-cell epitope binds to the major histocompatibility complex (MHC) on surface of antigen-presenting cells (APCs) and MHC presents the antigen to the T-cell receptor (TCR) on T-cells [59]. The major histocompatibility complex (MHC) or human leukocyte antigen (HLA) is the gene family that helps the immune system to identify and destroy the foreign substance [61].

Vaccines have proven to be useful tools to control various viral diseases like influenza, smallpox, polio, hepatitis and rotavirus. The conventional methods of developing vaccines include attenuated or killed whole pathogen that improves immunity to a specific disease and involve only experimental methods of epitope identification. Vaccine development takes a long time with conventional methods because of the time consuming experimental screening of huge number of potential candidates [62]. The fact that only few AA residues are detected by B- and T-cells instead of whole pathogen is leveraged for vaccine development, understanding disease etiology, disease diagnosis and immune monitoring [58, 59]. Moreover, with advances in next-generation sequencing methods, proteomics, and transcriptomics

**Table 4** Illustrative examples for SVM applications using spectroscopy

Sr. no.	Reference	Brief description of work	Attributes and attribute selection
1.	Detection of cucumber green mottle mosaic virus-infected watermelon seeds using a near-infrared (NIR) hyperspectral imaging system: Application to seeds of the “Sambok Honey” Cultivar [55]	Classification of infected and healthy watermelon seeds using a near-infrared (NIR) hyperspectral imaging system. Hyperspectral imaging data 51 healthy & 45 infected samples were used.	Near infrared spectrum
2.	PCA criterion for SVM (MLP) classifier for flavivirus biomarker from salivary SERS spectra at febrile stage [51]	Classification of flavivirus biomarker, NS1 molecule, from Surface Enhanced Raman Spectroscopic (SERS) spectra of saliva. SERS spectra of 64 NS1 adulterated dataset and 64 control dataset were used.	Spectral data with 1801 features per spot per sample.
3.	Analysis of hepatitis B virus infection in blood sera using Raman spectroscopy and machine learning [52]	Analysis of hepatitis B virus (HBV) infection in human blood serum using Raman spectroscopy. Serum samples of 119 confirmed HBV infected patients and 84 healthy volunteers were used.	Raman spectral features
4.	Noninvasive assessment of hepatic fibrosis in patients with chronic hepatitis C using serum Fourier transform infrared spectroscopy [53]	Non-invasive differentiation of chronic hepatitis C (CHC) patients with extensive hepatic fibrosis from those without fibrosis using Fourier transform infrared (FTIR) spectroscopy of serum. Serum samples of 12 patients with no hepatic fibrosis and 11 patients with extensive fibrosis were used.	Fourier transform infrared spectral profiles.
5.	Fast detection of tobacco mosaic virus infected tobacco using laser induced breakdown spectroscopy [56]	Detect TMV-infected tobacco based on laser-induced breakdown spectroscopy (LIBS).	Full spectrum and observed emission lines of laser-induced breakdown spectroscopy (LIBS) for fresh & dried leaves.
6.	Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM) [54]	Classification of dengue suspected human blood sera; use of Raman spectroscopy combined by deciphering spectral differences between dengue positive and normal sera. Raman spectra of 31 were dengue positive and 53 were negative were used.	Features obtained from Raman Spectra

as well as ever increasing immune system data and databases, epitopes can be identified in few years. Once the epitopes are predicted using computational methods, the peptides can be experimentally tested for its binding affinity and ability to elicit desired immune response. Immunoinformatics involves the development of bioinformatics tools that analyses data to predict B- and T-cell epitopes which can stimulate immune response. In-silico prediction methods of epitope prediction can be beneficial to decrease the number of potential epitopes for experimental confirmation, develop epitope-based vaccines for hypervariable viruses and develop chimeric vaccines [59, 62]. Epitope based-vaccines can be safer and less expensive than conventional methods [62].

Predicted epitopes should take into account the desirable features of epitopes such as they should be conserved in different parts of viral lifecycle, their binding affinity and efficacy, they should bind to more than one allele of immune system molecules and most of them are proteins [59, 62]. Most epitope prediction methods are based on proteins and their different descriptors including physicochemical properties related profiles of proteins, evolutionary data, sequence motifs and quantitative matrices (QM) [58, 59]. SVM has been one of the most popular methods used for both B-cell & T-cell epitope prediction.

#### 9.4.1 T-Cell Epitope Prediction

T-cell epitopes are processed within a cell, linked with MHC & presented on T-cell surface to be recognized by T-cell receptor. Each of these steps decide the immunogenicity of T-cell epitopes. However, most of the T-cell epitopes focus on the step where a peptide is linked with MHC-I & MHC-II [59]. MHC-I binds to peptides of length 9–11 AA s and its pockets prefers peptides with certain physicochemical properties. Hence, peptide-MHC-I binding prediction methods work on peptide sequences of 9 AA residues. On the other hand, MHC-II binds to longer peptides but the prediction methods focus on peptide part that binds to the MHC-II groove. Large number of databases like IEDB, EPIMHC and AntiJen, store epitopes verified through experimental approaches [59]. These have served as rich sources of positive examples for several prediction methods.

Different computational methods/models have been used to predict epitopes like use of Sequence Motif, motif matrix, quantitative affinity matrices (QAM) etc. . However, machine learning (ML) methods have proven to be the most robust method for prediction [63]. With high dimensionality of the data and the limited observations, SVM comes as a better method. In a study, 36 stimulatory peptides and 167 non-stimulatory peptides were gathered, and physical properties of 20 AA s were used to develop models from Artificial Neural Network, Decision Tree & SVM. SVM proved to outperform prediction of stimulatory peptides with maximum sensitivity of 0.76 [64].

MHC2Pred is one of the freely available tools based on SVM to predict MHC-II binding peptides [65]. To develop a model for MHC2Pred, binding & non-binding peptides, based on IC50, were collected from MHCBN and JenPep database.

Peptides with less than 9 AA residues were discarded and rest of the peptides were looked for 9 AA s that would bind the MHC-II groove using Matrix Optimization Techniques (MOT) package. A vector of length 20 was created for each AA in 9-mer peptide where binders were given +1 and non-binders a -1. Each peptide was thus represented by 180 ( $9 \times 20$ ) length vectors. This data was used to develop SVM model which was later validated using fivefold cross validation and got an overall accuracy of method is >78% [65].

SVMHC is another tool for prediction of both MHC class I and class II binding peptides [66]. For MHC-I prediction model, peptides of length 8–10 were represented by a binary sparse encoding. For MHC-II peptide binding prediction, matrices by Sturniolo et al. [67] were used. These matrices represent HLA-DR peptide binding specificity where HLA-DR is an MHC-II cell surface receptor [67] (see sr. no. 1 of Table 6).

Predicting immunogenicity of epitopes can help in vaccine design and POPISK is a tool that predicts reactivity of T-cells to peptides and identify positions that are recognized by TCR [68]. POPISK uses SVM model with a weighted degree string kernel (see sr. no. 2 of Table 6).

#### 9.4.2 B-Cell Epitope Prediction

B-cell epitopes can be predicted based on physicochemical properties like hydrophilicity, flexibility, polarity, and exposed surface as well as secondary & 3D structures [62]. There are 566 AA indices that represents physicochemical properties of AA s listed in AAindex [69].

Linear epitopes can be predicted using antigen sequences by calculating AA propensity scales based on physicochemical properties. AA Propensities (AAP) calculation considers an overlapping window of length k AA s in a protein sequence and for each window, average propensity value of AA s is calculated, where propensity value can be hydrophilicity, accessibility, flexibility, polarity, antigenicity, beta-turn, surface exposed scale, etc. The average value is assigned to the AA in middle of the window. AA s residues that passes the threshold are considered as potential epitopes. A combination of different propensity values can be used with specific weights [70].

Due to poor performance of AA propensity scales, Machine learning (ML) methods were later adopted to distinguish B-cell epitopes from non-epitopes. BCPREDS and SVMtrip [71] are epitope prediction tools based on Support Vector Machine (SVM) [59]. More information on SVMtrip is provided in sr. no. 3 of Table 6.

Conformational B-cell epitopes can be predicted using features related to the structure of the proteins. One of the studies have used combination of physicochemical features, evolutionary PSSM features and structural features as protrusion index (PI), accessible surface area (ASA), relative accessible surface area (RSA) and B-factor [72] (see sr. no. 4 of Table 6). Physicochemical properties of AA s were derived from AAIndex. PSSM represents the attributes extracted from repeated multiple sequence alignment of sequences that can be generated using PSI-BLAST

with specific number of iterations. It is a scoring matrix where each position in the multiple sequence alignment is given an AA substitution scores. PSSM is used to incorporate evolutionary information of a peptide [73–75]. Another study by Ansari et al. [76] on conformational B-cell epitope uses 3 types of features namely binary profile of pattern (BPP), physicochemical profile of patterns (PPP) and composition profile of patterns (CPP) (see sr. no. 5 of Table 6). In this study, patterns of different lengths were created from the sequences. Then for each pattern 3 feature vectors were created, (1) BPP, a vector of length 21 based on binary number for occurrence and non-occurrence of AA, (2) PPP, a vector of length 5 based on 5 physicochemical properties named Hydrophobicity, Flexibility, Polarity\_Grantham, Polarity\_Ponnuswami, Antigenicity and (3) CPP based on composition of patterns. CBTope server uses this method for predicting B-cell epitopes [76].

Listed in Table 5 are some freely available T-cell & B-cell epitope prediction web servers based on SVM.

Information on some more SVM based epitope prediction studies have been provided in Table 6.

## 9.5 Applications of SVM Involving Protein-Protein Interaction in Virology

Proteins are the workhorses of a cell that carry out majority of the functions in a cell. Eighty percent of proteins are not functional in isolated forms but they operate in complexes by interacting with other molecules [77, 78]. Protein-protein interaction (PPI) is the physical & functional interactions of proteins that controls wide range of molecular processes in a cell, like signal transduction, cell-cell communication, transcription, replications etc. [79]. PPIs can be responsible for altering kinetic properties of enzyme, modifying proteins activity, changing specificity of protein binding, constructing new binding sites and regulatory function. Alteration or malfunction of these interactions can lead to diseases [79]. The collection of all the protein-protein interaction of cell or an organism is called interactome. The study of PPIs can help in predicting a biological process involving protein of unknown function, fasten the pace of understanding functional pathways or to know biochemistry

**Table 5** List of freely available epitope prediction servers

Sr. no.	Server	Reference link	Epitope predicted
1.	MHC2Pred	<a href="http://www.imtech.res.in/raghava/mhc2pred/">http://www.imtech.res.in/raghava/mhc2pred/</a>	T-cell epitope
2.	SVMHC	<a href="http://abi.inf.uni-tuebingen.de/Services/SVMHC">http://abi.inf.uni-tuebingen.de/Services/SVMHC</a>	T-cell epitope
3.	SVRMHC	<a href="http://svrmhc.biolead.org/">http://svrmhc.biolead.org/</a>	T-cell epitope
4.	BCPRED	<a href="http://ailab.ist.psu.edu/bcpred/">http://ailab.ist.psu.edu/bcpred/</a>	B-cell epitope
5.	SVMTriP	<a href="http://sysbio.unl.edu/SVMTriP/prediction.php">http://sysbio.unl.edu/SVMTriP/prediction.php</a>	B-cell epitope
6.	EPSVR	<a href="http://sysbio.unl.edu/EPSVR/">http://sysbio.unl.edu/EPSVR/</a>	B-cell epitope
7.	CBTOPE	<a href="http://crdd.osdd.net/raghava/cbtope/">http://crdd.osdd.net/raghava/cbtope/</a>	B-cell epitope

**Table 6** Illustrative examples of epitope prediction based on SVM

Sr. no	Reference	Brief description of work	Attributes and attribute selection
1.	SVMHC: A server for prediction of MHC-binding peptides [66]	<b>Purpose:</b> Identification of MHC-I and MHC-II binding peptides <b>Dataset:</b> MHC-binding peptides of different lengths were extracted from the MHCPEP and SYFPEITHI databases.	For MHC-I – Binary sparse encoding of 8–10 k-mer length of AA s For MHC-II – Matrices representing HLA-DR peptide binding specificity
2.	POPISK: T-cell reactivity prediction using support vector machines and string kernels [68]	<b>Purpose:</b> Predict immunogenicity of peptides by predicting T-cell reactivity i.e. if a peptide is immunogenic or non-immunogenic using SVM with a weighted degree string kernel. <b>Dataset:</b> Extracted peptide binders of length 9 along with their associated human MHC class I alleles and immunogenicity from three databases, MHCPEP, SYFPEITHI and IEDB. Negatively annotated peptides were used as non-immunogenic peptides Final dataset – 558 immunogenic and 527 non-immunogenic peptides	Matched sub-sequences of length p at a position in 2 sequences
3.	SVMTriP A Method to Predict Antigenic Epitopes Using Support Vector Machine [71]	<b>Purpose:</b> Predict linear B-cell epitopes using SVM with RBF Kernel <b>Dataset:</b> Dataset constructed by extracting non-redundant linear B-cell epitopes (10AA, 12AA, 14AA, 16AA, 18AA, and 20AA) from IEDB. For negative dataset, non-epitope part of corresponding antigen used. Final dataset: 4925 non-redundant epitope sequences each for positive and negative dataset.	Tripeptide similarity using Blosum62 matrix and propensity scores
4.	Positive-unlabeled learning for the prediction of conformational B-cell epitopes [73]	<b>Purpose:</b> PUPre (Positive-Unlabeled Prediction) method to – (1) identify non-epitope residues using weighted SVM and (2) model to distinguish epitope and non-epitope residues <b>Dataset:</b> 2123 residues labelled as epitopes and 16,615 unlabeled residues by processing data from PDB	Feature vector of 239 features including 205 physico-chemical features collected from AAIndex 21 evolutionary PSSM features 13 structural features. <b>Attribute selection:</b> Wilcoxon rank-sum test was applied to select informative features and resulted in 89 selected features

(continued)

**Table 6** (continued)

Sr. no	Reference	Brief description of work	Attributes and attribute selection
5.	Identification of conformational B-cell epitopes in an antigen from its primary sequence [76]	<b>Purpose:</b> Use SVM with RBF Kernel to identify conformational B-cell epitopes <b>Dataset:</b> 187 antigenic protein chains having 2261 amino acid residues that were antibody interacting and 107,414 amino acid residues as non-antibody interacting	Binary profile of patterns (BPP), Physico-chemical profile of patterns (PPP), composition profile of patterns (CPP) Explanation of features in Sect. 9.4.2
6.	SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction [100]	<b>Purpose:</b> Identification of linear B-cell epitopes using SVM string kernel prediction model <b>Dataset:</b> Linear B-cell epitopes of lengths 12- to 20-mers extracted from EL-Manzalawy dataset [101]	Sequences encoded in bi-profile manner where they have attributes from 2 pools – positive position-specific and negative position-specific profiles <b>Feature extraction:</b> Bayes Feature Extraction (BFE)
7.	Application of support vector machines for T-cell epitopes prediction [64]	<b>Purpose:</b> T-cell epitope prediction with an MHC I restricted T-cell clone. <b>Dataset:</b> 36 stimulatory peptides and 167 non-stimulatory peptides which were further divided into positive and negative set by random sampling	188 physical properties of 20 AA s <b>Attribute selection:</b> Ten factors extraction from 188 physical properties of 20 AA s

of a cell [77, 79, 80]. Knowledge of specific PPI can also help in identification of drug targets [79].

PPI data can be mapped to large scale networks where nodes represent proteins and edges represent their physical or functional interactions. These networks are known as PPI networks (PIN) [77, 79]. PPI networks can be used to extract various information like functionality of a protein based on its placement in the network as the closely linked proteins can have similar biological activity. PPI can also be used to decipher which complex a protein belongs to and the diseases related to a protein [79]. The knowledge that is encapsulated in the PPI can help improve the biological and biomedical applications [77].

Virus-host proteins interactions are key to viral infection and subsequent pathogenesis. Many PPIs are involved between virus and host during a viral infection where the virus proteins take over the host transcriptional machinery [78]. It has been believed that viral proteins bind to the host protein that are highly connected [81]. Endogenous interface, with respect to virus-host systems, are responsible for interactions in their own system i.e. host-host PPI and virus-virus PPI. On the other hand, exogenous interfaces are responsible of virus-host interactions. Both virus and host compete for endogenous and exogenous interfaces [81]. Mutations at protein interfaces can reduce or increase their binding affinities by changing protein electrostatics and structural properties. Virus and host proteins change their surface

resides through mutations as an evolutionary result to compete for binding partner. However, host tends to be less variable than viruses. Viruses diversify through various modes of molecular evolution, including conservation, horizontal gene transfer, gene duplication and molecular mimicry [81]. Viral proteins constantly inhibit host-host interactions and therefore, blocking such interactions between virus & host can aid in biomedical applications by identification of drug targets and developing anti-viral therapies [81]. For e.g. a drug, Maraviroc, binds the cellular co-receptor CCR5, a receptor on white blood cells involved in immune system, preventing it from interacting with GP120 of HIV1 which is essential for entry of HIV-1 in host [82]. As viruses pose a global threat, understanding of virus & human PPIs can help in development of vaccines for treatment.

Comprehensive PPI networks have been generated using experimental methods. These experimental methods employ different techniques like tandem affinity purification, affinity chromatography, coimmunoprecipitation, protein arrays, protein fragment complementation, phage display, X-ray crystallography, and NMR spectroscopy [79]. However, due to the huge PPI data and the time consuming experimental methods, computational methods are increasingly becoming popular to analyse the PPI networks and find out the functions of unexplored proteins. Computational methods of PPI detection are based on sequences, structure of molecules, gene fusion, phylogenetic tree and gene expression [79].

Detection of virus-host interactions using machine learning methods have proved to be very useful. Several SVM models have been developed for the same purpose; known PPIs as positive set, are used to train the models to predict whether two proteins interact or not. Positive set data can be extracted from experimental data available in the databases. Selecting negative dataset is complicated. Negatome, a database of negative interactions developed using text mining, can be used to gather negative data set [83, 84].

Emamjomeh et al. [85], developed SVM model to predict PPI interactions between human and hepatitis C virus (HCV) [D32]. In this study, SVM was combined with other learning methods like random forest (RF), Naïve Bayes (NB) and multilayer perceptron (MLP). Feature vectors were generated for HCV & human proteins which included six different AA composition (ACC), pseudo AA composition (PAC), PSSM as evolutionary information feature, network centrality measures, tissue information and post-translational modification (PTM) information [85]. AA composition is the simplest descriptor used to represent a protein sequence. However, with this descriptor the sequence order of AA s is lost and hence, pseudo AA is used which involves AA composition as well as sequence order-based features [5] (see sr. no. 1 of Table 7).

Cui et al. [86] developed an SVM model for prediction of virus-host PPI for 2 viruses, human papillomaviruses (HPV) and hepatitis C virus (HCV). This SVM model is based on relative frequency of AA triplets (RFAT) between virus & host AA sequences and GO annotations of protein. RFAT generates fixed length for variable length proteins and enables models to achieve a better accuracy. In this study, a vector based on AA triplets & biochemical similarity is generated. Based on biochemical properties of AA residues, 6 categories are defined as {IVLM}, {FYW}, {HKR}, {DE}, {QNTP}, and {ACGS}. Using this classification of AA s,

**Table 7** Illustrative examples Protein-Protein interaction studies based on SVM

Sr. no	Reference	Brief description of work	Attributes and attribute selection
1.	Predicting protein–protein interactions between human and hepatitis C virus via an ensemble learning method [85]	<b>Purpose:</b> Predicting PPI between human and hepatitis C virus using SVM combined with RF, NB and MLP. <b>Dataset:</b> 657 positive interactions from human-HCV PPI from IntAct database and 2910 negative interactions	AA composition, pseudo AA composition, PSSM, network centrality feature, tissue information feature, 31 PTM types
2.	Prediction of protein-protein interactions between viruses and human by an SVM model [86]	<b>Purpose:</b> Prediction of protein- protein interactions using SVM binary classifier. <b>Dataset:</b> Training dataset had 500 positive and negative interactions. Test set had 195positive and negative interactions. Positive dataset was extracted from the infection mapping project (I-MAP) whereas negative from HPRD by random selection of human proteins	Feature vector of relative frequency of 216 AA triplets Details of attribute Sect. 9.5
3.	An improved method of predicting interactions between virus [89]	<b>Purpose:</b> An improved method of predicting interactions between virus and proteins including human papillomaviruses (HPV) and hepatitis C virus (HCV), using SVM with RBF kernel.	Features used: Relative frequency of AA triplets (RFAT), the frequency difference of AA triplets (FDAT) between virus and host proteins, and AA composition (AC). RFAT feature generation–clustered 20 AA s into 4 groups based on chemical properties of side chain of the AA s yields 64 AA triplets
4.	A generalized approach to predicting PPI [78]	<b>Purpose:</b> Prediction of PPI between virus and host using SVM with RBF kernel. Additionally, a generic model to predict PPI of any virus & host <b>Dataset:</b> Multiple training and test datasets used	Features used – RFAT, FDAT, AC, normalized frequency of each AA group, transition & distribution. RFAT feature generation – 20 AAs into 7 groups based on dipoles & volumes of side chains of AA s yielding 343 possible AA triplets. A vector of $343 + 343 = 686$ was generated for virus-host pair

(continued)

**Table 7** (continued)

Sr. no	Reference	Brief description of work	Attributes and attribute selection
5.	Supervised learning and prediction of physical interactions between human and HIV proteins [88]	<b>Purpose:</b> Prediction of human-HIV PPI using SVM with a linear kernel. <b>Dataset:</b> 1028 human–HIV PPIs from four public databases, Biomolecular Interaction Network Database, the Database of Interacting Proteins, IntAct, and Reactome Negative dataset was generated by randomly pairing human and HIV proteins.	Four-mers sequence, protein domains responsible for interactions and PPI network information. Four-mers sequence and 7 categories of AA, were used to generate feature vector of 4802 ( $7^4 * 2$ )

there are  $6 \times 6 \times 6 = 216$  possible AA triplets [86]. The protein sequence is converted into AA triplets and the vector of 216 length is created that contains the frequency of each category in sequences of variable length. LIBSVM [87] was used to generate model with the radial basis function (RBF) as a kernel function. For dataset, HCV & human interaction data was extracted from the infection mapping project (I-MAP) whereas for HPV, data was extracted from NCBI Bio Systems Database. For HCV accuracy of 85.1 was achieved whereas for HPV it was 87.5 [86] (see sr. no. 2 of Table 7).

RFAT has been used in many studies with different combinations of categories and k-mer. In a study of HIV and human PPI [88], four-mer sequences were used instead of triplet. With 7 categories and 4-mer sequences, RFAT vector of 4802 ( $7^4 * 2$ ) length was generated (see sr. no. 5 of Table 7).

Kim et al. [89] used 4 categories based on chemical properties of side chain of the AA s making 64 AA triplets combination (see sr. no. 3 of Table 7).

In another study of PPI by Zhou et al. [78] and Shen et al. [90], a similar feature vector of triplets is produced but 7 categories of AA residues are used instead of 6 and these categories are based on dipole and volumes of the side chains of AA s. With 7 categories 343 ( $7 \times 7 \times 7$ ) AA triplets are possible. RFAT feature vector had 686 elements i.e. 343 for host and 343 for virus. Zhou et al. [78] uses more features as frequency difference of AA triplets (FDAT) between virus and host proteins, AA composition (AC) in each pair of host and virus proteins, normalized frequency of each AA group, transition and distribution of AA groups. As a result of these 6 features, a feature vector of length 1175 was created. Again, LIBSVM [87] with RBF was used to develop model. Best performance was obtained with combination of all these 6 features with accuracy of 85.64% (see sr. no. 4 of Table 7).

Most of the prediction methods are specific to a virus-host combination. However, there are SVM based methods that are generic enough to predict PPIs of virus and host that were not used for training set. The approach by Zhou et al. [78] is one of such methods i.e. it does not require model for each host-virus pair. Another method called DeNovo, is a generic method that can predict novel PPIs. This method is based on SVM that trains on different virus-host PPIs [91].

Table 7 shows some studies on SVM model that are used for protein-protein interaction of virus-host.

## 10 Miscellaneous Examples

Apart from above examples, there are some noticeable studies employing other approaches to address problems in virology. Microarray is a method that uses microscopic chip where each spot-on chip has a DNA/cDNA sequence attached. These sequences bind to the complementary unknown sequences & thereby detects gene expressions of thousands of genes. In Virology, Microarray is used to screen viruses for which genomes are available in GenBank by looking at the conserved viral sequences. Microarray gene expression profiles are also used to detect the immune response that can further help in classifying disease caused by viruses, that is conventionally done using quantitative real time PCR (qPCR). SVM can be used to detect immune response by using microarray gene expression data. Due to big size of microarray data, important features are extracted using feature selection methods.

In a study [92], the authors have reported that DNA microarray technology can be used as a high-throughput method to analyse polymorphisms within a short region of the FMDV genome encoding relevant functions in antigenicity and receptor recognition. Their SVM based methodology classifies the samples based on their hybridization signal. This prediction methodology has wide ranging applications to fine genotyping including studies of heterogeneous viral populations, genetic changes in virus, bacteria, and genes of rapidly evolving cells, such as tumor cells.

Predicting the hosts of newly discovered viruses is important for pandemic surveillance of infectious diseases. Li and SUN [93] investigated the use of alignment-based and alignment-free methods and support vector machine using mononucleotide frequency and dinucleotide bias to predict the hosts of viruses, and applied these approaches to three datasets: rabies virus, coronavirus, and influenza A virus [93] also showed that SVM predicts the hosts of viruses with a high degree of accuracy.

The phosphorylation of virus proteins by host kinases is linked to viral replication leading to an inhibition of normal host-cell functions. Unravelling of phosphorylation mechanisms in virus proteins can aid in drug design and treatment. In this study [94] a two-layered Support Vector Machines (SVMs) was applied to train a predictive model for identification of phosphorylation sites.

Replication of their DNA genomes is a central step in the reproduction of many viruses. [V4] proposes a novel least-squares support vector machines (LS-SVMs) model with viruses of herpes family along with data sets involving a collection of caudoviruses coming from three viral families under the order of caudovirales. The LS-SVM approach provides superior performance as compared to those given by the previous methods. Ensembled with previously proposed methods, the LS-SVM approach further improves the prediction accuracy for the herpesvirus replication origins. Recursive feature elimination was used to extract the most informative attri-

butes and provides important domain knowledge in terms of the most significant features of the data sets [95] further conclude LS-SVMs can potentially be a very reliable and robust tool for viral replication origin prediction.

## 11 Web Server

SVM has been used in a variety of studies on viruses across different data types. Some of the tools mentioned in these studies are available as standalone tools whereas others are used in the backend of freely available web-servers. Web servers are user friendly and more intuitive making it easy for user to input data and analyse the output. Table 8 shows some of the web servers based on SVM models that are used in virology.

**Table 8** Examples of SVM based web servers SVM for Virology Studies

Sr. no	Reference	Brief description of work	Attributes and attribute selection
1.	A genotypic method for determining HIV-2 coreceptor usage enables epidemiological studies and clinical decision support [99]	<p><b>Purpose:</b> Geno2pheno is a web service to ensure that the virus can use only the CCR5 coreceptor (R5) and cannot evade the drug by using the CXCR4 coreceptor (X4-capable) using V3 loop of the HIV-2 glycoprotein</p> <p><b>Link:</b> <a href="https://www.geno2pheno.org/">https://www.geno2pheno.org/</a></p> <p><b>Dataset:</b> To build model, 126 pairs of HIV-2 amino-acid sequences and phenotypic coreceptor usage as R5 or X4-capable</p>	V3 loop region of the HIV-2 glycoprotein
2.	AVCpred: An integrated web server for prediction and design of antiviral compounds [102]	<p><b>Purpose:</b> AVCpred is a web server for prediction of antiviral compounds (AVC) for HIV, HCV, HBV, HHV &amp; 26 other viruses with QSAR-based model</p> <p><b>Link:</b> <a href="http://crdd.osdd.net/servers/avcpred">http://crdd.osdd.net/servers/avcpred</a></p> <p><b>Dataset:</b> Antiviral compounds extracted from ChEMBL bioactivity database – 389 compounds for HIV, 467 for HCV, 124 for HHV, 112 for HBV, and 1391 for other 26 viruses</p>	18,000 chemical descriptors (1D, 2D, and 3D) using PaDEL <b>Attribute selection:</b> Filter named ‘RemoveUseless’ followed by ClassifierSubsetEval (attribute evaluator) with BestFirst (search method) module available in Weka package

(continued)

**Table 7** (continued)

Sr. no	Reference	Brief description of work	Attributes and attribute selection
3.	Prediction of linear B-cell epitopes of hepatitis C virus for vaccine development [75]	List of web servers based on SVM for B-cell epitope prediction: BCPred: <a href="http://ailab.ist.psu.edu/bcpred/">http://ailab.ist.psu.edu/bcpred/</a> SVMTriP: <a href="http://sysbio.unl.edu/SVMTriP/prediction.php">http://sysbio.unl.edu/SVMTriP/prediction.php</a> Bcell-HCV: <a href="http://e045.life.nctu.edu.tw/BcellHCV">http://e045.life.nctu.edu.tw/BcellHCV</a>	Features used for prediction: BCPred: AAP propensity SVMTriP: Tri-peptide Bcell-HCV: Physicochemical properties
4.	SVMHC: a server for prediction of MHC-binding peptides [66]	<b>Purpose:</b> Identification of MHC-binding peptides <b>Link:</b> <a href="http://abi.inf.uni-tuebingen.de/Services/SVMHC">http://abi.inf.uni-tuebingen.de/Services/SVMHC</a> <b>Dataset:</b> MHC-binding peptides extracted from the MHCPEP and SYFPEITHI databases of varying length.	Binary sparse encoding of 8–10 k-mer length of AAs

## 12 Concluding Remarks

In this review, we illustrated the use of Support Vector Machines as a tool for building learning models in viral biology. SVM plays a vital role in building Quantitative structure activity relationship models. The robustness and accuracy of SVM models based rigorously on statistical learning theory has paved the way for quicker, faster and reliable methods of identification of potent molecules in drug design. SVM models have also enabled development of tools for rational design of novel vaccines. Recent advances in NGS technology could also be easily incorporated with SVM for building models with increased performance. We have also listed large number of case studies and examples in different areas of viral biology where SVM has been deployed with productive results

## References

1. Solomatine DP. Data-driven modelling: paradigm, methods, experiences. In: Proceedings of the 5th international conference on hydroinformatics; 2002 July 1. p. 1–5.
2. Mika S, Schölkopf B, Smola AJ, Müller KR, Scholz M, Rätsch G. Kernel PCA and de-noising in feature spaces. In: Advances in neural information processing systems; 1999. p. 536–542.
3. Vapnik VN. An overview of statistical learning theory. IEEE Trans Neural Netw. 1999;10(5):988–99.

4. Cristianini N, Shawe-Taylor J. An introduction to support vector machines and other Kernel-based learning methods. Cambridge: Cambridge university press; 2000.
5. Chou KC. Pseudo amino acid composition and its applications in bioinformatics, proteomics and system biology. *Curr Proteomics*. 2009;6(4):262–74.
6. Saes Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics*. 2007;23(19):2507–17.
7. Ma S, Huang J. Penalized feature selection and classification in bioinformatics. *Brief Bioinform*. 2008;9(5):392–403.
8. Srivastava A, Ghosh S, Anantharaman N, Jayaraman VK. Hybrid biogeography based simultaneous feature selection and MHC class I peptide binding prediction using support vector machines and random forests. *J Immunol Methods*. 2013;387(1–2):284–92.
9. Weston J, Watkins C. Support vector machines for multi-class pattern recognition. In: Esann 1999 April 21, vol 99. p. 219–224.
10. Suykens JA, Vandewalle J. Least squares support vector machine classifiers. *Neural Process Lett*. 1999;9(3):293–300.
11. Tax DM, Duin RP. Support vector domain description. *Pattern Recogn Lett*. 1999;20(11–13):1191–9.
12. Schölkopf B, Smola AJ, Williamson RC, Bartlett PL. New support vector algorithms. *Neural computation*. 2000 May 1;12(5):1207–45.
13. Dudek AZ, Arodz T, Gálvez J. Computational methods in developing quantitative structure-activity relationships (QSAR): a review. *Comb Chem High Throughput Screen*. 2006;9(3):213–28.
14. <http://www.unaids.org/en/resources/campaigns/HowAIDSchangedeverything/factsheet>.
15. Qureshi A, Rajput A, Kaur G, Kumar M. HIVprot: an integrated web based platform for prediction and design of HIV proteins inhibitors. *J Chem*. 2018;10(1):12.
16. <http://bioinfo.imtech.res.in/manojk/hivproti>.
17. Hdoufane I, Bjji I, Soliman M, Tadjer A, Villemain D, Bogdanov J, Cherqaoui D. In silico SAR studies of HIV-1 inhibitors. *Pharmaceuticals*. 2018;11(3):69.
18. Mei J, Zhao J. Prediction of HIV-1 and HIV-2 proteins by using Chou's pseudo amino acid compositions and different classifiers. *Sci Rep*. 2018;8(1):2359.
19. Richardson AM, Lidbury BA. Enhancement of hepatitis virus immunoassay outcome predictions in imbalanced routine pathology data by data balancing and feature selection before the application of support vector machines. *BMC Med Inform Decis Mak*. 2017;17(1):121.
20. Qasim MK, Algamal ZY, Ali HM. A binary QSAR model for classifying neuraminidase inhibitors of influenza A viruses (H1N1) using the combined minimum redundancy maximum relevancy criterion with the sparse support vector machine. *SAR QSAR Environ Res*. 2018;29(7):517–27.
21. Huang CC, Chang CC, Chen CW, Ho SY, Chang HP, Chu YW. PClass: protein quaternary structure classification by using bootstrapping strategy as model selection. *Genes*. 2018;9(2):91.
22. Metzker ML. Sequencing technologies—the next generation. *Nat Rev Genet*. 2010;11(1):31.
23. Ansorge WJ. Next-generation DNA sequencing techniques. *New Biotechnol*. 2009;25(4):195–203.
24. Medini D, Serruto D, Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R. Microbiology in the post-genomic era. *Nat Rev Microbiol*. 2008;6(6):419.
25. Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol*. 2008;26(10):1135.
26. MacLean D, Jones JD, Studholme DJ. Application of 'next-generation' sequencing technologies to microbial genetics. *Nat Rev Microbiol*. 2009;7(4):287.
27. Cox-Foster DL, Conlan S, Holmes EC, Palacios G, Evans JD, Moran NA, Quan PL, Briese T, Hornig M, Geiser DM, Martinson V. A metagenomic survey of microbes in honey bee colony collapse disorder. *Science*. 2007;318(5848):283–7.
28. Palacios G, Druce J, Du L, Tran T, Birch C, Briese T, Conlan S, Quan PL, Hui J, Marshall J, Simons JF. A new arenavirus in a cluster of fatal transplant-associated diseases. *N Engl J Med*. 2008;358(10):991–8.

29. Towner JS, Sealy TK, Khristova ML, Albariño CG, Conlan S, Reeder SA, Quan PL, Lipkin WI, Downing R, Tappero JW, Okware S. Newly discovered ebola virus associated with hemorrhagic fever outbreak in Uganda. *PLoS Pathog.* 2008;4(11):e1000212.
30. Wong K, Fong TT, Bibby K, Molina M. Application of enteric viruses for fecal pollution source tracking in environmental waters. *Environ Int.* 2012;45:151–64.
31. Bibby K, Viau E, Peccia J. Viral metagenome analysis to guide human pathogen monitoring in environmental samples. *Lett Appl Microbiol.* 2011;52(4):386–92.
32. Ge X, Li Y, Yang X, Zhang H, Zhou P, Zhang Y, Shi Z. Metagenomic analysis of viruses from bat fecal samples reveals many novel viruses in insectivorous bats in China. *J Virol.* 2012;86(8):4620–30.
33. Nakamura S, Yang CS, Sakon N, Ueda M, Tougan T, Yamashita A, Goto N, Takahashi K, Yasunaga T, Ikuta K, Mizutani T. Direct metagenomic detection of viral pathogens in nasal and fecal specimens using an unbiased high-throughput sequencing approach. *PLoS One.* 2009;4(1):e4219.
34. de Vries M, Deijs M, Canuti M, van Schaik BD, Faria NR, van de Garde MD, Jachimowski LC, Jebbink MF, Jakobs M, Luyf AC, Coenjaerts FE. A sensitive assay for virus discovery in respiratory clinical samples. *PLoS One.* 2011;6(1):e16118.
35. Yozwiak NL, Skewes-Cox P, Stenglein MD, Balmaseda A, Harris E, DeRisi JL. Virus identification in unknown tropical febrile illness cases using deep sequencing. *PLoS Negl Trop Dis.* 2012;6(2):e1485.
36. Schowalter RM, Pastrana DV, Pumphrey KA, Moyer AL, Buck CB. Merkel cell polyomavirus and two previously unknown polyomaviruses are chronically shed from human skin. *Cell Host Microbe.* 2010;7(6):509–15.
37. Woo PJ, Reifman J. A quantitative quasispecies theory-based model of virus escape mutation under immune selection. *Proc Natl Acad Sci.* 2012;109(32):12980–5.
38. Bartolini B, Chillemi G, Abbate I, Bruselles A, Rozera G, Castrignanò T, Paoletti D, Picardi E, Desideri A, Pesole G, Capobianchi MR. Assembly and characterization of pandemic influenza A H1N1 genome in nasopharyngeal swabs using high-throughput pyrosequencing. *Microbiol Q J Microbiol Sci.* 2011;34(4):391.
39. Selleri M, Piralla A, Rozera G, Giombini E, Bartolini B, Abbate I, Campanini G, Rovida F, Dossena L, Capobianchi MR, Baldanti F. Detection of haemagglutinin D222 polymorphisms in influenza A (H1N1) pdm09-infected patients by ultra-deep pyrosequencing. *Clin Microbiol Infect.* 2013;19(7):668–73.
40. Jia JA, Liang X, Chen S, Wang H, Li H, Fang M, Bai X, Wang Z, Wang M, Zhu S, Sun F. Next-generation sequencing revealed divergence in deletions of the preS region in the HBV genome between different HBV-related liver diseases. *J Gen Virol.* 2017;98(11):2748–58.
41. Bai X, Jia JA, Fang M, Chen S, Liang X, Zhu S, Zhang S, Feng J, Sun F, Gao C. Deep sequencing of HBV pre-S region reveals high heterogeneity of HBV genotypes and associations of word pattern frequencies with HCC. *PLoS Genet.* 2018;14(2):e1007206.
42. Xu B, Tan Z, Li K, Jiang T, Peng Y. Predicting the host of influenza viruses based on the word vector. *PeerJ.* 2017;5:e3579.
43. Wokaun A, Schrader B. Infrared and Raman spectroscopy-methods and applications. VCH, Weinheim; 1995, DM 298,-, ISBN 3-527-26446-9. Berichte der Bunsengesellschaft für physikalische Chemie. 1996;100(7):1268-.
44. Gremlich HU, Yan B. Infrared and Raman spectroscopy of biological materials. Boca Raton: CRC Press; 2000.
45. Wartewig S, Neubert RH. Pharmaceutical applications of Mid-IR and Raman spectroscopy. *Adv Drug Deliv Rev.* 2005;57(8):1144–70.
46. Vandenebeele P. Practical Raman spectroscopy: an introduction. Chichester, United Kingdom: Wiley; 2013 Jul 3.
47. Blanch EW, Hecht L, Barron LD. Vibrational Raman optical activity of proteins, nucleic acids, and viruses. *Methods.* 2003;29(2):196–209.

48. Tsuoboi M, Kubo Y, Ikeda T, Overman SA, Osman O, Thomas GJ. Protein and DNA residue orientations in the filamentous virus Pf1 determined by polarized Raman and polarized FTIR spectroscopy. *Biochemistry*. 2003;42(4):940–50.
49. Benevides JM, Juuti JT, Tuma R, Bamford DH, Thomas GJ. Characterization of subunit-specific interactions in a double-stranded RNA virus: Raman difference spectroscopy of the φ6 procapsid. *Biochemistry*. 2002;41(40):11946–53.
50. Tuma R, Thomas GJ Jr. Mechanisms of virus assembly probed by Raman spectroscopy: the icosahedral bacteriophage P22. *Biophys Chem*. 1997;68(1–3):17–31.
51. Radzol AR, Lee KY, Mansor W, Omar IS. PCA criterion for SVM (MLP) classifier for flavivirus biomarker from salivary SERS spectra at febrile stage. In: 2016 38th annual international conference of the IEEE Engineering in Medicine and Biology Society (EMBC); 2016 August 16, p. 6206–6209. IEEE.
52. Khan S, Ullah R, Khan A, Ashraf R, Ali H, Bilal M, Saleem M. Analysis of hepatitis B virus infection in blood sera using Raman spectroscopy and machine learning. *Photodiagn Photodyn Ther*. 2018;23:89–93.
53. Scaglia E, Sockalingum GD, Schmitt J, Gobinet C, Schneider N, Manfait M, Thiéfin G. Noninvasive assessment of hepatic fibrosis in patients with chronic hepatitis C using serum Fourier transform infrared spectroscopy. *Anal Bioanal Chem*. 2011;401(9):2919.
54. Khan S, Ullah R, Khan A, Wahab N, Bilal M, Ahmed M. Analysis of dengue infection based on Raman spectroscopy and support vector machine (SVM). *Biomed Opt Express*. 2016;7(6):2249–56.
55. Lee H, Kim MS, Lim HS, Park E, Lee WH, Cho BK. Detection of cucumber green mottle mosaic virus-infected watermelon seeds using a near-infrared (NIR) hyperspectral imaging system: application to seeds of the “Sambok Honey” cultivar. *Biosyst Eng*. 2016;148:138–47.
56. Peng J, Song K, Zhu H, Kong W, Liu F, Shen T, He Y. Fast detection of tobacco mosaic virus infected tobacco using laser-induced breakdown spectroscopy. *Sci Rep*. 2017;7:44551.
57. Liang TC. Epitopes. <https://www.sciencedirect.com/topics/immunology-and-microbiology/epitope>.
58. Desai DV, Kulkarni-Kale U. T-cell epitope prediction methods: an overview. In: *Immunoinformatics*. New York: Humana Press; 2014. p. 333–64.
59. Sanchez-Trincado JL, Gomez-Perez M, Reche PA. Fundamentals and methods for T-and B-cell epitope prediction. *J Immunol Res*. 2017;2017:1–14.
60. Mukonyora M. A review of important discontinuous B-cell epitope prediction tools. *J Clin Cell Immunol*. 2015;6:358–62.
61. Genetics Home Reference. Human leukocyte antigens. <https://ghr.nlm.nih.gov/primer/genefamily/hla>.
62. Soria-Guerra RE, Nieto-Gomez R, Govea-Alonso DO, Rosales-Mendoza S. An overview of bioinformatics tools for epitope prediction: implications on vaccine development. *J Biomed Inform*. 2015;53:405–14.
63. Backert L, Kohlbacher O. Immunoinformatics and epitope prediction in the age of genomic medicine. *Genome Med*. 2015;7(1):119.
64. Zhao Y, Pinilla C, Valmori D, Martin R, Simon R. Application of support vector machines for T-cell epitopes prediction. *Bioinformatics*. 2003;19(15):1978–84.
65. MHC2Pred: SVM based method for prediction of promiscuous MHC Class II binders. <http://crdd.osdd.net/raghava/mhc2pred/info.html>.
66. Dönnies P, Kohlbacher O. SVMHC: a server for prediction of MHC-binding peptides. *Nucleic Acids Res*. 2006;34(suppl\_2):W194–7.
67. Sturniolo T, Bono E, Ding J, Raddrizzani L, Tuereci O, Sahin U, Braxenthaler M, Gallazzi F, Protti MP, Sinigaglia F, Hammer J. Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol*. 1999;17(6):555.
68. Tung CW, Ziehm M, Kämper A, Kohlbacher O, Ho SY. POPISK: T-cell reactivity prediction using support vector machines and string kernels. *BMC Bioinf*. 2011;12(1):446.

69. Kawashima S, Ogata H, Kanehisa M. AAindex: amino acid index database. *Nucleic Acids Res.* 1999;27(1):368–9.
70. Su CH, Pal NR, Lin KL, Chung IF. Identification of amino acid propensities that are strong determinants of linear B-cell epitope using neural networks. *PLoS One.* 2012;7(2):e30617.
71. Yao B, Zhang L, Liang S, Zhang C. SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity. *PLoS One.* 2012;7(9):e45152.
72. Bhagwat M, Aravind L. Psi-blast tutorial. In: Comparative genomics. Totowa: Humana Press; 2007. p. 177–86.
73. Ren J, Liu Q, Ellis J, Li J. Positive-unlabeled learning for the prediction of conformational B-cell epitopes. *BMC Bioinf.* 2015;16(18):S12.
74. PSSM Viewer. [https://www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm\\_viewer.cgi](https://www.ncbi.nlm.nih.gov/Class/Structure/pssm/pssm_viewer.cgi).
75. Huang WL, Tsai MJ, Hsu KT, Wang JR, Chen YH, Ho SY. Prediction of linear B-cell epitopes of hepatitis C virus for vaccine development. *BMC Med Genet.* 2015;8(4):S3.
76. Ansari HR, Raghava GP. Identification of conformational B-cell epitopes in an antigen from its primary sequence. *Immunome Res.* 2010;6(1):6.
77. Safari-Alighiarloo N, Taghizadeh M, Rezaei-Tavirani M, Goliae B, Peyvandi AA. Protein-protein interaction networks (PPI) and complex diseases. *Gastroenterol Hepatol Bed Bench.* 2014;7(1):17.
78. Zhou X, Park B, Choi D, Han K. A generalized approach to predicting protein-protein interactions between virus and host. *BMC Genomics.* 2018;19(6):165.
79. Rao VS, Srinivas K, Sujini GN, Kumar GN. Protein-protein interaction detection: methods and analysis. *Int J Proteomics.* 2014;2014:147648.
80. Gonzalez MW, Kann MG. Protein interactions and disease. *PLoS Comput Biol.* 2012;8(12):e1002819.
81. Brito AF, Pinney JW. Protein–protein interactions in virus–host systems. *Front Microbiol.* 2017;8:1557.
82. MacArthur RD, Novak RM. Maraviroc: the first of a new class of antiretroviral agents. *Clin Infect Dis.* 2008;47(2):236–41.
83. Mei S, Zhu H. A novel one-class SVM based negative data sampling method for reconstructing proteome-wide HTLV-human protein interaction networks. *Sci Rep.* 2015;5:8034.
84. Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, Ruepp A, Frishman D. Negatome 2.0: a database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis. *Nucleic Acids Res.* 2013;42(D1):D396–400.
85. Emamjomeh A, Goliae B, Zahiri J, Ebrahimpour R. Predicting protein–protein interactions between human and hepatitis C virus via an ensemble learning method. *Mol BioSyst.* 2014;10(12):3147–54.
86. Cui G, Fang C, Han K. Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinf.* 2012;13(7):S5. BioMed Central.
87. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. *ACM Trans Intell Syst Technol (TIST).* 2011;2(3):27.
88. Dyer MD, Murali TM, Sobral BW. Supervised learning and prediction of physical interactions between human and HIV proteins. *Infect Genet Evol.* 2011;11(5):917–23.
89. Kim B, Alguwaizani S, Zhou X, Huang DS, Park B, Han K. An improved method for predicting interactions between virus and human proteins. *J Bioinforma Comput Biol.* 2017;15(01):1650024.
90. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. Predicting protein–protein interactions based only on sequences information. *Proc Natl Acad Sci.* 2007;104(11):4337–41.
91. Eid FE, ElHefnawi M, Heath LS. DeNovo: virus-host sequence-based protein–protein interaction prediction. *Bioinformatics.* 2015;32(8):1144–50.
92. Martín V, Perales C, Abia D, Ortíz AR, Domingo E, Briones C. Microarray-based identification of antigenic variants of foot-and-mouth disease virus: a bioinformatics quality assessment. *BMC Genomics.* 2006;7(1):117.

93. Li H, Sun F. Comparative studies of alignment, alignment-free and SVM based approaches for predicting the hosts of viruses based on viral sequences. *Sci Rep.* 2018;8(1):10032.
94. Huang KY, Lu CT, Bretaña NA, Lee TY, Chang TH. ViralPhos: incorporating a recursively statistical method to predict phosphorylation sites on virus proteins. *BMC Bioinf.* 2013;14(16):S10.
95. Cruz-Cano R, Chew DS, Choi KP, Leung MY. Least-squares support vector machine approach to viral replication origin prediction. *INFORMS J Comput.* 2010;22(3):457–70.
96. Shatabda S, Saha S, Sharma A, Dehzangi A. iPHLoc-ES: identification of bacteriophage protein locations using evolutionary and structural features. *J Theor Biol.* 2017;435:229–37.
97. Qin Z, Wang M, Yan A. QSAR studies of the bioactivity of hepatitis C virus (HCV) NS3/4A protease inhibitors by multiple linear regression (MLR) and support vector machine (SVM). *Bioorg Med Chem Lett.* 2017;27(13):2931–8.
98. Liu Z, Lv H, Han J, Liu R. A computational model for predicting transmembrane regions of retroviruses. *J Bioinforma Comput Biol.* 2017;15(03):1750010.
99. Döring M, Borrego P, Büch J, Martins A, Friedrich G, Camacho RJ, Eberle J, Kaiser R, Lengauer T, Taveira N, Pfeifer N. A genotypic method for determining HIV-2 coreceptor usage enables epidemiological studies and clinical decision support. *Retrovirology.* 2016;13(1):85.
100. Wee LJ, Simarmata D, Kam YW, Ng LF, Tong JC. SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction. *BMC Genomics.* 2010;11(4):S21. BioMed Central.
101. EL-Manzalawy Y, Dobbs D, Honavar V. Predicting linear B-cell epitopes using string kernels. *J Mol Recognit Interdiscip J.* 2008;21(4):243–55.
102. Qureshi A, Kaur G, Kumar M. AVC pred: an integrated web server for prediction and design of antiviral compounds. *Chem Biol Drug Des.* 2017;89(1):74–83.

# Eliminating Cervical Cancer: A Role for Artificial Intelligence



Lynette J. Menezes, Lianet Vazquez, Chilukuri K. Mohan,  
and Charurut Somboonwit

**Abstract** Cervical cancer caused by infection with the human papillomavirus (HPV), is the leading cause of mortality among women in many low-resource countries and the fourth leading cause of mortality globally. Decades of cervical cytology screening with subsequent treatment have resulted in marked declines in incidence and mortality of cervical cancer in developed regions, but resource-deprived regions lag behind because of suboptimal access to screening and treatment. Artificial intelligence (AI) technologies are showing promising results in early detection and prediction of cervical cancer progression with potential for future integration into screening and treatment modalities. This chapter begins with an overview of HPV infection, examines HPV pathogenesis and cervical cancer epidemiology; discusses the effectiveness of current primary and secondary prevention strategies and explores the potential role of AI technologies in improving cervical screening, diagnosis and treatment with the goal of eliminating cervical cancer.

**Keywords** Cervical cancer · Human papillomavirus · HPV · Cervical screening · HPV testing · HPV vaccination · Artificial intelligence · Artificial neural networks · Pap test · Machine learning

---

L. J. Menezes (✉) · C. Somboonwit

Division of Infectious Disease & International Medicine, Department of Internal Medicine,  
Morsani College of Medicine, University of South Florida, Tampa, FL, USA  
e-mail: [lmenezes@health.usf.edu](mailto:lmenezes@health.usf.edu); [csomboon@health.usf.edu](mailto:csomboon@health.usf.edu)

L. Vazquez  
Harvard Medical School, Boston, MA, USA  
e-mail: [lianet\\_vazquez@hms.harvard.edu](mailto:lianet_vazquez@hms.harvard.edu)

C. K. Mohan  
Department of Electrical Engineering & Computer Science, Center for Science  
and Technology, Syracuse University, Syracuse, NY, USA  
e-mail: [ckmohan@syr.edu](mailto:ckmohan@syr.edu)

**Key Phrases**

HPV and cervical carcinogenesis, cervical cytology screening, digital cervicography, HPV DNA testing, HPV vaccine impact, cervical screening guidelines, cervical cancer low-resource settings, cervical cancer and AI, cervical cancer screening and ANNs, cervical cancer diagnosis and AI

## 1 Introduction

Elimination of cervical cancer remains an enormous global health challenge. As the fourth most common cancer affecting women, cervical cancer accounted for an estimated 569,800 cases and 311,400 deaths worldwide in 2018 [1, 2]. Despite the substantial decline in cervical cancer mortality in developed countries, cervical cancer remains the leading cause of cancer mortality among women in 42 low-resource countries [2]. Infection by high-risk genotypes of the human papillomavirus (HPV) causes nearly all cases of cervical cancer [3]. HPV is the most common sexually transmitted infection worldwide with an estimated 80% of sexually active men and women expected to acquire an anogenital HPV infection in their lifetime [4]. Persistent high-risk HPV infection has been implicated in cancers at multiple anatomic sites. Nearly 100% of invasive cervical cancers, 88% of anal cancers, 70% of oropharyngeal and vaginal cancers, 43% of vulvar cancers, and 50% of penile cancers have been attributed to HPV infection [5, 6].

Primary and secondary prevention efforts through timely HPV vaccination and regular cervical screening followed by prompt treatment are essential to eliminate cervical cancer. Because of strong and widespread cervical screening implementation by well-developed public health systems, cervical cancer has declined dramatically in high-resource regions. Numerous challenges, however, impede the scaling up of HPV immunization and cervical screening efforts, particularly in low-resource settings with the largest burden of the disease. Newer technologies utilizing artificial intelligence (AI) show promising initial results in optimizing screening, diagnosis and treatment of cervical cancer. Accompanied by the widespread uptake of the HPV vaccine, these AI tools could become the “disruptive technology” needed to eliminate cervical cancer in high-resource countries. The urgent question is whether these innovations can be translated into effective low-cost interventions for the early detection and treatment of cervical precancer in low-resource settings. This chapter will describe the current epidemiology of cervical cancer globally including its pathogenesis and natural history; examine current primary and secondary prevention efforts; provide an overview of artificial intelligence technologies such as machine learning and neural networks; and explore the prospects of these AI technologies in expanding cervical screening coverage, diagnosis and treatment.

## 2 HPV and Cervical Carcinogenesis

HPV is highly transmissible through microabrasions of the mucosal or cutaneous epithelial site [7]. Of more than 100 genotypes identified, 13 in the alpha genus, HPV16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59 and 68 have been designated high-risk based on their degree of carcinogenicity [7]. HPV16 is the most carcinogenic causing more than 50% of all cervical cancers followed by HPV18 representing 15% of all cases including a high proportion of adenocarcinomas [8, 9]. Viral genome integration into the host DNA is a crucial step in the progression to neoplasia and has been frequently found in malignant tumors caused by HPV genotypes 16, 18 and 45 [10]. In a significant proportion of cancers caused by HPV31 and 33, however, tumor development precedes viral genome integration [10]. Of critical importance is the continual expression of the HPV encoded E6 and E7 oncoproteins to maintain cell proliferation [7].

## 3 Epidemiology and Natural History of Cervical Cancer

Recent global estimates show a consistent decline in cervical cancer incidence and mortality in developed regions but a rise in both incidence and mortality in the less developed regions of Southern and Eastern Africa [2]. Overall, the global incidence rate declined slightly from an estimated 14 per 100,000 women in 2012 [11] to 13.1 per 100,000 women in 2018, but low and middle-income countries (LMICs) continue to bear a disproportionate burden of cervical cancer cases and deaths. Swaziland in Southern Africa has the highest age-standardized incidence at 73 cases per 100,000 women with an incidence of 43 per 100,000 women in the region, followed by East Africa at 40 per 100,000 women [2]. More than 85% of cervical cancer deaths occur in less developed regions [12]; Malawi has the highest mortality at 53 per 100,000 women [2].

Large-scale prospective studies have convincingly demonstrated the temporal relationship between acquisition and persistence of carcinogenic HPV infection and the development of graded cervical intraepithelial neoplasia (CIN1, CIN2, CIN3) [13, 14]. These natural history studies have also shown that among HIV-negative women, most HPV infections are transient and that the low-grade squamous epithelial lesions (LSIL) or CIN1 that may arise, regress spontaneously within 1–2 years [15, 16]. A small fraction of these HPV infections persist and can progress to CIN3 rapidly within 2–3 years in young women [14]. However, the time to progression from CIN3 to invasive cervical cancer may take decades [7] presenting an ideal opportunity to screen all adult women and halt the progression of early neoplastic disease to cancer.

Conversely, women infected with the human immunodeficiency virus (HIV) experience a 2- to 25-fold risk for cervical cancer [17]. HPV incidence and prevalence are high among HIV-infected women with a significant proportion harboring

multiple genotypes than HIV-negative women [18–20]. HPV infection in these women persists longer with a rapid progression to cervical neoplasia and invasive cancer at a younger age compared to HIV-negative women [21, 22]. Despite treatment, HIV-infected women may experience recurrent cervical disease [23]. Because of this higher risk for HIV-positive women, cervical cancer is designated as an AIDS-defining illness by the Centers for Disease Control (CDC) [24].

## 4 Secondary Prevention

### 4.1 Cervical Cytology Screening

Despite the advent of HPV immunization in 2006, cervical screening remains the cornerstone strategy for the secondary prevention of cervical cancer among adult women. The most widely used screening modality is cytology testing using the conventional Papanicolaou (Pap) test or the newer liquid-based thin layer preparation, the standard of care in developed regions. Guidelines promoting repeated cytology testing at intervals of 1–3 years resulted in a considerable decline in cervical cancer incidence and mortality in high-income countries. However, its success has been limited in low-resource settings because of poor laboratory infrastructure, lack of trained personnel to collect specimens, prepare and stain smears, absence of or inadequately trained cytopathologists to accurately detect early neoplastic changes in the epithelium, and delays in processing results [25, 26]. Where available, the limited referral facilities for colposcopy-directed biopsy by highly-skilled clinicians, and subsequently histological diagnosis by trained pathologists becomes even more challenging [26, 27]. Beyond these, for most impoverished women, distance to clinic sites, poor transportation, low literacy, communication issues and repeated visits are major barriers to accessing treatment, following a positive screening result [25, 28]. To prevent loss to follow-up and based on all the accumulated evidence on comparative screening tests, the WHO 2013 guidelines recommend screen-and-treat approaches using HPV testing where feasible or visual inspection with acetic acid (VIA) where resources are insufficient [29] (See Table 1).

### 4.2 Visual Inspection of the Cervix

In low-resource settings, visual inspection with acetic acid (VIA) involves applying 3–5% dilute acetic acid to the cervix with a cotton swab or spray and inspecting the cervix for acetowhite lesions with the naked eye using bright light 1 minute after application [33]. Based on numerous studies evaluating VIA in LMICs, some propose that it be considered an interim, alternative for both HIV-positive and HIV-negative women as part of a “screen-and-treat” protocol in such settings [25, 26, 33].

**Table 1** Summary of cervical cancer screening guidelines for women by the United States Preventive Services Task Force (USPSTF) [30], American College of Gynecology and Obstetrics (ACOG) [31], American Society for Colposcopy and Cervical Pathology (ASCCP) [32] and World Health Organization (WHO) [29]

Population	USPSTF	ACOG and ASCCP	WHO <sup>a</sup>
Age <21 years	No screening	No screening	No screening for women <30 years unless HIV+ or living in high HIV prevalence area
Age 21–29 years	Cervical cytology screening (Pap smear) alone every 3 years.	Cervical cytology screening every 3 years	No screening for women <30 years unless HIV+ or living in high HIV prevalence area
Age 30–65 years	Cervical cytology screening (Pap smear) alone every 3 years hrHPV co-testing alone or Cervical cytology and hrHPV co-testing every 5 years	Cervical cytology and HPV co-testing every 5 years (preferred) Cervical cytology screening alone every 3 years (acceptable)	Prioritize women 30–49 years <sup>a</sup> Screen and treat with VIA/cytology & colposcopy Screen and treat with VIA where no resources available Screening interval with HPV testing should be minimum 5 years (not less) Screening interval with VIA/cytology should be 3–5 years
Age >65 years	No screening, if adequate prior negative screening and not high-risk for cervical cancer	No screening, if previous history of negative screening Women with prior history of $\geq$ CIN2 must continue routine screening 3–5 years as in women 30–65 years for at least 20 years	No screening for women without cervix and previous negative screening for CIN2
After hysterectomy	No screening for women without cervix and no history of $\geq$ CIN2 or cervical cancer	Same as age-specific recommendations for unvaccinated women	
HPV vaccinated			

VIA visual inspection with acetic acid, CIN2 cervical intraepithelial neoplasia grade 2, HPV human papillomavirus, hrHPV high-risk HPV

<sup>a</sup>WHO guidelines for resource-limited settings with no organized screening efforts

VIA requires limited equipment, and primary care providers including community-level health workers can be trained to perform it easily [25, 26, 33]. However, its low sensitivity and specificity, high intra- and inter-observer variation in diagnosis, inaccurate results in menopausal women and quality control issues [26, 28, 33] make it an inadequate test to achieve the goal of reducing the cervical cancer burden in impoverished countries. Several studies have evaluated digital cervicography as a tool to improve VIA sensitivity and specificity [34, 35]. Digital cervicography utilizes a camera to produce a cervical picture or cervigram of the cervix after application of acetic acid. Two studies among HIV-positive women in Zambia and Johannesburg, South Africa found a slight improvement in sensitivity but not specificity [34, 35].

### 4.3 HPV Nucleic Acid Testing

Convincing evidence indicates that HPV nucleic acid testing might be a superior and cost-effective screening strategy than cytology testing or VIA to prevent cervical cancer among women 30 years and older [36–38]. In developed regions, the algorithm of HPV DNA testing in combination with cytology has effectively increased screening intervals to 5 years for HIV-negative women and to 3 years for HIV-positive women with normal cytology and a negative HPV test [30, 32, 39] (See Table 1). More recently, the newer HPV tests such as cobas® HPV test and Aptima® HPV assay with their ability to stratify risk of precancer, allow their utilization as a primary screening test among women 25 years and older [39]. Currently, the FDA has approved five HPV tests to be used alone (cobas® only) or in conjunction with Pap cytology for women 30 years and older: Hybrid Capture 2® (Qiagen, Germantown, MD), Cervista™ HPV HR & Cervista™ HPV 16/18 (Hologic Inc., Madison, WI), cobas® HPV test (Roche Molecular Systems Inc., Branchburg, NJ), Aptima® HPV assay and Aptima® HPV 16 18/45 genotype assay (Hologic Inc., San Diego, CA) and BD Onclarity™ HPV assay (Becton Dickinson and Company, Sparks, MD) [40]. Aptima® is the only test that detects HPV mRNA whereas the other four tests detect HPV DNA. Other promising tests include the OncoE6™Cervical Test for detecting the virally encoded E6 oncoprotein [41] and the Trovagene HPV test to detect HPV in the urine [42].

Modeling studies have clearly demonstrated that screening women 30 years and older with a point-of-care high-risk HPV test—with its high sensitivity, ease of specimen collection by provider or self-sampling—is the single best investment with treatment, to reduce cervical cancer risk in high-burden settings [43, 44]. For resource-constrained settings, Xpert® HPV (Cepheid, Sunnyvale, CA) [45] *careHPV* (Qiagen, Gaithersburg, MD) [46] and H13 (Hybribio, Hongkong) [28] are three lower-cost HPV tests that require limited training and infrastructure and can be used as first-line screening for women 30 years and older in LMICs. The infrastructure required, such as human resources, laboratory equipment as well as HPV test kit cost, is still not affordable. In the case of H13 and *careHPV*, to be affordable, the kits

require 48–96 samples with a processing time of 2–3 hours, a significant impediment to widespread implementation. Despite the challenges of cost and overtreatment because of moderate specificity, screening with an HPV test alone remains the best option in a “screen and treat” model compared to visual inspection of the cervix using acetic acid (VIA) or an HPV-VIA triage combination [43, 44].

#### **4.4 Newer Cervical Screening Methods**

To counter the overtreatment of HPV positive women without high-grade neoplasia, newer tests that assess immunohistochemical biomarkers such as, Ki-67 and p16<sup>ink4a</sup> [47, 48] and HPV methylation status [49, 50] that provide risk discrimination beyond what current HPV tests deliver may improve current algorithms for accurate detection of high-grade disease. An RT-qPCR assay to assess p16-mRNA expression in addition to visual reading of p16 immunohistochemistry (IHC) improved the accuracy in diagnosing CIN from 232 HPV positive cervical tissue specimens. However, the digital reading of the p16 stained sections was less sensitive and specific [47]. In van Zummeren et al.’s study, a combined immunoscore for graded overexpression of Ki-67 and p16<sup>ink4a</sup> in IHC-stained cervical tissue specimens allowed for an objective method for pathologists to accurately diagnose CIN3 and CIN1 compared with interpretation of standard H & E stained specimens [48]. A blinded case-control study within the HPV FOCAL cervical screening trial found that among HPV positive women aged 25–65, S5 HPV methylation testing had a high sensitivity and positive predictive value for CIN3, generating risk stratification comparable to an algorithm using cytology to triage HPV positive women [49]. Detecting advanced disease in women at high risk for rapid progression to cervical cancer is an area where methylation markers might prove valuable in the near future [49, 50].

### **5 Impact of Primary Prevention with the HPV Vaccine**

More than 10 years after the quadrivalent vaccine Gardasil® (HPV 6, 11, 16 and 18) and the bivalent vaccine Cervarix® (HPV 16 and 18), were first introduced in the US, numerous studies have convincingly proven their safety and efficacy in decreasing HPV prevalence, as well as incidence of CIN [51–53]. Currently, the nine-valent Gardasil®9 (HPV 6, 11, 16, 18, 31, 33, 45, 52 and 58) is the only vaccine available for use in the US, although Cervarix® and Gardasil® are available in other countries [54]. All three vaccines with efficacy ranging from 96.7% up to 100%, are recommended at 2 doses for girls and boys starting at age 9 and before their 15th birthday and at 3 doses after the 15th birthday up to age 26 years [54].

As a primary prevention tool, widespread dissemination of the HPV vaccine globally, particularly the nine-valent vaccine, has the potential to eliminate cervical cancer and reduce the burden of other HPV-associated cancers. Recent reports from

developed regions that implemented the bivalent and quadrivalent vaccine early on reveal significant reduction in the prevalence of hrHPV infection [51, 55], advanced cervical neoplasia [52, 53, 56], and genital warts among vaccinated women [53, 57]; and evidence of herd protection in unvaccinated men and women [53, 55, 57, 58]. Post-hoc analyses of vaccine trials among women with a history of CIN show reduction in HPV prevalence and recurrent disease [53]. Although the HPV vaccine has been licensed in more than 100 countries, the coverage of the vaccine in many developed and developing regions is low to modest at best [59, 60]. Barriers to implementation and uptake of the vaccine persist both in the US and globally [61, 62]. Promoting the HPV vaccine as a cancer prevention vaccine and incorporating it into national immunization programs in GAVI eligible countries, while assuring its provision with other age-appropriate vaccinations is key to increasing vaccine coverage.

Given the latency period before onset of invasive cervical cancer and the low uptake of the vaccine, the full-benefits of HPV vaccination are decades away. In this scenario, cervical screening remains an effective strategy to reduce cervical cancer among older women in high-burden low-resource settings. However, the lack of skilled pathologists makes it imperative to find objective tests to improve diagnostic precision. Screening algorithms that incorporate more objective technologies—with the goal of providing additional risk discrimination for early and accurate detection—can lead to improved algorithms for the management of precancerous lesions. Cost-effectively translating these technologies—for application in low-resource settings—is key to achieving the International Papillomavirus Society's (IPVS) goal of eliminating cervical cancer as a public health problem [63]. The next section will describe the essential concepts in artificial intelligence (AI) and the prospect of AI technologies in addressing the need for improved cervical screening tools as well as its potential application in low-resource settings, thereby accelerating the process of achieving cervical cancer elimination.

## 6 Essential Concepts of AI, Expert Systems, Machine Learning, Neural Networks

In the context of medical applications, *Artificial Intelligence (AI)* refers to a collection of algorithms and their implementation, some of which are drawn from our understanding of the principles underlying *natural* (human) intelligence. In particular, AI includes algorithms developed to solve complex pattern recognition tasks while circumventing the computational obstacles of exhaustive search. Carefully designed knowledge representation formalisms and inference methods constitute the arsenal of an AI researcher, applied often to assist in medical diagnosis, in addition to the tasks of prognosis and decision-making.

The earliest medical applications of AI included programs such as MYCIN [64]. MYCIN was a *rule-based expert system* developed to assist physicians in selecting

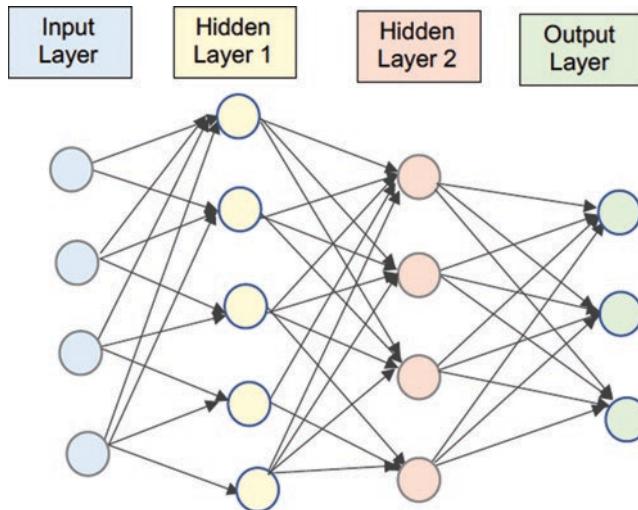
the most appropriate antimicrobial against an infectious disease [64]. The algorithm was built using knowledge elicitation techniques [65] from expert data, then represented using a set of *production rules*. Considerable uncertainty often accompanies medical diagnosis and prognosis. Aleatoric variability (randomness) accompanies the lack of determinism between conditions and effects. Inferences must be made although some relevant variables are unknown or unknowable. Limitations in the measurement process or the size of the observed data may inhibit learning the relations between input and output variables in a model.

Imprecision may also be inherent in the language used by medical experts, e.g. subjective judgments (whether a lesion is low-grade or high-grade) by pathologists when evaluating cytology specimens or biopsied tissue. This is also true of colposcopy evaluations which are equally subjective. Rule-based medical *expert systems* or *decision support systems* have addressed uncertainty using different approaches, such as Conditional Probabilities, Fuzzy Logic, Certainty Factors, and Belief Intervals. Another approach has been to capture chunks of knowledge in *Case Based Reasoning (CBR)* systems [66], in which abstractions of actual cases (e.g. linking observable evidence to diagnoses) are directly represented. The decision corresponding to a new case is based on evaluating its similarity to existing cases; the most similar case in the case library is determined, and its conclusion modified using *adaptation rules* learnt specifically for this task.

Both rule-based and case-based systems attempt to represent knowledge using (mostly) independent chunks. In recent years, medical AI systems have been using more complex knowledge representation approaches that use *trees* or *graphs*. *Machine Learning* [67] is a subfield of AI, focused on the development of algorithms to learn the architecture and parameter values of models that best explain the processes that generated observable data. Much of the current enthusiasm for AI (*big data, data mining or computational intelligence*) is due to the successful implementation of machine learning algorithms that develop models to answer questions by analyzing data, seeking patterns and formulating models without human intervention.

Decision trees, for instance, involve asking a series of questions, where the algorithm learns from preceding answers to determine the follow-up questions. Learning algorithms (such as CART, ID3, and C4.5) are applied to identify the best sets of questions to ask at each step, resulting in the construction of the decision tree. In recent years, a popular approach has been to use *random forests* [68] in which multiple decision trees are quasi-randomly generated, evaluated, and their answers combined to yield a final result.

Much more attention has focused on (artificial) *Neural Networks (NNs)* [69], inspired by biological neural systems. The most popular NN models are *multilayer perceptrons* that contain one or two *hidden layers of neurons* (See Fig. 1). Each neuron computes a nonlinear function (such as the hyperbolic tangent) of the weighted sum of inputs from all the neurons in the preceding layer. In a typical medical application, symptoms, biomarkers or test results are supplied to the input layer, and the outermost layer's outputs are interpreted as identifying diagnostic classes.



**Fig. 1** Graphic representation of a neural network composed of an input, four-neuron layer, two hidden layers, and a final output layer. Data entered into the input layer is processed in the first hidden layer. The output from this first hidden layer becomes the input for the second hidden layer which further processes the data to provide the final output. Applied to cervical cancer, the input layer could be data related to cervical disease biomarkers, including but not limited to cytology, p16INK4a/Ki-67, S5 methylation status, HPV 16/18 or other hrHPV test results. The two hidden layers of neurons perform successive nonlinear transformations of the input data, facilitating the final prediction or classification step performed by the output layer; in this example, discriminating between three diagnostic classes such as Normal, CIN1, and CIN2+

Learning occurs by repeatedly presenting *training data* to the NN and making incremental changes to the weights in such a way that accuracy improves at each step. Additional *test data* help evaluate whether the learning process has been successful.

Recently, *Deep NNs* [70] with many layers have been used for problems involving the analysis of images. The earliest *convolutional* layers extract features (e.g. detecting the occurrence of circular arcs), while the final layers are used for prediction. Practical concerns with the application of such models include the difficulty of interpreting the learned NN models, the potential for overfitting (especially when too many parameter values are to be learnt), and imbalance. For example, there may be too many examples of one class (e.g. normal cytology) and too few examples of the other class (e.g. CIN3).

## 7 Cervical Cancer Screening and AI

The urgent need to eliminate cervical cancer has stimulated a growing interest in the scientific community to assess the utility of AI techniques in screening, diagnosing, and treating cervical cancer. Several recent studies have employed AI techniques

with the goals of: reducing the workload of cytopathologists by identifying high-grade lesions from normal cytology; reducing observer bias and variability in classification; providing a more cost-effective mechanism of classification; and personalizing screenings to reduce false positive rates.

William et al. [71] conducted a comprehensive review of AI applications to cervical cancer screenings using cytology. Technologies have been developed and optimized overtime to increase accuracy, sensitivity and specificity of these techniques, particularly in the areas of segmentation, feature extraction, and classification. The purpose of segmentation is to isolate the cell nucleus, where dysplasia could first be noticed. It is commonly done using a thresholding technique, which has proved useful in identifying nuclei in overlapping cells. Feature extraction further analyzes the selected area and depending on the program, it can focus on structure, texture, or both. Classification integrates the data obtained from segmentation and feature extraction and categorizes the sample as normal, dysplastic, or neoplastic, depending on the algorithm. Some studies have been able to establish a binary classification of normal versus abnormal, while others have identified different degrees of dysplasia. The best binary classifiers from the literature are K-nearest-neighbors and support vector machines algorithms, but the majority of the existing algorithms yield an accuracy of approximately 93.78%, which is still considered low [71].

Although the review documented a number of promising advances in the field, the authors concluded that weaknesses continue to exist in classifying certain classes of cells [71]. Further optimization is needed in the fields of segmentation, feature extraction and classification, such as using hybrid segmentation and multi-level classifiers that combine K-nearest-neighbors algorithms with support vector machines, and pixel level classifications [71]. Bora et al. looked at 1320 smear images using an ensemble classification technique comprised of three individual classifiers: Least Square Support Vector Machine (LSSVM), Multilayer Perceptron (MLP) and Random Forest (RF) [72]. The ensemble classifier focused on analyzing shape, texture and color features, in order to automate the process of segmentation and selection of region of interest. The device classified cervical dysplasia into two-level (normal and abnormal) and three-level classes (Negative for Intraepithelial Lesion or Malignancy (NILM), Low-grade Squamous Intraepithelial Lesion (LSIL) and High-grade Squamous Intraepithelial Lesion (HSIL)). The ensemble classifier outperformed the individual classifiers, with an accuracy of 98.11% and precision of 98.38% [72].

AI could also be used to personalize cervical cancer screenings based on medical history and genetic profile. Information gathered through these methods can be incorporated into a machine-learning-based risk prediction tool to tailor screening frequency based on individual risk profile. Wagholikar et al. built a computerized clinical decision support system (CDSS) for cervical cancer screening based on chart review of Pap reports and HPV testing results [73]. The CDSS assesses various risk profiles from these reports and develops patient-specific recommendations for future screening and management. Free-text narrative data posed a challenge for automated processing because of the lack of standardization and difficulty in extracting relevant information [73]. Text-mining tools could help analyze narrative data and incorporate

it into the automated clinical decision support system, as accomplished by Weegar et al. [74]. Their study employed text-mining tools to analyze health records and built an algorithm to detect early symptoms of cervical cancer. Subsequently, they combined the Clinical Entity Finder (CEF) and NegEx tools to perform entity recognition and negation detection respectively. The study concluded that further optimization was needed, but early results appear promising [74].

More recent studies have attempted to optimize machine-learning-based risk prediction tools. Kyrgiou et al. used artificial neural networks to integrate cytology, HPV DNA typing, E6&E7mRNA and p16<sup>INK4a</sup> biomarkers' results into a clinical decision support scoring system (DSSS) for individualized management of cervical abnormalities in women [75]. Although this ANN based DSSS succeeded in predicting women with the highest risk for precancer, the need for three additional biomarkers does not make it feasible in most settings. One group built a machine-learning model to examine the screening histories of women from their medical charts and stratify them into high-risk and low-risk groups [76]. The algorithm's performance was inadequate but this approach continues to be explored [76].

## 8 Cervical Cancer Diagnosis and AI

The current methods to diagnose cervical cancer include obtaining and analyzing a biopsy using histological grading, as well as incorporating other diagnostic techniques such as CT scans, MRIs, and PET scans. These methods have several drawbacks such as high cost, complex sample preparation, long processing time, as well as subjective interpretation of results [77].

AI could prove instrumental in making the diagnostic process more efficient and cost effective. Long et al. [78] established a deep learning-based model utilizing cervical cancer biomarkers and genetic panels to aid diagnosis and prognosis of disease. They investigated the specific biomarkers correlated with initiation, invasion, and progression of cervical cancers, and identified several genes that were differentially expressed. Their 168-gene deep learning differentiation model diagnosed patients with cervical cancer with an accuracy of 97.96% (99.01% sensitivity and 95.65% specificity) [78]. More recently, others used laser-induced breakdown spectroscopy (LIBS) along with the combined methods of principal component analysis (PCA) and support vector machine (SVM) to distinguish normal from cervical cancer tissue [77]. The potential of this technology would allow real-time cervical cancer diagnosis with an identification accuracy of 94.4% [77].

## 9 Cervical Cancer Treatment and AI

Staging of cervical cancer is important in order to assess the type of treatment needed—the choice of surgery, radiotherapy, chemotherapy, or a combination. The current staging system involves a myriad of inspection methods in the form of

imaging, excretory urography, cystoscopy, and proctoscopy, often found to be subjective in interpretation [79]. Mu et al. set out to optimize cervical cancer staging through automatic classification using a support vector machine classifier, focusing on PET image texture analysis [79]. The algorithm was successful in differentiating early stage from advanced stage cervical cancer with high accuracy [79].

No fully autonomic delineation method yet exists for cervical cancer, but there's been ongoing exploration of this approach. Torheim et al. attempted to develop an automatic tumor segmentation and delineation method for locally advanced cervical cancers [80]. They used Fisher's Linear Discriminant Analysis (LDA) to classify DCE-MRI images, following calibration by two experienced radiologists. The goal was to build probability maps based on imaging to better guide radiotherapy dose delivery – the dose administered would be proportional to tumor probability [80].

## 10 AI Applications to Cervical Cancer Screening in Resource-Limited Settings

AI technologies to assist in diagnosing cervical lesions in resource-limited settings are needed. Song et al. [81] proposed using digital cervicography combined with a data-driven computer algorithm to identify cancerous lesions. Because of its low cost and low threshold for technical expertise, digital cervicography if optimized, could be a valued alternative to VIA in resource-limited settings. Similarly, Hu et al. developed an automated visual evaluation algorithm based on region-based convolutional neural network analysis to evaluate cervigrams [82]. The algorithm was tasked with two main functions: correctly identifying the location of the cervix in the input image, and estimating the probability that the image represents a case of normal tissue, dysplasia, or neoplasia. This method identified cancerous lesions with greater accuracy than previous cervigram interpretation methods or conventional cytology [82].

Although traditional cervicography is now obsolete, the application of automated visual evaluation using contemporary digital images could increase both sensitivity and specificity of traditional cervical cancer screenings in resource-limited settings. The dissemination strategy would entail minimal training of health workers to take well-lit, in-focus images of the cervix, while requiring little equipment: a digital camera, acetic acid, a disposable specula, and the algorithm software [82].

## 11 Conclusion

Early detection and prompt treatment of cervical cancer remain a challenge to the elimination of cervical cancer both in developed and developing regions. Compelling scientific evidence confirms that HPV vaccination in combination with regular cervical screening using HPV testing are the two most cost-effective approaches to prevent precancerous lesions and ultimately cervical cancer. Further triage with cytology in high-resource countries or VIA in low-resource settings for HPV-positive women

can provide additional risk discrimination prior to treatment. Yet, investment in national, organized cervical screening and HPV immunization programs in most resource-deprived regions is lacking. Advances in AI applications to cervical cancer such as image-based modules, text mining and artificial neural networks are promising and encouraging in potentially augmenting clinical decision making by increasing accuracy of detecting advanced cervical disease and improving its management. Further investment in refining and evaluating these technologies to improve screening and treatment in high-burden low-resource settings is warranted. Meanwhile, national governments must invest in both national immunization programs to institute and increase HPV vaccination coverage as well as establish and maintain organized cervical screening programs utilizing the best available scientific evidence.

**Acknowledgments** None.

**Potential Conflicts of Interest** The authors have no potential conflicts of interest to report.

## References

1. Ferlay J, Colombet M, Soerjomataram I, et al. Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. *Int J Cancer*. 2019;144(8):1941–53.
2. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*. 2018;68(6):394–424.
3. Walboomers JM, Jacobs MV, Manos MM, et al. Human papillomavirus is a necessary cause of invasive cervical cancer worldwide. *J Pathol*. 1999;189(1):12–9.
4. Chesson HW, Dunne EF, Hariri S, Markowitz LE. The estimated lifetime probability of acquiring human papillomavirus in the United States. *Sex Transm Dis*. 2014;41(11):660–4.
5. de Martel C, Ferlay J, Franceschi S, et al. Global burden of cancers attributable to infections in 2008: a review and synthetic analysis. *Lancet Oncol*. 2012;13(6):607–15.
6. Chaturvedi AK, Engels EA, Pfeiffer RM, et al. Human papillomavirus and rising oropharyngeal cancer incidence in the United States. *J Clin Oncol*. 2011;29(32):4294–301.
7. International Agency for Research on Cancer (IARC). Human papillomaviruses. In: IARC Working Group on the Evaluation of Carcinogenic Risks to Humans, editor. Biological agents. Volume 100 B. A review of human carcinogens. Lyon: IARC; 2012. p. 255–313.
8. Clifford GM, Smith JS, Plummer M, Munoz N, Franceschi S. Human papillomavirus types in invasive cervical cancer worldwide: a meta-analysis. *Br J Cancer*. 2003;88(1):63–73.
9. Smith JS, Lindsay L, Hoots B, et al. Human papillomavirus type distribution in invasive cervical cancer and high-grade cervical lesions: a meta-analysis update. *Int J Cancer*. 2007;121(3):621–32.
10. Vinokurova S, Wentzensen N, Kraus I, et al. Type-dependent integration frequency of human papillomavirus genomes in cervical lesions. *Cancer Res*. 2008;68(1):307–13.
11. Ferlay J, Soerjomataram I, Dikshit R, et al. Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *Int J Cancer*. 2015;136(5):E359–86.
12. World Health Organization (WHO). Human papillomavirus (HPV) and cervical cancer. WHO. Fact sheets Web site. [https://www.who.int/en/news-room/fact-sheets/detail/human-papillomavirus-\(hpv\)-and-cervical-cancer](https://www.who.int/en/news-room/fact-sheets/detail/human-papillomavirus-(hpv)-and-cervical-cancer). Published 2019. Updated January 24, 2019. Accessed 24 Mar 2019.

13. Rodriguez AC, Schiffman M, Herrero R, et al. Longitudinal study of human papillomavirus persistence and cervical intraepithelial neoplasia grade 2/3: critical role of duration of infection. *J Natl Cancer Inst.* 2010;102(5):315–24.
14. Winer RL, Kiviat NB, Hughes JP, et al. Development and duration of human papillomavirus lesions, after initial infection. *J Infect Dis.* 2005;191(5):731–8.
15. Winer RL, Hughes JP, Feng Q, et al. Early natural history of incident, type-specific human papillomavirus infections in newly sexually active young women. *Cancer Epidemiol Biomark Prev.* 2011;20(4):699–707.
16. Moscicki AB, Shibuski S, Hills NK, et al. Regression of low-grade squamous intra-epithelial lesions in young women. *Lancet.* 2004;364(9446):1678–83.
17. De Vuyst H, Lillo F, Broutet N, Smith JS. HIV, human papillomavirus, and cervical neoplasia and cancer in the era of highly active antiretroviral therapy. *Eur J Cancer Prev.* 2008;17(6):545–54.
18. Clifford GM, Goncalves MA, Franceschi S. Human papillomavirus types among women infected with HIV: a meta-analysis. *AIDS.* 2006;20(18):2337–44.
19. Hawes SE, Critchlow CW, Sow PS, et al. Incident high-grade squamous intraepithelial lesions in Senegalese women with and without human immunodeficiency virus type 1 (HIV-1) and HIV-2. *J Natl Cancer Inst.* 2006;98(2):100–9.
20. Menezes LJ, Poongulali S, Tommasino M, et al. Prevalence and concordance of human papillomavirus infection at multiple anatomic sites among HIV-infected women from Chennai, India. *Int J STD AIDS.* 2016;27(7):543–53.
21. Hagensee ME, Cameron JE, Leigh JE, Clark RA. Human papillomavirus infection and disease in HIV-infected individuals. *Am J Med Sci.* 2004;328(1):57–63.
22. Gichangi PB, Bwayo J, Estambale B, et al. Impact of HIV infection on invasive cervical cancer in Kenyan women. *AIDS.* 2003;17(13):1963–8.
23. Huchko MJ, Leslie H, Maloba M, Zakaras J, Bukusi E, Cohen CR. Outcomes up to 12 months after treatment with loop electrosurgical excision procedure for cervical intraepithelial neoplasia among HIV-infected women. *J Acquir Immune Defic Syndr.* 2015;69(2):200–5.
24. Centers for Disease Control and Prevention. AIDS-defining conditions. *MMWR Morb Mortal Wkly Rep.* 2008;57(RR-10):9.
25. Denny L. The prevention of cervical cancer in developing countries. *BJOG.* 2005;112(9):1204–12.
26. Wright TC, Kuhn L. Alternative approaches to cervical cancer screening for developing countries. *Best Pract Res Clin Obstet Gynaecol.* 2012;26(2):197–208.
27. Denny L. Prevention of cervical cancer. *Reprod Health Matters.* 2008;16(32):18–31.
28. Fokom Domgue J, Schiffman M, Wentzensen NH, et al. Assessment of a new lower-cost real-time PCR assay for detection of high-risk human papillomavirus: useful for cervical screening in limited-resource settings? *J Clin Microbiol.* 2017;55(8):2348–55.
29. World Health Organization (WHO). WHO guidelines for screening and treatment of precancerous lesions for cervical cancer prevention. 2013. [https://apps.who.int/iris/bitstream/handle/10665/94830/9789241548694\\_eng.pdf?sequence=1](https://apps.who.int/iris/bitstream/handle/10665/94830/9789241548694_eng.pdf?sequence=1). Accessed 15 Apr 2019.
30. Curry SJ, Krist AH, Owens DK, et al. Screening for cervical cancer: US preventive services task force recommendation statement. *JAMA.* 2018;320(7):674–86.
31. Saslow D, Solomon D, Lawson HW, et al. American Cancer Society, American Society for Colposcopy and Cervical Pathology, and American Society for Clinical Pathology screening guidelines for the prevention and early detection of cervical cancer. *Am J Clin Pathol.* 2012;137(4):516–42.
32. The American College of Obstetricians and Gynecologists. Practice bulletin no. 168: cervical cancer screening and prevention. *Obstet Gynecol.* 2016;128(4):e111–30.
33. Sankaranarayanan R, Nessa A, Esmy PO, Dangou J-M. Visual inspection methods for cervical cancer prevention. *Best Pract Res Clin Obstet Gynaecol.* 2012;26(2):221–32.
34. Bateman AC, Parham GP, Sahasrabuddhe VV, et al. Clinical performance of digital cervicography and cytology for cervical cancer screening in HIV-infected women in Lusaka, Zambia. *J Acquir Immune Defic Syndr.* 2014;67(2):212–5.

35. Firnhaber C, Mao L, Levin S, et al. Evaluation of a cervicography-based program to ensure quality of visual inspection of the cervix in HIV-infected women in Johannesburg, South Africa. *J Low Genit Tract Dis.* 2015;19:1.
36. Sankaranarayanan R, Nene BM, Shastri SS, et al. HPV screening for cervical cancer in rural India. *N Engl J Med.* 2009;360(14):1385–94.
37. Flores YN, Bishai DM, Lorincz A, et al. HPV testing for cervical cancer screening appears more cost-effective than Papanicolaou cytology in Mexico. *Cancer Causes Control.* 2011;22(2):261–72.
38. Arbyn M, Ronco G, Anttila A, et al. Evidence regarding human papillomavirus testing in secondary prevention of cervical cancer. *Vaccine.* 2012;30(Supplement 5):F88–99.
39. Huh WK, Ault KA, Chelmow D, et al. Use of primary high-risk human papillomavirus testing for cervical cancer screening: interim clinical guidance. *Obstet Gynecol.* 2015;125(2):330–7.
40. Demarco M, Carter-Pokras O, Hyun N, et al. Validation of a Human Papillomavirus (HPV) DNA cervical screening test that provides expanded HPV typing. *J Clin Microbiol.* 2018;56(5):e01910–7.
41. Torres KL, Marino JM, Pires Rocha DA, et al. Self-sampling coupled to the detection of HPV 16 and 18 E6 protein: a promising option for detection of cervical malignancies in remote areas. *PLoS One.* 2018;13(7):e0201262.
42. Cuzick J, Cadman L, Ahmad AS, et al. Performance and diagnostic accuracy of a urine-based human papillomavirus assay in a referral population. *Cancer Epidemiol Biomark Prev.* 2017;26(7):1053–9.
43. Campos NG, Jeronimo J, Tsu V, Castle PE, Mvundura M, Kim JJ. The cost-effectiveness of visual triage of human papillomavirus-positive women in three low- and middle-income countries. *Cancer Epidemiol Biomark Prev.* 2017;26(10):1500–10.
44. Campos NG, Lince-Deroche N, Chibwesha CJ, et al. Cost-effectiveness of cervical cancer screening in women living with HIV in South Africa: a mathematical modeling study. *J Acquir Immune Defic Syndr.* 2018;79(2):195–205.
45. Cuschieri K, Geraets D, Cuzick J, et al. Performance of a cartridge-based assay for detection of clinically significant Human Papillomavirus (HPV) infection: lessons from VALGENT (Validation of HPV Genotyping Tests). *J Clin Microbiol.* 2016;54(9):2337–42.
46. Qiao YL, Sellors JW, Eder PS, et al. A new HPV-DNA test for cervical-cancer screening in developing regions: a cross-sectional study of clinical accuracy in rural China. *Lancet Oncol.* 2008;9(10):929–36.
47. Vasiljevic N, Carter PD, Reuter C, et al. Role of quantitative p16(INK4A) mRNA assay and digital reading of p16(INK4A) immunostained sections in diagnosis of cervical intraepithelial neoplasia. *Int J Cancer.* 2017;141(4):829–36.
48. van Zummeren M, Leeman A, Kremer WW, et al. Three-tiered score for Ki-67 and p16(ink4a) improves accuracy and reproducibility of grading CIN lesions. *J Clin Pathol.* 2018;71(11):981–8.
49. Cook DA, Krajden M, Brentnall AR, et al. Evaluation of a validated methylation triage signature for human papillomavirus positive women in the HPV FOCAL cervical cancer screening trial. *Int J Cancer.* 2019;144(10):2587–95.
50. Brentnall AR, Vasiljevic N, Scibior-Bentkowska D, et al. HPV33 DNA methylation measurement improves cervical pre-cancer risk estimation of an HPV16, HPV18, HPV31 and \textit{EpB41L3} methylation classifier. *Cancer Biomark.* 2015;15(5):669–75.
51. Ahrlund-Richter A, Cheng L, Hu YOO, et al. Changes in cervical Human Papillomavirus (HPV) prevalence at a youth clinic in Stockholm, Sweden, a decade after the introduction of the HPV vaccine. *Front Cell Infect Microbiol.* 2019;9:59.
52. Wright TC Jr, Parvu V, Stoler MH, et al. HPV infections and cytologic abnormalities in vaccinated women 21–34 years of age: results from the baseline phase of the Onclarity trial. *Gynecol Oncol.* 2019;153(2):259–65.
53. Brotherton JML, Bloem PN. Population-based HPV vaccination programmes are safe and effective: 2017 update and the impetus for achieving better global coverage. *Best Pract Res Clin Obstet Gynaecol.* 2018;47:42–58.

54. Meites E, Kempe A, Markowitz LE. Use of a 2-dose schedule for human papillomavirus vaccination – updated recommendations of the advisory committee on immunization practices. MMWR Morb Mortal Wkly Rep. 2016;65(49):1405–8.
55. Spinner C, Ding L, Bernstein DI, et al. Human papillomavirus vaccine effectiveness and herd protection in young women. Pediatrics. 2019;143(2) <https://doi.org/10.1542/peds.2018-1902>.
56. McClung NM, Gargano JW, Park IU, et al. Estimated number of cases of high-grade cervical lesions diagnosed among women – United States, 2008 and 2016. MMWR Morb Mortal Wkly Rep. 2019;68(15):337–43.
57. Patel C, Brotherton JM, Pillsbury A, et al. The impact of 10 years of human papillomavirus (HPV) vaccination in Australia: what additional disease burden will a nonavalent vaccine prevent? Euro Surveill. 2018;23:41.
58. Palmer T, Wallace L, Pollock KG, et al. Prevalence of cervical disease at age 20 after immunisation with bivalent HPV vaccine at age 12–13 in Scotland: retrospective population study. BMJ. 2019;365:i1161.
59. Simms KT, Steinberg J, Caruana M, et al. Impact of scaled up human papillomavirus vaccination and cervical screening and the potential for global elimination of cervical cancer in 181 countries, 2020–99: a modelling study. Lancet Oncol. 2019;20(3):394–407.
60. Franco M, Mazzuzza S, Padek M, Brownson RC. Going beyond the individual: how state-level characteristics relate to HPV vaccine rates in the United States. BMC Public Health. 2019;19(1):246.
61. Pierce Campbell CM, Menezes LJ, Paskett ED, Giuliano AR. Prevention of invasive cervical cancer in the United States: past, present, and future. Cancer Epidemiol Biomark Prev. 2012;21(9):1402–8.
62. Gallagher KE, LaMontagne DS, Watson-Jones D. Status of HPV vaccine introduction and barriers to country uptake. Vaccine. 2018;36(32 Pt A):4761–7.
63. International Papillomavirus Society. IPVS Statement: Moving towards Elimination of Cervical Cancer as a Public Health Problem. <https://ipvsoc.org/wp-content/uploads/2018/02/IPVs-statement-on-elimination.pdf>. Published 2018. Updated February 2018. Accessed 24 Apr 2019.
64. Buchanan BG, Shortliffe EH. Rule based expert systems: the mycin experiments of the stanford heuristic programming project (the Addison-Wesley series in artificial intelligence). Reading, MA: Addison-Wesley Longman Publishing Co., Inc.; 1984.
65. Firlej M, Hellens D. Knowledge elicitation: a practical handbook. Englewood Cliffs, NJ: Prentice-Hall; 1991.
66. Kolodner J. Case-based reasoning. Sanfranciso, CA: Morgan Kaufmann Publishers Inc.; 1993.
67. Bishop CM. Pattern recognition and machine learning (information science and statistics). New York, NY: Springer; 2006.
68. Smith C. Decision trees and random forests: a visual introduction for beginners. Blue Windmill Media; 2017.
69. Mehrotra K, Mohan CK, Ranka S. Elements of artificial neural networks. Cambridge, MA: MIT Press; 1997.
70. Aggarwal CC. Neural networks and deep learning. New York, NY: Springer; 2018.
71. William W, Ware A, Basaza-Ejiri AH, Obungoloch J. A review of image analysis and machine learning techniques for automated cervical cancer screening from pap-smear images. Comput Methods Programs Biomed. 2018;164:15–22.
72. Bora K, Chowdhury M, Mahanta LB, Kundu MK, Das AK. Automated classification of Pap smear images to detect cervical dysplasia. Comput Methods Programs Biomed. 2017;138:31–47.
73. Wagholikar KB, MacLaughlin KL, Henry MR, et al. Clinical decision support with automated text processing for cervical cancer screening. J Am Med Inform Assoc. 2012;19(5):833–9.
74. Weegar R, Kvist M, Sundstrom K, Brunak S, Dalianis H. Finding cervical cancer symptoms in swedish clinical text using a machine learning approach and NegEx. AMIA Annu Symp Proc. 2015;2015:1296–305.

75. Kyrgiou M, Pouliakis A, Panayiotides JG, et al. Personalised management of women with cervical abnormalities using a clinical decision support scoring system. *Gynecol Oncol*. 2016;141(1):29–35.
76. Baltzer N, Sundstrom K, Nygard JF, Dillner J, Komorowski J. Risk stratification in cervical cancer screening by complete screening history: applying bioinformatics to a general screening population. *Int J Cancer*. 2017;141(1):200–9.
77. Wang J, Li L, Yang P, et al. Identification of cervical cancer using laser-induced breakdown spectroscopy coupled with principal component analysis and support vector machine. *Lasers Med Sci*. 2018;33(6):1381–6.
78. Long NP, Jung KH, Yoon SJ, et al. Systematic assessment of cervical cancer initiation and progression uncovers genetic panels for deep learning-based early diagnosis and proposes novel diagnostic and prognostic biomarkers. *Oncotarget*. 2017;8(65):109436–56.
79. Mu W, Chen Z, Liang Y, et al. Staging of cervical cancer based on tumor heterogeneity characterized by texture features on (18)F-FDG PET images. *Phys Med Biol*. 2015;60(13):5123–39.
80. Torheim T, Malinen E, Hole KH, et al. Autodelineation of cervical cancers using multiparametric magnetic resonance imaging and machine learning. *Acta Oncol*. 2017;56(6):806–12.
81. Song D, Kim E, Huang X, et al. Multimodal entity coreference for cervical dysplasia diagnosis. *IEEE Trans Med Imaging*. 2015;34(1):229–45.
82. Hu L, Bell D, Antani S, et al. An observational study of deep learning and automated evaluation of cervical images for cancer screening. *J Natl Cancer Inst*. 2019; <https://doi.org/10.1093/jnci/djy225>.

# HIV and Injection Drug Use: New Approaches to HIV Prevention



Charurut Somboonwit, Lianet Vazquez, and Lynette J. Menezes

**Abstract** Injection drug use has become a major public health problem. Its emerging significance is demonstrated globally in the dual HIV and HCV epidemics among people who inject drugs (PWID). Despite the advent of effective antivirals against HIV and HCV, PWID face multiple barriers to access and adherence to such treatments. Additionally, the lack of infrastructure for medication-assisted therapy for opioid addiction, inadequate treatment for underlying mental health disorders, and poor access to needle-syringe exchange programs and HIV pre-exposure prophylaxis pose grave challenges to control these epidemics. In this chapter, we focus on the impact of the global injection drug use epidemic as well as new approaches on HIV prevention and the HIV care continuum for people who inject drugs.

**Keywords** HIV · PWID · Artificial intelligence · Machine learning · Injection drug use · Care continuum · Social media · Phylodynamics · Phylogenetic · Digital technologies

## Key Phrases

HIV epidemic among PWID, early detection of HIV infection in PWID, HIV prevention in PWID, social media, HIV prevention and artificial intelligence

---

C. Somboonwit (✉) · L. J. Menezes

Division of Infectious Disease & International Medicine, Department of Internal Medicine,  
Morsani College of Medicine, University of South Florida, Tampa, FL, USA  
e-mail: [csomboon@health.usf.edu](mailto:csomboon@health.usf.edu); [lmenezes@health.usf.edu](mailto:lmenezes@health.usf.edu)

L. Vazquez  
Harvard Medical School, Boston, MA, USA  
e-mail: [lianet\\_vazquez@hms.harvard.edu](mailto:lianet_vazquez@hms.harvard.edu)

## 1 Introduction

Infection with the human immunodeficiency virus (HIV) among people who inject drugs (PWID) is increasingly becoming a global health problem. Recent estimates suggest that over 35 million people are living with HIV globally, with 1.8 million new cases and 970,000 deaths annually [1]. Injection drug use in low-and middle-income countries (LMICs) has increased, with an estimated 15.6 million PWID worldwide [2, 3]. In addition, 10–17% of HIV infections worldwide and 30% of those outside of Africa are attributed to injection drug use [3]. PWID are also often burdened with other concurrent blood-borne infections such as Hepatitis C virus (HCV). Globally, more than 1.1 million PWID are co-infected with HIV and HCV [2].

Injection drug use not only contributes to heightened transmission risk, but also to disease progression. HIV typically enters the central nervous system (CNS) and establishes a reservoir within a few days of infection by interacting with microglial cells, circulating macrophages, lymphocytes (mainly CD4+ T helper cells) and astrocytes. On-going viral replication and expression of HIV proteins such as Tat and gp120, elicit an immune response that further deteriorates the CNS [4]. These processes are enhanced by common illicit substances such as morphine/heroin, cocaine and amphetamine/methamphetamine by fostering glial cell activation, promoting HIV replication, augmenting transcription of viral proteins, depleting CD4+ cells, and contributing to the deterioration of the blood brain barrier (BBB) [5, 6].

Despite the advent of effective antivirals against HIV, PWID continue to face multiple barriers to access and adherence to such treatments, often secondary to their substance use disorder. The lack of infrastructure for opioid agonist therapy (OAT), inadequate treatment for underlying mental health disorders, and poor access to needle and syringe exchange programs (NSEP) and pre-exposure prophylaxis (PrEP) all contribute to considerable challenges to the containment of the HIV epidemic within this population [7]. The purpose of this chapter is to discuss the impact of the global injection drug use epidemic on HIV prevention and treatment as well as to assess the potential of using digital technology and artificial intelligence (AI) to reduce injection drug use and HIV transmission.

## 2 Epidemiology of HIV Among PWID

Each year approximately 13 million people inject drugs worldwide [3]. In the United States, the recent opioid crisis can be traced back to the early 1990s from over-prescription of opioid medications. Among patients who were prescribed opioids for chronic pain, an estimated 21–29% misused them, 8–12% subsequently developed opioid use disorder, and 4–6% transitioned to using heroin [8–10]. Misuse of such medications, along with heroin and fentanyl use has contributed to a countrywide epidemic that mostly impacts white, rural America. In 2016, an

estimated 2.1 million people suffered from an opioid use disorder, with 886,000 using heroin. In 2017, of the 70,237 deaths due to drug overdose, 47,600 were related to opioid use [11].

Injection drug use is in turn associated with higher risk of HIV infection. Globally, roughly 1.7 million people who inject drugs (PWID) are also infected with HIV [3]. Previous studies estimate that PWID have a 22-fold risk of acquiring HIV (UNAIDS) [12]. In January 2015, an outbreak of HIV infection among PWID in rural Indiana caused approximately 4400 new HIV diagnoses. Epidemiologic studies showed that the outbreak began in 2011, became rampant in mid-2014, and that early transmission was associated with condomless sexual activity and injection drug use. The study emphasized the need for early and sustained efforts to detect infections, and prevent or interrupt rapid transmission within networks of uninfected PWID [13].

### **3 HIV Care Continuum in PWID – Challenges to Prevention and Treatment**

The World Health Organization (WHO) recommends initiation of antiretroviral therapy (ART) for all people living with HIV, regardless of disease stage. ART not only provides clinical benefit to the individual living with HIV, but also successful viral suppression that is paramount in preventing transmission [14, 15]. The United Nations Program on HIV and AIDS (UNAIDS) 90-90-90 approach set the goal that by 2020, 90% of all people living with HIV will know their HIV status, 90% of people with an HIV diagnosis will receive ART, and 90% of people receiving ART will achieve viral suppression, thus pioneering a global initiative to combat the HIV epidemic [16].

In spite of the global push for raising awareness about prevention and treatment of HIV, a plethora of challenges remain, particularly concerning PWID. HIV-positive PWID are not only subject to structural barriers preventing access to care, but they are also burdened with additional challenges posed by their substance use disorder, including stigma, homelessness, being underinsured or lacking insurance altogether, concomitant need for treatment for their substance use disorder that is often denied due to criminalization of drug use, HIV-specific criminal laws and divide between HIV care and OAT [17].

A study in New York City showed that more than half of HIV transmissions occurred among PWID, who were unaware of their HIV status. More than one-third of those HIV transmissions were due to known HIV-infected PWID who were not receiving ART. The findings highlighted the aforementioned for PWID to maintain viral suppression [18]. Female drug users are among the most marginalized groups that experience excessive stigmatization and disparities in access to treatment. Indeed, PWID, especially female drug users have hardly benefitted from ART and neither are they actively participating in HIV care [19].

Evidence-based studies have asserted the efficacy of combining harm reduction programs with HIV treatment in prevention and disease progression. A recent controlled trial examined the utility of a comprehensive treatment approach encompassing OAT and ART therapy, along with social interventions (e.g. psychosocial counseling, and monetary assistance applied towards transportation and medication costs). The study showed high retention rates and adherence to OAT and ART treatment in the intervention arm, demonstrating the need for a holistic approach to the management of HIV-positive PWID [20]. Despite the results of this trial, and the proven efficacy of OAT on quality of life, mortality and HIV related transmission and treatment, only 8% of PWID receive OAT worldwide [21]. In Malaysia, buprenorphine and methadone were introduced in 2002 and 2005 respectively, yet only about 10% of PWID benefit from these therapies [22, 23]. Similarly, needle and syringe exchange program (NSEP) coverage is low according to WHO indicators (<100 needle-syringes distributed per PWID per year; <20 OAT recipients per PWID per year). At the global level, less than 1% of PWID live in countries with high coverage of both NSEP and OAT (defined as >200 needle-syringes distributed per PWID and >40 OAT recipients per 100 PWID). These areas also lack adequate HIV and HCV prevention interventions for PWID [24].

In addition to lack of access to harm reduction programs, there's a problem of coordination between harm reduction initiatives and HIV continuum of care. In cases where OAT and NSEP programs are available, access may not translate into HIV testing and ART initiation. A study from India found that only 42.3% of PWID engaged in harm reduction programs reported prior HIV testing, while only 57.9% of those ART eligible started therapy [25]. In the US, HIV screening among PWID is also suboptimal, with only 58% of those affected having received HIV testing [26]. Furthermore, even though PrEP is acceptable among PWID, awareness and uptake remain low [27]. Roth et al. showed that of the PWID surveyed, only 12.4% of PWID were PrEP-aware, while only 2.6% reported receiving a prescription [28].

Concomitant OAT and ART treatment is also challenging from a pharmacological standpoint. The pharmacokinetics (PK) of antiretrovirals (ARVs), direct-acting antivirals (DAAs), and addiction treatment medications are all influenced by a number of factors, which make drug-drug interactions of ARVs, DAAs, buprenorphine and methadone difficult to manage. With infections, such as HIV and HCV, changes to physiological transporters and metabolic functions ensue, increasing the need for pharmacologic modeling techniques to allow for the prediction of drug PK in specific organs and the plasma compartment [29]. Recent studies have examined the effect of single-nucleotide polymorphisms (SNPs) for the gene encoding the delta-opioid receptor, and the association of the genetic effect with buprenorphine treatment outcome in men and women. Treatment outcomes based on SNP variants were found to be worse in women [30].

HIV status disclosure poses another set of challenges. Although disclosure is associated with increased social support and protective behaviors against HIV transmission, lack of disclosure in the face of persistent societal stigma, remains a contributing factor to disease transmission. Nasarruddin found that although most PWID (64–86%) disclose their HIV-positive status to trusted individuals (i.e. family

members and intimate sexual partners), disclosure to non-intimate sexual partners and fellow injection drug users is relatively lower. In such cases, the protective effects of disclosure in terms of risk reduction continue to be effaced by prevailing stigma [31].

In light of the aforementioned challenges, it is important to develop interventions that increase access and promote adherence. Governments can play a role by choosing treatment of PWID over criminalization, to reduce stigma and promote positive dynamics between law enforcement, the judiciary and PWID. They can create flexible policies that facilitate OAT access and resource re-allocation, as accomplished in some middle-income countries such as Malaysia and Ukraine [21, 32]. For example, in the Ukraine, incarceration contributed to up to half of all new HIV infections among PWID. The government intervened by making OAT available in prisons and offering post-imprisonment continuity of care and access to OAT, which ultimately led to a reduction of HIV transmission within this population. Other countries such as Armenia, Kyrgyzstan, and Moldova have successfully introduced the UN HIV prevention strategies including OAT with methadone and NSEP, albeit more needs to be done [33]. Increasing the availability of OAT programs with adequate supervision is paramount, as well as increasing awareness and access to PrEP treatment. It is also critical to enhance the HIV continuum for female drug users by promoting social, economic and legal policies that overhaul the many years of discrimination and stigmatization faced by female drug users worldwide.

## 4 Newer Approaches to HIV Prevention in PWID

### 4.1 *Digital Technologies and HIV Prevention in PWID*

As PWID bear a significant portion of the global HIV burden, increased prevention, treatment and care of this population is of utmost importance. Some of the programs that have been suggested focus on NSEP, OAT, HIV testing and counseling, HIV treatment and care, condom programs, behavioral interventions, prevention and treatment of comorbid infections and mental health conditions, sexual health interventions, and overdose prevention [3]. Yet these efforts are insufficient, for reasons discussed earlier in the chapter. Integration of more innovative approaches is needed in order to maximize the impact of these interventions.

Social media is a promising approach to overcome barriers to prevention and treatment of HIV. Increasing technological sophistication and internet accessibility makes technology-based interventions more cost-effective and easily scalable after initial production costs. It bolsters expansion of clinic-based strategies beyond individual care and can have a positive impact on hard-to-reach populations, especially those with stigmatized behaviors in healthcare settings, such as PWID and MSM. Digital platforms ensure consistency of content delivery and facilitate intervention by minimally trained personnel [34, 35].

Globally, 70% of youth (ages 15–24) are online, with mobile cellular subscriptions having reached 7 billion and mobile broadband subscription growing at a 20% annual rate [36]. Media platforms provide a sustainable and cost-effective method for reaching hidden populations, bridging communication among a wide range of users in various geographic and social contexts, disseminating health information and promoting awareness and medication adherence, providing social support, as well as intervening in the setting of risky behaviors [37]. This strategy is not challenge free, with one of the main concerns being the need for anonymity and confidentiality in communication about HIV prevention and treatment because of HIV-related stigma [37]. Dissemination and implementation also remain as looming challenges, including the lack of resources for the high costs of design and programming, issues of government monitoring and censorship of internet content especially in regions where homosexuality is punishable by death, along with the funding timelines that falter behind the acceleration of technological advances [34]. Yet, it remains important to determine how to harness the power of social media and related technologies to understand HIV transmission dynamics, and strategically target vulnerable, high-risk groups with appropriate messaging [38].

Social media interventions have been used more to promote safe sexual behaviors, HIV testing than safe drug use. A randomized controlled trial using Facebook friendship networks showed increased safe sex and injection practices, along with high HIV testing uptake following exposure to sexual health messaging [39]. Another study employed a live-chat platform to target substance use among young MSM in the US. The results showed an overall decrease in drug and alcohol consumption, paired with an increase in condom use [40]. Other digital media interventions have similarly shown delayed initiation of sex, higher prevalence of condom self-efficacy and abstinence attitudes, increased awareness of HIV and other sexually transmitted infections, and more widespread discussions around family planning [35]. Examples of digital media interventions include: [HealthMpowerment.org](#) which particularly targets African American youth at risk of acquiring HIV; SiHLEWeb, a web-based, culturally-tailored HIV/STI prevention program (for young African American females); as well as HIV Prevention England (HPE), which campaigned for condom use and HIV testing via several social medial platforms such as Facebook, Twitter, Grindr, Gaydar, etc.) [41]. Novel eHealth interventions also create opportunities for health promotion along the continuum of HIV care and prevention. They can be used on multiple platforms such as smartphones or social media, with individualized approaches and real-time assessments. Using eHealth, mHealth and “Web 2.0” social media strategies can effectively reach and engage key populations in HIV prevention across the testing, treatment, and care continuum [34].

The HOPE Study in particular showed encouraging results through a peer-led HIV prevention *initiative* using social media (Facebook and Facebook Messenger) targeting African American and Latino MSM [42]. In this study, peer leaders were trained using a tailored curriculum composed of discussion and role-playing exercises that highlight basic knowledge of HIV/AIDS, awareness of contemporary sociocultural HIV/AIDS challenges in the digital age, and effective communication training for peer leaders while using interactive social media-based HIV prevention techniques [42]. Ethical issues related to Facebook and health interventions are discussed throughout the sessions. This curriculum is available to use as a health promotion tool using social networking sites such as Facebook. More efforts should be made to utilize these sites for HIV prevention in broader risk groups [42].

## 4.2 Phylogenetic and Phylodynamic Analyses

Phylogenetics allows the study of evolutionary relationships between species. Phylogenetic analysis has successfully used sequencing data from available national drug resistance data to delineate evolutionary relationships between HIV viral strains associated with onward transmission of HIV in diverse geographic settings [43]. Brenner et al. emphasize the value of combining viral phylogenetics—which tracks the link between viral variants—and epidemiological data to trace the disease and emergent transmission networks real-time at the population-level [43]. This is of special benefit to identify outbreaks among high-risk groups that are unaware of their HIV status and addressing drug resistance. Additionally, phylogenetic analysis can identify patterns in emerging drug resistance against the 23 anti-retroviral drugs used in HIV-1 management in low-, middle- and high-income settings, as well as patterns in transmitted drug-resistant viral strains [43]. The authors make a compelling argument for the use of integrase inhibitors given the increasing resistance to current first-line regimens in LMICs [43].

Most recently, the CDC and other partners used phylodynamics, to understand HIV transmission patterns during an outbreak of undiagnosed HIV infections in rural Indiana [13]. The outbreak began in 2011 but was only recognized in 2015 [13]. Novel methods were employed that used demographic, behavioral and molecular data to generate transmission networks [13]. They built genetic distance networks by measuring distances between polymerase (*pol*) sequences of all patients diagnosed with an HIV infection and incorporated data on their reported high-risk contacts to generate transmission networks [13]. By integrating epidemiological and laboratory data, subgroups were identified that shared epidemiologic and genetic commonalities. The earliest infections were then traced to one of these sub-groups. Results from this phylodynamic analysis further suggest that the majority of the early infections were likely associated with sexual activity and then introduced into the injection drug use community [13]. These new approaches if used timely can inform early detection of outbreaks and prevent rapid transmission of HIV among PWID and other at-risk groups [13, 43].

### **4.3 Artificial Intelligence and HIV Prevention in PWID**

Artificial intelligence (AI) technologies such as machine learning (ML), have recently gained popularity for a number of medical applications. By using complex algorithms to recognize patterns in data, machine learning is being widely used as a prediction tool [44]. In HIV research, ML can be of particular value in making predictions from large amounts of data gathered from patient demographics, genetic and clinical markers, drug resistance patterns, ART adherence and treatment outcomes. As such, this section examines the utility of AI technologies as it applies to HIV prevention among PWID.

Machine learning is comprised of wide-ranging models and algorithms typically categorized into supervised and unsupervised learning [44]. A supervised learning approach involves the ability to predict an output variable based on a range of predictor/input variables. The data used are called labeled data and may consist of a wide array of information including text, biomarkers, DNA sequences, screening results and such [44]. Supervised learning approaches have been commonly used in medical research and practice and generally include multiple linear regression, artificial neural networks, decision trees, random forest, support vector machines, among others [44–48]. In contrast, unsupervised learning analyzes unlabeled data—no differentiation of input from the output data. Its goal is to identify patterns and correlations in the given data using clustering and dimensionality reduction, although its computational complexity remains a major disadvantage [49].

Random forest (RF), a machine learning method has been used to predict factors influencing receipt of HIV testing among participants from a substance use disorder treatment program. Pan et al., used data from 1281 participants who were either HIV-negative or of unknown status from 12 US community-based substance use disorder treatment programs as part of a National Institute on Drug Abuse Clinical Trials Network HIV testing and counseling study (CTN-0032) [50]. Participants were randomized into three treatment groups. RF analysis allowed the assessment of 109 factors in the prediction model. RF also delivered much higher prediction accuracy than logistic regression [50]. Treatment group was the key predictor among all covariates of receiving HIV testing. Participants reporting a higher number of condomless sex acts were less likely to receive their HIV test results compared to those who reported use of condoms or no sexual activity [50]. Random forest is a promising analytic tool to cull the most important predictors from a large number of covariates efficiently and may be valuable in future HIV behavioral research evaluating prevention and treatment interventions [50, 51].

Some interventions have used AI technologies to help HIV-infected individuals build self-efficacy and communication skills to support HIV status disclosure to an intimate partner. HIV disclosure for young MSM, a group with the highest-risk for HIV is a very challenging social and emotional experience because of the negative impact of stigma and discrimination [52]. Muessig et al. designed the Tough Talks Virtual Reality (VR) program using VR characters for young HIV-positive MSM to practice HIV status disclosure in a safe and confidential manner [52]. Through the emerging VR character utterances and diverse disclosure dialogues developed from participants, three scenarios were created (neutral, sympathetic and negative) [52]. The AI system was trained with the diverse VR utterances and diverse disclosure dialogues. When using a fully automated disclosure scenario with neutral response, the AI system replied appropriately to 71% of participant utterances. Most study participants agreed the program was easy to use and would be willing to use the system frequently. This study was able to demonstrate the feasibility of using virtual reality with automation to practice HIV status disclosure and could be used with any high-risk group such as PWID who face similar stigma and discrimination [52].

Surveillance systems that capture large amounts of data through collaboration between health care providers, laboratories, and the population might benefit from the use of machine learning methods. One of the challenges encountered in such complex systems is the administrative delay in reports of diagnosed cases. A recent study of the Portuguese surveillance system utilized machine learning methodologies including multilayer artificial neural networks (MLP), naive Bayesian classifiers, support vector machines, and the k-nearest neighbor algorithm to identify factors associated with reporting delays [53]. The multilayer artificial neural network was the model with the highest classification accuracy among the aforementioned methodologies at 63%. Most cases had a time lag of less than 3 months for reporting the HIV diagnosis. Being heterosexual and injecting drugs were among the two leading risk factors; however, cases with reported delays were distributed among both groups. The authors concluded that the reduced classification accuracy of MLP was directly related to the poor quality of data entered into the input layer of neurons [53].

A summary of the aforementioned approaches is included in Table 1.

**Table 1** New approaches to HIV prevention in high-risk groups

New Approaches	Description of the approach	Application
Digital technology and social media	Bull et al. 2012 [39] used Facebook messaging networks as a medium for health education. The study found that exposure to sexual health messaging led to increased condom use. Lelutiu-Weinberger et al. 2015 [40] used a live-chat platform targeting substance use in young MSM. The study showed an overall decrease in drug and alcohol consumption, paired with an increase in condom use.  Guse et al. 2012 [35] reviewed several digital media interventions, including web-based, text messaging and social networking sites. They found delayed initiation of sex, higher prevalence of condom use, increased awareness about sexually transmitted infections, and more widespread discussions around family planning following implementation of selected interventions.  Jaganath et al. 2012 [42] trained peer leaders, through adaptation of the Community Popular Opinion Leader (C-POL) model, and an HIV/AIDS-centered curriculum, along with strategies for effective use of digital platforms for the promotion of health literacy and HIV prevention practices.  Tso LS et al. 2016 [41] conducted a review of additional social media interventions to prevent HIV, which included educational website development and targeted messaging through social medial platforms such as Facebook, Twitter, Grindr, and Gaydar.  Muessig et al. [34] reviewed novel eHealth interventions that can be used on multiple platforms such as smartphones or social media, with individualized approaches and real-time assessments.	Health promotion, HIV prevention, safe sex and injection practices  Drug use reduction, safe sex practices  Safe sex practices, family planning  Health education  HIV prevention  Health promotion, HIV prevention
Phylogenetic and phylodynamic analysis	Brenner et al. 2017 [43] combined viral phylogenetics and epidemiological data to trace the emergent HIV transmission networks real-time at the population-level. Phylogenetic analysis could also be used to identify drug resistance patterns to anti-retroviral therapy, in order to optimize first line treatment.  Campbell et al. 2017 [13] used demographic, behavioral and molecular data to generate genetic distance HIV transmission networks, identifying the principal source of the outbreak.	Early detection of HIV transmission and identification of ART resistance  Early detection of HIV outbreaks
Artificial intelligence	Pan et al. 2017 [50] used data from 1281 participants from a substance use disorder program to predict factors influencing receipt of HIV testing. Higher risk groups were found to be less likely to receive HIV testing.  Muessig et al. 2018 [52] developed the Tough Talks Virtual Reality (VR) program, which uses VR characters for young HIV-positive MSM to practice HIV status disclosure in a safe and confidential manner.  Oliviera et al. 2017 [53] used multilayer artificial neural network analysis to identify factors associated with reporting delays in HIV status.	Research  Promoting safe HIV status disclosure in the setting of stigma and discrimination  Research

## 5 Conclusion

The HIV epidemic among PWID is a significant on-going global health problem despite efforts in harm reduction (including NSEP, OAT) and care continuum improvements. Barriers to these endeavors include criminalization of drug use, discrimination against PWID, a dearth of access to harm reduction, lack of other health services and ineffective policies and laws. The use of newer technologies including AI to tackle this growing epidemic is promising. It can help to identify cases and pattern of HIV infection among PWID, connecting patients, modifying behaviors by various interventions, developing clinical applications and prediction of disease outcomes. Despite the potential of low-cost applications, these technologies are vastly premature and unaffordable to apply on a global scale. More data from future research studies and timely funding support will help improve and implement these technologies to translate research into clinical application in HIV prevention and management of PWID.

**Conflicts of Interest** The authors report no potential conflicts of interest.

## References

1. World Health Organization (WHO). HIV/AIDS. <https://www.who.int/news-room/fact-sheets/detail/hiv-aids>. Published 2018. Updated July 19, 2018. Accessed 15 Apr 2019.
2. Degenhardt L, Peacock A, Colledge S, et al. Global prevalence of injecting drug use and sociodemographic characteristics and prevalence of HIV, HBV, and HCV in people who inject drugs: a multistage systematic review. *Lancet Glob Health.* 2017;5(12):e1192–207.
3. World Health Organization (WHO). HIV/AIDS: People who inject drugs. <https://www.who.int/hiv/topics/idu/en/>. Published 2019. Accessed 15 Apr 2019.
4. Shapshak P, Kanguane P, Fujimura RK, et al. Editorial neuroAIDS review. *AIDS* (London, England). 2011;25(2):123–41.
5. Tyagi M, Bukrinsky M, Simon GL. Mechanisms of HIV transcriptional regulation by drugs of abuse. *Curr HIV Res.* 2016;14(5):442–54.
6. Tyagi M, Weber J, Bukrinsky M, Simon GL. The effects of cocaine on HIV transcription. *J Neurovirol.* 2016;22(3):261–74.
7. Csete J, Kamarulzaman A, Kazatchkine M, et al. Public health and international drug policy. *Lancet* (London, England). 2016;387(10026):1427–80.
8. Cicero TJ, Ellis MS, Surratt HL, Kurtz SP. The changing face of heroin use in the United States: a retrospective analysis of the past 50 years. *JAMA Psychiat.* 2014;71(7):821–6.
9. Carlson RG, Nahhas RW, Martins SS, Daniulaityte R. Predictors of transition to heroin use among initially non-opioid dependent illicit pharmaceutical opioid users: a natural history study. *Drug Alcohol Depend.* 2016;160:127–34.
10. National Institute on Drug Abuse. Opioid overdose crisis. <https://www.drugabuse.gov/drugs-abuse/opioids/opioid-overdose-crisis#nine>. Published 2019. Updated January 2019. Accessed 15 Apr 2019.
11. Scholl L, Seth P, Kariisa M, Wilson N, Baldwin G. Drug and opioid-involved overdose deaths – United States, 2013–2017. *MMWR Morb Mortal Wkly Rep.* 2018;67(5152):1419–27.

12. Joint United Nations Programme on HIV/AIDS (UNAIDS). Miles to go. Closing gaps, breaking barriers, righting injustices. 2018. [https://www.unaids.org/sites/default/files/media\\_asset/miles-to-go\\_en.pdf](https://www.unaids.org/sites/default/files/media_asset/miles-to-go_en.pdf). Accessed 16 Apr 2019.
13. Campbell EM, Jia H, Shankar A, et al. Detailed transmission network analysis of a large opiate-driven outbreak of HIV infection in the United States. *J Infect Dis.* 2017;216(9):1053–62.
14. Lundgren JD, Babiker AG, Gordin F, et al. Initiation of antiretroviral therapy in early asymptomatic HIV infection. *N Engl J Med.* 2015;373(9):795–807.
15. WHO. Guideline on when to start antiretroviral therapy and on pre-exposure prophylaxis. Geneva: World Health Organization.
16. Joint United Nations Programme on HIV/AIDS (UNAIDS). Fast-track: ending the AIDS epidemic by 2030. 2014. [https://www.unaids.org/sites/default/files/media\\_asset/JC2686\\_WAD2014report\\_en.pdf](https://www.unaids.org/sites/default/files/media_asset/JC2686_WAD2014report_en.pdf). Accessed 18 Apr 2019.
17. Centers for Disease Control (CDC). HIV among people who inject drugs. <https://www.cdc.gov/hiv/group/hiv-idu.html>. Published 2019. Updated March 15, 2019. Accessed 18 Apr 2019.
18. Escudero DJ, Lurie MN, Mayer KH, et al. The risk of HIV transmission at each step of the HIV care continuum among people who inject drugs: a modeling study. *BMC Public Health.* 2017;17(1):614.
19. Metsch L, Philbin MM, Parish C, Shiu K, Frimpong JA, Giangle M. HIV testing, care, and treatment among women who use drugs from a global perspective: progress and challenges. *J Acquir Immune Defic Syndr.* 2015;69(Suppl 2):S162–8.
20. Miller WC, Hoffman IF, Hanscom BS, et al. A scalable, integrated intervention to engage people who inject drugs in HIV care and medication-assisted treatment (HPTN 074): a randomised, controlled phase 3 feasibility and efficacy study. *Lancet (London, England).* 2018;392(10149):747–59.
21. Bachireddy C, Weisberg DF, Altice FL. Balancing access and safety in prescribing opioid agonist therapy to prevent HIV transmission. *Addiction.* 2015;110(12):1869–71.
22. Bruce RD, Govindasamy S, Sylla L, Haddad MS, Kamarulzaman A, Altice FL. Case series of buprenorphine injectors in Kuala Lumpur, Malaysia. *Am J Drug Alcohol Abuse.* 2008;34(4):511–7.
23. Degenhardt LMB, Wirtz AL, Wolfe D, Kamarulzaman A, Carrieri MP, Strathdee SAM-SK, Kazatchkine M, Beyer C. What has been achieved in HIV prevention, treatment and care for people who inject drugs, 2010–2012? A review of the six highest burden countries. *Int J Drug Policy.* 2014;25(1):8.
24. Larney S, Peacock A, Leung J, et al. Global, regional, and country-level coverage of interventions to prevent and manage HIV and hepatitis C among people who inject drugs: a systematic review. *Lancet Glob Health.* 2017;5(12):e1208–20.
25. Smith MK, Graham M, Latkin CA, Go VL. Using contact patterns to inform HIV interventions in persons who inject drugs in Northern Vietnam. *J Acquir Immune Defic Syndr.* 2018;78(1):1–8.
26. Tempalski B, Cooper HLF, Kelley ME, et al. Identifying which place characteristics are associated with the odds of recent HIV testing in a large sample of people who inject drugs in 19 US metropolitan areas. *AIDS Behav.* 2019;23(2):318–35.
27. Smith DK, Van Handel M, Grey J. Estimates of adults with indications for HIV pre-exposure prophylaxis by jurisdiction, transmission risk group, and race/ethnicity, United States, 2015. *Ann Epidemiol.* 2018;28:850–857.e9.
28. Roth A, Tran N, Piecara B, Welles S, Shinefeld J, Brady K. Factors associated with awareness of pre-exposure prophylaxis for HIV among persons who inject drugs in Philadelphia: national HIV behavioral surveillance, 2015. *AIDS Behav.* 2019;23(7):1833–40.
29. Bednarsz CJ, Venuto CS, Ma Q, Morse GD. Pharmacokinetic considerations for combining antiretroviral therapy, direct-acting antiviral agents for hepatitis C virus, and addiction treatment medications. *Clin Pharmacol Drug Dev.* 2017;6(2):135–9.

30. Clarke TK, Crist RC, Ang A, et al. Genetic variation in OPRD1 and the response to treatment for opioid dependence with buprenorphine in European-American females. *Pharmacogenomics J.* 2014;14(3):303–8.
31. Nasaruddin AM, Saifi RA, Othman S, Kamarulzaman A. Opening up the HIV epidemic: a review of HIV seropositive status disclosure among people who inject drugs. *AIDS Care.* 2017;29(5):533–40.
32. Thomson N, Moore T, Crofts N. Assessing the impact of harm reduction programs on law enforcement in Southeast Asia: a description of a regional research methodology. *Harm Reduct J.* 2012;9:23.
33. Altice FL, Azbel L, Stone J, et al. The perfect storm: incarceration and the high-risk environment perpetuating transmission of HIV, hepatitis C virus, and tuberculosis in Eastern Europe and Central Asia. *Lancet (London, England).* 2016;388(10050):1228–48.
34. Muessig KE, Nekkanti M, Bauermeister J, Bull S, Hightow-Weidman LB. A systematic review of recent smartphone, Internet and Web 2.0 interventions to address the HIV continuum of care. *Curr HIV/AIDS Rep.* 2015;12(1):173–90.
35. Guse K, Levine D, Martins S, et al. Interventions using new digital media to improve adolescent sexual health: a systematic review. *J Adolesc Health.* 2012;51(6):535–43.
36. International Telecommunication Union (ITU). ICT facts and figures 2017. 2017. <https://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2017.pdf>. Accessed 19 Apr 2019.
37. Taggart T, Grewe ME, Conserve DF, Gliwa C, Roman Isler M. Social media and HIV: a systematic review of uses of social media in HIV communication. *J Med Internet Res.* 2015;17(11):e248.
38. Latkin CA, Davey-Rothwell MA, Knowlton AR, Alexander KA, Williams CT, Boodram B. Social network approaches to recruitment, HIV prevention, medical care, and medication adherence. *J Acquir Immune Defic Syndr (1999).* 2013;63(Suppl 1):S54–8.
39. Bull SS, Levine DK, Black SR, Schmiege SJ, Santelli J. Social media-delivered sexual health intervention: a cluster randomized controlled trial. *Am J Prev Med.* 2012;43(5):467–74.
40. Lelutiu-Weinberger C, Pachankis JE, Gamarel KE, Surace A, Golub SA, Parsons JT. Feasibility, acceptability, and preliminary efficacy of a live-chat social media intervention to reduce HIV risk among Young men who have sex with men. *AIDS Behav.* 2015;19(7):1214–27.
41. Tso LS, Tang W, Li H, Yan HY, Tucker JD. Social media interventions to prevent HIV: a review of interventions and methodological considerations. *Curr Opin Psychol.* 2016;9:6–10.
42. Jaganath D, Gill HK, Cohen AC, Young SD. Harnessing Online Peer Education (HOPE): integrating C-POL and social media to train peer leaders in HIV prevention. *AIDS Care.* 2012;24(5):593–600.
43. Brenner BG, Ibanescu RI, Hardy I, Roger M. Genotypic and phylogenetic insights on prevention of the spread of HIV-1 and drug resistance in “real-world” settings. *Viruses.* 2017;10(1):10.
44. Bisaso KR, Anguzu GT, Karungi SA, Kiragga A, Castelnovo B. A survey of machine learning applications in HIV clinical research and care. *Comput Biol Med.* 2017;91:366–71.
45. Singh Y, Mars NNM. Applying machine learning to predict patient-specific current CD 4 cell count in order to determine the progression of human immunodeficiency virus (HIV) infection. *Afr J Biotechnol.* 2013;12(23):11.
46. Larder B, Wang D, Revell A. Application of artificial neural networks for decision support in medicine. *Methods Mol Biol (Clifton, NJ).* 2008;458:123–36.
47. Li Y, Rapkin B. Classification and regression tree uncovered hierarchy of psychosocial determinants underlying quality-of-life response shift in HIV/AIDS. *J Clin Epidemiol.* 2009;62(11):1138–47.
48. Munoz-Moreno JA, Perez-Alvarez N, Munoz-Murillo A, et al. Classification models for neuropsychological impairment in HIV infection based on demographic and clinical variables. *PLoS One.* 2014;9(9):e107625.
49. Choi I, Chung AW, Suscovich TJ, et al. Machine learning methods enable predictive modeling of antibody feature:function relationships in RV144 vaccinees. *PLoS Comput Biol.* 2015;11(4):e1004185.

50. Pan Y, Liu H, Metsch LR, Feaster DJ. Factors associated with HIV testing among participants from substance use disorder treatment programs in the US: a machine learning approach. *AIDS Behav.* 2017;21(2):534–46.
51. Isabelle Guyon AE. An introduction to variable and feature selection. *Mach Learn Res.* 2003;3:25.
52. Muessig KE, Knudtson KA, Soni K, et al. “I didn’t tell you sooner because I didn’t know how to handle it myself”: developing a virtual reality program to support HIV-status disclosure decisions. *Digit Cult Educ.* 2018;10:22–48.
53. Oliveira A, Faria BM, Gaio AR, Reis LP. Data mining in HIV-AIDS surveillance system: application to Portuguese data. *J Med Syst.* 2017;41(4):51.

# Innovative Technologies for Advancement of WHO Risk Group 4 Pathogens Research



James Logue, Jeffrey Solomon, Brian F. Niemeyer, Kambez H. Benam,  
Aaron E. Lin, Zach Bjornson, Sizun Jiang, David R. McIlwain,  
Garry P. Nolan, Gustavo Palacios, and Jens H. Kuhn

**Abstract** Risk Group 4 pathogens are a group of often lethal human viruses for which there are no widely available vaccines or therapeutics. These viruses are endemic to specific geographic locations and typically cause relatively infrequent,

---

J. Logue · J. H. Kuhn (✉)

Integrated Research Facility at Fort Detrick, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Fort Detrick, Frederick, MD, USA

Integrated Research Facility at Fort Detrick (IRF-Frederick), Division of Clinical Research (DCR), National Institute of Allergy and Infectious Diseases (NIAID), National Institutes of Health (NIH), Frederick, MD, USA

e-mail: [kuhnjens@mail.nih.gov](mailto:kuhnjens@mail.nih.gov)

J. Solomon

Integrated Research Facility at Fort Detrick, Clinical Monitoring Research Program Directorate, Frederick National Laboratory for Cancer Research sponsored by the National Cancer Institute, Fort Detrick, Frederick, MD, USA

B. F. Niemeyer

Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO, USA

K. H. Benam

Division of Pulmonary Sciences and Critical Care Medicine, Department of Medicine, University of Colorado, Anschutz Medical Campus, Aurora, CO, USA

Department of Bioengineering, University of Colorado Denver, Aurora, CO, USA

A. E. Lin

Harvard Program in Virology, Harvard Medical School, Boston, MA, USA

FAS Center for Systems Biology, Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA

Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, MA, USA

Z. Bjornson · S. Jiang · D. R. McIlwain · G. P. Nolan

Baxter Laboratory for Stem Cell Biology, Department of Microbiology and Immunology, Stanford University School of Medicine, Stanford, CA, USA

G. Palacios

United States Army Medical Research Institute of Infectious Diseases, Fort Detrick, Frederick, MD, USA

self-limiting, but often devastating human disease outbreaks (e.g. Ebola virus, Kyasanur Forest disease virus, Lassa virus). The overall rarity of disease outbreaks with the associated lack of clinical data and the requirement for research on Risk Group 4 pathogens to be performed in maximum (biosafety level 4) containment necessarily impede progress in medical countermeasure development. Next-generation technologies may aid to bridge the current gaps of knowledge by increasing the amount of useful data that can be gleaned from individual diagnostic samples, possibly even at point-of-care; enable personalized medicine approaches through genomic virus characterization in the clinic; refine our comprehension of pathogenesis by using *ex vivo* technologies such as organs-on-chips or organoids; identify novel correlates of protection or disease survival that could inform novel medical countermeasure development; or support patient and treatment response monitoring through non-invasive techniques such as medical imaging. This chapter provides an overview of a subset of such technologies and how they may positively impact the field of Risk Group 4 pathogen research in the near future.

**Keywords** AI · Artificial intelligence · Biosafety level 4 · BS-4 · CODEX · CRISPR · CyTOF · *In silico* · Medical imaging · MIBI · Next-generation sequencing · Organoid · Organs-on-chips · Pathology · Risk group 4 · Single-cell sequencing · Third generation sequencing · Transparent animals

## 1 Introduction

World Health Organization (WHO) Risk Group (RG-) 4 pathogens are a relatively small group of high-consequence viral pathogens that can cause serious or life-threatening disease in humans or other animals and for which effective medical countermeasures (MCMs) are usually not available [1]. Handling replicative forms of these pathogens typically requires maximum (biosafety level 4 [BSL-4]) containment facilities (Table 1), of which there are only around three dozen globally [2]. Notorious examples of RG-4 pathogens include viruses that are associated with acute disease outbreaks, such as Ebola virus (EBOV), which recently caused a human disease outbreak in Western Africa encompassing more than 28,000 cases and more than 11,000 deaths [3] or Lassa virus, which in 2018 infected  $\approx$ 1,500 people in Nigeria [4, 5], and viruses that cause temporally isolated small case clusters, such as Kyasanur Forest disease virus (9,594 human infections from 1957 to 2017) [6]. Several of these viruses are considered potential source material for the development of biological weapons [7, 8] and are therefore considered research priorities within national public health and biodefense programs [9, 10]. Accelerated and increasingly focused efforts to develop MCMs for the prevention and/or treatment of RG-4 pathogens are undertaken to alleviate potential community suffering and the associated socioeconomic impact.

**Table 1** Examples of Risk Group 4 pathogens requiring maximum (biosafety level 4) containment in the US [32]

Family	Virus (abbreviation)
<i>Arenaviridae</i>	Chapare virus (CHAPV)
	Guanarito virus (GTOV)
	Junín virus (JUNV)
	Lassa virus (LASV)
	Lujo virus (LUJV)
	Machupo virus (MACV)
	Sabiá virus (SBAV)
<i>Filoviridae</i>	Bundibugyo virus (BDBV)
	Ebola virus (EBOV)
	Marburg virus (MARV)
	Ravn virus (RAVV)
	Sudan virus (SUDV)
	Taï Forest virus (TAFV)
<i>Flaviviridae</i>	Alkhurma hemorrhagic fever virus (AHFV)
	Kyasanur Forest disease virus (KFDV)
	Omsk hemorrhagic fever virus (OHFV)
	Tick-borne encephalitis virus (TBEV)
<i>Herpesviridae</i>	Herpes B virus (BV)
<i>Nairoviridae</i>	Crimean-Congo hemorrhagic fever virus (CCHFV)
<i>Paramyxoviridae</i>	Hendra virus (HeV)
	Nipah virus (NiV)
<i>Poxviridae</i>	Variola virus (VARV)

Research on RG-4 pathogens generally involves cell culture methods (*in vitro*) with live viruses at BSL-4 or surrogate systems (e.g., minigenomes, virion-like particles, recombinant expression of individual viral proteins, virion pseudotyping) at BSL-2/3 and/or animal models (*in vivo*) to investigate viral pathogenesis and host responses to infection. Cell culture has been used as a simple tool to research specific aspects of viral infection, such as screening candidate therapeutics for antiviral activity [11–21] or quantifying host immune responses, such as virus-neutralizing antibody titers [22, 23]. However, as current common cell culture methods cannot model the complexities of a body system, animal BSL-4 (ABSL-4) models, such as rodents or nonhuman primates, are generally used to model disease [24–28].

Next to increased security measures to prevent unauthorized entry or agent misuse or theft [29–31], (A)BSL-4 laboratories have multiple layers of redundant safety precautions, including positive-pressure suits, Class III biological safety cabinets, and validated methods to inactivate pathogens to protect the laboratory worker from accidental, and potentially lethal, infections [32–35]. The enhanced regulatory and biosafety environment can encumber research physically and limit the talent pool of researchers that are permitted to work at the (A)BSL-4 facilities.

Therefore, technological advancements in RG-4 pathogens research become especially important to maximize data output and lessen the time required for research.

Additionally, research performed at field laboratories in outbreak and virus-endemic zones with permission from internal review boards in the affected countries has provided important insights into disease course associated with human RG-4 pathogen infections. Monitoring of patients has refined pathogenic key events, including serum chemical and hematological value aberrations during disease, thereby providing guidance to clinicians and researchers of disease progression and disease model development, respectively [36–41]. Other research has focused on MCM testing for some RG-4 pathogens and occasionally shown considerable promise in clinical trials or ring vaccinations [42, 43]. However, as RG-4 pathogen disease outbreaks often occur in underdeveloped countries and/or geographically remote areas, research can be hampered by limited access to resources, transportation, or a skilled and local technician pool. For these reasons, advancements in research tools and the simplification of test methodology could help to alleviate the challenges associated with performing research.

## 2 Medical Imaging

### 2.1 *Infectious Disease Imaging and Artificial Intelligence*

Advanced imaging modalities, such as computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET), single photon emission computed tomography (SPECT), and ultrasound (US), are being developed at one active BSL-4 facility to study immune and other host system responses to infection with RG-4 pathogens [44–47]. Imaging has the advantage of being non-invasive and can detect signs of infectious disease at earlier timepoints than through clinical signs alone. Findings from qualitative radiology reports can be quantified according to standardized methods, including longitudinal image registration and organ/lesion segmentation, which measure morphological and physiological changes due to disease. However, these quantitative methods often require time consuming manual tracing of regions of interest, and tracings are subjective. Therefore, development of automated methods is needed to decrease time requirements, reduce variability, and increase accuracy, and such methods are now on the horizon.

For instance, artificial intelligence (AI) has considerable promise in the advancement of the medical imaging field. AI algorithms that learn from data can be unsupervised or supervised (e.g., “deep learning”), with the latter algorithm trained prior to use on large pools of data. However, imaging data alone are not sufficient to train supervised deep learning algorithms of neural networks as imaging data must be labeled (e.g., pneumonia vs. normal x-ray, lesions vs. normal tissue) for proper algorithm identification. This labeling process is often quite time-consuming and prone to human error.

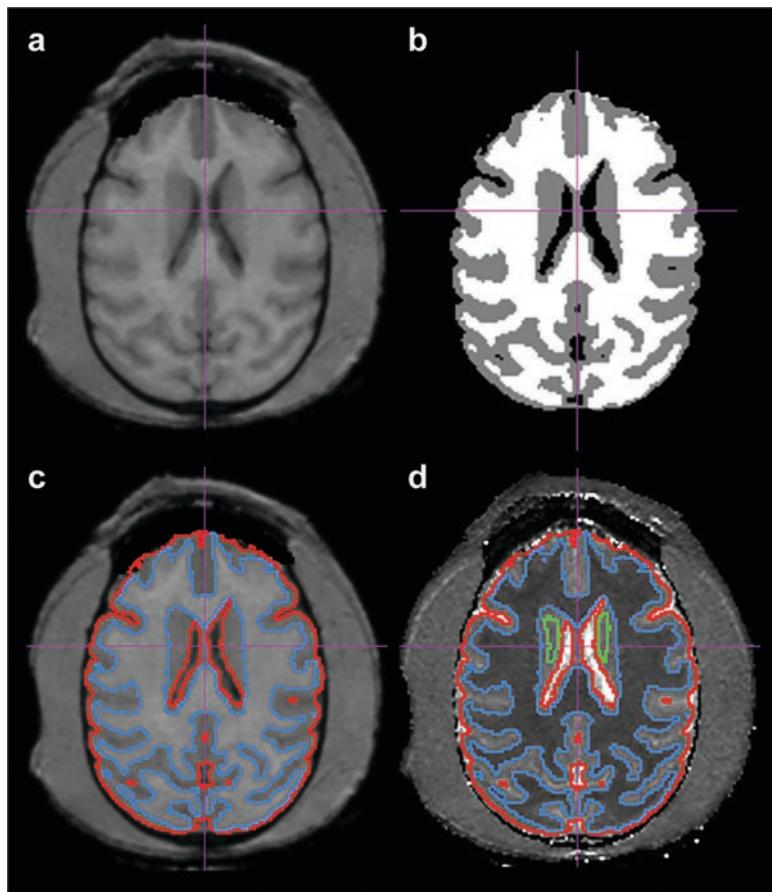
To avoid these errors, another field of AI, natural language processing, utilizes text-based reports generated by radiologists to associate findings with images for training of deep learning algorithms [48]. Alternatively, if the quantity of data are insufficient to train AI algorithms, data augmentation methods, such as rotation, horizontal flips, or random crops can be performed [49]. In addition, a neural network architecture called generative adversarial networks (GANs) can be used to generate synthetic images [50]. Because acquisition of adequate training data for all pathologic imaging phenotypes can be challenging, a one-class classification neural network has been developed that trains only on normal images and can detect abnormal images [51].

A type of deep learning neural network called a convolutional neural network integrates image feature extraction within artificial neural networks containing many hidden layers to both classify and segment images. These deep learning neural networks can be designed in various ways and, currently, the VGG16, GoogleNet, Inception4, Inception\_Resnet are popular architectures for image classification, whereas 2D/3D U-net and V-net are popular architectures for image segmentation [52, 53].

## 2.2 *Infectious Disease Imaging and Artificial Intelligence in a BSL-4 Environment*

In one active BSL-4 facility, medical imaging can be performed on RG-4 pathogen-infected animals within containment [44–47]. Whereas AI is not yet used to analyze all available imaging modalities, an unsupervised AI algorithm called “fuzzy c-means clustering” is already utilized in MRI to segment brain tissues into grey matter, white matter, and cerebrospinal fluid. In combination with digital brain atlas registration, the brain is robustly segmented into multiple sub-regions that are co-located over longitudinal scans (Fig. 1).

MR images are composed of a three-dimensional grid of voxels that are assigned signal values throughout the image. Since voxel values in MRI scans have arbitrary units (unlike CT scans, which are characterized by pixel values that are directly proportional to the density of the imaged tissue), parametric maps, which assign a quantity to each voxel, are created from multiple MRI sequences. Parametric maps to provide images, for example, in the form of T1 and T2 relaxometry maps [54]. These physical quantities, which represent the relaxation of proton spins in tissue, change with tissue composition. As an example, if brain edema results from a viral infection, the T1 values will increase due to the addition of fluid in the tissue. The regions delineated with “fuzzy c-means clustering” and “atlas-based registration” can be directly applied to these parametric maps, and changes in disease status, such as accumulation of fluid or blood in the brain, can be predicted from abnormal findings. MRI of the brain of experimental animals will likely be exceptionally useful to refine the sequence of pathogenetic events in diseases caused by encephalitic RG-4



**Fig. 1** Brain Segmentation. (a) Synthetic T1-weighted axial MRI scan of a nonhuman primate brain. (b) “Fuzzy c-means clustering” analysis with voxels classified as cerebrospinal fluid (black), grey matter (grey), and white matter (white). (c) Contours representing grey matter (red) and white matter (blue) overlaid on the original MRI scan. (d) Contours of grey and white matter and digital atlas-based contour of caudate (green) overlaid on the quantitative T1 map

pathogens, such as Hendra virus, Nipah virus, or tick-borne encephalitis virus. For example, brain MRI scans of patients infected with Nipah virus have shown acute encephalitis with multiple lesions visible using T2-weighted and fluid attenuated inversion recovery (FLAIR) images [55, 56]. These same MR imaging sequences performed in human studies may be performed and refined in animal models using clinical MR scanners within the BSL-4 environment.

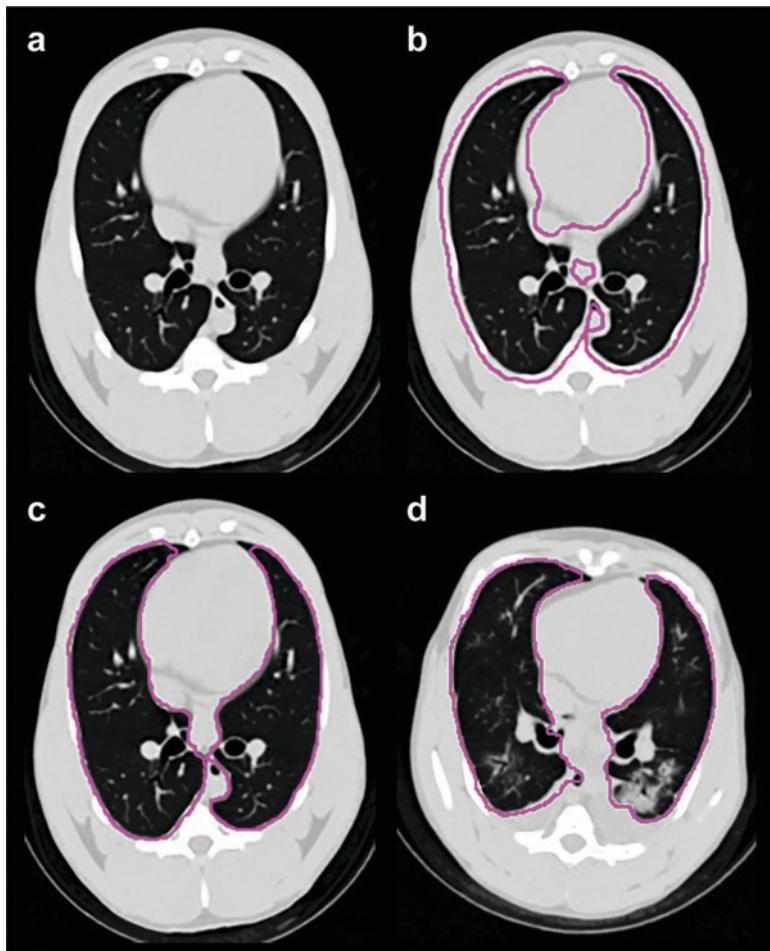
Many RG-4 pathogens diffusely affect multiple organs within the human body, including lung abnormalities caused by Nipah virus infection [57], liver damage, and disseminated intravascular coagulopathy during EBOV infection [58]. Therefore, segmenting liver, kidney, spleen, lungs, and lymph nodes in collected images is needed to detect structural or physiologic changes. Deep learning convolutional

neural networks can automatically segment multiple abdominal organs in CT and MRI modalities using the processes described above. Given this robust segmentation, radiometric features, such as texture and histogram analysis of voxels within the region of interest, can be used to classify stages of the disease process and correlate these features with clinical parameters, such as liver enzyme concentrations or virus titers. For instance, a method to quantify lung abnormalities in various disease models is now available [59]. Initially, the lungs are automatically segmented from a chest CT scan. Segmenting the lung field without any pathologic condition is done with standard image analysis methods, such as region growing, which starts with the initialization of a seed point within the region of interest and includes areas in the vicinity of the seed point based on whether the signal intensity is within a given threshold. This iterative process ends when values of neighboring voxels are not within the threshold. However, the region growing algorithm will fail in the initial lung segmentation when hyper-dense voxels in the lung are outside the threshold, requiring manual correction. Recently, an AI method was established to fully automate the lung field segmentation process when the lungs contain hyper-dense pathological areas [60]. This deep learning algorithm was trained with thousands of normal and abnormal human CT lung images to segment images of lung fields. Post-processing with morphological operators such as erosions (removal of areas) and dilations (inclusion of areas) is then needed to finalize the segmentation process. By modifying the post-processing parameters, accurate lung field segmentation was achieved in a nonhuman primate (Fig. 2), whereas the lung field was originally overestimated.

Current methods of analyses also involve correlations between imaging biomarkers and clinical measures, such as cytokine profiles, viral DNA/RNA concentrations, and blood composition testing. In the future, AI methods could be applied to integrate the collected imaging and clinical data to generate predictive models of disease outcome. Machine-learned features in images may be used to predict abnormal status prior to clinical manifestations of these abnormalities even though they may not be visible to the human eye. Integrating imaging and non-imaging measures may predict survival and efficacy of novel vaccines or therapies.

## 2.3 Molecular Imaging Probe Development

Molecular imaging is used to gain an understanding of cellular and molecular status compared with anatomic imaging, such as standard CT and MRI, which provide structural information on a larger scale. Molecular probes, such as fluorodeoxyglucose (<sup>18</sup>F-FDG; a marker of cellular glycolytic activity), can indicate increased cell metabolic activity in organs during infection with RG-4 pathogens [61]. Unfortunately, probes such as <sup>18</sup>F-FDG are not entirely specific, and development of agent/disease-specific probes is urgently needed. Examples of more specific probes that have been investigated to study host responses in infectious disease imaging include PET fluorine radioisotopes such as fluoro-thymidine (<sup>18</sup>F-FLT) [62],



**Fig. 2** Lung field segmentation. **(a)** Axial slice of nonhuman primate chest CT scan with no apparent lung abnormalities. **(b)** Contour of the lung field segmented by a “deep learning” algorithm trained on human CT scans (lung field is over-estimated). **(c)** Contour of a nonhuman primate lung field more accurately segmented after modification with post-processing methods. **(d)** Accurately segmented nonhuman primate lung field from a CT scan with apparent lung abnormalities

fluorine-18 radio-labeled serum albumin ( $^{18}\text{F}$ -albumin) [63], and  $^{18}\text{F}$ -*N,N*-diethyl-2-[4-(2-fluoroethoxy)phenyl]-5,7-dimethylpyrazolo[1,5-*a*]pyrimidine-3-acetamide ( $^{18}\text{F}$ -DPA)-714 [64].  $^{18}\text{F}$ -FLT can be used to investigate cellular proliferation during cancers such as lymphoma or during infectious diseases.  $^{18}\text{F}$ -albumin can be used to detect vessel leakage.  $^{18}\text{F}$ -DPA-714 is a selective ligand for the translocator protein (TSPO) to investigate over-expression of activated macrophages and serves as a biomarker for neuroinflammation. This marker could prove useful in the study of

disease caused by RG-4 pathogens such as Ebola virus as macrophage activation and increased vessel leakage are key pathogenic events of Ebola virus disease. In contrast to host-specific probes such as those mentioned above, probes that can attach to virions or reporter-encoding open reading frames that can be inserted into a RG-4 pathogen genome could directly localize a virus to specific areas of the body. Development of such reporter viruses has just begun. For instance, a gene encoding the solute carrier family 5 member 5 (SLC5A5, aka sodium/iodide symporter) was inserted into the Middle Eastern respiratory syndrome coronavirus genome (a RG-3 pathogen) and resulted in viable virus [65]. In the future, SLC5A5 and other imaging reporters could be incorporated into a variety of viral vectors to obtain *in vivo* visualization of location and aid in the evaluation of vaccine and therapeutic development. However, a major hurdle to overcome is virus attenuation after reporter gene insertion.

### 3 Pathology: Tissue and Pathogen Imaging

System-wide responses required to overcome exposure to RG-4 pathogens involve complex interactions between resident tissue cells and infiltrating immune cells, yet the identification of specific cell types in tissue sections is hindered by the limitations of traditional immunofluorescence. Spectral overlap of fluorophores typically restricts immunofluorescence studies to a maximum of around four antibody channels, thereby precluding simultaneous identification of multiple highly specialized cell types and invading pathogens in a single tissue section. Though the development of multiple multiplexed imaging modalities [66–69] has been vital in overcoming these limitations, we will focus on only a few new advancements in pathological imaging.

#### 3.1 Fluorescence-Based Multiplexed Tissue Imaging Tools

A new technique called CO-detection by indexing (CODEX) bypasses the limits of immunofluorescent antibody channels by using antibodies labeled with indexed DNA tags. With this technology, a cocktail of upwards of 50 DNA-indexed antibodies can stain a tissue section prior to iterative fluorescent visualization cycles to assemble a single 50+ parameter image [70]. CODEX is a highly effective multiplexing technique because a single antibody binding step eliminates much of the signal degradation that would otherwise be associated with stripping and re-staining of antibodies. The commercially available CODEX instrument automatically exchanges buffers needed to accomplish iterative imaging cycles. This instrument has a relatively small footprint and may be practical for use inside BSL-4 containment or after optimization of reagents to use with inactivated samples in RG-4 pathogen studies.

### 3.2 Metal Tag-Based Multiplexed Tissue Imaging Tools

Another technique called multiplexed ion beam imaging (MIBI) utilizes secondary ion mass spectrometry to generate high-dimensional images through mass spectrometry analysis of lanthanide-labeled antibodies on a pixel-by-pixel level [71]. This commercially available technology has thus far been leveraged for deep spatial understanding of archival breast cancer tissues [72]. A key feature of metal-tagged tissue imaging is the highly stable nature of the isotopes. Labeled samples can be archived theoretically indefinitely, for instance allowing reacquisition of target sample regions after analysis or reimaging with higher resolution instruments years later. In the MIBI workflow, inactivated tissues (e.g., formalin-fixed paraffin-embedded [FFPE]) are processed following conventional immunohistochemistry (IHC) protocols with the exception of the antibody cocktail. Routine tissue staining consists of 40 or more lanthanide-tagged antibodies, compared to the conventional one or two antibodies in IHC. A parallel method, termed Imaging Mass Cytometry (IMC), utilizing laser ablation coupled to a cytometry by time of flight (CyTOF) mass cytometer is also commercially available [73]. The antibodies and reagents for sample preparation are mostly cross-compatible.

### 3.3 Pathogen Detection in Tissue Sections

Current methods for the detection of pathogens in tissues can be divided into (1) antibody-based detection and (2) nucleic acid (NA)-based detection. Antibody-based methods (IHC) are severely limited by the availability of specific antibodies clones and by the conservation of the targeted epitope. Although NA-based methods, such as *in situ* hybridization (ISH), are ideal for identification of sequence-specific targets, these methods also have disadvantages, such as necessary signal amplification of targets, challenging experimental protocols, and complex probe design to achieve specificity and sensitivity. These disadvantages have largely been circumvented by the development of next generation ISH methods and probe design software [74–81]. RNAscope, an example of next-generation ISH, has been successfully implemented for the detection of RG-4 viruses (e.g., MARV, EBOV [82, 83]). RNAscope has also been used to follow single-integration events of simian immunodeficiency virus in tissues [76], demonstrating the sensitivity of the technology. Alternative enzymatic-based, virus-specific ISH has also been adapted for the surveillance of hepatitis delta virus [76, 84]. Currently, RNAscope as a method has demonstrated an ability to work on the IMC for the detection of highly abundant copies of RNA in FFPE tissues [81]. However, even with advances in these technologies, IHC- and ISH-based methods continue to suffer from the limits to the number of markers that can be examined simultaneously even if they are combined.

Future work to couple multiplexed imaging techniques (e.g., IMC, MIBI, CODEX) with sensitive ISH technologies will be instrumental for the mechanistic

dissection of virus-infected cells in the context of their tissue microenvironment and of viral reservoirs in the broader environment. Such coupling will increase understanding of the dynamics of viral infection and replication in RG-4 pathogen studies.

### ***3.4 Pathological Imaging in Transparent Animal Models***

Whereas the current and futuristic technologies described earlier are useful for targeted imaging of tissue sections, sectioning of tissues makes it difficult to unbiassedly investigate pathogen distribution and the effects of infection on an organism as a whole. Sectioning is required, however, as mammalian tissues are naturally opaque, impeding any imaging deeper into the tissue than a few hundred micrometers [85]. To combat this limitation, researchers have begun developing chemical methods to render tissues transparent (tissue clearing), including entire adult mouse bodies following skin removal [86–90], and image these transparent tissues optically. For example, tissue clearing was used in combination with an antibody signal-boosting technique to produce high-resolution neuronal projection maps of adult mouse brains [91]. Additionally, these techniques were used in conjunction to detect cancer metastases in a transparent mouse and assess therapeutic antibody targeting of these cancer cells at the single-cell level [92].

Though yet to be realized for RG-4 pathogens, research using the combination of tissue clearing and optical imaging could be used to investigate RG-4 pathogen infection, host response, and treatment efficacy. Optical reporters, such as fluorescent proteins inserted into RG-4 pathogen genomes [19, 93–95] or optically labeled antibody systems that bind to pathogens [96–100] or specific immune cells (e.g., antibodies for flow cytometry), can be used to investigate pathogen or immune cell distribution throughout the host at varying time-points during disease. Additionally, treatment efficacy could be assessed against RG-4 pathogens similarly to the cancer treatment assessment described previously [92].

## **4 Cell Marker Analysis in Solution**

### ***4.1 Multiplexed Analysis of Cells in Solution***

Flow cytometry analysis of single cells in solution has been the cornerstone of advances in our understanding of immune responses to infection over the past few decades. However, spectral overlap of the fluorescent marker-tagged antibodies used in flow cytometry limits simultaneous examination of a large number of cellular features. Replacing these fluorescent tags with metal ion antibody tags enabled the development of CyTOF, which overcomes traditional multiplexing limitations in blood or other dissociated cell profiling [101]. CyTOF has been

applied to monitor immune responses to disease and vaccination [102]. Expanding the multiplexing capabilities of single-cell measurements offers exponential returns for profiling the complexity of immune cells [103]. Whereas flow cytometry is rarely performed with more than 12 parameters, limiting analysis to a subset of cell types or signaling readouts, CyTOF comfortably identifies 40 or more parameters on single cells. A single antibody panel can identify all major immune cell subsets in a blood sample, in addition to quantifying activation status, cytokine production, or signaling states of each of those cell types. In the case of RG-4 pathogen studies, sample volumes are frequently in limited supply as they are either sourced from rare human disease outbreaks or from the relatively few experimentally infected animals in the limited number of BSL-4 facilities. A fringe benefit of highly multiplexed technologies, including CyTOF, is that these small samples can now produce a greater number of measurements for evaluation of more hypotheses simultaneously. For instance, the gap between existing experimental and clinical data for RG-4 pathogens might be narrowed by carefully planned CyTOF studies in which rare patient samples are examined simultaneously for multiple cellular features observed in experimental models.

The dividends of CyTOF in shedding light on infectious disease are underscored by the findings of a number of recent studies [104–107], but use of CyTOF has yet to be realized for RG-4 pathogen research. One of the challenges for use of CyTOF in RG-4 studies is the currently unsuitable design of components of the CyTOF instrument for operation inside maximum containment environments (compressed gas and high-volume exhaust requirements, glassware, and superheated components). To address this challenge, workflows are currently under development that will enable CyTOF analysis of virus-inactivated samples derived from RG-4 pathogen studies. A set of CyTOF reagents for direct comparison of immune system responses in humans, laboratory mice, and non-human primate animal models frequently used in RG-4 pathogens research is already available [108, 109]. Similar reagents should be considered for development of guinea pig, hamster, ferret, and other models. This expanding toolset will likely contribute towards a framework for future in-depth RG-4 pathogen studies across different animal species, thereby also informing the choice of which animal model to use for divergent scientific questions.

## 4.2 Computational Tools for Analysis of High-Throughput and High-Dimensional Data

CyTOF, single-cell sequencing, and high-dimensional imaging all yield large datasets that are time-consuming to analyze exhaustively. The same general analysis principle applies to most of these datasets: partition of cells by phenotype and subsequent analysis of their functions, behaviors and/or relationships. A large number of tools have been developed or adapted from other fields to perform these tasks in automated or semi-automated fashions [110]. In RG-4 pathogen research, a key benefit of computational tools for multiparameter single-cell data is the possibility

to identify biomarkers consisting of unanticipated combinations of parameters that would be missed by manual approaches (e.g., gating of cytometry data).

In many cases, tools used for CyTOF can be directly applied to segmented single-cell data from multi-parameter imaging. However, the addition of spatial position to multiplexed single-cell data created from imaging results in an enormously higher depth of information. Tools to address how the structure of cellular neighborhoods within tissues impact health and disease are now being developed [70, 72, 111].

## 5 Virus and Patient Sequencing

Widespread adoption of next-generation sequencing (NGS) has revolutionized virtually every facet of molecular biology and human health, including the study of RG-4 pathogens. Determination of the first human genome sequence took a decade and was performed predominantly by Sanger technology [112, 113]. Since then, NGS instrumentation has blossomed, and NGS data output grows exponentially every year. Generating 20 billion reads in a single sequencing run is now possible. Each currently available NGS platform has advantages and disadvantages [114–116], but large sequencing centers now regularly generate a single human genome sequence every few minutes using many of these platforms. In contrast to the Sanger method, NGS can sequence millions of DNA strands at the single-molecule level, which also allows researchers to obtain accurate sequence information from smaller genomes such as RG-4 pathogens. In the future, extensively cataloguing human and pathogen genomic variation will provide better insight into host-pathogen interactions and thereby enable personalized medicine approaches even in exotic disease outbreaks.

Researchers studying human biology have devised upstream workflows that take advantage of the ability of NGS to sequence millions of reads and then mine these data for answers to new scientific questions. Those seeking to study RG-4 pathogens can leverage most of these NGS-based assays. Though replication-competent RG-4 pathogens must be handled in BSL-4 facilities, NGS protocols can be performed at a lower BSL (e.g., BSL-2) if samples were appropriately inactivated and their nucleic acids (NAs) extracted. From a plethora of NGS applications, we describe three broad categories of NGS-based assays: (See Section 5.1) detecting unknown NAs; (See Section 5.2) marking unknown features of NAs (structure, modified bases) and mapping their locations; and (See Section 5.3) quantifying biomolecules in a sample (See Section 5.1).

### 5.1 *Detection of Unknown Pathogens*

Metagenomic NGS (mNGS) is a powerful tool for identifying pathogens in clinical or environmental samples [117, 118]. Diseases caused by many RG-4 pathogens are generally challenging to diagnose, as patients with these diseases often present with

non-specific (“influenza-like”) clinical signs and even highly replicative viruses typically comprise a minority (<1%) of the NAs in a sample. Rather than testing for every pathogen individually, scientists can use mNGS to sequence millions of molecules from all NAs in a sample, revealing any low frequency NAs (from pathogens or non-pathogenic organisms). Supplemental methods for manipulating NAs can further enhance sensitivity and cost-effectiveness of mNGS, including multiplex polymerase chain reaction (PCR) [119–121], hybrid capture [122, 123], and clustered regularly interspaced short palindromic repeats (CRISPR)-based methods [124, 125], in ways that are not currently possible for other biomolecules like proteins or lipids.

Recent uses of mNGS for RG-4 pathogens includes analyzing the sequences from the 2013–2016 Ebola virus disease (EVD) epidemic in Western Africa [126–132], the two most recent EVD outbreaks in the Democratic Republic of the Congo [133, 134], and recent Lassa fever outbreaks in Nigeria [4, 5]. In each case, multiple research groups collaborated with African partners to collect clinical samples, inactivate them with guanidinium-based reagents, extract viral RNA, reverse transcribe RNA to cDNA, and sequence by mNGS. In these cases, as specific causal pathogens were suspected, it was possible to perform multiplex PCR to enrich (concentrate) EBOV or LASV content and then to sequence on the universal serial bus (USB)-sized Oxford Nanopore minION device in the field [120]. In contrast, other groups have used non-targeted amplification methods [126], which can reveal intriguing co-infection dynamics between pathogens of interest and other pathogens that would normally be removed by the enrichment process (e.g., EBOV and *Plasmodium* [135] or EBOV and GB virus C [136]).

Decisions on appropriate public health responses can also be greatly informed by the collection and cataloguing of hundreds or thousands of viral genome sequences during an outbreak. Molecular epidemiology (i.e., use of viral sequencing data to identify disease transmission chains) can provide key insights on outbreak information such as animal-to-human or human-to-human transmission [4, 5, 126, 137], instances of suspected “super spreading” [138], and transmission from persistently infected disease survivors [139–141]. Sequencing data also inform models of factors that influence outbreak scale and severity [132] and/or viral evolutionary rate [142], and facilitate identification of high frequency mutations that have functional impact on viral infectivity [143, 144].

## 5.2 Mapping Features of Nucleic Acid Sequences

In addition to identifying unknown sequences, NGS can also map the locations of unknown functional features of NAs, such as epigenomic and epitranscriptomic characteristics including interactions between nucleic acids and proteins or complex secondary and tertiary structure formations [145, 146]. Though these features have been more thoroughly studied in human NAs, viral NAs also fold into complex

structures for replication [147] and/or harbor modifications that can dampen immunity [148].

Though physical methods such as crystallography and nuclear magnetic resonance remain powerful tools for elucidating NA form and function, chemical and enzymatic methods aid in identifying features in a high-throughput manner when coupled with NGS [149]. Enzymes that cleave specific NA features or chemicals that modify specific bases can terminate reverse transcription or produce errors at structured regions [150] or biomarkers [151, 152] that are sensitively and accurately quantified by NGS. Though evaluation of NA features of RG-4 pathogens have been less frequent, these studies can be very informative. One study used NGS to map the RNA structure of an EBOV minigenome and identified that the trailer non-coding region bound to heat-shock protein A8 promoted minigenome replication [153]; better understanding of host-virus interactions essential for replication could identify new targets for MCMs. Though live EBOV was not used in this study, another group used a similar NGS approach for RG-2/3 viruses, Sindbis virus and Venezuelan equine encephalitis virus. Although it was expected that the genomes of these two alphaviruses fold into different RNA structures, it is noteworthy that these two structures directly led to differing viral infectivity [154]. In addition to these technologies, “third generation sequencing” technologies (see Section 5.5) have the potential to facilitate sequencing of DNA and RNA base modifications directly [155–157], which can distinguish unmodified from modified bases by measuring changes in electrical currents.

### 5.3 Biomolecule Quantification

Aside from identifying and characterizing virus genomes directly, NGS can also be used to read DNA and DNA barcodes for a range of functional assays ranging from large screens to single-molecule and single-cell sequencing. Human genome scientists have developed genome-wide protein knock-out or knock-down (e.g., CRISPR, RNA interference, small hairpin RNA, small molecule) screens and massively parallel reporter assays [158–160] to screen thousands of DNAs simultaneously. For example, researchers generated a vesicular stomatitis Indiana virus expressing LASV glycoprotein [161, 162]. They then created cells with randomly knocked-out genes using a retroviral gene-trap vector and exposed these cells to the recombinant virus to identify host entry factors for LASV. By NGS of retroviral insertion sites of uninfected cells, they found that genes critical for glycosylating α-dystroglycan (the major LASV cell-surface receptor) were also required for LASV entry during infection.

One promising technology that is starting to be used for virology is single-cell RNA sequencing (scRNA-seq) [163–167]. During scRNA-seq, individual cells are isolated, and unique DNA barcodes are applied to each cell’s RNA followed by NGS to associate each RNA with its cognate cell. Home-brewed methods like

Drop-seq [168] and commercial options like 10X Genomics are becoming increasingly popular because scientists can profile thousands of cells in mixed populations, such as peripheral blood mononuclear cells (PBMCs) or even dissociated tissues. Researchers can also use scRNA-seq to measure heterogeneity in viral replication. For instance, scientists sequenced 3,000–4,000 single cells infected at low multiplicity of infection with influenza A virus (FLUAV). Most cells contained <1% FLUAV mRNA, but a number of cells had ≈50% FLUAV reads, indicating extreme heterogeneity of infection, partly attributed to variability of the FLUAV replicative machinery [164]. However, as Drop-seq and 10X Genomics are droplet-based and require specialized equipment, platforms of alternative methods that are microchip- [169] or microwell-based [170], such as Seq-Well [171, 172], are advantageous. These platforms are portable, have minimal equipment requirements, and can be easily decontaminated and discarded. A new technique, Slide-seq, allows spatially resolved single-cell RNA sequencing after transferring RNA from tissue sections onto a new surface covered in DNA-barcoded beads. Using Slide-seq, cell types and their activation states can be directly determined using standard histological work-up [173]. Seq-Well and related technologies could therefore be used in BSL-4 laboratories or in the field, thereby facilitating functional and single-cell studies for RG-4 pathogens.

## 5.4 *Databases and Bioinformatics for Sequencing*

The new wave of NGS technologies has spurred new public databases and bioinformatic tools that comprehensively and quickly analyze millions of short reads or thousands of long reads. Because RG-4 pathogens are relatively rare, NGS data generation and sharing are critical. In the US, the NIH supports the National Center for Biotechnology Information's sequence read archive for raw sequencing data, GenBank for consensus sequences from humans and viruses, and a range of other databases for processed data. The NIH also supports the Virus Pathogen Resource [174], which collates sequence data and experiments from host factor assays. Smaller data portals like virological.org and nextstrain.org [175] also exist and can rapidly disseminate pre-publication data. Advances in algorithms and computing power, described in other chapters of this book, will certainly facilitate searching [176], classifying [177], and processing these massive data sets. These data will require advanced modeling methods in conjunction with basic molecular biology and public health efforts.

## 5.5 “Third Generation Sequencing” Methods

Some of the newest NGS technologies, often dubbed third-generation sequencing, have been driven by nanotechnology and possess unique properties that further expand the molecular biology toolkit [114–116]. In contrast to Illumina, Roche 454,

and Ion Torrent short-read methodologies that involve cleaving genomic material and sequencing-by-synthesis in cycles ( $\leq 600$  bases per read), Pacific Biosciences and Oxford Nanopore methodologies both rely on nanoscale pores to processively read entire DNA strands, producing reads up to hundreds of kilobases long. Though the error rate is often much higher than that of short-read methodologies, long reads are particularly useful for *de novo* genome assembly, i.e., the assembly of an unknown genome sequence [178, 179] and reconstruction of large haplotypes with multiple mutations or variants [179]. Nanoscale pores also possess other unique benefits. Using Pacific Biosciences technology, a researcher can continuously sequence the same DNA strand in a circle, creating a circular consensus sequence, and thus reducing the error rate [180, 181]. Using Oxford Nanopore technology, a researcher can sequence RNA directly [155–157] and also identify DNA and RNA base modifications [182]. Additionally, the direction of the pores can also be reversed [183–185], and the same DNA strand is read twice for an improved consensus sequence. Perhaps the biggest advantages of Oxford Nanopore devices are their small sizes, typically resembling a USB drive, and the ease of their use [5, 120]. Continued developments in nanotechnology will further reduce instrumentation size and sample requirements and improve the error rate and selectivity of NGS.

At present, a Star Trek-style tricorder device for universal diagnosis remains the unobtainable, yet holy grail for assigning etiological agents to fevers of unknown origin is within sight. In many cases, access to technology, rather than the technology itself, is the limiting factor, and point-of-care devices are increasingly sought. For example, paper-based lateral flow assays are standard for pregnancy and antibody/antigen testing [186]. Numerous isothermal methods are under development as alternatives to PCR and are continuously improved because DNA/RNA tests offer a complementary approach for detecting pathogens [187–189]. NGS equipment, in particular Oxford Nanopore, is already portable and has been utilized during numerous outbreaks around the world [120, 121, 190, 191] and even in space aboard the International Space Station [192].

## 6 Disease Modeling

Accurate model systems are of the utmost importance for studying normal physiology and pathobiology of human diseases. For the past several decades, animal models have been the standard systems used to emulate human disease processes, with conventional two-dimensional (2D) *in vitro* systems complementing animal models by reducing system complexity and increasing throughput. If no suitable animal model is yet identified, research is often limited to *in vitro* systems. However, no one model system is truly capable of reproducing the complex biological processes observed in humans. Translating findings from animal models to human subjects can be quite challenging as large biological differences, altered disease severity, and altered susceptibility to pathogens exists between humans and other animals [193–195]. In addition, conventional 2D *in vitro* systems often only recreate cell-cell

interactions and fail to maintain the complexity of tissue-tissue and organ-organ communication, which is of critical importance to disease processes *in vivo*.

The following sections highlight the use of organs-on-chips and organoids as models of complex disease states, as screening platforms for new biomarkers, and as advantageous systems for the study of infectious diseases and therapeutic interventions. Whereas the examples outlined below primarily represent work with RG-2 pathogens, organs-on-chips and organoids could relatively easily be utilized for the study of RG-4 agents. Importantly, with these systems, the study of highly pathogenic agents using human tissues in a complex, dynamic setting could closely resemble relevant *in vivo* systems.

## 6.1 *Organs-on-Chips*

To bridge the gap between animal models and basic *in vitro* systems, advances in microengineering and microfluidics were channeled to create organ-on-chip technology. Organs-on-chips are microfluidic cell-culture devices, fabricated using soft lithography from inert, gas permeable, polymers [196]. These biomimetic systems recreate tissue-tissue interfaces and biophysical properties of organs, including mechanical torsion (e.g., cyclic “breathing” motion associated with expansion and contraction of alveolar and capillary interfaces) and shear force from blood flow [196–198].

Originally used to recreate the lung alveolus, organs-on-chips have been adapted to recreate the human small airway, liver, intestine, kidney, bone, blood vessels, bone marrow, neuronal tissue, cardiac muscle, and cornea [199]. Given their flexibility in design, well-defined architecture, and wide range of sources for cellular materials, organs-on-chips represent an excellent and adaptable model system to study a wide array of diseases, including chronic obstructive pulmonary disease (COPD), asthma, liver disease, cardiovascular disease, and malignancies [200–205].

Another growing application of this technology is modeling the pathogenesis of infectious diseases. In particular, research in the area of respiratory infections has been greatly propelled using lung-on-a-chip and small airway-on-a-chip technologies [200, 201, 206–210]. For instance, a small airway-on-a-chip was used to model respiratory infection through use of a toll-like receptor 3 agonist, poly-inosinic-poly-cytidylic acid (poly-IC), thereby mimicking cellular events during viral infection of lung epithelial cells [201]. This model replicated complex disease states, such as viral exacerbation of disease in patients suffering from COPD and asthma, and helped identify new potential biomarkers for COPD exacerbation, such as macrophage colony-stimulating factor [201]. Meanwhile, additional lung model systems have been used to study fungal and bacterial infections in the lung. For instance, a multi-compartment human bronchiole was created to investigate the production of inflammatory cytokines during colonization with an eurotiomycete (*Aspergillus fumigatus*) and a gammaproteobacterium (*Pseudomonas aeruginosa*) [211]. Interestingly, this work showed that colonization of

the artificial bronchiole with less virulent *A. fumigatus* strain results in increased production of inflammatory cytokines and recruitment of leukocytes, a finding that would be less likely made if the experiments were performed on cell monolayers *in vitro* [211]. Moreover, inflammatory cytokine production differed when the bronchioles were exposed to volatile compounds produced from co-cultures of *P. aeruginosa* and *A. fumigatus* compared to monocultures of either microbe [211].

As described above, lung-on-chip and related models have been widely used to study respiratory infections [200, 201, 206–211]. However, organ-on-chip technology is not restricted to the lung and has been applied to study infection of other organ systems including the liver, the central nervous system, and the intestine. For instance, primary human hepatocytes were used in combination with organ-on-chip technology to facilitate the study of hepatitis B virus infection *in vitro* [212]. The hepatitis B virus life cycle and host immune responses (e.g., cytokine responses) to infection were successfully modeled. In addition, cutting-edge micro-extrusion three-dimensional (3D) printing techniques were adapted to develop a “3D nervous system-on-a-chip” for the study of viral infections of the central nervous system [213]. Using this system, it was found that Schwann cells were refractory to pseudorabies virus infection, but that these cells still nevertheless participated in axon-to-cell spread of the virus. Additionally, infection with a human enterovirus, coxsackievirus B1, was successfully modeled using a human gut-on-chips. The virus infected and replicated within intestinal epithelium and stimulated inflammatory cytokine release in a polarized fashion [214].

Organ-on-chip systems are highly applicable for bridging gaps left between animal and 2D *in vitro* models, particularly with respect to the drug discovery process. During the drug discovery process, the highest leading cause of candidate drug attrition during clinical trials are failures in drug efficacy and safety [215, 216]. Organ-on-chip platforms provide more biologically complex environments that are better suited than conventional 2D systems for testing drug activities prior to clinical trials [199, 217]. Further, by mimicking several of the complex characteristics of whole organ systems, use of organs-on-chips can help reduce the use of animal models in the drug discovery process, which would serve to reduce the cost of the drug discovery process.

## 6.2 *Organoids*

Organs-on-chips are clearly advantageous for modeling human infectious disease. However, these chips are not the only advanced micro-engineered system suitable for this task. Organoids are 3D organ structures consisting of organ-specific cells grown from (induced, embryonic, or adult) stem cells via self-organizing mechanisms [218, 219]. One of the advantages of organoids is the capacity to mimic some of the complexities and functions of natural organs [219]. Given the structural and functional similarities to natural organs, organoids have been used extensively to study infectious disease with human samples [199, 217, 220]. Human respiratory syncytial virus, *Helicobacter pylori*, hepatitis C virus, and Zika virus

infection have all been modeled using organoids derived from the lung, gastric, liver, and nervous systems [220]. Recently, human airway organoids were used as a screening platform to study the infectivity of emerging FLUAVs [221]. FLUAV strains known to be highly infectious for humans were associated with higher replication rates in the organoids compared to strains known to poorly infect humans [221].

Similar to organ-on-chip technology, organoid utility surpasses simple disease modeling as they have been extensively used in the drug discovery process [217, 220]. A cerebral organoid was used to model Zika virus infection and to identify potential therapeutic compounds to abrogate damage associated with infection [222]. Three compounds that were previously identified as having protective effects during flavivirus infection (oxytetracycline, ivermectin, and azithromycin) limited Zika virus infection of the organoids and reduced tissue damage, suggesting these compounds may be good candidates for limiting Zika virus infection and associated damage *in vivo* [222].

Overall, while helpful in reproducing several salient features of tissue and organ pathophysiology, there is still room for improvement in both organoids and organs-on-chips. For instance, organoids often exhibit high variability in size and shape, do not experience naturally occurring mechanical forces (e.g., breathing airflow in the lung airways or rhythmic expansion-retraction of alveoli during inhalation-exhalation), and lack microvascular blood-like flow for circulation of immune cells and continuous nutrient and oxygen supply. In addition, accurately accessing the luminal content of organoids for biochemical analysis is challenging if not impossible. Integrating emerging genetic engineering tools such as CRISPR-associated protein 9 (Cas9) or transcription activator-like effector nucleic acid (TALENs) with organoids or organs-on-chips would be of high value. With this integration, researchers can introduce new sensitivity to pathogens (when absent) or to dissect underlying mechanisms of host protection or organ injury during host-pathogen interactions (e.g., through CRISPR-Cas9 mediated deletion of genes proposed to play roles in protection and susceptibility of infection with pathogens).

### **6.3 Improved Design of Experimental and In Silico Studies**

Technological advancements in disease modeling, including organs-on-chips and organoids, in part aim to increase experimental productivity and reduce the number of animals required to meet research goals. Computationally-aided improvements through design of experiments (DOE) is another approach with the potential to improve the efficiency and yield from studies. Although also significantly reducing the burden of research, improved efficiency is especially relevant given the expense and general difficulty associated with performing research in BSL-4 facilities. Although not a new field, DOE is generally applied to chemistry and pharmaceutical development more than biology, and its core concepts have recently been

enhanced by machine learning concepts [223]. For example, fractional factorial DOE could be used for multivariate drug screens to reduce the number of actual conditions required to be measured, without reducing the experimental yield. The same technique could possibly be used to reduce the number of animals required in *in vivo* studies. Inference-based and machine learning-based methods for drug repurposing are improving (reviewed in [224]). Given the low incidence of most RG-4 pathogen infections, drug repurposing is an important source of potential therapeutics.

## 7 Conclusions: Incorporating Futuristic Technologies into Risk Group 4 Research

Advanced research tools, such as those described here, are constantly under development and provide new and exciting ways to investigate RG-4 pathogens. The opportunity to further define aspects of disease pathogenesis and host response to infection can help to tease out new therapeutic and vaccine targets to combat these diseases. However, although some of these technologies already constitute a marked advancement in RG-4 pathogens research, their use in a BSL-4 environment comes with a few noteworthy complications.

BSL-4 laboratories require yearly maintenance, at a minimum, to replace air filters and to perform required servicing for laboratory infrastructure components. If these futuristic technologies are housed within the BSL-4 laboratory spaces, required machinery will have to be able to withstand repeated, yearly decontamination processes (e.g., paraformaldehyde gassing with subsequent neutralization, Microchem Plus surface decontamination) as the laboratory space is prepared for servicing. Additionally, as use of a BSL-4 facility requires extensive training and registration, outside technicians do not generally have the ability or permission to enter laboratory spaces and service instruments. Instruments, therefore, should not be too complex so that a BSL-4 scientist could troubleshoot them effectively. Alternatively, these instruments can also be housed in a lower biosafety level if test samples can be safely inactivated and removed from the BSL-4 laboratory. Researchers have used this method as part of multiple techniques, including many diagnostic tests that begin with an inactivation step and sample removal from containment before testing [225–227]. As technologies are not generally developed with inactivation methods in mind, a viral inactivation method for safe handling of samples in a lower BSL laboratory that maintains the integrity of sample components will need to be identified. After identification, the effects of these inactivation methods on test integrity (e.g., sample dilutions, signal loss) will require further study.

Additionally, RG-4 research is not only confined to the controlled environment of a BSL-4 laboratory space. Deployment of these technologies to outbreak regions provides the opportunity to further characterize human disease progression and correlates of disease outcome, without the caveats associated with disease models. Use of CyTOF and single-cell sequencing on patient samples collected during disease,

for example, could provide valuable data on immune responses that lead to disease survival. Additionally, as autopsies of deceased humans infected with RG-4 pathogens are rare, utilizing CODEX and MIBI could glean valuable information about human disease, even with a single autopsy. However, as outbreaks generally occur in developing countries, challenges to overcome may include lack of infrastructure (including electricity), transportation, and/or staffing. Newly developed technologies will need to be robust enough to counter changes in humidity, temperature, and many other potential stressors occurring in an outbreak setting. Given these complications, the futuristic technologies described in this chapter provide the opportunity to advance the understanding of highly pathogenic and consequential viruses and the diseases they cause.

**Acknowledgements** We thank Laura Bollinger and Jiro Wada (both NIH/NIAID Integrated Research Facility at Fort Detrick, Frederick, MD, USA) for critically editing the manuscript and figure development, respectively. This work was supported in part through Battelle Memorial Institute's prime contract with the US National Institute of Allergy and Infectious Diseases (NIAID) under Contract No. HHSN272200700016I (J.L., J.H.K.) and with federal funds from the National Cancer Institute (NCI), National Institutes of Health (NIH), under Contract No. HHSN261200800001 (J.S.) and by the US FDA under Contract No. HHSF223201610018C (D.R.M., Z.B., S.J., G.P.N.). The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the views or official policies, either expressed or implied, of the US Department of the Army, the US Department of Defense, the US Department of Health and Human Services, or of the institutions and companies affiliated with the authors. Mention of trade names, commercial products or services, or organizations does not imply endorsement by the U.S. Government. In no event shall any of these entities have any responsibility or liability for any use, misuse, inability to use, or reliance upon the information contained herein.

## References

1. World Health Organization. Laboratory biosafety manual. 3rd ed. Geneva: World Health Organization; 2004.
2. Rusek B, Sharples FE, Hottes AK. Biosecurity challenges of the global expansion of high-containment biological laboratories: summary of a workshop. Washington, DC: National Academies Press; 2011.
3. Bullard SG. A day-by-day chronicle of the 2013-2016 Ebola outbreak. Cham: Springer; 2018.
4. Siddle KJ, Eromon P, Barnes KG, Mehta S, Oguzie JU, Odia I, et al. Genomic analysis of Lassa virus during an increase in cases in Nigeria in 2018. *N Engl J Med.* 2018;379(18):1745–53.
5. Kafetzopoulou LE, Pullan ST, Lemey P, Suchard MA, Ehichioya DU, Pahlmann M, et al. Metagenomic sequencing at the epicenter of the Nigeria 2018 Lassa fever outbreak. *Science.* 2019;363(6422):74–7.
6. Chakraborty S, Andrade FCD, Ghosh S, Uelmen J, Ruiz MO. Historical expansion of Kyasanur forest disease in India from 1957 to 2017: a retrospective analysis. *GeoHealth.* 2019;3(2):44–55.
7. Borio L, Inglesby T, Peters CJ, Schmaljohn AL, Hughes JM, Jahrling PB, et al. Hemorrhagic fever viruses as biological weapons: medical and public health management. *JAMA.* 2002;287(18):2391–405.

8. Leitenberg M, Zilinskas RA, Kuhn JH. The Soviet biological weapons program: a history. Cambridge: Harvard University Press; 2012.
9. US Centers for Disease Control and Prevention. Bioterrorism agents/diseases. <https://emergency.cdc.gov/agent/agentlist-category.asp>. 2018.
10. US National Institute of Allergy and Infectious Diseases. NIAID emerging infectious diseases/pathogens. <https://www.niaid.nih.gov/research/emerging-infectious-diseases-pathogens>. 2016.
11. Postnikova E, Cong Y, DeWald LE, Dyall J, Yu S, Hart BJ, et al. Testing therapeutics in cell-based assays: Factors that influence the apparent potency of drugs. PLoS One. 2018;13(3):e0194880.
12. Bolken TC, Laquerre S, Zhang Y, Bailey TR, Pevear DC, Kickner SS, et al. Identification and characterization of potent small molecule inhibitor of hemorrhagic fever New World arenaviruses. Antiviral Res. 2006;69(2):86–97.
13. Larson RA, Dai D, Hosack VT, Tan Y, Bolken TC, Hruby DE, et al. Identification of a broad-spectrum arenavirus entry inhibitor. J Virol. 2008;82(21):10768–75.
14. Lee AM, Rojek JM, Spiropoulou CF, Gundersen AT, Jin W, Shaginian A, et al. Unique small molecule entry inhibitors of hemorrhagic fever arenaviruses. J Biol Chem. 2008;283(27):18734–42.
15. Radoshitzky SR, Kuhn JH, de Kok-Mercado F, Jahrling PB, Bavari S. Drug discovery technologies and strategies for Machupo virus and other New World arenaviruses. Expert Opin Drug Discov. 2012;7(7):613–32.
16. Lo MK, Shi P-Y, Chen Y-L, Flint M, Spiropoulou CF. In vitro antiviral activity of adenosine analog NITD008 against tick-borne flaviviruses. Antiviral Res. 2016;130:46–9.
17. Flint M, McMullan LK, Dodd KA, Bird BH, Khristova ML, Nichol ST, et al. Inhibitors of the tick-borne, hemorrhagic fever-associated flaviviruses. Antimicrob Agents Chemother. 2014;58(6):3206–16.
18. Focher F, Lossani A, Verri A, Spadari S, Maioli A, Gambino JJ, et al. Sensitivity of monkey B virus (*Cercopithecine herpesvirus 1*) to antiviral drugs: role of thymidine kinase in antiviral activities of substrate analogs and acyclonucleosides. Antimicrob Agents Chemother. 2007;51(6):2028–34.
19. Welch SR, Scholte FEM, Flint M, Chatterjee P, Nichol ST, Bergeron É, et al. Identification of 2'-deoxy-2'-fluorocytidine as a potent inhibitor of Crimean-Congo hemorrhagic fever virus replication using a recombinant fluorescent reporter virus. Antiviral Res. 2017;147:91–9.
20. Zivcec M, Metcalfe MG, Albariño CG, Guerrero LW, Pegan SD, Spiropoulou CF, et al. Assessment of inhibitors of pathogenic Crimean-Congo hemorrhagic fever virus strains using virus-like particles. PLoS Negl Trop Dis. 2015;9(12):e0004259.
21. Hotard AL, He B, Nichol ST, Spiropoulou CF, Lo MK. 4'-Azidocytidine (R1479) inhibits henipaviruses and other paramyxoviruses with high potency. Antiviral Res. 2017;144:147–52.
22. Robinson JE, Hastie KM, Cross RW, Yenni RE, Elliott DH, Rouelle JA, et al. Most neutralizing human monoclonal antibodies target novel epitopes requiring both Lassa virus glycoprotein subunits. Nat Commun. 2016;7:11544.
23. Saphire EO, Schendel SL, Fusco ML, Gangavarapu K, Gunn BM, Wec AZ, et al. Systematic analysis of monoclonal antibodies against Ebola virus GP defines features that contribute to protection. Cell. 2018;174(4):938–52 e13.
24. de Wit E, Munster VJ. Animal models of disease shed light on Nipah virus pathogenesis and transmission. J Pathol. 2015;235(2):196–205.
25. Mendoza EJ, Warner B, Safronetz D, Ranadheera C. Crimean-Congo haemorrhagic fever virus: past, present and future insights for animal modelling and medical countermeasures. Zoonoses Public Health. 2018;65(5):465–80.
26. Siragam V, Wong G, Qiu X-G. Animal models for filovirus infections. Zool Res. 2018;39(1):15–24.
27. Smith DR, Holbrook MR, Gowen BB. Animal models of viral hemorrhagic fever. Antiviral Res. 2014;112:59–79.

28. Zivcic M, Safronetz D, Feldmann H. Animal models of tick-borne hemorrhagic fever viruses. *Pathogens*. 2013;2(2):402–21.
29. Blaine JW. Establishing a national biological laboratory safety and security monitoring program. *Biosecur Bioterror*. 2012;10(4):396–400.
30. Shelby BD, Cartagena D, McClee V, Gangadharan D, Weyant R. Transfer of select agents and toxins: 2003–2013. *Health Secur*. 2015;13(4):256–66.
31. Henkel RD, Miller T, Weyant RS. Monitoring select agent theft, loss and release reports in the United States—2004–2010. *Appl Biosaf*. 2012;17(4):171–80.
32. US Department of Health and Human Services, US Centers for Disease Control and Prevention, US National Institutes of Health. Biosafety in microbiological and biomedical laboratories (BMBL). 5th ed. Washington, DC: HHS Publication No. (CDC) 93-8395, US Government Printing Office; 2009.
33. Bressler DS, Hawley RJ. Safety considerations in the biosafety level 4 maximum-containment laboratory. In: Wooley D, Byers K, editors. *Biological safety: principles and practices*. 5th ed. Washington, DC: American Society of Microbiology; 2017. p. 695–717.
34. Janosko K, Holbrook MR, Adams R, Barr J, Bollinger L, Newton JT, et al. Safety precautions and operating procedures in an (A)BSL-4 laboratory: 1. biosafety level 4 suit laboratory suite entry and exit procedures. *J Vis Exp*. 2016;116:e52317.
35. Mazur S, Holbrook MR, Burdette T, Joselyn N, Barr J, Pusl D, et al. Safety precautions and operating procedures in an (A)BSL-4 laboratory: 2. general practices. *J Vis Exp*. 2016;116:e53600.
36. Grove JN, Branco LM, Boisen ML, Muncy IJ, Henderson LA, Schieffelin JS, et al. Capacity building permitting comprehensive monitoring of a severe case of Lassa hemorrhagic fever in Sierra Leone with a positive outcome: case report. *Virol J*. 2011;8:314.
37. Hunt L, Gupta-Wright A, Simms V, Tamba F, Knott V, Tamba K, et al. Clinical presentation, biochemical, and haematological parameters and their association with outcome in patients with Ebola virus disease: an observational cohort study. *Lancet Infect Dis*. 2015;15(11):1292–9.
38. Ruzeck D, Avšič Županc T, Borde J, Chrdle A, Eyer L, Karganova G, et al. Tick-borne encephalitis in Europe and Russia: review of pathogenesis, clinical features, therapy, and vaccines. *Antiviral Res*. 2019;164:23–51.
39. Weigler BJ. Biology of B virus in macaque and human hosts: a review. *Clin Infect Dis*. 1992;14(2):555–67.
40. Mourya DT, Viswanathan R, Jadhav SK, Yadav PD, Basu A, Chadha MS. Retrospective analysis of clinical information in Crimean-Congo haemorrhagic fever patients: 2014–2015. *India Indian J Med Res*. 2017;145(5):673–8.
41. Goh KJ, Tan CT, Chew NK, Tan PS, Kamarulzaman A, Sarji SA, et al. Clinical features of Nipah virus encephalitis among pig farmers in Malaysia. *N Engl J Med*. 2000;342(17):1229–35.
42. Chowell G, Kiskowski M. Modeling ring-vaccination strategies to control Ebola virus disease epidemics. In: *Mathematical and statistical modeling for emerging and re-emerging infectious diseases*. Basel: Springer International Publishing; 2016. p. 71–87.
43. Kennedy SB, Bolay F, Kieh M, Grandits G, Badio M, Ballou R, et al. Phase 2 placebo-controlled trial of two vaccines to prevent Ebola in Liberia. *N Engl J Med*. 2017;377(15):1438–47.
44. Keith L, Chefer S, Bollinger L, Solomon J, Yellayi S, Seidel J, et al. Preclinical imaging in BSL-3 and BSL-4 environments: imaging pathophysiology of highly pathogenic infectious diseases. In: *Pharmaco-imaging in drug and biologics development, AAPS advances in the pharmaceutical sciences series*, vol 8. New York: Springer; 2014. p. 271–90.
45. Byrum R, Keith L, Bartos C, St Claire M, Lackemeyer MG, Holbrook MR, et al. Safety precautions and operating procedures in an (A)BSL-4 laboratory: 4. medical imaging procedures. *J Vis Exp*. 2016;116:e53601.
46. de Kok-Mercado F, Kutlak FM, Jahrling PB. The NIAID Integrated Research Facility at Fort Detrick. *Appl Biosaf*. 2011;16(2):58–66.
47. Jahrling PB, Keith L, St Claire M, Johnson RF, Bollinger L, Lackemeyer MG, et al. The NIAID Integrated Research Facility at Frederick, Maryland: a unique international resource

- to facilitate medical countermeasure development for BSL-4 pathogens. *Pathog Dis.* 2014;71(2):213–9.
48. Banerjee I, Ling Y, Chen MC, Hasan SA, Langlotz CP, Moradzadeh N, et al. Comparative effectiveness of convolutional neural network (CNN) and recurrent neural network (RNN) architectures for radiology text report classification. *Artif Intell Med.* 2019;97:79–88.
  49. Hussain Z, Gimenez F, Yi D, Rubin D. Differential data augmentation techniques for medical imaging classification tasks. *AMIA Annu Symp Proc.* 2017;2017:979–84.
  50. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial nets. In: Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ, editors. *Advances in neural information processing systems 27 (NIPS 2014)*. Montréal: Palais des Congrès de Montréal; 2014. p. 2672–80.
  51. Perera P, Patel VM. Learning deep features for one-class classification. *arXiv.* 2018;1801.05365. <https://arxiv.org/abs/1801.05365>
  52. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. *arXiv.* 2015;1409.556. <https://arxiv.org/abs/1409.1556>
  53. Ronneberger O, Fischer P, Brox T. U-net: convolutional networks for biomedical image segmentation. In: *Medical image computing and computer-assisted intervention – MICCAI 2015*. 9351: Springer, Cham; 2015. p. 234–41.
  54. Eminian S, Hajdu SD, Meuli RA, Maeder P, Hagmann P. Rapid high resolution T1 mapping as a marker of brain development: normative ranges in key regions of interest. *PLoS One.* 2018;13(6):e0198250.
  55. Sarji SA, Abdullah BJJ, Goh KJ, Tan CT, Wong KT. MR imaging features of Nipah encephalitis. *AJR Am J Roentgenol.* 2000;175(2):437–42.
  56. Lim CCT, Sitoh YY, Hui F, Lee KE, Ang BSP, Lim E, et al. Nipah viral encephalitis or Japanese encephalitis? MR findings in a new zoonotic disease. *AJNR Am J Neuroradiol.* 2000;21(3):455–61.
  57. Cong Y, Lentz MR, Lara A, Alexander I, Bartos C, Bohannon JK, et al. Loss in lung volume and changes in the immune response demonstrate disease progression in African green monkeys infected by small-particle aerosol and intratracheal exposure to Nipah virus. *PLoS Negl Trop Dis.* 2017;11(4):e0005532.
  58. Baseler L, Chertow DS, Johnson KM, Feldmann H, Morens DM. The pathogenesis of Ebola virus disease. *Annu Rev Pathol.* 2017;12:387–418.
  59. Solomon J, Douglas D, Johnson R, Hammoud D. New image analysis technique for quantitative longitudinal assessment of lung pathology on CT in infected rhesus macaques. In: *2014 IEEE 27th International symposium on computer-based medical systems*. Piscataway: Institute of Electrical and Electronics Engineers (IEEE); 2014. p. 169–72.
  60. Harrison AP, Xu Z, George K, Lu L, Summers RM, Mollura DJ. Progressive and multi-path holistically nested neural networks for pathological lung segmentation from CT images. *Med Image Comput Comput Assist Interv – MICCAI 2017 MICCAI 2017 Lect Notes in Comput Sci.* 2017;10435:621–9.
  61. Vaidyanathan S, Patel CN, Scarsbrook AF, Chowdhury FU. FDG PET/CT in infection and inflammation—current and emerging clinical applications. *Clin Radiol.* 2015;70(7):787–800.
  62. Shi KY, Li ZL, Yousefi B, Liu Z, Herz M, Huang SC, et al. Hierarchical dual-tracer modeling for rapid [18F] FLT and [18F] FDG PET scans on mice with lymphoma tumors. *J Nucl Med.* 2015;56(Suppl 3):374.
  63. Niu G, Lang L, Kiesewetter DO, Ma Y, Sun Z, Guo N, et al. In vivo labeling of serum albumin for PET. *J Nucl Med.* 2014;55(7):1150–6.
  64. Sinharay S, Papadakis G, Tu T-W, Kovacs Z, Reid W, Frank J, et al. Imaging neuroinflammation using 18F-DPA-714 PET after disrupting the blood brain barrier with pulsed focused ultrasound. *J Nucl Med.* 2017;58(Suppl 1):207.
  65. Chefer S, Seidel J, Cockrell AS, Yount B, Solomon J, Hagen KR, et al. The human sodium iodide symporter as a reporter gene for studying Middle East respiratory syndrome coronavirus pathogenesis. *mSphere.* 2018;3(6):e00540–18.

66. Zhang W, Hubbard A, Jones T, Racolta A, Bhaumik S, Cummins N, et al. Fully automated 5-plex fluorescent immunohistochemistry with tyramide signal amplification and same species antibodies. *Lab Invest*. 2017;97(7):873–85.
67. Lin J-R, Fallahi-Sichani M, Sorger PK. Highly multiplexed imaging of single cells using a high-throughput cyclic immunofluorescence method. *Nat Commun*. 2015;6:8390.
68. Schubert W, Bonnekoh B, Pommer AJ, Philipsen L, Böckelmann R, Malykh Y, et al. Analyzing proteome topology and function by automated multidimensional fluorescence microscopy. *Nat Biotechnol*. 2006;24(10):1270–8.
69. Gannot G, Tangrea MA, Erickson HS, Pinto PA, Hewitt SM, Chuaqui RF, et al. Layered peptide array for multiplex immunohistochemistry. *J Mol Diagn*. 2007;9(3):297–304.
70. Goltsev Y, Samusik N, Kennedy-Darling J, Bhate S, Hale M, Vazquez G, et al. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. *Cell*. 2018;174(4):968–81 e15.
71. Angelo M, Bendall SC, Finck R, Hale MB, Hitzman C, Borowsky AD, et al. Multiplexed ion beam imaging of human breast tumors. *Nat Med*. 2014;20(4):436–42.
72. Keren L, Bosse M, Marquez D, Angoshtari R, Jain S, Varma S, et al. A structured tumor-immune microenvironment in triple negative breast cancer revealed by multiplexed ion beam imaging. *Cell*. 2018;174(6):1373–87 e19.
73. Giesen C, Wang HAO, Schapiro D, Zivanovic N, Jacobs A, Hattendorf B, et al. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat Methods*. 2014;11(4):417–22.
74. Beliveau BJ, Joyce EF, Apostolopoulos N, Yilmaz F, Fonseka CY, McCole RB, et al. Versatile design and synthesis platform for visualizing genomes with Oligopaint FISH probes. *Proc Natl Acad Sci U S A*. 2012;109(52):21301–6.
75. Beliveau BJ, Kishi JY, Nir G, Sasaki HM, Saka SK, Nguyen SC, et al. OligoMiner provides a rapid, flexible environment for the design of genome-scale oligonucleotide *in situ* hybridization probes. *Proc Natl Acad Sci U S A*. 2018;115(10):E2183–E92.
76. Deleage C, Wietgrefe SW, Del Prete G, Morcock DR, Hao XP, Piatak M Jr, et al. Defining HIV and SIV reservoirs in lymphoid tissues. *Pathog Immun*. 2016;1(1):68–106.
77. Frei AP, Bava F-A, Zunder ER, Hsieh EWY, Chen S-Y, Nolan GP, et al. Highly multiplexed simultaneous detection of RNAs and proteins in single cells. *Nat Methods*. 2016;13(3):269–75.
78. Kishi JY, Lapan SW, Beliveau BJ, West ER, Zhu A, Sasaki HM, Saka SK, Wang Y, Cepko CL, Yin P (2019) SABER amplifies FISH: enhanced multiplexed imaging of RNA and DNA in cells and tissues. *Nature Methods* 16 (6):533–544. <https://doi.org/10.1038/s41592-019-0404-0>
79. Raj A, van den Bogaard P, Rifkin SA, van Oudenaarden A, Tyagi S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Methods*. 2008;5(10):877–9.
80. Rouhanifard SH, Mellis IA, Dunagin M, Bayatpour S, Jiang CL, Dardani I, et al. ClampFISH detects individual nucleic acid molecules using click chemistry-based amplification. *Nat Biotechnol*. 2019;37(1):84.
81. Wang F, Flanagan J, Su N, Wang LC, Bui S, Nielson A, et al. RNAscope: a novel *in situ* RNA analysis platform for formalin-fixed, paraffin-embedded tissues. *J Mol Diagn*. 2012;14(1):22–9.
82. Zeng X, Blangett CD, Koistinen KA, Schellhase CW, Bearss JJ, Radoshitzky SR, et al. Identification and pathological characterization of persistent asymptomatic Ebola virus infection in rhesus monkeys. *Nat Microbiol*. 2017;2:17113.
83. Coffin KM, Liu J, Warren TK, Blangett CD, Kuehl KA, Nichols DK, et al. Persistent Marburg virus infection in the testes of nonhuman primate survivors. *Cell Host Microbe*. 2018;24(3):405–16.e3.
84. Winer BY, Shirvani-Dastgerdi E, Bram Y, Sellau J, Low BE, Johnson H, et al. Preclinical assessment of antiviral combination therapy in a genetically humanized mouse model for hepatitis delta virus infection. *Sci Transl Med*. 2018;10(447):eaap9328.
85. Tuchin VV, Tuchin V. Light scattering methods and instruments for medical diagnosis. Bellingham: SPIE Press; 2007.

86. Yang B, Treweek JB, Kulkarni RP, Deverman BE, Chen C-K, Lubeck E, et al. Single-cell phenotyping within transparent intact tissue through whole-body clearing. *Cell*. 2014;158(4):945–58.
87. Tainaka K, Kubota SI, Suyama TQ, Susaki EA, Perrin D, Ukai-Tadenuma M, et al. Whole-body imaging with single-cell resolution by tissue decolorization. *Cell*. 2014;159(4):911–24.
88. Jing D, Zhang S, Luo W, Gao X, Men Y, Ma C, et al. Tissue clearing of both hard and soft tissue organs with the PEGASOS method. *Cell Res*. 2018;28(8):803–18.
89. Pan C, Cai R, Quacquarelli FP, Ghasemigharagoz A, Lourbopoulos A, Matryba P, et al. Shrinkage-mediated imaging of entire organs and organisms using uDISCO. *Nat Methods*. 2016;13(10):859–67.
90. Kubota SI, Takahashi K, Nishida J, Morishita Y, Ehata S, Tainaka K, et al. Whole-body profiling of cancer metastasis with single-cell resolution. *Cell Rep*. 2017;20(1):236–50.
91. Cai R, Pan C, Ghasemigharagoz A, Todorov MI, Försterer B, Zhao S, et al. Panoptic imaging of transparent mice reveals whole-body neuronal projections and skull-meninges connections. *Nat Neurosci*. 2019;22(2):317–27.
92. Pan C, Schoppe O, Parra-Damas A, Cai R, Todorov MI, Gondi G, et al. Deep learning reveals cancer metastasis and therapeutic antibody targeting in whole body. *bioRxiv*. 2019:541862. <https://www.biorxiv.org/content/10.1101/541862v1>
93. Lo MK, Nichol ST, Spiropoulou CF. Evaluation of luciferase and GFP-expressing Nipah viruses for rapid quantitative antiviral screening. *Antiviral Res*. 2014;106:53–60.
94. Cai Y, Iwasaki M, Beitzel BF, Yú S, Postnikova EN, Cubitt B, et al. Recombinant Lassa virus expressing green fluorescent protein as a tool for high-throughput drug screens and neutralizing antibody assays. *Viruses*. 2018;10(11):655.
95. Welch SR, Guerrero LW, Chakrabarti AK, McMullan LK, Flint M, Bluemling GR, et al. Lassa and Ebola virus inhibitors identified using minigenome and recombinant virus reporter systems. *Antiviral Res*. 2016;136:9–18.
96. Liu DX, Perry DL, Evans DeWald L, Cai Y, Hagen KR, Cooper TK, et al. Persistence of Lassa virus associated with severe systemic arteritis in convalescing guinea pigs (*Cavia porcellus*). *J Infect Dis*. 2019;219(11):1818–22.
97. Panchal RG, Kota KP, Spurges KB, Ruthel G, Tran JP, Boltz RC, et al. Development of high-content imaging assays for lethal viral pathogens. *J Biomol Screen*. 2010;15(7):755–65.
98. Thangamani S, Hermance ME, Santos RI, Slovak M, Heinze D, Widen SG, et al. Transcriptional immunoprofiling at the tick-virus-host interface during early stages of tick-borne encephalitis virus transmission. *Front Cell Infect Microbiol*. 2017;7:494.
99. Oestereich L, Rieger T, Neumann M, Bernreuther C, Lehmann M, Krasemann S, et al. Evaluation of antiviral efficacy of ribavirin, arbidol, and T-705 (favipiravir) in a mouse model for Crimean-Congo hemorrhagic fever. *PLoS Negl Trop Dis*. 2014;8(5):e2804.
100. Baseler L, de Wit E, Scott DP, Munster VJ, Feldmann H. Syrian hamsters (*Mesocricetus auratus*) oronasally inoculated with a Nipah virus isolate from Bangladesh or Malaysia develop similar respiratory tract lesions. *Vet Pathol*. 2015;52(1):38–45.
101. Bjornson ZB, Nolan GP, Fanti WJ. Single-cell mass cytometry for analysis of immune system functional states. *Curr Opin Immunol*. 2013;25(4):484–94.
102. Reeves PM, Sluder AE, Paul SR, Scholzen A, Kashiwagi S, Poznansky MC. Application and utility of mass cytometry in vaccine development. *FASEB J*. 2018;32(1):5–15.
103. Bendall SC, Nolan GP, Roederer M, Chattopadhyay PK. A deep profiler's guide to cytometry. *Trends Immunol*. 2012;33(7):323–32.
104. Bekele Y, Lakshminanth T, Chen Y, Mikes J, Nasi A, Petkov S, et al. Mass cytometry identifies distinct CD4+ T cell clusters distinguishing HIV-1-infected patients according to antiretroviral therapy initiation. *JCI Insight*. 2019;4(3):e125442.
105. Coindre S, Tchitchek N, Alaoui L, Vaslin B, Bourgeois C, Goujard C, et al. Mass cytometry analysis reveals the landscape and dynamics of CD32a<sup>+</sup> CD4<sup>+</sup> T cells from early HIV infection to effective cART. *Front Immunol*. 2018;9(1217):1217.

106. Gossez M, Rimmelé T, Andrieu T, Debord S, Bayle F, Malcus C, et al. Proof of concept study of mass cytometry in septic shock patients reveals novel immune alterations. *Sci Rep.* 2018;8(1):17296.
107. Hamlin RE, Rahman A, Pak TR, Maringer K, Mena I, Bernal-Rubio D, et al. High-dimensional CyTOF analysis of dengue virus-infected human DCs reveals distinct viral signatures. *JCI Insight.* 2017;2(13):e92424.
108. Bjornson-Hooper ZB, Fragiadakis GK, Spitzer MH, Madhireddy D, Hu K, Lundsten K, et al. Cell type-specific monoclonal antibody cross-reactivity screening in non-human primates and development of comparative immunophenotyping panels for CyTOF. *bioRxiv.* 2019:577759. <https://www.biorxiv.org/content/10.1101/577759v1>
109. Bjornson-Hooper ZB, Fragiadakis GK, Spitzer MH, Madhireddy D, McIlwain D, Nolan GP. A comprehensive atlas of immunological differences between humans, mice and non-human primates. *bioRxiv.* 2019:574160. <https://www.biorxiv.org/content/10.1101/574160v1>
110. Kimball AK, Oko LM, Bullock BL, Nemenoff RA, van Dyk LF, Clambey ET. A beginner's guide to analyzing and visualizing mass cytometry data. *J Immunol.* 2018;200(1):3–22.
111. Schapiro D, Jackson HW, Raghuraman S, Fischer JR, Zanotelli VRT, Schulz D, et al. his-toCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat Methods.* 2017;14(9):873–6.
112. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature.* 2001;409(6822):860–921.
113. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. *Science.* 2001;291(5507):1304–51.
114. Metzker ML. Sequencing technologies – the next generation. *Nat Rev Genet.* 2010;11(1):31–46.
115. Levy SE, Myers RM. Advancements in next-generation sequencing. *Annu Rev Genomics Hum Genet.* 2016;17:95–115.
116. Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. *Nature.* 2017;550(7676):345–53.
117. Houldcroft CJ, Beale MA, Breuer J. Clinical and biological insights from viral genome sequencing. *Nat Rev Microbiol.* 2017;15(3):183–92.
118. Gu W, Miller S, Chiu CY. Clinical metagenomic next-generation sequencing for pathogen detection. *Annu Rev Pathol.* 2019;14:319–38.
119. Worobey M, Watts TD, McKay RA, Suchard MA, Granade T, Teuwen DE, et al. 1970s and 'patient 0' HIV-1 genomes illuminate early HIV/AIDS history in North America. *Nature.* 2016;539(7627):98–101.
120. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. *Nature.* 2016;530(7589):228–32.
121. Quick J, Grubaugh ND, Pullan ST, Claro IM, Smith AD, Gangavarapu K, et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat Protoc.* 2017;12(6):1261–76.
122. Matranga CB, Andersen KG, Winnicki S, Busby M, Gladden AD, Tewhey R, et al. Enhanced methods for unbiased deep sequencing of Lassa and Ebola RNA viruses from clinical and biological samples. *Genome Biol.* 2014;15(11):519.
123. Metsky HC, Siddle KJ, Gladden-Young A, Qu J, Yang DK, Brehio P, et al. Capturing sequence diversity in metagenomes with comprehensive and scalable probe design. *Nat Biotechnol.* 2019;37(2):160–8.
124. Quan J, Langelier C, Kuchta A, Batson J, Teyssier N, Lyden A, Caldera S, McGeever A, Dimitrov B, King R, Wilheim J, Murphy M, Ares LP, Travisano KA, Sit R, Amato R, Mumbengegwi DR, Smith JL, Bennett A, Gosling R, Mourani PM, Calfee CS, Neff NF, Chow ED, Kim PS, Greenhouse B, DeRisi JL, Crawford ED (2019) FLASH: a next-generation CRISPR diagnostic for multiplexed detection of antimicrobial resistance sequences. *Nucleic Acids Res* 47 (14):e83–e83. <https://doi.org/10.1093/nar/gkz418>
125. Gu W, Crawford ED, O'Donovan BD, Wilson MR, Chow ED, Retallack H, et al. Depletion of abundant sequences by hybridization (DASH): using Cas9 to remove unwanted high-

- abundance species in sequencing libraries and molecular counting applications. *Genome Biol.* 2016;17:41.
- 126. Gire SK, Goba A, Andersen KG, Sealfon RSG, Park DJ, Kanneh L, et al. Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science.* 2014;345(6202):1369–72.
  - 127. Carroll MW, Matthews DA, Hiscox JA, Elmore MJ, Pollakis G, Rambaut A, et al. Temporal and spatial analysis of the 2014–2015 Ebola virus outbreak in West Africa. *Nature.* 2015;524(7563):97–101.
  - 128. Ladner JT, Wiley MR, Mate S, Dudas G, Prieto K, Lovett S, et al. Evolution and spread of Ebola virus in Liberia, 2014–2015. *Cell Host Microbe.* 2015;18(6):659–69.
  - 129. Tong Y-G, Shi W-F, Liu D, Qian J, Liang L, Bo X-C, et al. Genetic diversity and evolutionary dynamics of Ebola virus in Sierra Leone. *Nature.* 2015;524(7563):93–6.
  - 130. Park DJ, Dudas G, Wohl S, Goba A, Whitmer SLM, Andersen KG, et al. Ebola virus epidemiology, transmission, and evolution during seven months in Sierra Leone. *Cell.* 2015;161(7):1516–26.
  - 131. Simon-Loriere E, Faye O, Faye O, Koivogui L, Magassouba N, Keita S, et al. Distinct lineages of Ebola virus in Guinea during the 2014 West African epidemic. *Nature.* 2015;524(7563):102–4.
  - 132. Dudas G, Carvalho LM, Bedford T, Tatem AJ, Baele G, Faria NR, et al. Virus genomes reveal factors that spread and sustained the Ebola epidemic. *Nature.* 2017;544(7650):309–15.
  - 133. Mbala-Kingebeni P, Aziza A, Paola ND, Wiley MR, Makiala-Mandanda S, Caviness K, et al. Medical countermeasures during the 2018 Ebola virus disease outbreak in the North Kivu and Ituri Provinces of the Democratic Republic of the Congo: a rapid genomic assessment. *Lancet Infect Dis.* 2019;19(6):648–57.
  - 134. Mbala-Kingebeni P, Pratt CB, Wiley MR, Diagne MM, Makiala-Mandanda S, Aziza A, et al. Near real-time genomic assessment of medical countermeasures under outbreak conditions: a case study for Ebola virus variant “Tumba” (Democratic Republic of the Congo, May–July 2018). *Lancet Infect Dis.* 2019;19(6):641–7.
  - 135. Carroll MW, Haldenby S, Rickett NY, Pályi B, Garcia-Dorival I, Liu X, et al. Deep sequencing of RNA from blood and oral swab samples reveals the presence of nucleic acid from a number of pathogens in patients with acute Ebola virus disease and is consistent with bacterial translocation across the gut. *mSphere.* 2017;2(4):e00325–17.
  - 136. Lauck M, Bailey AL, Andersen KG, Goldberg TL, Sabeti PC, O’Connor DH. GB virus C coinfections in west African Ebola patients. *J Virol.* 2015;89(4):2425–9.
  - 137. Andersen KG, Shapiro BJ, Matranga CB, Sealfon R, Lin AE, Moses LM, et al. Clinical sequencing uncovers origins and evolution of Lassa virus. *Cell.* 2015;162(4):738–50.
  - 138. Lau MSY, Dalziel BD, Funk S, McClelland A, Tiffany A, Riley S, et al. Spatial and temporal dynamics of superspreading events in the 2014–2015 West Africa Ebola epidemic. *Proc Natl Acad Sci U S A.* 2017;114(9):2337–42.
  - 139. Mate SE, Kugelman JR, Nyenswah TG, Ladner JT, Wiley MR, Cordier-Lassalle T, et al. Molecular evidence of sexual transmission of Ebola virus. *N Engl J Med.* 2015;373(25):2448–54.
  - 140. Diallo B, Sissoko D, Loman NJ, Bah HA, Bah H, Worrell MC, et al. Resurgence of Ebola virus disease in Guinea linked to a survivor with virus persistence in seminal fluid for more than 500 days. *Clin Infect Dis.* 2016;63(10):1353–6.
  - 141. Arias A, Watson SJ, Asogun D, Tobin EA, Lu J, Phan MVT, et al. Rapid outbreak sequencing of Ebola virus in Sierra Leone identifies transmission chains linked to sporadic cases. *Virus Evol.* 2016;2(1):vew016.
  - 142. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol.* 2018;4(1):vey016.
  - 143. Diehl WE, Lin AE, Grubaugh ND, Carvalho LM, Kim K, Kyaw PP, et al. Ebola virus glycoprotein with increased infectivity dominated the 2013–2016 epidemic. *Cell.* 2016;167(4):1088–98 e6.

144. Urbanowicz RA, McClure CP, Sakuntabhai A, Sall AA, Kobinger G, Müller MA, et al. Human adaptation of Ebola virus during the West African outbreak. *Cell*. 2016;167(4):1079–87 e5.
145. Li S, Mason CE. The pivotal regulatory landscape of RNA modifications. *Annu Rev Genomics Hum Genet*. 2014;15:127–50.
146. Kennedy EM, Courtney DG, Tsai K, Cullen BR. Viral epitranscriptomics. *J Virol*. 2017;91(9):e02263–16.
147. Rausch JW, Sztuba-Solinska J, Le Grice SFJ. Probing the structures of viral RNA regulatory elements with SHAPE and related methodologies. *Front Microbiol*. 2017;8:2634.
148. Gonzales-van Horn SR, Sarnow P. Making the mark: the role of adenosine modifications in the life cycle of RNA viruses. *Cell Host Microbe*. 2017;21(6):661–9.
149. Illumina. NGS library prep methods. <https://www.illumina.com/techniques/sequencing/ngs-library-prep/library-prep-methods.html>. 2019.
150. Jayaraman D, Kenyon JC. New windows into retroviral RNA structures. *Retrovirology*. 2018;15(1):11.
151. Dirks RAM, Stunnenberg HG, Marks H. Genome-wide epigenomic profiling for biomarker discovery. *Clin Epigenetics*. 2016;8:122.
152. Li X, Xiong X, Yi C. Epitranscriptome sequencing technologies: decoding RNA modifications. *Nat Methods*. 2016;14(1):23–31.
153. Sztuba-Solinska J, Diaz L, Kumar MR, Kolb G, Wiley MR, Jozwick L, et al. A small stem-loop structure of the Ebola virus trailer is essential for replication and interacts with heat-shock protein A8. *Nucleic Acids Res*. 2016;44(20):9831–46.
154. Kutchko KM, Madden EA, Morrison C, Plante KS, Sanders W, Vincent HA, et al. Structural divergence creates new functional features in alphavirus genomes. *Nucleic Acids Res*. 2018;46(7):3657–70.
155. Kilianski A, Roth PA, Liem AT, Hill JM, Willis KL, Rossmaier RD, et al. Use of unamplified RNA/cDNA-hybrid nanopore sequencing for rapid detection and characterization of RNA viruses. *Emerg Infect Dis*. 2016;22(8):1448–51.
156. Keller MW, Rambo-Martin BL, Wilson MM, Ridenour CA, Shepard SS, Stark TJ, et al. Direct RNA sequencing of the coding complete influenza A virus genome. *Sci Rep*. 2018;8(1):14408.
157. Depledge DP, Srinivas KP, Sadaoka T, Bready D, Mori Y, Placantonakis DG, et al. Direct RNA sequencing on nanopore arrays redefines the transcriptional complexity of a viral pathogen. *Nat Commun*. 2019;10(1):754.
158. Kheradpour P, Ernst J, Melnikov A, Rogov P, Wang L, Zhang X, et al. Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Res*. 2013;23(5):800–11.
159. Mogno I, Kwasnieski JC, Cohen BA. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res*. 2013;23(11):1908–15.
160. Tewhey R, Kotliar D, Park DS, Liu B, Winnicki S, Reilly SK, et al. Direct identification of hundreds of expression-modulating variants using a multiplexed reporter assay. *Cell*. 2018;172(5):1132–4.
161. Carette JE, Guimaraes CP, Wuethrich I, Blomen VA, Varadarajan M, Sun C, et al. Global gene disruption in human cells to assign genes to phenotypes by deep sequencing. *Nat Biotechnol*. 2011;29(6):542–6.
162. Jae LT, Raaben M, Riemersma M, van Beusekom E, Blomen VA, Velds A, et al. Deciphering the glycosylome of dystroglycanopathies using haploid screens for lassa virus entry. *Science*. 2013;340(6131):479–83.
163. Rato S, Golumbeanu M, Telenti A, Ciuffi A. Exploring viral infection using single-cell sequencing. *Virus Res*. 2017;239:55–68.
164. Russell AB, Trapnell C, Bloom JD. Extreme heterogeneity of influenza virus infection in single cells. *Elife*. 2018;7:e32303.
165. Zanini F, Pu S-Y, Bekerman E, Einav S, Quake SR. Single-cell transcriptional dynamics of flavivirus infection. *Elife*. 2018;7:e32942.

166. Steuerman Y, Cohen M, Pesheh-Yaloz N, Valadarsky L, Cohn O, David E, et al. Dissection of influenza infection *in vivo* by single-cell RNA sequencing. *Cell Syst.* 2018;6(6):679–91 e4.
167. Doğanay S, Lee MY, Baum A, Peh J, Hwang S-Y, Yoo J-Y, et al. Single-cell analysis of early antiviral gene expression reveals a determinant of stochastic *IFNB1* expression. *Integr Biol (Camb).* 2017;9(11):857–67.
168. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell.* 2015;161(5):1202–14.
169. Dura B, Choi J-Y, Zhang K, Damsky W, Thakral D, Bosenberg M, et al. scFTD-seq: freeze-thaw lysis based, portable approach toward highly distributed single-cell 3' mRNA profiling. *Nucleic Acids Res.* 2019;47(3):e16.
170. Yuan J, Sims PA. An automated microwell platform for large-scale single cell RNA-Seq. *Sci Rep.* 2016;6:33883.
171. Gierahn TM, Wadsworth MH 2nd, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. *Nat Methods.* 2017;14(4):395–8.
172. Ordovas-Montanes J, Dwyer DF, Nyquist SK, Buchheit KM, Vukovic M, Deb C, et al. Allergic inflammatory memory in human respiratory epithelial progenitor cells. *Nature.* 2018;560(7720):649–54.
173. Rodrigues SG, Stickels RR, Goeva A, Martin CA, Murray E, Vanderburg CR, et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science.* 2019;363(6434):1463–7.
174. Pickett BE, Sadat EL, Zhang Y, Noronha JM, Squires RB, Hunt V, et al. ViPR: an open bioinformatics database and analysis resource for virology research. *Nucleic Acids Res.* 2012;40(Database issue):D593–8.
175. Hadfield J, Megill C, Bell SM, Huddleston J, Potter B, Callender C, et al. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics.* 2018;34(23):4121–3.
176. Bradley P, den Bakker HC, Rocha EPC, McVean G, Iqbal Z. Ultrafast search of all deposited bacterial and viral genomic data. *Nat Biotechnol.* 2019;37(2):152–9.
177. Menegaux R, Vert JP. Continuous embeddings of DNA sequencing reads and application to metagenomics. *J Comput Biol.* 2019;26(6):509–18.
178. Loman NJ, Quick J, Simpson JT. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nat Methods.* 2015;12(8):733–5.
179. Jain M, Koren S, Miga KH, Quick J, Rand AC, Sasani TA, et al. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol.* 2018;36(4):338–45.
180. Westbrook CJ, Karl JA, Wiseman RW, Mate S, Koroleva G, Garcia K, et al. No assembly required: full-length MHC class I allele discovery by PacBio circular consensus sequencing. *Hum Immunol.* 2015;76(12):891–6.
181. Dilernia DA, Chien J-T, Monaco DC, Brown MP, Ende Z, Deymier MJ, et al. Multiplexed highly-accurate DNA sequencing of closely-related HIV-1 variants using continuous long reads from single molecule, real-time sequencing. *Nucleic Acids Res.* 2015;43(20):e129.
182. Smith AM, Jain M, Mulroney L, Garalde DR, Akeson M. Reading canonical and modified nucleotides in 16S ribosomal RNA using nanopore direct RNA sequencing. *bioRxiv.* 2017;132274. <https://www.biorxiv.org/content/10.1101/132274v1>
183. Cherf GM, Lieberman KR, Rashid H, Lam CE, Karplus K, Akeson M. Automated forward and reverse ratcheting of DNA in a nanopore at 5-Å precision. *Nat Biotechnol.* 2012;30(4):344–8.
184. Loose M, Malla S, Stout M. Real-time selective sequencing using nanopore technology. *Nat Methods.* 2016;13(9):751–4.
185. Kubota T, Lloyd K, Sakashita N, Minato S, Ishida K, Mitsui T. Clog and release, and reverse motions of a DNA in a nanopore. *Polymers.* 2019;11(1):84.
186. Dhillon RS, Kelly JD, Srikrishna D, Garry RF. Overlooking the importance of immunoassays. *Lancet Infect Dis.* 2016;16(10):1109–10.
187. Pardee K, Green AA, Ferrante T, Cameron DE, DaleyKeyser A, Yin P, et al. Paper-based synthetic gene networks. *Cell.* 2014;159(4):940–54.

188. Yen CW, de Puig H, Tam JO, Gomez-Marquez J, Bosch I, Hamad-Schifferli K, et al. Multicolored silver nanoparticles for multiplexed disease diagnostics: distinguishing dengue, yellow fever, and Ebola viruses. *Lab Chip*. 2015;15(7):1638–41.
189. Myhrvold C, Freije CA, Gootenberg JS, Abudayeh OO, Metsky HC, Durbin AF, et al. Field-deployable viral diagnostics using CRISPR-Cas13. *Science*. 2018;360(6387):444–8.
190. Quick J, Ashton P, Calus S, Chatt C, Gossain S, Hawker J, et al. Rapid draft sequencing and real-time nanopore sequencing in a hospital outbreak of *Salmonella*. *Genome Biol*. 2015;16:114.
191. Naveca FG, Claro I, Giovanetti M, de Jesus JG, Xavier J, Iani FCM, et al. Genomic, epidemiological and digital surveillance of chikungunya virus in the Brazilian Amazon. *PLoS Negl Trop Dis*. 2019;13(3):e0007065.
192. Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins KH, McIntyre ABR, et al. Nanopore DNA sequencing and genome assembly on the International Space Station. *Sci Rep*. 2017;7(1):18022.
193. Coers J, Starnbach MN, Howard JC. Modeling infectious disease in mice: co-adaptation and the role of host-specific IFN $\gamma$  responses. *PLoS Pathog*. 2009;5(5):e1000333.
194. Greek R, Menache A. Systematic reviews of animal models: methodology versus epistemology. *Int J Med Sci*. 2013;10(3):206–21.
195. O'Donnell CD, Subbarao K. The contribution of animal models to the understanding of the host range and virulence of influenza A viruses. *Microbes Infect*. 2011;13(5):502–15.
196. Huh D, Hamilton GA, Ingber DE. From three-dimensional cell culture to organs-on-chips. *Trends Cell Biol*. 2011;21(12):745–54.
197. Bhatia SN, Ingber DE. Microfluidic organs-on-chips. *Nat Biotechnol*. 2014;32(8):760–72.
198. Huh DD. A human breathing lung-on-a-chip. *Ann Am Thorac Soc*. 2015;12(Suppl 1):S42–4.
199. Niemeyer BF, Zhao P, Tuder RM, Benam KH. Advanced microengineered lung models for translational drug discovery. *SLAS Discov*. 2018;23(8):777–89.
200. Benam KH, Novak R, Nawroth J, Hirano-Kobayashi M, Ferrante TC, Choe Y, et al. Matched-comparative modeling of normal and diseased human airway responses using a microengineered breathing lung chip. *Cell Syst*. 2016;3(5):456–66 e4.
201. Benam KH, Villenave R, Lucchesi C, Varone A, Huber C, Lee H-H, et al. Small airway-on-a-chip enables analysis of human lung inflammation and drug responses *in vitro*. *Nat Methods*. 2016;13(2):151–7.
202. Gori M, Simonelli MC, Giannitelli SM, Businaro L, Trombetta M, Rainer A. Investigating nonalcoholic fatty liver disease in a liver-on-a-chip microfluidic device. *PLoS One*. 2016;11(7):e0159729.
203. Lee J, Choi B, No DY, Lee G, Lee S-R, Oh H, et al. A 3D alcoholic liver disease model on a chip. *Integr Biol (Camb)*. 2016;8(3):302–8.
204. Ribas J, Sadeghi H, Manbachi A, Leijten J, Brinegar K, Zhang YS, et al. Cardiovascular organ-on-a-chip platforms for drug discovery and development. *Appl In Vitro Toxicol*. 2016;2(2):82–96.
205. Sontheimer-Phelps A, Hassell BA, Ingber DE. Modelling cancer in microfluidic human organs-on-chips. *Nat Rev Cancer*. 2019;19(2):65–81.
206. Benam KH, Denney L, Ho L-P. How the respiratory epithelium senses and reacts to influenza virus. *Am J Respir Cell Mol Biol*. 2019;60(3):259–68.
207. Benam KH, Ingber DE. Commendation for exposing key advantage of organ chip approach. *Cell Syst*. 2016;3(5):411.
208. Benam KH, Königshoff M, Eickelberg O. Breaking the *in vitro* barrier in respiratory medicine. Engineered microphysiological systems for chronic obstructive pulmonary disease and beyond. *Am J Respir Crit Care Med*. 2018;197(7):869–75.
209. Benam KH, Mazur M, Choe Y, Ferrante TC, Novak R, Ingber DE. Human lung small airway-on-a-chip protocol. *Methods Mol Biol*. 2017;1612:345–65.
210. Benam KH, Vladar EK, Janssen WJ, Evans CM. Mucociliary defense: emerging cellular, molecular, and animal models. *Ann Am Thorac Soc*. 2018;15(Suppl 3):S210–S5.

211. Barkal LJ, Procknow CL, Álvarez-García YR, Niu M, Jiménez-Torres JA, Brockman-Schneider RA, et al. Microbial volatile communication in human organotypic lung models. *Nat Commun.* 2017;8(1):1770.
212. Ortega-Prieto AM, Skelton JK, Wai SN, Large E, Lussignol M, Vizcay-Barrena G, et al. 3D microfluidic liver cultures as a physiological preclinical tool for hepatitis B virus infection. *Nat Commun.* 2018;9(1):682.
213. Johnson BN, Lancaster KZ, Hogue IB, Meng F, Kong YL, Enquist LW, et al. 3D printed nervous system on a chip. *Lab Chip.* 2016;16(8):1393–400.
214. Villenave R, Wales SQ, Hamkins-Indik T, Papafragkou E, Weaver JC, Ferrante TC, et al. Human gut-on-a-chip supports polarized infection of Coxsackie B1 virus *in vitro*. *PLoS One.* 2017;12(2):e0169412.
215. Mullard A. Parsing clinical success rates. *Nat Rev Drug Discov.* 2016;15(7):447.
216. Arrowsmith J, Miller P. Trial watch: phase II and phase III attrition rates 2011–2012. *Nat Rev Drug Discov.* 2013;12(8):569.
217. Fang Y, Eglen RM. Three-dimensional cell cultures in drug discovery and development. *SLAS Discov.* 2017;22(5):456–72.
218. Clevers H. Modeling development and disease with organoids. *Cell.* 2016;165(7):1586–97.
219. Lancaster MA, Knoblich JA. Organogenesis in a dish: modeling development and disease using organoid technologies. *Science.* 2014;345(6194):1247125.
220. Xu H, Jiao Y, Qin S, Zhao W, Chu Q, Wu K. Organoid technology in disease modelling, drug development, personalized treatment and regeneration medicine. *Exp Hematol Oncol.* 2018;7:30.
221. Zhou J, Li C, Sachs N, Chiu MC, Wong BH-Y, Chu H, et al. Differentiated human airway organoids to assess infectivity of emerging influenza virus. *Proc Natl Acad Sci U S A.* 2018;115(26):6822–7.
222. Watanabe M, Butth JE, Vishlaghi N, de la Torre-Ubieta L, Taxidis J, Khakh BS, et al. Self-organized cerebral organoids with human-specific features predict effective drugs to combat Zika virus infection. *Cell Rep.* 2017;21(2):517–32.
223. Elazazy MS. Factorial design and machine learning strategies: impacts on pharmaceutical analysis. In: Sharmin E, Zafar F, editors. Spectroscopic analyses – developments and applications. London: IntechOpen; 2017. p. 443.
224. Vanhaelen Q, Mamoshina P, Aliper AM, Artemov A, Lezhnina K, Ozerov I, et al. Design of efficient computational workflows for *in silico* drug repurposing. *Drug Discov Today.* 2017;22(2):210–22.
225. Centers for Disease Control and Prevention. Ebola virus NP real-time RT-PCR assay. <https://www.fda.gov/downloads/medicaldevices/safety/emergencysituations/ucm436307.pdf>. 2016.
226. Cepheid. Xpert® Ebola assay – Instructions for use: for use under an Emergency Use Authorization (EUA) only. <https://www.fda.gov/downloads/MedicalDevices/Safety/EmergencySituations/UCM439578.pdf>. 2015.
227. Jensen KS, Adams R, Bennett RS, Bernbaum J, Jahrling PB, Holbrook MR. Development of a novel real-time polymerase chain reaction assay for the quantitative detection of Nipah virus replicative viral RNA. *PLoS One.* 2018;13(6):e0199534.

# Space Exploration and Travel, Future Technologies for Inflight Monitoring and Diagnostics



Jean-Pol Fripiat

**Abstract** Astronauts have been found to suffer from weakened immune systems. This phenomenon is frequently associated to latent virus reactivation. Maintaining crew health is therefore a concern to enable future long-term space missions such as those to Mars and beyond. Indeed, a Mars mission will imply 6 months of travel each way plus the surface stay. Thus, future space exploration needs innovative technologies to monitor health and perform personalized diagnostic and medicine. This chapter reviews some of the effects of spaceflight on the immune system and microorganisms, describes Earth-based models used to study the effects of such harsh environment, presents constraints associated to inflight analyses, biosensors for astronauts' health inflight monitoring and some high-throughput "omics" technologies that are sufficiently developed to be ready for deployment and use soon onboard a spacecraft.

**Keywords** Immunosuppression · Virus · Omics · Personalized medicine · Diagnostic · Biosensor · Miniaturization · Automatic devices · Extreme conditions · Spaceflight

## 1 Introduction

Since Yuri Gagarin became the first human to leave the confines of Earth in 1961, an increasing number of humans have traveled to space and permanent inhabited space stations (Mir and then the International Space Station – ISS) have been constructed. Studies performed on humans or animals sent to these stations, or subjected to ground-based models used to simulate space conditions, have revealed that these missions induce physiological dysregulations such as muscle atrophy,

---

J.-P. Fripiat (✉)

Stress Immunity Pathogens Laboratory, EA 7300, Faculty of Medicine,  
Université de Lorraine, Vandœuvre-lès-Nancy, France  
e-mail: [jean-pol.fripiat@univ-lorraine.fr](mailto:jean-pol.fripiat@univ-lorraine.fr)

bone demineralization, cardiovascular and metabolic dysfunctions, impaired cognitive processes and reduced immunological competence.

This last effect is known since a long time as it was noted that 15 of the 29 astronauts involved in Apollo missions developed bacterial or viral infections during, immediately after, or within 1 week of landing [1]. Later studies confirmed that all compartments of the immune system are affected by extreme conditions encountered during such missions (for review see [2, 3]). It was also demonstrated that immune system dysregulation occur during a flight and persist during 6-month orbital spaceflight [4, 5]. Furthermore, a recent study revealed that about 50% of the astronauts who spent 6 months onboard the ISS faced immunological problems [6] thereby confirming in-flight dysregulation distinct from the influences of landing and readaptation following deconditioning [4, 6, 7].

Concerning innate immunity, it was shown, for example, that astronaut's monocytes exhibit phenotypic and cytokine-production deregulations, a reduced ability to engulf *E. coli*, elicit an oxidative burst and degranulate [8–10]. Reactive oxygen species production by macrophages [11] as well as neutrophils phagocytosis and oxidative burst capacities are also significantly reduced [12].

Regarding acquired immunity, several studies reported a reduction of T-cell activation under low gravity conditions (see for example [13–15]). Numerous studies investigated this phenomenon [13–15] and highlighted that almost all cellular parameters can be affected such as: (i) gene expression, as shown by lower expressions of Interleukin-2 (IL-2) and IL-2 receptor alpha chain [16]; (ii) cell-cell interactions and cytoskeleton structure, as T lymphocytes were found to be highly motile under microgravity while the motility of monocytes was severely reduced and the structure of their cytoskeleton was modified [17–21]; (iii) signal transduction, as PKA and NF-κB signaling pathways were shown to contribute to T cell dysfunction under altered gravity [22–24] and (iv) disturbed expression of cell cycle regulatory proteins [25].

As to humoral immunity, it was shown using the urodele amphibian *Pleurodeles waltl* as animal model [26], that spaceflight affects antibody production in response to an antigenic stimulation [27–29]. Hypergravity and simulated microgravity did also impair the proliferative responses of murine B-lymphocytes [30, 31]. The maturation of immune cells belonging to the myeloid [32–36], B- [37–39] and T- [40–42] lineages were also shown to be reduced under low gravity conditions.

Low natural killer cell cytotoxicity, providing immunological resistance and defense not only against foreign microorganisms but also against body cells transformed because of a virus infection or malignancy, and a delay in responses to hypersensitivity skin tests were observed [43, 44]. Reactivation of latent herpes viruses has also frequently been reported. For example, during and immediately after spaceflight, Varicella zoster virus VZV DNA was detected in saliva, indicative of subclinical reactivation while no VZV DNA was detected prior spaceflight [45]. Furthermore, it was demonstrated that saliva contained infectious VZV particles [46]. Additional studies revealed that 50% of the astronauts shed live infectious VZV in their saliva asymptotically during short duration spaceflights [47] and

that this number increases to 65% during long-duration missions [5]. Importantly, a few cases resulted in clinical disease manifesting as herpes zoster [45]. Taken together, these studies indicate that reactivation of VZV, particularly during longer duration spaceflights, can potentially lead to clinical disease. These data explain why Flight Medicine Clinic at Johnson Space Center, NASA, has initiated vaccination of all crewmembers with Zostavax, a vaccine to prevent shingles before spaceflight. These reactivations can be correlated with a drop in interferon production and elevated levels of stress hormones, suggesting that spaceflight-associated stressors may be responsible for these reactivations [48–53]. Thus, latent virus reactivation can be considered as a good biomarker of spaceflight-induced weakening of cell-mediated immunity and viral load measurements can be used to assess the functionality of the immune system.

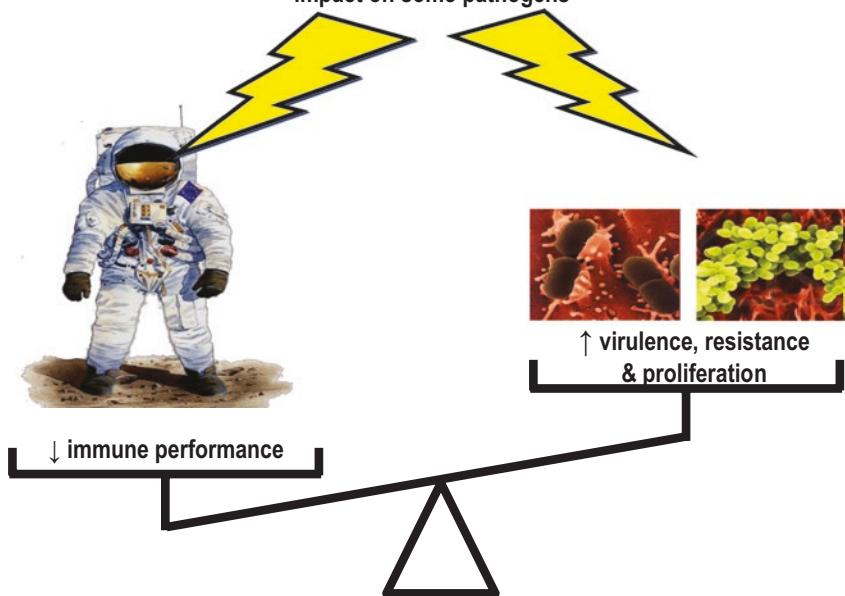
In parallel of this immunological weakening, changes in microbial growth characteristics and pathogenicity have been observed for several microorganisms [54, 55]. Depending on studied bacteria, enhanced or decrease virulence [56, 57], altered antimicrobial susceptibility [58, 59] and/or increased biofilm formation [57, 59, 60] have been described due to modulation of gene expression regulation [55–57, 61, 62]. Furthermore, some data suggest that antibiotics could be less effective in space [63]. Concerning viral infection, it was shown that microgravity leads to a downregulation of BALB/c mice resistance to herpes simplex virus type 1 (HSV-1) infection related to modulation of central nervous system and cytokines produced by the immune system [64]. This observation sticks well with reactivations of latent herpes viruses noted in astronauts (see above). As for bacteria, it was shown that high energy (HZE) particles, high energy hadrons and cosmic radiation cause mutations in the genome of the T4 bacteriophage [65, 66]. Finally, it is important to keep in mind that, as the duration of space missions will increase, the potential for infectious diseases to arise during flight may become a critical issue because the probability of cross-contamination between crew members will increase.

Thus microbial changes, coupled to dysregulations of the immune system, could explain the higher susceptibility to infection noted in astronauts [6] (Fig. 1). Both are a concern for all space agencies and must be seriously investigated and understood to be able to preserve astronauts' health during future deep-space exploration missions such as deployment of a Lunar station followed by multiple Mars flyby missions.

## 2 How to Study the Effects of Spaceflight?

Up to now most studies performed on astronauts were conducted on samples (peripheral blood, urine, saliva) taken before and after spaceflight. Studies analyzing samples collected inflight are less frequent. Such studies were conducted on samples frozen in the space station and analyzed once sent back to Earth. An important progress was achieved recently with rare opportunities of sending back to Earth,

**Stressors encountered during spaceflight (microgravity, radiation, chronic stress, fluid shear, hydrostatic pressure, etc.) weaken the immune system and can have a positive impact on some pathogens**



**Fig. 1** Numerous dysregulations of the immune system have been reported during and following spaceflight. In parallel, spaceflight has been shown to increase the virulence, antibiotic resistance, and proliferation of some pathogens suggesting that susceptibility to infections could be enhanced during space missions. (Reproduced from Frippiat et al. [3] with permission from Nature Publishing Group)

within 48 h, few milliliters of astronaut blood in ACD (acid-citrate-dextrose) collection tubes containing both anticoagulant and nutrients designed to maintain cellular viability [4].

However, despite this major advance, serious limitations in the availability and the experimental protocols that can be carried out with samples from astronauts as well as in the possibility to perform *in situ* analyses remain. Consequently, ground-based models have been developed to mimic the effects of spaceflight conditions on an organism. The most commonly used to reduce gravity constraint are head-down tilt bed rest for humans [67] and anti-orthostatic tail suspension for rodents [68]. Exposure to anti-orthostatic tail suspension has been shown to result in alterations of the immune system similar to those observed after spaceflight [69]. This treatment increases the susceptibility of female Swiss/Webster mice to D variant of encephalomyocarditis virus infection while they are normally resistant to this pathogen [70]. This correlated with a drop in interferon production and elevated levels of stress hormones suggesting once again that stress may be responsible of latent virus reactivations [70, 71]. Indeed, it is known that hormones such as catecholamine, which are

released during stressful situations, regulate immune functions through adrenergic receptors located on immune cells, particularly  $\beta 2$ -type receptors. Additionally, glucocorticoids produced in response to stress affect both innate and acquired immunity. However, more experiment will be required to precise the effects of anti-orthostatic tail suspension on immune surveillance. It would, for example, be interesting to submit latently infected mice to this model to precise by which mechanisms viruses reactivate under these conditions. In the same way, it was demonstrated that anti-orthostatic tail suspension impairs organism defenses against Gram-negative bacteria (*P. aeruginosa* or *K. pneumonia*) [72, 73]. Finally, several papers have shown that simulated microgravity negatively impacts, as real microgravity encountered during spaceflight, the production of murine B- and T-cells [38–40, 42] which very likely contribute too to lower host defenses. Regarding head-down tilt bed rest, studies did not reveal significant latent viral reactivations (EBV, VZV, CMV) [74, 75] but a decrease of the levels of virus-specific T cells was noted during one study [74]. Interestingly, when stress was applied during bed rest, reactivation of EBV and VZV viruses could be noted [76, 77] in agreement with the neuromodulation described above.

### 3 Constraints Associated to Inflight Analyses

Inflight diagnostic and monitoring of infection is currently limited. Major limiting points are (i) the size of diagnostic devices as space is limited in the ISS, (ii) limited manpower because astronauts are very busy with all the tasks they have to do in the space station, (iii) limited power supply and (iv) the impossibility to perform real-time analysis of data. Data downloading from the ISS for on-Earth analysis is possible but increasing hardware efficiency and future implementation of “omics” technologies will significantly increase the amounts of data and consequently complicate this downloading process. Miniaturization and automatization of diagnostic devices, associated to artificial intelligence, are therefore required. Indeed, we might expect significant reduction in hardware size and energy consumption thanks to the miniaturization of various components, and automatization would reduce manpower requirement.

### 4 Towards Inflight Personalized Monitoring and Diagnostic

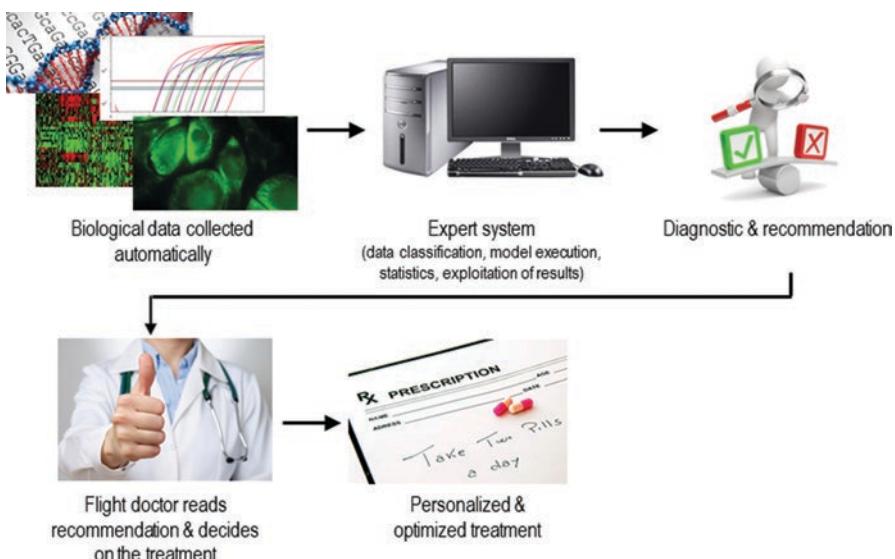
The capacity to analyze cellular and molecular targets from minimal amounts of body fluids is required to implement precision medicine. Indeed, access to relevant bio-molecular information is necessary to allow adequate diagnostic aimed at protecting astronauts against the detrimental effects of space exploration. Therefore, the development of biosensors, tools for rapid on-site automated preparation and analysis of small sample volumes, is needed.

Space biotechnology is a nascent field whose aim is to exploit *in situ* state-of-the-art high throughput techniques for amplifying and sequencing DNA, measuring levels of transcripts, proteins and metabolites (“omics” techniques). A number of efforts are ongoing to develop instruments to carry out such techniques in the International Space Station [78]. This new generation of automated and miniaturized machines takes advantage, for example, of microfluidic technology, on-chip DNA amplification and of the possibility of quantifying cytokines and stress markers using luminescence- or fluorescence-based assays. Such new devices would also be of interest to detect microorganisms in spacecraft. Currently, these microbiological investigations are done using classic sampling methods that take time to generate results and might be impossible on a mission to Mars.

These biosensors are essential to advance biomedical and physiological studies to control and reduce the effects of space-related stressors on living systems, as illustrated by a market survey of commercially available portable diagnostics performed by NASA [79].

Furthermore, carrying out measurements *in situ* provides considerable advantages over traditional post-flight data analysis. Indeed, *in-situ* high-throughput capabilities will allow flight surgeon to monitor crew health in real time, permitting adaptive and responsive diagnostics and treatment regimens (Fig. 2).

Several inflight biosensors able to perform analyses of samples to provide semi-quantitative or quantitative results in few minutes and with minimum resource consumption in terms of instrumentation’s weight, volume, storage conditions and power consumption do already exist.



**Fig. 2** Artificial intelligence, by its ability to treat an extremely high amount of data, such as those generated by “omics” techniques, will be an important aid to medical decision. It will contribute to a more personalized medicine

Indeed, NASA and University of Colorado developed a VZV detection kit for use with saliva that can be used in space as well as on Earth [80, 81]. This point-of-care diagnostic device based on chemiluminescence is rapid, simple to use, non-invasive, inexpensive (does not require expensive equipment) and detects the VZV virus in less than 1 h, thereby allowing early diagnosis and treatment. Additionally, as stated above, such kit could also be used to monitor the functionality of the immune system as latent virus reactivation is a good biomarker of spaceflight-induced weakening of cell-mediated immunity.

Pharmacological treatment is the first-line therapy for a disease. However, previous study suggested significant alterations in the pharmacodynamics and pharmacokinetics of drugs under spaceflight conditions [82, 83]. Consequently, real-time biosensors that can measure drug concentration, or other biomarkers, in blood (or other biological fluids) have been developed. An example of biosensor that was successfully used to evaluate changes in pharmacokinetics and pharmacodynamics in the ISS is the Reflotron IV biochemical analyzer [84]. Another example is proved by I-STAT (produced by Abbott, US) which is used on the International Space Station to perform comprehensive blood analysis (blood gas, chemistry, coagulation and cardiac marker). IMMUNOLAB is a third example of biosensor allowing analyses to be performed on blood, urine or saliva samples onboard the ISS [85]. Inflight quantification of cytokines is the major task of this device performing both sample preparation using commercial kits and target detection *via* fluorescence microscopy. IMMUNOLAB allows a reproducible inflight analysis of samples, with high precision and an easy operation, to enable the investigation of short- and long-term effects on the immune system of the crew. Microflow1 complements IMMUNOLAB capabilities. This fiber-optic fluorescence-based and portable flow cytometry platform was used onboard the ISS for immunophenotyping and microbead-based multiplexed immunoassays. Microflow1's performance was comparable to that of a commercial flow cytometer in a standard laboratory environment, demonstrating that its fiber-optic cytometer technology is compatible with space environment [86]. rHEALTH, for 'reusable Handheld Electrolyte and Lab Technology for Humans', is another example. It is a reusable microfluidic device that performs rapid, low cost cell counts and measurements of electrolytes, proteins, and other biomarkers [87]. The rHEALTH sensor is a compact portable device that employs cutting-edge fluorescence detection optics, innovative microfluidics, and nanostrip reagents to perform a suite of hematology, chemistry, and biomarker assays from a single drop of blood or body fluid. Developed to monitor astronaut health on the ISS and during long-term space flight, terrestrial applications for this ground-breaking technology include point-of-care diagnostics at a patient's bedside, in a doctor's office or hospital. Recently, a compact biosensor which astronauts could use inflight to measure clinical biomarkers in saliva or other biological fluids in order to monitor their health was developed by the Italian Space Agency in collaboration with NASA [88, 89]. This biosensor is based on the chemiluminescence lateral flow immunoassay technique, an immunoanalytical approach based on the use of a nitrocellulose membrane on which the immunoreagents are immobilized in specific areas and relying on highly-sensitive chemiluminescence detection. The biosensor was designed to

measure the levels of cortisol, a chronic stress marker, in a minimal volume of saliva. This biosensor was successfully used in 2017 onboard the ISS during the VITA mission, thereby demonstrating the feasibility of performing sensitive immunological clinical chemistry analyses directly onboard the ISS.

Blood collection is also easier than before as the US Food and Drug Administration (FDA) recently approved a needle-free collection device developed by Tasso Inc. (Seattle, US) for capillary blood that is promising for inflight use.

Biosensors allowing DNA amplification, quantification and sequencing have also been developed. RT-qPCR, used to monitor viral load, detect pathogens, perform genotyping, diagnostic infectious diseases and evaluate immune responses, has recently been miniaturized using microfluidics [90]. Additionally, an aqueous ready-to-use PCR mixture that remains stable at 20 °C for over 60 days and is stable indefinitely when refrigerated (4 °C) or frozen (-20 °C) has been developed by Crews' lab. PCR devices for space biology have also been developed and successfully deployed on the ISS for technical validation (Fig. 3).

Another major advance is the possibility to implement next generation sequencing (NGS). This technology allows determining, for example, how viruses adapt/mutate in an individual over time depending on interactions with immune functions and environment. This is important because microbial mutation rates increase in microgravity. Solar and cosmic radiation met during space missions, with a cumulative dose obviously increasing with mission duration, could also contribute to the appearance of mutations potentially associated to resistance. These mutations could impinge on the capacity to treat effectively the infections that will doubtless arise during such long and stressful endeavor as interplanetary missions. A promising example is the Personal Genome Machine (PGM) created by Ion Torrent™, a small device allowing fast sequencing at low cost and the rapid and sensitive detection of mutations [91]. However, sample preparation will have to be optimized to reduce crew contribution. Another promising venue is nanopore technology. Nanopores are molecular-scale sensors that are transforming the field of sequencing as they can



**Fig. 3** Wetlab-2 System developed at NASA's Ames Research Center to perform RT-qPCR onboard the ISS. This system allows obtaining real-time gene expression data from samples processed and analyzed onboard the space station. It comprises the Sample Preparation Module (a) designed to lyse cells and extract RNA, the Sample De-bubbler and Pipette Loading Device (b) that removes air bubbles from fluids and transfers liquid samples into PCR reaction tubes, and a commercial PCR instrument (Cepheid SmartCycler®) that can perform up to 16 PCR reactions in parallel (c). (Figure adapted from <https://www.nasa.gov/ames/research/space-biosciences/wetlab-2>. Image Credits: NASA/Dominic Hart)

electrically detect single biological molecules such as proteins or DNA with precision. A nice example of nanopore application is MinION, a portable real-time device for DNA and RNA sequencing, weighting less than 100 g, developed by Oxford Nanopore Technologies, UK. It has been successfully used on the ISS [92] thereby demonstrating the feasibility of DNA sequencing in space conditions. Recently, a novel nanopore DNA device has been designed that can quickly and precisely detect disease biomarkers at the point-of-care, and which could be a major advance in personalized diagnostic medicine [93]. This solid-state nano-filtered device transforms the identity of individual biomolecules into an electrical signal to allow more accurate measurements of single DNA molecules.

## 5 Conclusion and Perspectives

Some inflight biosensor diagnostic devices are already available to monitor astronauts' health but progresses are still required to ensure safe long-duration missions. Biosensors able to accept different types of sample (e.g. human blood, saliva, urine) and enable a multiplex approach, where different types of assay may be performed together on the same sample (e.g. quantitative detection of proteins and nucleic acids), would be of paramount interest. This will require technological development, but rapid progresses in the conception of compact and efficient devices, coupled to artificial intelligence able to assess and analyze big data, will for sure contribute to the development of these forthcoming biosensors required for efficient inflight monitoring and diagnostic leading to significant progress in space medicine to protect astronauts from diseases and mitigate detrimental effects of space-related stressors. These advances will also be of paramount importance to defeat diseases on Earth through telemedicine improvement [94].

**Acknowledgments** JPF and his team acknowledge support from the French Space Agency (CNES), the European Space Agency (ESA), the French Ministry of Higher Education and Research, the Université de Lorraine, the Région Lorraine and the “Impact Biomolecules” project of the “Lorraine Université d’Excellence” (Investissements d’avenir – ANR).

## References

1. Kimzey SL. Hematology and immunology studies. In: Johnson RS, Dietlein LF, editors. Biomedical Results from Skylab. Washington D.C.: National Aeronautics and Space Administration, U.S. Goverment Printing Office; 1977. p. 249–82.
2. Guéguinou N, Huin-Schohn C, Bascove M, Bueb JL, Tschirhart E, Legrand-Frossi C, et al. Could spaceflight-associated immune system weakening preclude the expansion of human presence beyond Earth’s orbit? *J Leukoc Biol.* 2009;86:1027–38. <https://doi.org/10.1189/jlb.0309167>.

3. Frippiat JP, Crucian BE, de Quervain DJF, Grimm D, Montano N, Praun S, et al. Towards human exploration of space: the THESEUS review series on immunology research priorities. *NPJ Microgravity.* 2016;2:16040. <https://doi.org/10.1038/npjmgav.2016.40>.
4. Crucian B, Stowe RP, Mehta S, Quiriarte H, Pierson D, Sams C. Alterations in adaptive immunity persist during long-duration spaceflight. *NPJ Microgravity.* 2015;1:15013. <https://doi.org/10.1038/npjmgav.2015.13>.
5. Mehta SK, Laudenslager ML, Stowe RP, Crucian BE, Feiveson AH, Sams CF, et al. Latent virus reactivation in astronauts on the international space station. *NPJ Microgravity.* 2017;3:11. <https://doi.org/10.1038/s41526-017-0015-y>.
6. Crucian B, Babiak-Vazquez A, Johnston S, Pierson DL, Ott CM, Sams C. Incidence of clinical symptoms during long-duration orbital spaceflight. *Int J Gen Med.* 2016;9:383–91. <https://doi.org/10.2147/IJGM.S114188>.
7. Crucian B, Johnston S, Mehta S, Stowe R, Uchakin P, Quiriarte H, et al. A case of persistent skin rash and rhinitis with immune system dysregulation onboard the International Space Station. *J Allergy Clin Immunol Pract.* 2016;4:759–62. <https://doi.org/10.1016/j.jaip.2015.12.021>.
8. Kaur I, Simons ER, Kapadia AS, Ott CM, Pierson DL. Effect of spaceflight on ability of monocytes to respond to endotoxins of gram-negative bacteria. *Clin Vaccine Immunol.* 2008;15:1523–8. <https://doi.org/10.1128/CVI.00065-08>.
9. Rykova MP, Antropova EN, Larina IM, Morukov BV. Humoral and cellular immunity in cosmonauts after the ISS missions. *Acta Astronaut.* 2008;63:697–705. <https://doi.org/10.1016/j.actaastr.2008.03.016>.
10. Crucian B, Stowe R, Quiriarte H, Pierson D, Sams C. Monocyte phenotype and cytokine production profiles are dysregulated by short-duration spaceflight. *Aviat Space Environ Med.* 2011;82:857–62.
11. Brungs S, Kolanus W, Hemmersbach R. Syk phosphorylation – a gravisensitive step in macrophage signalling. *Cell Commun Signal.* 2015;13:9. <https://doi.org/10.1186/s12964-015-0088-8>.
12. Kaur I, Simons ER, Castro VA, Ott CM, Pierson DL. Changes in neutrophil functions in astronauts. *Brain Behav Immun.* 2004;18:443–50. <https://doi.org/10.1016/j.bbi.2003.10.005>.
13. Cogoli A, Tschopp A, Fuchs-Bislin P. Cell sensitivity to gravity. *Science.* 1984;225:228–30.
14. Cogoli A. The effect of space flight on human cellular immunity. *Environ Med.* 1993;37:107–6.
15. Gridley DS, Slater JM, Luo-Owen X, Rizvi A, Chapes SK, Stodieck LS, et al. Spaceflight effects on T lymphocyte distribution, function and gene expression. *J Appl Physiol.* 2009;106:194–202. <https://doi.org/10.1152/japplphysiol.91126.2008>.
16. Walther I, Pippia P, Meloni MA, Turrini F, Mannu F, Cogoli A. Simulated microgravity inhibits the genetic expression of interleukin-2 and its receptor in mitogen-activated T lymphocytes. *FEBS Lett.* 1998;436:115–8. [https://doi.org/10.1016/S0014-5793\(98\)01107-7](https://doi.org/10.1016/S0014-5793(98)01107-7).
17. Sciola L, Cogoli-Greuter M, Cogoli A, Spano A, Pippia P. Influence of microgravity on mitogen binding and cytoskeleton in Jurkat cells. *Adv Space Res.* 1999;24:801–5. [https://doi.org/10.1016/S0273-1177\(99\)00078-2](https://doi.org/10.1016/S0273-1177(99)00078-2).
18. Cogoli-Greuter M. Effect of gravity changes on the cytoskeleton in human lymphocytes. *Gravit Space Biol Bull.* 2004;17:27–37.
19. Meloni MA, Galleri G, Camboni MG, Pippia P, Cogoli A, Cogoli-Greuter M. Modeled microgravity affects motility and cytoskeletal structures. *J Gravit Physiol.* 2004;11:197–8.
20. Meloni MA, Galleri G, Pippia P, Cogoli-Greuter M. Cytoskeleton changes and impaired motility of monocytes at modelled low gravity. *Protoplasma.* 2006;229:243–9. <https://doi.org/10.1007/s00709-006-0210-2>.
21. Meloni MA, Galleri G, Pani G, Saba A, Pippia P, Cogoli-Greuter M. Space flight affects motility and cytoskeletal structures in human monocyte cell line J-111. *Cytoskelet Hoboken NJ.* 2011;68:125–37. <https://doi.org/10.1002/cm.20499>.
22. Boonyaratanaornkit JB, Cogoli A, Li CF, Schopper T, Pippia P, Galleri G, et al. Key gravity-sensitive signaling pathways drive T-cell activation. *FASEB J.* 2005;19:2020–2. <https://doi.org/10.1096/fj.05-3778fje>.

23. Chang TT, Walther I, Li CF, Boonyaratankornkit J, Galleri G, Meloni MA, et al. The Rel/NF- $\kappa$ B pathway and transcription of immediate early genes in T cell activation are inhibited by microgravity. *J Leukoc Biol.* 2012;92:1133–45. <https://doi.org/10.1189/jlb.0312157>.
24. Martinez EM, Yoshida MC, Candelario TL, Hughes-Fulford M. Spaceflight and simulated microgravity cause a significant reduction of key gene expression in early T-cell activation. *Am J Physiol Regul Integr Comp Physiol.* 2015;308:R480–8. <https://doi.org/10.1152/ajpregu.00449.2014>.
25. Thiel CS, Paulsen K, Bradacs G, Lust K, Tauber S, Dumrese C, et al. Rapid alterations of cell cycle control proteins in human T lymphocytes in microgravity. *Cell Commun Signal.* 2012;10:1. <https://doi.org/10.1186/1478-811X-10-1>.
26. Frippiat JP. Contribution of the urodele amphibian *Pleurodeles waltl* to the analysis of spaceflight-associated immune system deregulation. *Mol Immunol.* 2013;56:434–41. <https://doi.org/10.1016/j.molimm.2013.06.011>.
27. Boxio R, Dournon C, Frippiat JP. Effects of a long-term spaceflight on immunoglobulin heavy chains of the urodele amphibian *Pleurodeles waltl*. *J Appl Physiol.* 2005;98:905–10.
28. Bascove M, Huin-Schohn C, Guéguinou N, Tschirhart E, Frippiat JP. Spaceflight-associated changes in immunoglobulin VH gene expression in the amphibian *Pleurodeles waltl*. *FASEB J.* 2009;23:1607–15. <https://doi.org/10.1096/fj.08-121327>.
29. Bascove M, Guéguinou N, Schaerlinger B, Gauquelin-Koch G, Frippiat JP. Decrease in antibody somatic hypermutation frequency under extreme, extended spaceflight conditions. *FASEB J.* 2011;25:2947–55. <https://doi.org/10.1096/fj.11-185215>.
30. Guéguinou N, Bojados M, Jamon M, Derradji H, Baatout S, Tschirhart E, et al. Stress response and humoral immune system alterations related to chronic hypergravity in mice. *Psychoneuroendocrinology.* 2012;37:137–47. <https://doi.org/10.1016/j.psyneuen.2011.05.015>.
31. Gaignier F, Schenten V, De Carvalho Bittencourt M, Gauquelin-Koch G, Frippiat JP, Legrand-Frossi C. Three weeks of murine hindlimb unloading induces shifts from B to T and from Th to Tc splenic lymphocytes in absence of stress and differentially reduces cell-specific mitogenic responses. *PLoS One.* 2014;9:e92664. <https://doi.org/10.1371/journal.pone.0092664>.
32. Vacek A, Michurina TV, Serova LV, Rotkovská D, Bartonícková A. Decrease in the number of progenitors of erythrocytes (BFUe, CFUe), granulocytes and macrophages (GM-CFC) in bone marrow of rats after a 14-day flight onboard the Cosmos-2044 Biosatellite. *Folia Biol.* 1991;37:35–41.
33. Davis TA, Wiesmann W, Kidwell W, Cannon T, Kerns L, Serke C, et al. Effect of spaceflight on human stem cell hematopoiesis: suppression of erythropoiesis and myelopoiesis. *J Leukoc Biol.* 1996;60:69–76.
34. Ichiki AT, Gibson LA, Jago TL, Strickland KM, Johnson DL, Lange RD, et al. Effects of spaceflight on rat peripheral blood leukocytes and bone marrow progenitor cells. *J Leukoc Biol.* 1996;60:37–43.
35. Ortega MT, Pecaut MJ, Gridley DS, Stodieck LS, Ferguson V, Chapes SK. Shifts in bone marrow cell phenotypes caused by spaceflight. *J Appl Physiol.* 2009;106:548–55. <https://doi.org/10.1152/japplphysiol.91138.2008>.
36. Sotnezova EV, Markina EA, Andreeva ER, Buravkova LB. Myeloid precursors in the bone marrow of mice after a 30-day space mission on a Bion-M1 biosatellite. *Bull Exp Biol Med.* 2017;162:496–500. <https://doi.org/10.1007/s10517-017-3647-8>.
37. Huin-Schohn C, Guéguinou N, Schenten V, Bascove M, Gauquelin-Koch G, Baatout S, et al. Gravity changes during animal development affect IgM heavy-chain transcription and probably lymphopoiesis. *FASEB J.* 2013;27:333–41. <https://doi.org/10.1096/fj.12-217547>.
38. Lescale C, Schenten V, Djeghloul D, Bennabi M, Gaignier F, Vandamme K, et al. Hind limb unloading, a model of spaceflight conditions, leads to decreased B lymphopoiesis similar to aging. *FASEB J.* 2015;29:455–63. <https://doi.org/10.1096/fj.14-259770>.
39. Tascher G, Gerbaix M, Maes P, Chazarin B, Ghislain S, Antropova E, et al. Analysis of femurs from mice embarked on board BION-M1 biosatellite reveals a decrease in immune cell development, including B cells, after 1 wk of recovery on Earth. *FASEB J.* 2019; <https://doi.org/10.1096/fj.201801463R>.

40. Woods CC, Banks KE, Gruener R, DeLuca D. Loss of T cell precursors after spaceflight and exposure to vector-averaged gravity. *FASEB J.* 2003;17:1526–8. <https://doi.org/10.1096/fj.02-0749fje>.
41. Woods CC, Banks KE, Lebsack TW, White TC, Anderson G, Maccallum T, et al. Use of a microgravity organ culture dish system to demonstrate the signal dampening effects of modeled microgravity during T cell development. *Dev Comp Immunol.* 2005;29:565–82. <https://doi.org/10.1016/j.dci.2004.09.006>.
42. Ghislain S, Ouzren-Zarhloul N, Kaminski S, Frippiat JP. Hypergravity exposure during gestation modifies the TCR $\beta$  repertoire of newborn mice. *Sci Rep.* 2015;5:9318. <https://doi.org/10.1038/srep09318>.
43. Taylor GR, Janney RP. In vivo testing confirms a blunting of the human cell-mediated immune mechanism during space-flight. *J Leukoc Biol.* 1992;51:129–32. <https://doi.org/10.1002/jlb.51.2.129>.
44. Meshkov D, Rykova M. The natural cytotoxicity in cosmonauts on board space stations. *Acta Astronaut.* 1995;36:719–26. [https://doi.org/10.1016/0094-5765\(95\)00162-X](https://doi.org/10.1016/0094-5765(95)00162-X).
45. Mehta SK, Cohrs RJ, Forghani B, Zerbe G, Gilden DH, Pierson DL. Stress- induced sub-clinical reactivation of varicella zoster virus in astronauts. *J Med Virol.* 2004;72:174–9. <https://doi.org/10.1002/jmv.10555>.. 196.
46. Cohrs RJ, Mehta SK, Schmid DS, Gilden DH, Pierson DL. Asymptomatic reactivation and shed of infectious varicella zoster virus in astronauts. *J Med Virol.* 2008;80:1116–22. <https://doi.org/10.1002/jmv.21173>.
47. Mehta SK, Laudenslager ML, Stowe RP, Crucian BE, Sams CF, Pierson DL. Multiple latent viruses reactivate in astronauts during space shuttle missions. *Brain Behav Immun.* 2014;41:210–7. <https://doi.org/10.1016/j.bbi.2014.05.014>.. 198.
48. Meehan R, Whitton P, Sams C. The role of Psychoneuroendocrine factors on spaceflight-induced immunological alterations. *J Leukoc Biol.* 1993;54:236–44.
49. Crucian BE, Cubbage ML, Sams CF. Altered cytokine production by specific human peripheral blood cell subsets immediately following space flight. *J Interf Cytokine Res.* 2000;20:547–56. <https://doi.org/10.1089/10799900050044741>.
50. Mehta SK, Stowe RP, Feiveson AH, Tyring SK, Pierson DL. Reactivation and shedding of cytomegalovirus in astronauts during spaceflight. *J Infect Dis.* 2000;182:1761–4. <https://doi.org/10.1086/317624>.
51. Mehta SK, Crucian BE, Stowe RP, Simpson RJ, Ott CM, Sams CF, et al. Reactivation of latent viruses is associated with increased plasma cytokines in astronauts. *Cytokine.* 2013;61:205–9. <https://doi.org/10.1016/j.cyto.2012.09.019>.
52. Stowe RP, Mehta SK, Ferrando AA, Feeback DL, Pierson DL. Immune responses and latent herpesvirus reactivation in spaceflight. *Aviat Space Environ Med.* 2001;72:884–91.
53. Stowe RP, Pierson DL, Barrett ADT. Elevated stress hormone levels relate to Epstein-Barr virus reactivation in astronauts. *Psychosom Med.* 2001;63:891–5.
54. Horneck G, Klaus DM, Mancinelli RL. Space microbiology. *Microbiol Mol Biol Rev.* 2010;74:121–56. <https://doi.org/10.1128/MMBR.00016-09>.
55. Zea L, Prasad N, Levy SE, Stodieck L, Jones A, Shrestha S, et al. A molecular genetic basis explaining altered bacterial behavior in space. *PLoS One.* 2016;11:e0164359. <https://doi.org/10.1371/journal.pone.0164359>.
56. Rosenzweig JA, Ahmed S, Eunson J, Chopra AK. Low-shear force associated with modeled microgravity and spaceflight does not similarly impact the virulence of notable bacterial pathogens. *Appl Microbiol Biotechnol.* 2014;98:8797–807. <https://doi.org/10.1007/s00253-014-6025-8>.
57. Cervantes JL, Hong BY. Dysbiosis and immune dysregulation in outer space. *Int Rev Immunol.* 2015;35:67–82. <https://doi.org/10.3109/08830185.2015.1027821>.
58. Klaus DM, Howard HN. Antibiotic efficacy and microbial virulence during space flight. *Trends Biotechnol.* 2006;24:131–6. <https://doi.org/10.1016/j.tibtech.2006.01.008>.

59. Lynch SV, Mukundakrishnan K, Benoit MR, Ayyaswamy PS, Matin A. Escherichia coli biofilms formed under low-shear modeled microgravity in a ground-based system. *Appl Environ Microbiol.* 2006;72:7701–10. <https://doi.org/10.1128/AEM.01294-06>.
60. Kim H, Bhunia AK. Secreted *Listeria* adhesion protein (Lap) influences Lap-mediated *Listeria monocytogenes* paracellular translocation through epithelial barrier. *Gut Pathog.* 2013;5:16. <https://doi.org/10.1186/1757-4749-5-16>.
61. Crabbe A, Schurr MJ, Monsieurs P, Morici L, Schurr J, Wilson JW, et al. Transcriptional and proteomic responses of *Pseudomonas aeruginosa* PAO1 to spaceflight conditions involve Hfq regulation and reveal a role for oxygen. *Appl Environ Microbiol.* 2011;77:1221–30. <https://doi.org/10.1128/AEM.01582-10>.
62. Shi J, Wang Y, He J, Li P, Jin R, Wang K, et al. Intestinal microbiota contributes to colonic epithelial changes in simulated microgravity mouse model. *FASEB J.* 2017;31:3695–709. <https://doi.org/10.1096/fj.201700034R>.
63. Juergensmeyer MA, Juergensmeyer EA, Guikema JA. Long-term exposure to spaceflight conditions affects bacterial response to antibiotics. *Microgravity Sci Technol.* 1999;12:41–7.
64. Fuse A, Sato T. Effect of microgravity changes on virus infection in mice. *J Gravit Physiol.* 2004;11:P65–6.
65. Yurov SS, Akoev IG, Akhmadieva AK, Livanova IA, Leont'eva GA, Marennii AM, et al. Genetic effects of cosmic radiation on bacteriophage T4Br+ (on materials of biological experiment "Soyuz-Apollo"). *Life Sci Space Res.* 1979;17:129–32.
66. Yurov SS, Akoev IG, Leont'eva GA. Effect of HZE particles and space hadrons on bacteriophages. *Adv Space Res.* 1983;3:51–60.
67. Hargens AR, Vico L. Long-duration bed rest as an analog to microgravity. *J Appl Physiol.* 2016;120:891–903. <https://doi.org/10.1152/japplphysiol.00935.2015>.
68. Globus RK, Morey-Holton E. Hindlimb unloading: rodent analog for microgravity. *J Appl Physiol.* 2016;120:1196–206. <https://doi.org/10.1152/japplphysiol.00997.2015>.
69. Sonnenfeld G. Use of animal models for space flight physiology studies, with special focus on the immune system. *Gravit Space Biol Bull.* 2005;18:31–5.
70. Gould CL, Sonnenfeld G. Enhancement of viral pathogenesis in mice maintained in an antiorthostatic suspension model: coordination with effects on interferon production. *J Biol Regul Homeost Agents.* 1987;1:33–6.
71. O'Donnell PM, Orshal JM, Sen D, Sonnenfeld G, Aviles HO. Effects of exposure of mice to hindlimb unloading on leukocyte subsets and sympathetic nervous system activity. *Stress.* 2009;12:82–8. <https://doi.org/10.1080/10253890802049269>.
72. Belay T, Aviles H, Vance M, Fountain K, Sonnenfeld G. Effects of the hindlimb-unloading model of spaceflight conditions on resistance of mice to infection with *Klebsiella pneumoniae*. *J Allergy Clin Immunol.* 2002;110:262–8.
73. Aviles H, Belay T, Fountain K, Vance M, Sonnenfeld G. Increased susceptibility to *Pseudomonas aeruginosa* infection under hindlimb-unloading conditions. *J Appl Physiol.* 2003;95:73–80.
74. Crucian BE, Stowe RP, Mehta SK, Yetman DL, Leal MJ, Quirarte HD, et al. Immune status, latent viral reactivation, and stress during long-duration head-down bed rest. *Aviat Space Environ Med.* 2009;80(5 Suppl):A37–44.
75. Kelsen J, Bartels LE, Dige A, Hvas CL, Frings-Meuthen P, Boehme G, et al. 21 Days head-down bed rest induces weakening of cell-mediated immunity – some spaceflight findings confirmed in a ground-based analog. *Cytokine.* 2012;59:403–9. <https://doi.org/10.1016/j.cyto.2012.04.032>.
76. Mehta SK, Crucian B, Pierson DL, Sams C, Stowe RP. Monitoring immune system function and reactivation of latent viruses in the Artificial Gravity Pilot Study. *J Gravit Physiol.* 2007;14:P21–5.
77. Uchakin PN, Stowe RP, Paddon-Jones D, Tobin BW, Ferrando AA, Wolfe RR. Cytokine secretion and latent herpes virus reactivation with 28 days of horizontal hypokinesia. *Aviat Space Environ Med.* 2007;78:608–12.

78. Karouia F, Peyvan K, Pohorille A. Toward biotechnology in space: high-throughput instruments for *in situ* biological research beyond Earth. *Biotechnol Adv.* 2017;35:905–32. <https://doi.org/10.1016/j.biotechadv.2017.04.003>.
79. Nelson E, Chait A. Portable diagnostics technology assessment for space missions. NASA. 2010. <https://ntrs.nasa.gov/search.jsp?R=20100011008> 2019-02-28T16:23:06+00:00Z.
80. Cohrs RL. Rapid saline test for varicella zoster virus. *New Horiz Transl Med.* 2015;2:5. <https://doi.org/10.1016/j.nhtm.2014.11.036>.
81. Technology Transfer & Commercialization Office. Rapid detection of shingles (varicella zoster virus- VZV). NASA. 2012. [https://www.nasa.gov/centers/johnson/pdf/690989main\\_VZV%20TOPSheet.pdf](https://www.nasa.gov/centers/johnson/pdf/690989main_VZV%20TOPSheet.pdf).
82. Graebe A, Schuck EL, Lensing P, Putcha L, Derendorf H. Physiological, pharmacokinetic, and pharmacodynamic changes in space. *J Clin Pharmacol.* 2004;44:837–53. <https://doi.org/10.1177/0091270004267193>.
83. Kast J, Yu Y, Seubert CN, Wotring VE, Derendorf H. Drugs in space: pharmacokinetics and pharmacodynamics in astronauts. *Eur J Pharm Sci.* 2017;109S:S2–8. <https://doi.org/10.1016/j.ejps.2017.05.025>.
84. Goncharov IB. Research on the particulars of pharmacological effects during long-term space-flight. NASA. 2018. [https://www.nasa.gov/mission\\_pages/station/research/experiments/533.html](https://www.nasa.gov/mission_pages/station/research/experiments/533.html).
85. Stenzel C. Deployment of precise and robust sensors on board ISS for scientific experiments and for operation of the station. *Anal Bioanal Chem.* 2016;408:6517–36. <https://doi.org/10.1007/s00216-016-9789-0>.
86. Dubeau-Laramée G, Rivière C, Jean I, Mermut O, Cohen LY. Microflow1, a sheathless fiber-optic flow cytometry biomedical platform: demonstration onboard the International Space Station. *Cytometry A.* 2014;85:322–31. <https://doi.org/10.1002/cyto.a.22427>.
87. Office of Technology Partnerships and Planning. The rHEALTH sensor. NASA. 2011. [https://technology.grc.nasa.gov/documents/\\_6\\_Universalbiomedicalanalysissensor\\_SS-rHealth-2011.pdf](https://technology.grc.nasa.gov/documents/_6_Universalbiomedicalanalysissensor_SS-rHealth-2011.pdf).
88. Roda A, Mirasoli M, Guardigli M, Zangheri M, Caliceti C, Calabria D, Simoni P. Advanced biosensors for monitoring astronauts' health during long-duration space missions. *Biosens Bioelectron.* 2018;111:18–26. <https://doi.org/10.1016/j.bios.2018.03.062>.
89. Zangheri M, Mirasoli M, Guardigli M, Di Nardo F, Anfossi L, Baggiani C, et al. Chemiluminescence-based biosensor for monitoring astronauts' health status during space missions: results from the International Space Station. *Biosens Bioelectron.* 2019;129:260–8. <https://doi.org/10.1016/j.bios.2018.09.059>.
90. Crews N, Ameel T, Wittwer C, Gale B. Flow-induced thermal effects on spatial DNA melting. *Lab Chip.* 2008;8:1922–9. <https://doi.org/10.1039/b807034b>.
91. Zanella I, Merola F, Biasiotto G, Archetti S, Spinelli E, Di Lorenzo D. Evaluation of the Ion Torrent PGM sequencing workflow for the routine rapid detection of BRCA1 and BRCA2 germline mutations. *Exp Mol Pathol.* 2017;102:314–20. <https://doi.org/10.1016/j.yexmp.2017.03.001>.
92. Castro-Wallace SL, Chiu CY, John KK, Stahl SE, Rubins KH, McIntyre, et al. Nanopore DNA sequencing and genome assembly on the International Space Station. *Sci Rep.* 2017;7:18022. <https://doi.org/10.1038/s41598-017-18364-0>.
93. Briggs K, Madejski G, Magill M, Kastritis K, de Haan HW, McGrath JL, et al. DNA Translocations through nanopores under nanoscale preconfinement. *Nano Lett.* 2018;18:660–8. <https://doi.org/10.1021/acs.nanolett.7b03987>.
94. Shapshak P, Somboonwit C, Sinnott JT. Artificial intelligence and virology – quo vadis. *Bioinformation.* 2017;13:410–1. <https://doi.org/10.6026/97320630013410>.

# Futuristic Methods in Virus Genome Evolution Using the Third-Generation DNA Sequencing and Artificial Neural Networks



Hyunjin Shim

**Abstract** The Third-Generation in DNA sequencing has emerged in the last few years, using new technologies that allow the production of long-read sequences. Applications of Third-Generation sequencing enable real-time data production, changing the research paradigms in environmental, and facilitating medical sampling in virology. To take full advantage of the large-scale data generated from long-read sequencing, an innovation in downstream data analysis is necessary. Here, we discuss futuristic methods using machine learning approaches to analyze big genetic data. We discuss the future of twenty-first-century virology by presenting advanced approaches for virus studies using real-time data production and on-site data analysis with Third-Generation Sequencing and machine learning methods. We first introduce the basic concepts in conventional statistical models and methods in virology, building gradually into the necessity of innovating downstream data analysis to meet the advances in sequencing technologies. We argue that artificial neural networks can innovate downstream data analysis, as they can learn from big datasets without model assumptions nor feature specifications, as opposed to current data analysis in bioinformatics. Furthermore, we discuss how futuristic methods using artificial neural networks, combined with long-read sequences can revolutionize virus studies, using specific examples in supervised and unsupervised settings.

**Keywords** Artificial neural networks · Supervised learning · Unsupervised learning · Third-Generation DNA sequencing · Long-read DNA/RNA · Experimental evolution · Global virology · Likelihood-free · Model-free · Data-driven · Big data in virology

---

H. Shim (✉)

Department of Earth and Planetary Sciences, University of California, Berkeley,  
Berkeley, CA, USA

e-mail: [jinenstar@berkeley.edu](mailto:jinenstar@berkeley.edu)

### Key Concepts

Viruses evolve in complex cycles of transmission among hosts, making their evolutionary trajectory difficult to predict. Virus genome evolution can be studied on a population-scale and even on a global-scale using statistical models and simulation-based inference methods. However, these methods are limited in performance to the particular model used for inference. As long-read sequencing is revolutionizing the quantity and quality of genetic data, an innovation in the downstream analysis using machine learning methods is an urgent task to achieve in computational biology. Artificial Neural Networks (ANN) is a model-free approach that has great potential for applications to big genetic data. Futuristic approaches in virology envision achieving real-time data production and on-site data analysis using Third-Generation Sequencing and machine learning analysis. This data-driven research takes full advantage of the advances in data production by learning directly from the vast amount of data produced (unsupervised learning), and by predicting the output from a new sample using prior data (supervised learning).

## 1 Introduction: Virus as Model Organism

Viruses are the most abundant biological entities on Earth, which can infect all types of life forms, and only replicate inside cellular organisms. Viruses exploit diverse replication-expression strategies, whereas cellular organisms only use double-stranded DNA as the replicating form and single-stranded RNA as the transcribed form [1]. Viruses also display a plethora of genome architectures – single or double-stranded, linear or circular, DNA or RNA, monopartite or multipartite – with sizes spanning three orders of magnitude [1]. Many viral pathogens exhibit complex evolutionary trajectories during transmission between hosts and diversification within them: they may experience strong inter-host population bottlenecks and rapid intra-host population growth, as well as intense selective pressures due to host immunity and drug therapy [2, 3]. These phenomena have been investigated in a range of human pathogens impacting global health, including Human Immunodeficiency virus (HIV) [4] and Influenza A viruses (IAV) [5].

Viruses are highly appropriate model organisms for evolutionary biology, not only because they have broad implications for global health and biogeochemical processes, but also for addressing fundamental research and medical questions. Viruses have relatively simple and compact genomes due to the constraint of being dependent on cellular organisms for replication. Thus, their genomes are the simplest systems to model complex genome-wide interactions including epistasis and clonal interference [6]. Viruses also are the smallest microbial organisms that are at the boundary of living and non-living; thus, they provide key platforms to understanding

central biological questions, such as the origin of life, recombination, selfish elements, and host-pathogen coevolution [1, 7, 8].

Recent advances in data production and data analysis relating to the field of genetics have opened opportunities to investigate these questions using big genomic datasets and novel computational methods [9]. In this chapter, the potential of viruses as model organisms is explored from the perspective of experimental evolution to that of global virology.

## 1.1 *Experimental Evolution of Viruses*

Viruses evolve in complex cycles of transmission among hosts; thus, it is difficult to predict the evolutionary trajectory of these pathogens in nature. To better understand the evolution of pathogens on a population-scale (or even on a global-scale), experimental evolution procedures have been developed as the simplified version of the complex world. Since these procedures control the environment during evolution, the number of necessary assumptions is minimized [10–12]. Furthermore, since the genome evolution of these pathogens can be observed temporally, additional biological factors can be integrated into evolutionary models. For example, critical evolutionary processes such as drift, epistasis, clonal interference, and large offspring variance can be considered, factors that have mainly been ignored in large-scale pandemic models. These experimental studies may elucidate how pathogens evolve at larger and more complex scales in nature.

The standard procedure for experimental evolution of pathogens is to serially passage the microorganisms in cell cultures under the presence or absence of given treatments (e.g. drug/disinfectant agents). Samples of the pathogens are then collected and sequenced in a time-serial manner, and temporal allele trajectories of all nucleotides are generated. Due to the small size of virus genomes, population-level whole-genome sequences are accessible at low cost. These datasets are analyzed with the goal of estimating population genetic parameters such as effective population size ( $N_e$ ) and selection coefficients ( $s$ ). These parameters are essential in experimental evolution studies, as temporal changes in allele frequency can be deterministic due to fitness advantages (selection), or stochastic due to finite population sizes (genetic drift). Other model parameters such as mutation rate and recombination rate are important for understanding these processes, which increase genetic variations that enable viruses to adapt to new environments.

## 1.2 *Model Parameters in Virus Genome Evolution*

In virus genome evolution, several efforts have been made recently to infer parameters such as effective population size ( $N_e$ ), selection coefficients ( $s$ ), and mutation rate ( $\mu$ ) from time-sampled experimental evolution data [10–13]. These methods are

primarily based on the Wright-Fisher model (where population census size is assumed to be constant and discrete, utilizing random sampling of biallelic gene copies), with different approximations to infer population genetic parameters from temporal allele trajectories. These model parameters are the first steps to understanding the landscape of virus evolution, with the eventual goal of finding solutions to the problems of global viral epidemics. For example, the inference of effective population size ( $N_e$ ), selection coefficients ( $s$ ) and mutation rate ( $\mu$ ) in an experimental evolution setting may provide the initial clues of the potential threat of viral pathogens adapting to new drugs or treatments in natural settings [10–13].

### 1.2.1 Wright-Fisher Model

The Wright-Fisher model is one of the most fundamental models proposed by Sewall Wright and R.A. Fisher to represent the random process of allele frequency changes in finite populations [14, 15]. The model generalizes genetic drift of alleles from one generation to the next as a binomial distribution, in the absence of other evolutionary processes such as selection, gene flow, and mutation. The binomial distribution defines the probability of sampling  $k$  alleles in a sample of  $N$ :

$$\Pr(X = k) = \binom{N}{k} f^k (1-f)^{N-k} \quad (1)$$

where  $f$  is the allele frequency. For viruses,  $N$  represents the haploid population size. The assumptions made to simplify this process in virus population dynamics include discrete and non-overlapping generations, constant population sizes, and independent alleles. In the absence of other forces, the probability that an allele becomes fixed in the population under genetic drift is equivalent to its initial allele frequency. The general consequences of the Wright-Fisher model in modeling genetic drift in finite populations are: the change in allele frequency due to genetic drift is random, the magnitude of genetic drift is bigger in smaller populations, and an equilibrium state of a population under the influence of only genetic drift is fixation or loss of an allele.

### 1.2.2 Effective Population Size

The census population size  $N$  is the number of individuals in a population – however, not all individuals contribute to the changes in allele frequency in most biological populations. Thus, the genetic size of a population is defined as the effective population size  $N_e$ , which is proportional to the magnitude of genetic drift.  $N_e$  can be determined by comparing the rate of genetic drift in a census population with the rate of genetic drift in an idealized population under the assumptions of the Wright-Fisher model.

There are two ways to estimate the effective population size of a sample population: inbreeding effective population size ( $N_e^i$ ) and variance effective population size ( $N_e^v$ ).

The estimates from these two definitions of effective population sizes may differ when model assumptions for each definition are not met. For viruses, only the concept of the variance effective population size is relevant as viruses are haploid populations. The variance effective population size,  $Ne^v$ , can be estimated through expressing the changes of allele frequency over time in many replicates as a variance:

$$Var(\Delta f) = \frac{f_{t-1}(1-f)_{t-1}}{2Ne^v} \quad (2)$$

where  $f$  is the allele frequency of an allele of interest, and  $t$  is the number of generations.

When the population size fluctuates over time, as it does occur in viruses through inter-host population bottlenecks and intra-host population growth during infection, the effective population size can be calculated using the harmonic mean:

$$\frac{1}{Ne} = \frac{1}{t} \left[ \frac{1}{Ne_1} + \frac{1}{Ne_2} + \dots + \frac{1}{Ne_t} \right] \quad (3)$$

where  $t$  is the total number of generations over the fluctuating sizes. This harmonic mean gives greater weight to smaller values, due to the inverse of population size – thus, genetic bottlenecks are important factors in determining the strength of genetic drift in populations over time.

### 1.2.3 Selection Coefficient

Natural selection occurs when one phenotype causes a greater chance of survival and reproduction and the genotype that causes this phenotype increases in frequency over generations. The fitness of an individual can be defined by absolute fitness or relative fitness. The absolute fitness of the genotype A at any generation  $t$  is given as:

$$Absolute\ fitness\ of\ A = \frac{N_A(t)}{N_A(t-1)} \quad (4)$$

where  $N_A(t)$  is the population size of the genotype A at the generation  $t$ . Alternatively, the ratio of genotype-specific growth rates of A to B is given as the relative fitness:

$$\frac{N_B(t)}{N_A(t)} = w^t \frac{N_B(0)}{N_A(0)} \quad (5)$$

where  $w$  is the strength of natural selection and  $t$  is the number of generations. When  $w > 1$ , the genotype B is growing faster than the genotype A, and when  $w < 1$ ,

the genotype A is growing faster than the genotype B. Thus, the change in the frequency of the genotype A due to natural selection is given as:

$$\Delta f_A = \frac{f_t w_A}{f_t w_A + (1 - f_t) w_B} - f_t \quad (6)$$

$$s_A = 1 - w_A \quad (7)$$

where  $f_t$  is the frequency of the genotype A at the generation  $t$ . As shown above, fitness can be expressed in terms of selection coefficient,  $s_A$ , by taking the difference between the relative fitness and one. As viruses are entirely dependent on host cells for replication, they experience strong selective pressures during viral life cycles due to host immunity and/or drug treatment, both in experimental evolution and in nature.

#### 1.2.4 Mutation Rate

Since mutation is the ultimate source of genetic variation for all organisms, there has been much effort to understand the effect of mutation rate on the course of evolution. For host-dependent viruses, mutation is a vital mechanism for switching between hosts and evading host immune system. Particularly in experimental evolution, it is common to start serial passaging of viruses with a clonal population [10–12] – thus, the high mutation rate of viruses is responsible for creating the diversity of a virus population over the course of evolution. To calculate the mutation rate, the pairwise nucleotide diversity of an idealized population is assumed to be directly proportional to  $\mu N_e$ , where  $N_e$  is the effective population size, and  $\mu$  is the mutation rate, according to the neutral theory of molecular evolution [16].

Recent evidence at whole-genome and single-gene levels, shows that mutation rates tend to evolve by selection to improve replication fidelity, whose variation is in turn created under random genetic drift [17]. This theoretical work on the mutation-rate evolution is based on the fact that most mutations are deleterious [16], which is especially true for most viruses whose genome is compact and dedicated to protein-coding. A mutator that increases mutation rate (such as a DNA polymerase variant, or a DNA repair protein) is subjected to the opposing forces of mutation pressure (input) and selection/recombination (output) [17]. The degree to which selection can reduce the mutation rate is dependent on the effective population size ( $N_e$ ) that determines the strength of genetic drift. This drift-barrier hypothesis postulates that when the  $N_e$  is small – thus, with strong genetic drift – the efficiency of selection in removing mutators decreases. For example, when viruses are subjected to mutagenic agents that increase mutation rates above the natural equilibrium, the response of virus genomes to such rapid changes in this vital mechanism is found to influence the effectiveness of these treatments against viral pathogens [12].

### 1.2.5 Recombination Rate

Recombination is the exchange of genetic materials between chromosomes, and sexual reproduction is the production of new organisms by combining genetic information from two parents through recombination [18]. Recombination occurs in diploid eukaryotic organisms during reproduction, and also in prokaryotic cells and viruses when genetic material is transferred from a donor to a recipient. Recombination generates genetic variation and is also involved in other functions such as DNA repair. The ubiquity of recombination and sexual reproduction across the tree of life is one of the biggest questions in evolutionary biology. Evolutionary explanations for the advantages of recombination argue that it generates greater variability by breaking down genetic associations [19].

Viruses undergo genetic change by antigenic drift (point mutation) and antigenic shift (recombination or re-assortment). Genetic recombination is frequent in both RNA and DNA viruses and occurs at highly variable rates. For example, the rate of recombination per nucleotide in Retroviruses (ssRNA-RT) like HIV exceeds that of mutation, though it can vary from high to nonexistent in other RNA viruses [7]. Similarly, segmented viruses exhibit re-assortment rates ranging from low (e.g., Hantaviruses) to high (e.g., influenza A virus) [7].

Recombination/re-assortment allows viral genomes to undergo major genetic changes during evolution, such as increasing virulence and pathogenesis, evading host immunity, altering transmission tactics, expanding host ranges, evolving antiviral resistance, and potentially creating new viruses. There is an ongoing debate on whether recombination is a form of sexual reproduction for viruses, as is for cellular organisms [7]. Despite the extensive results that the benefits of recombination are responsible for the ubiquitous presence of sexual reproduction [19], the viral community asserts that there is little evidence of recombination being favored by natural selection for viruses and that it is a mechanistic by-product of other aspects of viral biology [7].

Recombination is a particularly important mechanism in viral epidemics, since it allows viral genomes to undergo major genetic changes. Furthermore, numerous viral genomes are almost entirely functional which makes most mutations deleterious. Recombination is therefore essential for breaking down negative genetic associations in these cases, as any advantageous allele is likely to be linked with other deleterious mutations. Other stochastic events such as genetic drift (due to small effective population sizes within hosts) and sweepstake events (due to stochastic variation in reproductive success) during host-to-host transmission are significant factors impacting the prediction of viral evolution [20].

## 1.3 *From Experimental Evolution to Global Virology*

We can build on the knowledge acquired from experimental evolution to study the complex and extensive network of viruses in nature, with the goal of predicting the evolutionary trajectories of viruses on a global scale. Furthermore, the population

genetic knowledge acquired from the experimental evolution studies may help investigate the role of viruses in the long-term evolution of various ecological systems, such as microbial communities and human microbiota, given the ubiquitous presence of viruses. Investigating viral biology on a global scale is important for the following reasons.

Firstly, viruses have the potential to become a threat to global public health as shown by the recent outbreaks of Ebola virus<sup>1</sup> and Zika virus,<sup>2</sup> aided by the ever-increasing mobility of human populations. Moreover, the possibility of air-borne and water-borne transmission shows how these pathogens can transmit rapidly among their hosts. The emergence of new viruses may occur unpredictably by forming new subtypes through antigenic drift and/or antigenic shift, and recombination. In viral epidemiology, recent advances in genetic data production and analysis are helping to predict evolutionary paths of viral pathogens, based on prior data and intricate models [21, 22].

Secondly, viruses play vital roles in biochemical processes of global microbiota. It is estimated that marine viruses are responsible for greater than one-third of host mortality per day, through lysing host cells, and over 1000 gene migrations per day through horizontal gene transfer [23]. In cold environments, the impact of virulence and viral-host evolution on soil biogeochemical cycling is yet to be verified – furthermore, the conditions in cold soils such as permafrost and glaciers are changing due to accelerated climate change (global warming). The impact of viruses on the ecosystem of microbiota in climate-sensitive areas continues to increase. Thus, investigating viral abundance and diversity to characterize the Earth's virome through metagenomic data is vital [24].

Thirdly, viruses have vast implications in the evolution of the Earth's biomes – the role of viruses is essential in evolving and regulating microbial communities, and the extent of its influence in ongoing research. Bacteriophages, commonly known as phages, are viruses that infect microbes, and virus-host interactions result in coevolution that often leads to biological novelties. Phages are selfish replicators that propagate their genomes through infecting microbes – thus, microbes evolve defense mechanisms such as CRISPR-Cas adaptive immune system against phage infections. Moreover, phages and hosts co-evolve to evade these mechanisms [25].

Viruses have two distinct states of reproduction: virulent (lytic) and temperate (lysogenic) [26]. Lytic phages affect the mortality of microbes, and this is the major force behind the organic matter cycling in various ecosystems. Lysogenic phages are integrated into the chromosome of microbes, and these prophages replicate together with their hosts. As the fitness of prophages is directly linked to their host, new dynamics between hosts and pathogens arise, such as inhibiting superinfection of the host by other phages, and introducing virulence factors into their hosts to invade new niches. Lysogeny can also gradually enable phages to evolve from lytic

---

<sup>1</sup> Ebola virus disease. (2019.02.01). Retrieved from <https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease>

<sup>2</sup> Zika virus. (2019.02.01). Retrieved from <https://www.who.int/news-room/fact-sheets/detail/zika-virus>

phages (complete selfishness) to cryptic phages (complete domestication), where they become unable to form infective particles and evolve as one selectable unit with the host.

## 1.4 Likelihood-Free Inference in Virus Genome Evolution

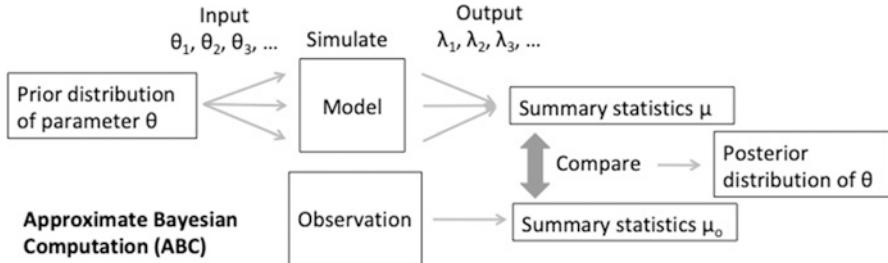
Advances in genomics allow the scientific community to move from small-scale inference such as experimental evolution to large-scale inference such as viral epidemics. In conventional evolutionary models, statistical inference with analytic approaches has been the focus of data analysis. However, it may be difficult to compute likelihood functions from models with many latent variables and/or complex architecture [27]. This problem is particularly prevalent in biological modeling since biological systems in nature tend to be very complex – such as phylogenetic trees, gene networks, and ecology systems, to name a few. Thus, numerous likelihood-free methods have been developed to overcome this limitation in model-based inference, including Approximate Bayesian Computation, where a likelihood function is replaced by simulation [28–30], and Variational Inference, where intractable integrals are approximated using standard probability distribution [31, 32]. This chapter discusses some of the popular simulation-based methods used to infer the parameters of a statistical model, with a particular focus on their application in virology.

### 1.4.1 Approximate Bayesian Computation

One of the optimal ways to describe observed data is by fitting a statistical model with known or/and unknown parameters. In Bayesian frameworks, the inference of model parameters from data is carried out by computing the posterior distribution  $P(\theta|Data)$  from the prior distribution  $P(\theta)$  and likelihood function  $P(Data|\theta)$  as follows:

$$P(\theta|Data) = \frac{P(Data|\theta)P(\theta)}{P(Data)} \quad (8)$$

However, when the model is very complex, the exact likelihood calculations can become intractable due to the presence of latent variables or heavy computation [27]. Approximate Bayesian Computation (ABC) is used to overcome this problem [33, 34]. ABC replaces likelihood calculations by simulating the model of interest directly with random input values from its prior distribution (Fig. 1). This allows very complex models to be inferred since simulating them with a set of input values is much easier and faster than calculating the likelihood of each parameter space [33]. These simulations are classified using summary statistics, which are



**Fig. 1** Illustration of the Approximation Bayesian Computation (ABC). ABC simulates a model with a set of input parameters and compares its output values to the observation to build a posterior distribution of parameters of interest

then compared to observed data to construct a posterior distribution for the parameter of interest (Fig. 1).

In ABC, summary statistics are used to summarize high-dimensional data such as a sample of multivariate measurements into low-dimensional statistics comparable to the observed values. However, as the number of summary statistics increases to describe complex data more accurately, a “curse of dimensionality” occurs in which most simulations get rejected due to increasing dimensionality. For example, if three summary statistics are accepted with 10% tolerance, 99.9% of the simulations will be rejected due to the three-dimensional tolerance space [27]. Furthermore, an unspecified loss of information by using insufficient summary statistics may lead to the problem of choosing the wrong model with ABC [35].

The ABC model choice is used to choose between two models:  $M_0$  as a null model, and  $M_1$  as an alternative model. The relative probability of  $M_1$  over  $M_0$  can be computed through the model posterior ratio as the Bayes factor  $B_{1,0}$  [34]:

$$\frac{p(M_1|D)}{p(M_0|D)} = \frac{p(D|M_1)p(M_1)}{p(D|M_0)p(M_0)} = B_{1,0} \frac{p(M_1)}{p(M_0)} \quad (9)$$

when the model prior  $p(M_0)$  is equal to  $p(M_1)$ . In practice, the model priors are made equal by producing the same number of simulations for each model and retaining the best simulations from the combined simulations. In ABC model choice, the posterior ratio is computed as the number of accepted simulations from  $M_1$  over those of  $M_0$  – giving the Bayes factor  $B_{1,0}$  which is an indicator of the support for a specific model.

As ABC has become an important tool for inferring parameters with complex models, many efforts are being made to improve the method [34], including integrations of Markov Chain Monte Carlo (MCMC) [30] and Hamiltonian Monte Carlo (HMC) [36]. These methods increase the power of ABC inference by improving the random sampling of input values from the prior distribution of a parameter. Markov Chain Monte Carlo (MCMC) is a widely used algorithm for sampling and attempts to solve the problem of sampling from a complex distribution.

Biological models are particularly complex, thus sampling from a complex distribution used to represent these models leads to the problem of the probability

of drawing a desired outcome very low. Thus, MCMC is introduced as a randomized algorithm that makes the sampling process more efficient by making the probability of sampling an output approximately equal to the target probability distribution. There are two steps in this algorithm: (1) Monte Carlo step which is a random walk taking a large sample, (2) Markov Chain step which estimates the expectation of the probability distribution. Examples of Monte Carlo methods include Metropolis-Hastings algorithm that uses a proposal density for the next move that can be accepted or rejected and Gibbs sampling that uses the conditional distributions of the target probability distribution. These methods are particularly efficient when sampling from high-dimensional distributions.

#### 1.4.2 Simulation-Based Inference in Virology

Simulation-based methods such as Approximate Bayesian Computation (ABC) are particularly useful in population genetics, where models to represent the forces of evolution (genetic drift, selection, gene flow, mutation) at a population-level become complex and dynamic. For this reason, the applications of ABC have been pioneered and widespread in population genetics – to name a few, for inferring growth rates and time of divergence from genetic data, and for model choice between competing models of human demographic history [27]. Other examples of ABC applications in the fields of evolution and ecology include phylogeography, systems biology and epidemiology [27].

In virology, simulation-based methods have been successfully applied to infer population genetic parameters such as effective population sizes and selection coefficients from the Wright-Fisher models. For example, according to the simulation studies comparing the performance of different time-serial methods for inferring population genetic parameters, the simulation-based method is shown to perform the best for virus populations that experience high selective pressures due to their dependency on host cells [37].

Despite their practicality, the performance of simulation-based methods is limited to the particular model used for inference. For example, the Wright-Fisher model used in these inference studies assumes mutations arising in the course of evolution of viruses are completely independent. This assumption is not valid even in the simplest organisms like viruses, as potential interactions between mutations are prevalent through compensation (epistasis) or competition (clonal interference). Thus, the conventional model-based inference, including simulation-based methods, has limited abilities to learn new biological features from data.

## 2 Artificial Neural Networks

Artificial Neural Networks (ANN) is a model-free approach that has great potential for applications to big genetic data. Deep learning using Artificial Neural Networks is a subfield of machine learning, where computers are able to extract patterns from

raw data and acquire their own knowledge of the real world. This task is done by representing data as a piece of information known as a feature – for example, the features about a patient such as the previous medical records, enable machine learning algorithms to predict the diagnosis of the patient. However, it is often difficult to manually extract the best features that might be useful for machine learning algorithms. This disadvantage is particularly problematic in complex tasks such as recognizing objects in pictures since these objects such as cars or dogs can take various shapes and colors. One solution is to use a machine learning technique called representation learning to learn features themselves instead of only mapping representation to output [38].

However, extracting high-level features from raw data is not an easy task. Deep learning tackles this problem by expressing representations in terms of simpler representations in a hierarchical manner. This advantage provides a flexible framework where features are easy to adapt and learn, without manual over-specification of features like other machine learning techniques. Historically, the fundamental ideas and architectures of deep learning have been inspired by biological neural networks in the brain; hence it goes by the name of Artificial Neural Networks or Multilayer Perceptron (MLP) or Feedforward Deep Network. In this chapter, the basic concepts of artificial neural networks are introduced, leading to the discussion of their potential applications in virology.

## 2.1 Feedforward Neural Networks

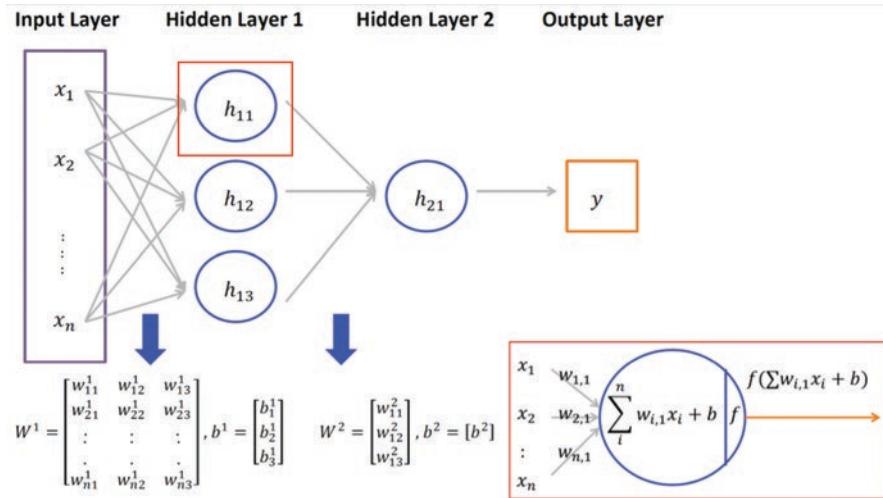
The goal of neural networks is to approximate a function  $f^*$ . A feedforward network defines a mapping  $y = f(x; \theta)$  and learns the parameters  $\theta$  that result in the best function approximation. In the feedforward neural networks, the information flows from the input  $x$  through the intermediate computations defining  $f$  to the output  $y$ , with no feedback (unlike recurrent neural networks). In mathematical terms, a multilayer perceptron is a function mapping input to output, which is composed of simpler functions. For instance, a multiplayer perceptron may be composed of a chain of three functions  $f^{(1)}$ ,  $f^{(2)}$ , and  $f^{(3)}$ :

$$f(x) = f^{(3)}\left(f^{(2)}\left(f^{(1)}(x)\right)\right) \quad (10)$$

where  $x$  is the input,  $f^{(1)}$  is the first hidden layer,  $f^{(2)}$  is the second hidden layer, and  $f^{(3)}$  is the output layer of the networks as shown in Fig. 2. The length of the chain is called the depth of the network, and it is where the concept of “deep learning” arises.

Within a neuron, an elementary unit shown as blue circles in Fig. 2, the input values in each hidden layer are multiplied by their corresponding weights ( $w$ ) and added by a constant bias term ( $b$ ):

$$f(x; w, b) = x^T w + b \quad (11)$$



**Fig. 2** Illustration of a fully-connected multilayer perceptron network

where the function used in hidden layers is called the activation function. Weights control the strength of connection and bias offsets the output while independent of data.

Depending on the particular neural networks used, the activation function can take diverse non-linear forms such as sigmoid  $f(x) = 1/(1 + e^{-x})$  or ReLU  $f(x) = \max(0, x)$  functions. During training, the data provide noisy and approximate examples  $x$  of  $f^*(x)$  to drive  $f(x)$  to match  $f^*(x)$ . The output layer must produce a value close to  $y$  without being told what intermediate layers should do (thus called hidden layers) – the learning algorithm implements an approximation of  $f^*$ .

## 2.2 Training Neural Networks

To train neural networks, we need to do the following steps:

1. Define a loss function that quantifies the inconsistency between predicted value  $\hat{y}$  and actual label  $y$ .
2. Find the parameters that minimize the loss function (optimization)

### 2.2.1 Loss Function

Given the dataset  $\{(x_i, y_i)\}_{i=1}^N$ , where  $x_i$  is the input and  $y_i$  is the output/label, the loss over the dataset is a sum of loss over the examples:

$$L = \frac{1}{N} \sum_i L(f(x_i, W), y_i) \quad (12)$$

Furthermore, the loss function has a regularization term (for example, L2 regularization tends to prefer smaller and spread out weights  $W$ ) that imposes a penalty on the complexity of a model to prevent overfitting.

$$R(W) = \sum_k \sum_l W_{k,l}^2 \quad (13)$$

Thus, the loss function is given as:

$$L(W) = \frac{1}{N} \sum_{i=1}^N L_i(f(x_i, W), y_i) + \lambda R(W) \quad (14)$$

where  $\lambda$  is the regularization strength, which is a hyperparameter to be chosen.

### 2.2.2 Optimization

In order to find the values of  $W$  that reduce a loss function, we can use the gradient descent:

$$\frac{df(x)}{dx} = \lim_{h \rightarrow 0} \frac{f(x+h) - f(x)}{h} \quad (15)$$

The computation of an analytic gradient is fast and exact but error prone. The computation of a numerical gradient is easy to write, but slow and approximate. To overcome the challenges of computing gradients over a huge neural network, the backpropagation algorithm computes the gradient cheaply by allowing the information from the cost to flow backward through the network. It works by splitting the gradient into local gradients with the chain rule to iteratively compute gradients for each layer.

## 2.3 Artificial Intelligence in Virology

The concept of Artificial Intelligence (AI) arises from the futuristic vision that machines can become intelligent by having the ability to learn and reason like humans. Two aspects of modern society are driving AI to be a thriving field – computational power and big data. Computational biologists also envision programmable computers that can mimic human intelligence to learn from biological data in the near future. The application of machine learning methods in biological systems allows model-free investigations, particularly through the rapid progress in deep learning, where neural networks are used to learn functional relationships from observed data [39]. Deep learning has the potential to solve complex problems such

as predicting viral epidemics using big genomic datasets, through multilayer neural networks, which were previously thought infeasible [39].

Statistical models in viral epidemics have to consider multiple factors and make strong assumptions, due to the complexity of viral biology, host biology, and host-pathogen interactions. As viruses only replicate within their hosts, they tend to evolve either a symbiotic relationship with their hosts that may be antagonistic (pathogenic), commensal (hitchhiking), and mutualistic (beneficial) [40]. This complexity in the viral life cycle makes the prediction of viral epidemics difficult, as predictive models have to incorporate events that occur at the micro-level within hosts and the macro-level between hosts. For instance, these models have to consider viral biology that is highly diverse between different virus types such as in virulence, replication, and recombination at the micro-level [1]. Furthermore, there are population genetic factors such as effective population size, mutation rate, and host defense systems that affect the evolutionary trajectories of viruses even within the same virus type. Additionally, these models have to incorporate factors that affect the transmission between hosts at the macro-level, such as migration patterns and transmission routes of their hosts.

Despite recent efforts, model-based methods have difficulties in incorporating these factors due to the complexity of solving or simulating these models [41]. Furthermore, these model-based methods do not take full advantage of hidden patterns in high dimensional biological data and/or big genetic data. To solve intricate problems such as viral epidemics, machine learning methods are promising tools to drive model-free and data-driven research to mine hidden patterns and make accurate predictions. Artificial neural networks are particularly useful for such problems, as they can extract features from data in a hierarchical manner [38].

We now have access to big genomic datasets that are essential resources for machine learning approaches. For example, Influenza Research Database ([fludb.org](https://www.fludb.org/)) currently has 661,000 segment sequences available,<sup>3</sup> which offers a great opportunity, but also an overwhelming challenge for traditional model-based approaches. Indeed, the scientific bottleneck no longer occurs at the level of genomic data production, but rather at the level of data analysis.

Moreover, other big data such as digital epidemiology and global human migration patterns are growing exponentially. Using artificial neural networks and genomic sequences from the Influenza Research Database, a predictive platform may be generated at the micro-level and/or macro-level. Using influenza virus (IAV) genome sequences and other epidemiologic data from the previous years as training examples, an algorithm may be designed to predict IAV strains likely to be prevalent in the following year. The current challenge is to curate different types of data as input and to design the architecture that can integrate both the events that happen at the micro-level within hosts and at the macro-level between hosts.

---

<sup>3</sup> Influenza Research Database. (2019.01.30). Retrieved from <https://www.fludb.org/brc/dataSummary.sp?decorator=influenza>

### 3 Third-Generation DNA Sequencing

In the field of biology, advances in technology and engineering are expanding the scope of questions that can be investigated with theoretical and computational tools. For example, DNA sequencing is becoming more efficient, accurate, and cost-effective each year, such that a human genome can be sequenced for only \$1000 at 30 $\times$  coverage, and at a speed of 18,000 genomes per year.<sup>4</sup> This innovation is astonishing, considering that the first full sequencing of the human genome was only completed in 2003. These rapid improvements in DNA sequencing allow more samples to be collected, from different individuals and/or at different time points.

Since the emergence of the Second Generation Sequencing in the late 1990s, lowered costs and improved efficacy of DNA sequencing is enabling the production of multiple time-point datasets [42], which are particularly revealing for the studies conducted in the domains of ancient DNA [42], experimental evolution [37], and clinical trials [43]. These datasets allow temporal observation of rapidly evolving organisms at different time-points in the past [10–12]. The studies of virus genome evolution benefit from the information on the temporal aspects of evolutionary forces, such as changes in selection, population sizes, or mutation rates [10, 12, 13]. Moreover, we can gain insights into the complex aspects of evolutionary trajectories, such as host-pathogen interactions or intra-/inter-species competition and cooperation.

Big genetic data generated from multiple individuals provide the means to investigate the genetic variants of a population, as shown by the success of genome-wide association studies (GWAS) [44]. For viruses, sampling multiple individuals at the whole-genome level is feasible due to the small genome size of viruses. However, most virus datasets have been produced with the technique of pooled sequencing, where genetic materials from several individuals are pooled for sequencing. While this method of pooled sequencing is cost-effective and time-efficient, individual haplotypes cannot be reconstructed as in individual sequencing [12]. The novel technologies of long-read sequencing are pioneering platforms to generate the genetic data from multiple individuals cost- and time-effectively, including viral pathogens. In this chapter, the recent advances in sequencing technologies and their impact on virus research are introduced.

---

<sup>4</sup> HiSeq X™ Series of Sequencing Systems. (2016.02.10). Retrieved from <http://www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet-hiseq-x-ten.pdf>

### ***3.1 DNA Sequencing Technologies***

#### **3.1.1 First-Generation Sequencing (1970)**

The structure of DNA was first modeled by James Watson and Francis Crick in 1953 [45], based on X-ray results by Rosalind Franklin. The first method of DNA sequencing was developed in the 1970s [46, 47], and the method developed by Frederick Sanger et al. in 1977 was introduced as one of the first commercially available techniques known as Sanger sequencing [47]. This method uses the selective incorporation of chain-terminating dideoxynucleotide (ddNTPs: ddATP, ddGTP, ddCTP, ddTTP) by DNA polymerase during *in vitro* DNA replication. Sanger sequencing was the most widely used sequencing method for the next 40 years.

#### **3.1.2 Second-Generation Sequencing (2000)**

By the year 2000, several new platforms were developed to achieve high-throughput sequencing method for DNA sequences. One of the most commercially successful products is from Illumina's Hi-Seq sequencers, which uses clonal amplification and sequencing by synthesis to allow parallel sequencing. The development of second-generation sequencing enabled more efficient and accurate DNA sequencing for a lower cost, for example, lowering the cost of sequencing a human genome from US\$1 million down to US\$4000 by 2013. Additionally, the rapid data production of DNA sequences has pushed the scientific community to develop new computational methods for sequence processing, storage, and analysis.

#### **3.1.3 Third-Generation Sequencing (2015)**

Most recently, another innovation in DNA sequencing has emerged using new methods that allow long-read sequences, as compared to the previous methods. The most notable techniques are Single-Molecule-Real-Time (SMRT) sequencing and Nanopore sequencing. SMRT sequencing is based on sequencing by synthesis, where DNA sequences are synthesized in small containers at the bottom of the well, and fluorescently labeled nucleotides are incorporated by DNA polymerase to be detected by the smallest light detector volume in the well. Nanopore sequencing detects the changes in ion current of a nanopore when a DNA sequence passes through it. Both technologies are revolutionizing DNA sequencing by the production of long-read DNA sequences (>10,000 nucleotides as compared to typically 100 nucleotides in the Second Generation Sequencing) in real-time.

### 3.2 Long-Read Sequencing for Virology

The Third-Generation DNA sequencing has emerged in the last few years using new technologies that allow the generation of substantially longer read sequences. In particular, the portable Nanopore sequencing device, MinION, can be used for de novo sequencing and metagenomics, and they are being tested for high-impact environmental research and pathogen analysis [48, 49].

Only a few years after the successful launch of the long-read sequencers, the scientific community is experiencing a rapid change in how genetic data are generated at the frontier research topics. For instance, since the first paper appeared in 2013 with the keyword “Nanopore” at the preprint server for biology,<sup>5</sup> the number of papers with the same keyword has been doubling every year until 2018 as shown in Table 1. At this trend, genetic studies with the Third-Generation sequencing are expected to become the norm of genetic research studies in a few years.

In virology, such innovations are also pervasive – various viruses are being sequenced with the long-read sequencer, including Ebola virus, chikungunya virus, dengue virus, yellow fever virus, and influenza A virus to name a few [49–53]. Since virus genomes are small and compact, long-read DNA/RNA sequences (>10,000 nucleotides) cover the complete genomes of many viral pathogens. Viral identification is made more efficient and accurate through such sequences in metagenomic samples. Other than reading long-read sequences, the Third-Generation sequencers such as Nanopore have further advantages for virus sampling. The portability and cost-effectiveness of these sequencers are enabling rapid and field-based analysis of pathogens on-site at resource-limited settings. For an epidemic response, early identification of the viral strain of a new outbreak is crucial in monitoring and containing the epidemic through effective vaccination [49, 53–55]. In the clinical setting, the capacity to sequence in real-time offers a revolutionary diagnostic tool for various viruses that can cause disease outbreaks in human [50, 52] and livestock [56, 57].

Furthermore, the ability to sequence DNA/RNA directly facilitates the characterization of genetic plasticity by eliminating the PCR step, for investigating complex topics such as epigenetic modification and transcriptome profiling [58–60]. For instance, long-read sequencing has greater efficiency in sequencing full-length transcripts and identifying RNA isoforms [60]. In metagenomics, these advantages

**Table 1** The number of papers with the keyword “Nanopore” at the preprint server for biology (BioRxiv) from 2013 to 2018

Year	2013	2014	2015	2016	2017	2018
BioRxiv “nanopore” papers	1	13	38	87	170	323

<sup>5</sup> BioRxiv: the preprint server for biology. (2019.02.07). Retrieved from <https://www.biorxiv.org/>

of highly parallel direct RNA sequencing also improve the assembly of viral genomes and the investigation of microdiversity [61]. As long-read sequencing is revolutionizing the quantity and quality of genetic data, an innovation in the downstream analysis using automated machine learning methods is an urgent task to achieve in computational biology.

## 4 Futuristic Methods in Virus Genome Evolution

The scientific community is experiencing a rapid change in the research paradigms, both through the advances in data production and data science. An ever-increasing growth of big data is evident in a wide variety of research fields, from computer vision, astronomy, weather forecasting, health care, to genetics. To process such data, researchers are pushing the boundaries of computational power and method. Novel computational approaches such as machine learning methods are being developed to learn from big datasets, and these advances in data science are inspiring researchers from other fields to apply those methods in their respective datasets. In computational biology, researchers attempt to apply these novel methods using machine learning to process and classify high dimensional biological data, to predict future events or to acquire new knowledge.

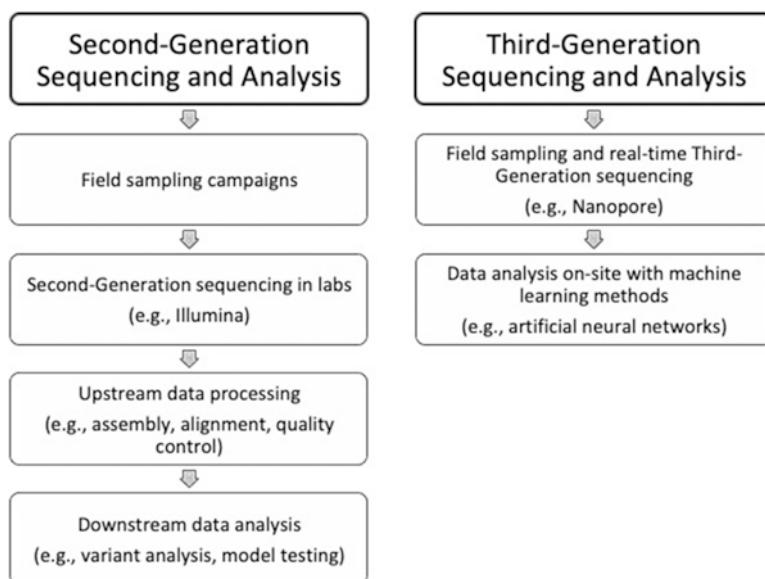
Model-free approaches such as machine learning methods eliminate the errors resulting from strong assumptions about underlying mechanisms by taking a data-driven approach [41]. The digitization of society has opened the age of big data, making machine learning methods more attractive by lightening the key burden of statistical estimation from small datasets [38]. This data-driven research takes full advantage of the advances in data production by learning directly from the vast amount of data produced (unsupervised learning). Furthermore, model-free methods may change the focus of study from inference to prediction, as machine learning methods can be designed to learn from previous input-output pairs to predict the output from a new sample (supervised learning). Unlike inference, the performance of predictive algorithms may be assessed directly with future data.

Advances in model-free approaches are particularly important for virus evolution. We aim to predict the course of viral pathogens efficiently and accurately from previous data, which is vital for critical societal and economic issues such as global public health and agricultural industry (supervised learning). Furthermore, we aim to acquire new knowledge on the ecology and evolution of viruses in various systems, through big genomic data of new viruses (unsupervised learning). In this chapter, the future of virology is discussed in conjunction with advancing sequencing technologies and data analyses, using specific examples in supervised learning and unsupervised learning setting.

#### 4.1 Futuristic Big Data Analysis for Virology

The rapid increase in genetic data leading to the current challenge of dealing with big data is expected to accelerate with the Third-Generation sequencing technologies. Applications of long-read sequencing enable real-time data production without an intermediate data processing step, as shown in Fig. 3. Until recently, considerable efforts have been devoted to the upstream data processing of the Second-Generation sequences – for assembly, alignment and quality control of short sequence contigs. In futuristic approaches, the conventional workflow of the Second-Generation Sequencing may be simplified into real-time data production and on-site data analysis using the Third-Generation Sequencing and machine learning analysis, respectively. This automatization is an important improvement to the diverse fields of virology such as experimental evolution as well as environmental and medical sampling campaigns. Long sequences from the Third-Generation sequencers eliminate the errors and uncertainties resulting from assembling short sequence contigs into a physical genome map in the Second-Generation sequencing. Thus, an innovation in the downstream data analysis step is necessary to analyze large-scale data generated from real-time long-read sequencing (Fig. 3).

It is of a future challenge in computational biology to achieve automated machine learning approaches that minimize human input to analyze real-time data from



**Fig. 3** Workflow of Second-Generation sequencing and analysis versus Third-Generation sequencing and analysis

long-read sequencers on-site. Machine learning combines pattern recognition and computational learning to perform predictive data analysis. As shown in Fig. 3, the application of neural networks innovates the downstream data analysis to process genetic data from the Third-Generation sequencing by big data analysis without model assumptions or feature specifications, as required in the current practice of bioinformatics.

## 4.2 Supervised Learning: Virus Classification

Supervised learning aims to predict an output from an input, based on previous input-output pairs. It infers a relation that maps input data to output data by learning a function that maps them through labeled training examples. Thus, supervised learning requires a big set of labeled training examples and a great deal of effort is put into gathering and labeling a training set from human experts or automated measurements. For artificial neural networks, the step to represent raw data as features is not necessary, but the architecture of neural networks (i.e., the number of neurons and the depth of layers) is a major factor that influences the performance of supervised learning. Then, the learning algorithm can be run on the labeled training set to increase the accuracy of the learned function for an optimized prediction. The performance of the learned function is measured through a test set that is unseen during the training set. The aforementioned advantages of artificial neural networks in learning non-linear and complex functions over several layers of neurons [38] apply to high dimensional genetic data.

In virology, the potential of supervised learning using artificial neural networks is significant. There are big databases of virus genomes readily available, which are essential resources to design a comprehensive platform for virus classification using supervised learning. Viruses display a vast diversity in form and function, with a vast functional diversity comparable to the corresponding microbial populations [62]. Applications of new sequencing technologies and advanced computational methods can improve the current knowledge of viruses by providing relevant data and techniques to test hypotheses.

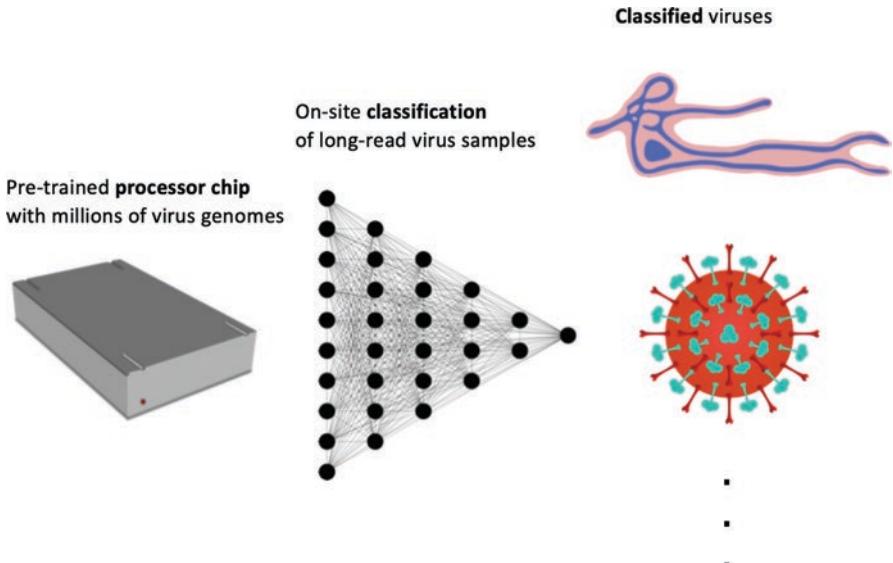
Despite their abundance, diversity and ecological significance in the biosphere, viruses are understudied in many ecological communities [63]. For example, the lack of knowledge on the impact of viruses is particularly relevant in rapidly changing ecosystems such as glaciers and permafrost [64–66]. Permafrost composes of approximately 25% of Earth’s land surface and stores almost 50% of global soil carbon in a frozen form. Near-surface permafrost across the Arctic region is rapidly thawing due to climate change [67]. Recent studies reveal that permafrost soils have an extensive presence and diversity of viruses – however, their impact on the microbial populations in this ecosystem of microorganisms is yet to be understood [68, 69]. On the other hand, glaciers cover approximately 10% of the Earth’s land surface and host diverse microbial communities [65]. Along with surprising micro-

bial diversity, viruses are also actively present in these cold ecosystems, with the rate of virus infection in these bacteria being one of the highest observed [66, 70]. Viruses in these permafrost- or glacier-covered soils experience strong selection pressures due to the extreme nature and isolation of these environments. Moreover, resident microorganisms have other key roles in global biogeochemical processes such as methane cycling, fermentation, degradation of complex molecules and carbon gas emissions. However, the difference in virus diversity and activity in these two rapidly changing ecosystems is yet to be studied and assessed. Given the ever-increasing amount of sequence data generated through high-throughput technologies, large-scale comparative analysis with metagenomics is an invaluable tool to investigate viral diversity and activity from such environmental samples. As an example of the futuristic approach to investigate the viral impact on the ecosystem of cold soils, field sampling campaigns in permafrost regions using novel portable and real-time devices for long-read sequencing may be used. After which, these samples may be analyzed on-site in real-time using machine learning methods for supervised learning. This will help us understand the virome in these climate sensitive regions and examine the impending impact of thawing on the global biogeochemical cycling and coupled viral-host evolution of microbial and viral communities.

Specifically, we can envision a portable processor chip for virus classification on-site of such sampling campaigns. In computer vision, one of the frontier fields in the application of neural networks, low-power and high-performance vision processor units including Movidius<sup>6</sup> are already commercially available. These are vision processor units that carry out inference of new examples from a trained network at very low power. As training takes most of the computational power and time, deep neural networks have already been trained with millions of labeled examples of images to learn highly non-linear and complex functions for image classification. Thus, these devices deploying deep learning can classify new examples quickly on-site, as inference/prediction using pre-trained neural networks is a computationally cheap task. For instance, Movidius may be applied for intelligent machine vision systems such as robotics and augmented & virtual reality to classify images as quickly as human vision is able to do. A processor chip with similar purposes is an example of futuristic methods in studying virus genome evolution – a device containing deep neural networks pre-trained with the vast amount of labeled virus genomes. These portable devices may be used along with long-read sequencing to provide an integrated pipeline of real-time data production and on-site data analysis, as shown in Fig. 4. A portable long-read sequencer and a portable processor chip of neural networks pre-trained with virus genomes may become sufficient for field sampling campaigns, from which diverse viruses can be sequenced and classified on-site. In the near future, such devices will revolutionize how sampling campaigns for clinical and environmental samples are undertaken, advancing the efficiency and pace of global virus studies for diagnostic purposes as well as for basic research.

---

<sup>6</sup>Movidius (2019.02.13). Retrieved from <https://www.movidius.com/>

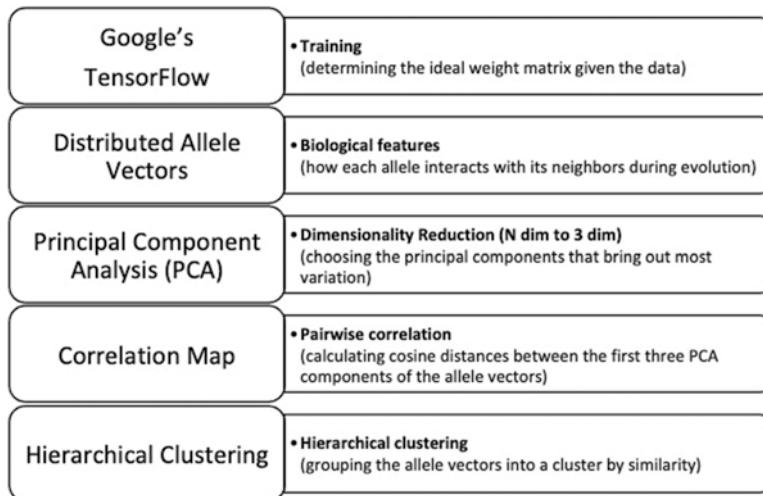


**Fig. 4** Supervised learning for on-site virus classification with a pre-trained processor chip

### 4.3 Unsupervised Learning: Virus Clustering

One of the principal advantages of artificial neural networks is its inherent capacity to extract features from raw data in a hierarchical manner [38]. The ability to extract features without strong assumptions or manual over-specifications makes artificial neural networks suitable for unsupervised learning. Unsupervised learning mines hidden patterns, structures, or features from unlabeled test data through methods such as cluster analysis. Such methods are useful at the exploratory stage of new studies without access to labeled datasets, to search for innovative questions and acquire pioneering knowledge.

In virology, unsupervised learning is a largely unexplored field despite its potential to solve challenges arising in the downstream analysis due to the increasing data production. Recently, a new approach has been developed to test the possibility of applying artificial neural networks to virus genetic data [6], based on the results from the simulation-based inference of population genetic parameters from an experimental evolution study [11]. In the study, a simple neural network was applied to the genetic datasets to execute unsupervised learning of virus genome evolution. This method was able to infer evolutionary distances from raw genetic datasets under the presence or absence of selective pressures in the experimental evolution setting. Figure 5 summarizes the workflow of unsupervised learning of virus genome evolution with the Nucleotide Skip-Gram Neural Network [6]. The first step is to train the Nucleotide Skip-Gram Neural Network using the deep learning software, where the genetic interactions between all significant mutations from the experimental evolution of



**Fig. 5** Workflow of unsupervised learning of virus genome evolution with the Nucleotide Skip-Gram Neural Network

virus [11] are learned as features of the neural networks. Subsequently, Principal Component Analysis (PCA) and hierarchical clustering are applied to produce a pairwise correlation map of these features, which reflects pair-wise evolutionary distances between all the mutations arising in the course of experimental evolution. Thus, this study demonstrates that mutations can be represented as distributed vectors that encode information of biological features and can be learned from data with artificial neural networks, rather than as discrete entities within classical population genetic models (e.g., Wright-Fisher model). This study shows that unsupervised learning with neural networks can achieve the same level of exploratory knowledge by clustering mutations of similar evolutionary trajectories automatically [6], as the human expertise had obtained from a series of experimental studies [11].

In viral metagenomics, the features to learn are biological or evolutionary similarities that exist between billions of sequences in raw metagenomic data from different sources. Unsupervised learning from artificial neural networks has the potential to offer considerable benefits for futuristic data exploration: from field sampling of natural populations using the real-time sequencer to pattern mining of raw metagenomic data. This novel pipeline eliminates the intermediate step of data processing through the Third-Generation sequencing and the necessity of model assumptions through artificial neural networks. Therefore, these proposed futuristic methods have broad implications of improving field sampling and genetic data analysis, as well as of investigating the evolution and ecology of viral and microbial communities.

**Acknowledgements** We thank Sunil Kumar Dogga, Ana K. Pitol, Hyun Jeong Shim for helpful discussions.

## References

1. Koonin EV, Dolja VV. A virocentric perspective on the evolution of life. *Curr Opin Virol.* 2013;3(5):546–57.
2. Tanaka MM, Valckenborgh F. Escaping an evolutionary lobster trap: drug resistance and compensatory mutation in a fluctuating environment. *Evolution (N Y).* 2011;65(5):1376–87.
3. Hall AR, Scanlan PD, Morgan AD, Buckling A. Host-parasite coevolutionary arms races give way to fluctuating selection. *Ecol Lett.* 2011;14(7):635–42.
4. Andrews SM, Rowland-Jones S. Recent advances in understanding HIV evolution. *F1000Research* [Internet]. Faculty of 1000 Ltd; 2017 [cited 2019 Jan 29];6:597. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/28529718>.
5. Schrauwen EJ, Fouchier RA. Host adaptation and transmission of influenza A viruses in mammals. *Emerg Microbes Infect* [Internet]. Nature Publishing Group; 2014 [cited 2019 Jan 29];3(1):1–10. Available from: <https://www.tandfonline.com/doi/full/10.1038/emi.2014.9>.
6. Shim H. Feature learning of virus genome evolution with the nucleotide skip-gram neural network. *Evol Bioinforma* [Internet]. SAGE PublicationsSage UK: London, England; 2019 [cited 2019 Jan 10];15:117693431882107. Available from: <http://journals.sagepub.com/doi/10.1177/1176934318821072>.
7. Simon-Loriere E, Holmes EC. Why do RNA viruses recombine? *Nat Rev Microbiol.* 2011;9:617–26.
8. Koonin EV., Dolja VV. Virus world as an evolutionary network of viruses and capsidless selfish elements. *Microbiol Mol Biol Rev* [Internet]. 2014;78(2):278–303. Available from: <http://mmbir.asm.org/lookup/doi/10.1128/MMBR.00049-13>.
9. Holmes EC. Viral evolution in the genomic age. *PLoS Biol* [Internet]. Public Library of Science; 2007 [cited 2016 Jun 24];5(10):e278. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/17914905>.
10. Foll M, Poh Y-P, Renzette N, Ferrer-Admetlla A, Bank C, Shim H, et al. Influenza virus drug resistance: a time-sampled population genetics perspective. *PLoS Genet* [Internet]. 2014 [cited 2014 Mar 20];10(2):e1004185. Available from: [http://www.ncbi.nlm.nih.gov/article/PLoS%20Genet%2010%20\(2\):e1004185.pdf?tool=pmcentrez&rendertype=abstract](http://www.ncbi.nlm.nih.gov/article/PLoS%20Genet%2010%20(2):e1004185.pdf?tool=pmcentrez&rendertype=abstract).
11. Zhong Q, Carratalà A, Shim H, Bachmann V, Jensen JD, Kohn T. Resistance of echovirus 11 to ClO<sub>2</sub> is associated with enhanced host receptor use, altered entry routes and high fitness. *Environ Sci Technol* [Internet]. American Chemical Society; 2017 [cited 2017 Sep 20];51(18):10746–55. Available from: <http://pubs.acs.org/doi/abs/10.1021/acs.est.7b03288>.
12. Carratalà Ripolles A, Shim H, Zhong Q, Bachmann V, Jensen JD, Kohn T. Experimental adaptation of human echovirus 11 to ultraviolet radiation leads to tolerance to disinfection and resistance to ribavirin. *Virus Evol* [Internet]. 2017 [cited 2017 Nov 4];3(November):1–11. Available from: [http://fdslive.oup.com/www.oup.com/pdf/production\\_in\\_progress.pdf](http://fdslive.oup.com/www.oup.com/pdf/production_in_progress.pdf).
13. Shim H, Laurent S, Matuszewski S, Foll M, Jensen JD. Detecting and quantifying changing selection intensities from time-sampled polymorphism data. *G3* [Internet]. 2016 [cited 2016 Apr 4];6(4):893–904. Available from: <http://www.g3journal.org/content/6/4/893.abstract>.
14. Wright S. Evolution in Mendelian populations. *Genetics* [Internet]. 1931;16(2):97–159. Available from: [http://www.ncbi.nlm.nih.gov/article/PLoS%20Genet%2010%20\(2\):e1004185.pdf?tool=pmcentrez&rendertype=abstract](http://www.ncbi.nlm.nih.gov/article/PLoS%20Genet%2010%20(2):e1004185.pdf?tool=pmcentrez&rendertype=abstract)
15. Fisher R. The genetical theory of natural selection. Oxford at the clarendon press. 1930.
16. Kimura M. Evolutionary rate at the molecular level. *Nature.* 1968;217:624–6.
17. Lynch M, Ackerman MS, Gout J-F, Long H, Sung W, Thomas WK, et al. Genetic drift, selection and the evolution of the mutation rate. *Nat Rev Genet* [Internet]. Nature Research; 2016 [cited 2016 Nov 15];17(11):704–14. Available from: [http://www.ncbi.nlm.nih.gov/article/PLoS%20Genet%2010%20\(2\):e1004185.pdf?tool=pmcentrez&rendertype=abstract](http://www.ncbi.nlm.nih.gov/article/PLoS%20Genet%2010%20(2):e1004185.pdf?tool=pmcentrez&rendertype=abstract)
18. Barton NH, Charlesworth B. Why sex and recombination? *Science (80- ).* 1998;281:1986–90.

19. Otto SP, Lenormand T. Resolving the paradox of sex and recombination. *Nat Rev Genet* [Internet]. Nature Publishing Group; 2002 [cited 2016 Jun 17];3(4):252–61. Available from: <http://www.nature.com/doifinder/10.1038/nrg761>.
20. Irwin KK, Laurent S, Matuszewski S, Vuilleumier S, Ormond L, Shim H, et al. On the importance of skewed offspring distributions and background selection in viral population genetics. Here [Internet]. Nature Publishing Group; 2016;1–7. Available from: <http://biorxiv.org/lookup/doi/10.1101/048975>.
21. Bedford T, Riley S, Barr IG, Broor S, Chadha M, Cox NJ, et al. Global circulation patterns of seasonal influenza viruses vary with antigenic drift. *Nature*. 2015;523:217–22.
22. Luksza M, Lässig M, Łuksza M, Lässig M. A predictive fitness model for influenza. *Nature* [Internet]. 2014 [cited 2014 Jul 10];507(7490):57–61. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24572367>.
23. Suttle CA. Viruses in the sea. *Nature*. 2005;437(7057):356–61.
24. Paez-Espino D, Eloe-Fadrosh EA, Pavlopoulos GA, Thomas AD, Huntemann M, Mikhailova N, et al. Uncovering Earth’s virome. *Nature*. 2016;536(7617):425–30.
25. Andersson AF, Banfield JF. Virus population dynamics and acquired virus resistance in natural microbial communities. *Science* (80–). 2008;320(5879):1047–50.
26. Argov T, Azulay G, Pasechnik A, Stadnyuk O, Ran-Sapir S, Borovok I, et al. Temperate bacteriophages as regulators of host behavior. *Curr Opin Microbiol* [Internet]. Elsevier Ltd; 2017;38:81–7. Available from: <https://doi.org/10.1016/j.mib.2017.05.002>.
27. Beaumont MA. Approximate Bayesian computation in evolution and ecology. *Annu Rev Ecol Evol Syst* [Internet]. Annual reviews; 2010 [cited 2014 Mar 20];41(1):379–406. Available from: <http://www.annualreviews.org/doi/abs/10.1146/annurev-ecolsys-102209-144621>.
28. Diggle PJ, Gratton RJ, Grattan RJ. Monte Carlo methods of inference for implicit statistical models. *J R Stat Soc Ser B* [Internet]. 1984 [cited 2016 Nov 10];46(2):193–227. Available from: <http://www.jstor.org/stable/2345504>.
29. Rubin DB. Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Ann Stat* [Internet]. Institute of Mathematical Statistics; 1984 [cited 2016 Nov 10];12(4):1151–72. Available from: <http://projecteuclid.org/euclid-aos/1176346785>.
30. Marjoram P, Molitor J, Plagnol V, Tavaré S. Markov chain Monte Carlo without likelihoods. *Proc Natl Acad Sci U S A* [Internet]. 2003 [cited 2014 Apr 29];100(26):15324–8. Available from: <http://www.pnas.org/content/100/26/15324.full>.
31. Minka TP. Expectation propagation for approximate Bayesian inference [Internet]. 2013 [cited 2017 Oct 17]. Available from: <https://arxiv.org/abs/1301.2294>.
32. Barthelmé S, Chopin N. Expectation propagation for likelihood-free inference. *J Am Stat Assoc* [Internet]. Taylor & Francis; 2014 2 [cited 2014 Apr 4];109(505):315–33. Available from: <https://doi.org/10.1080/01621459.2013.864178>.
33. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. *Genetics* [Internet]. 2002;162(4):2025–35. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1462356&tool=pmcentrez&rendertype=abstract>.
34. Sunnåker M, Busetto AG, Numminen E, Corander J, Foll M, Dessimoz C. Approximate Bayesian computation. *PLoS Comput Biol* [Internet]. 2013 [cited 2013 Oct 22];9(1):e1002803. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3547661&tool=pmcentrez&rendertype=abstract>.
35. Robert CP, Cornuet J-M, Marin J-M, Pillai NS. Lack of confidence in approximate Bayesian computation model choice. *Proc Natl Acad Sci U S A* [Internet]. 2011 [cited 2015 Aug 12];108(37):15112–7. Available from: <http://www.pnas.org/content/108/37/15112.full>.
36. Strathmann H, Sejdinovic D, Livingstone S, Szabo Z, Gretton A. Gradient-free Hamiltonian Monte Carlo with Efficient Kernel Exponential Families. In Cortes C, Lawrence ND, Lee DD, Sugiyama M, Garnett R (Eds). *Advances in Neural Information Processing Systems 28* (pp. 955–963). Curran Associates, Inc. 2015.
37. Foll M, Shim H, Jensen JD. WFABC: a Wright-Fisher ABC-based approach for inferring effective population sizes and selection coefficients from time-sampled data. *Mol Ecol Resour*. 2015;15(1):87–98.

38. Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016. 775.
39. Sheehan S, Song YS. Deep learning for population genetic inference. PLoS Comput Biol [Internet]. 2016;12(3):e1004845. Available from: <http://biarxiv.org/content/early/2015/10/02/028175.abstract>.
40. Roossinck MJ, Bazán ER. Symbiosis: viruses as intimate partners. Annu Rev Virol [Internet]. Annual reviews; 2017 [cited 2019 Feb 12];4(1):123–39. Available from: <http://www.annualreviews.org/doi/10.1146/annurev-virology-110615-042323>.
41. Angermueller C, Pärnamaa T, Parts L, Oliver S, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016;12(12):878.
42. Malaspinas A-S. Methods to characterize selective sweeps using time serial samples: an ancient DNA perspective. Mol Ecol [Internet]. 2016 [cited 2016 Nov 5];25(1):24–41. Available from: <http://doi.wiley.com/10.1111/mec.13492>.
43. Powney M, Williamson P, Kirkham J, Kolamunnage-Dona R. A review of the handling of missing longitudinal outcome data in clinical trials. Trials [Internet]. BioMed Central; 2014 [cited 2019 Jan 30];15:237. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/24947664>.
44. Hong EP, Park JW. Sample size and statistical power calculation in genetic association studies. Genomics Inform [Internet]. Korea Genome Organization; 2012 [cited 2019 Jan 30];10(2):117–22. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/23105939>.
45. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature [Internet]. 1953 [cited 2013 Oct 20];171(4356):737–8. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/13054692>.
46. Wu R. Nucleotide sequence analysis of DNA. Nat New Biol [Internet]. 1972 [cited 2017 Nov 4];236(68):198–200. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/4553110>.
47. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A [Internet]. 1977 [cited 2013 Oct 17];74(12):5463–7. Available from: <http://www.pubmedcentral.nih.gov/articlerender.fcgi?partid=431765&tool=pmcentrez&render type=abstract>.
48. Fuller CW, Kumar S, Porel M, Chien M, Bibillo A, Stranges PB, et al. Real-time single-molecule electronic DNA sequencing by synthesis using polymer-tagged nucleotides on a nanopore array. Proc Natl Acad Sci [Internet]. 2016;113(19):5233–8. Available from: <http://www.pnas.org/lookup/doi/10.1073/pnas.1601782113>.
49. Quick J, Loman NJ, Duraffour S, Simpson JT, Severi E, Cowley L, et al. Real-time, portable genome sequencing for Ebola surveillance. Nature [Internet]. Nature Publishing Group; 2016 [cited 2019 Feb 25];530(7589):228–32. Available from: <http://www.nature.com/articles/nature16996>.
50. Kafetzopoulou LE, Efthymiadis K, Lewandowski K, Crook A, Carter D, Osborne J, et al. Assessment of metagenomic Nanopore and Illumina sequencing for recovering whole genome sequences of chikungunya and dengue viruses directly from clinical samples. Eurosurveillance [Internet]. European Centre for Disease Prevention and Control; 2018 [cited 2019 Feb 7];23(50):1800228. Available from: <https://www.eurosurveillance.org/content/10.2807/1560-7917.ES.2018.23.50.1800228>.
51. Keller MW, Rambo-Martin BL, Wilson MM, Ridenour CA, Shepard SS, Stark TJ, et al. Direct RNA sequencing of the coding complete influenza A virus genome. Sci Rep [Internet]. Nature Publishing Group; 2018 [cited 2019 Feb 7];8(1):14408. Available from: <http://www.nature.com/articles/s41598-018-32615-8>.
52. Greninger AL, Naccache SN, Federman S, Yu G, Mbala P, Bres V, et al. Rapid metagenomic identification of viral pathogens in clinical samples by real-time nanopore sequencing analysis. Genome Med [Internet]. BioMed Central; 2015 [cited 2017 Oct 7];7(1):99. Available from: <http://genomemedicine.biomedcentral.com/articles/10.1186/s13073-015-0220-9>.
53. Faria NR, Kraemer MUG, Hill SC, Jesus JG de, Aguiar RS, Iani FCM, et al. Genomic and epidemiological monitoring of yellow fever virus transmission potential. Science (80-) [Internet]. American Association for the Advancement of Science; 2018 [cited 2019 Feb 7];361(6405):894–9. Available from: <http://science.sciencemag.org/content/361/6405/894>.

54. Mbala-Kingebeni P, Villabona-Arenas C-J, Vidal N, Likofata J, Nsio-Mbeta J, Makiala-Mandanda S, et al. Rapid confirmation of the Zaire Ebola virus in the outbreak of the Equateur province in the Democratic Republic of Congo: implications for public health interventions. *Clin Infect Dis* [Internet]. Oxford University Press; 2019 [cited 2019 Feb 8];68(2):330–3. Available from: <https://academic.oup.com/cid/article/68/2/330/5046756>.
55. Hansen S, Dill V, Shalaby MA, Eschbaumer M, Böhlken-Fascher S, Hoffmann B, et al. Serotyping of foot-and-mouth disease virus using oxford nanopore sequencing. *J Virol Methods* [Internet]. Elsevier; 2019 [cited 2019 Feb 8];263:50–3. Available from: <https://www.sciencedirect.com/science/article/pii/S0166093418303124?via%3Dihub=>.
56. Theuns S, Vanmechelen B, Bernaert Q, Deboutte W, Vandenhove M, Beller L, et al. Nanopore sequencing as a revolutionary diagnostic tool for porcine viral enteric disease complexes identifies porcine kobuvirus as an important enteric virus. *Sci Rep* [Internet]. Nature Publishing Group; 2018 [cited 2019 Feb 8];8(1):9830. Available from: <http://www.nature.com/articles/s41598-018-28180-9>.
57. Gallagher MD, Matejusova I, Nguyen L, Ruane NM, Falk K, Macqueen DJ. Nanopore sequencing for rapid diagnostics of salmonid RNA viruses. *Sci Rep* [Internet]. Nature Publishing Group; 2018 [cited 2019 Feb 8];8(1):16307. Available from: <http://www.nature.com/articles/s41598-018-34464-x>.
58. Prazsák I, Moldován N, Balázs Z, Tombácz D, Mogyeri K, Szűcs A, et al. Long-read sequencing uncovers a complex transcriptome topology in varicella zoster virus. *BMC Genomics* [Internet]. BioMed Central; 2018 [cited 2019 Feb 8];19(1):873. Available from: <https://biomedgenomics.biomedcentral.com/articles/10.1186/s12864-018-5267-8>.
59. Tombácz D, Prazsák I, Szűcs A, Dénes B, Snyder M, Boldogkői Z. Dynamic transcriptome profiling dataset of vaccinia virus obtained from long-read sequencing techniques. *Gigascience* [Internet]. Oxford University Press; 2018 [cited 2019 Feb 8];7(12). Available from: <https://academic.oup.com/gigascience/article/doi/10.1093/gigascience/giy139/5202462>.
60. Tombácz D, Balázs Z, Csabai Z, Snyder M, Boldogkői Z. Long-read sequencing revealed an extensive transcript complexity in herpesviruses. *Front Genet* [Internet]. Frontiers; 2018 [cited 2019 Feb 8];9:259. Available from: <https://www.frontiersin.org/article/10.3389/fgene.2018.00259/full>.
61. Garalde DR, Snell EA, Jachimowicz D, Sipos B, Lloyd JH, Bruce M, et al. Highly parallel direct RNA sequencing on an array of nanopores. *Nat Methods* [Internet]. Nature Publishing Group; 2018 [cited 2019 Feb 17];15(3):201–6. Available from: <http://www.nature.com/doifinder/10.1038/nmeth.4577>.
62. Dinsdale EA, Edwards RA, Hall D, Angly F, Breitbart M, Brulc JM, et al. Functional metagenomic profiling of nine biomes. *Nature* [Internet]. Nature Publishing Group; 2008 [cited 2018 Jan 27];452(7187):629–32. Available from: <http://www.nature.com/articles/nature06810>.
63. Roux S, Enault F, Hurwitz BL, Sullivan MB. VirSorter: mining viral signal from microbial genomic data. *PeerJ* [Internet]. PeerJ Inc.; 2015 [cited 2019 Feb 25];3:e985. Available from: <https://peerj.com/articles/985>.
64. Pratama AA, van Elsas JD. The ‘neglected’ soil virome – potential role and impact. *Trends Microbiol* [Internet]. Elsevier Ltd; 2018 [cited 2019 Feb 25];26(8):649–62. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/29306554>.
65. Bellas CM, Anesio AM, Barker G, et al. *Front Microbiol*. 2015;6(JUL):1–14.
66. Bellas CM, Anesio AM, Telling J, Stibal M, Tranter M, Davis S. Viral impacts on bacterial communities in Arctic cryoconite. *Environ Res Lett* [Internet]. IOP Publishing; 2013 [cited 2016 Sep 4];8(4):045021. Available from: <http://stacks.iop.org/1748-9326/8/i=4/a=045021?key=y=crossref.106c348767a8b18bcb0cbff9b3724b58>.
67. Schuur EAG, Abbott B. Climate change: high risk of permafrost thaw. *Nature* [Internet]. Nature Publishing Group; 2011 [cited 2018 Jan 29];480(7375):32–3. Available from: <http://www.nature.com/doifinder/10.1038/480032a>.
68. Colangelo-Lillis J, Eicken H, Carpenter SD, Deming JW. Evidence for marine origin and microbial-viral habitability of sub-zero hypersaline aqueous inclusions within permafrost near Barrow, Alaska. *FEMS Microbiol Ecol*. 2016;92(5):1–15.

69. Trubl G, Solonenko N, Chittick L, Solonenko SA, Rich VI, Sullivan MB. Optimization of viral resuspension methods for carbon-rich soils along a permafrost thaw gradient. PeerJ. 2016;4:e1999.
70. Bellas CM, Anesio AM. High diversity and potential origins of T4-type bacteriophages on the surface of Arctic glaciers. Extremophiles. 2013;17(5):861–70.

# Futuristic Methods for Treatment of HIV in the Nervous System



Allison Navis and Jessica Robinson-Papp

**Abstract** While many pathogens can seed the central nervous system (CNS) and cause severe morbidity and mortality, HIV presents a unique challenge. HIV is known to cross into the CNS early on in infection, and long-term infection leads to neurocognitive impairment. Antiretrovirals (ARVs) have decreased the complications related to HIV, and extended the lifespan for those living with the virus; however, they haven't lowered the overall prevalence of neurocognitive disorders. One theory is due to compartmentalization of the virus in the CNS, leading to low level viremia and inflammation. ARVs have variable penetrance into the brain, which may explain the compartmentalization. Many new techniques are being used to tackle this problem such as nanotechnology. Nanotechnology offers a means of modifying existing ARVs and increasing their CNS penetration as well as sustaining CNS concentrations by conjugating them with liposomes, or ligands that can bind to receptors on the surface of the blood-brain barrier. Further, techniques such as machine learning can help in understanding the pathogenesis of cognitive disorders by using computer algorithms to sort through hundreds of variables and determine pathological HIV genes or proteins that can be used to develop medications in the future. While these applications are still in the early-stages, they offer hope in tackling a longstanding problem.

**Keywords** Human immunodeficiency virus · Anti-retrovirals · Central nervous system · Blood-brain barrier · Nanotechnology · Machine learning

---

A. Navis (✉) · J. Robinson-Papp

Neuro-AIDS Division, Department of Neurology, Icahn School of Medicine at Mount Sinai,  
One Gustave L. Levy Place, New York, NY, USA  
e-mail: [Allison.navis@mountsinai.org](mailto>Allison.navis@mountsinai.org)

**Key Message**

- HIV infection of the central nervous system (CNS) leads to neurocognitive issues
- With the advent and widespread use of antiretrovirals (ARV) that suppress HIV in the serum, prevalence of HIV-associated dementia has decreased but rates of other cognitive changes have remained stable.
- The continued prevalence of HIV-associated neurocognitive disorders are thought to be partially due to low level HIV viremia in the CNS, which may be due to poor permeability of ARVs across the blood-brain barrier (BBB).
- Nanotechnology is being used to modify formulations of ARVs and increase their permeability across the BBB with sustained concentrations within the CNS.
- Machine learning is also being used to better understand the mechanisms of neurocognitive disorders so that new pharmaceuticals may be developed.
- While these new technologies show great promise, their utility in preventing cognitive changes still needs to be tested, as well as their safety in human populations

## 1 Introduction

Viruses that infect the central nervous system (CNS), do so through hematogenous spread from distant sites of infection elsewhere in the body. When this occurs, the resulting infection and inflammation can lead to meningitis, encephalitis, or myelitis depending on location of the infection in the nervous system. While the blood brain barrier (BBB) typically provides a protective barrier from toxins and pathogens that can infect the CNS, viruses have a variety of mechanisms that allow them to transverse this protective barrier. Viral infection mechanisms include intracellular transport through infected myeloid cells, entry through increased permeability of the BBB due to cytokines, and direct infection of endothelial cells of the BBB [1].

Once a virus enters the CNS, the innate immune system plays a critical role in pathogen recognition; this results in activation of signaling cascades that attack viral replication [1, 2]. Depending on the type of virus and resulting immune response, clinical outcomes can range from a self-limited aseptic meningitis [3] to long-lasting disability and death [4]. When there is no available treatment option for viral infections (e.g. West Nile virus and Enterovirus), the response of the innate immune system in controlling CNS neurovirulence is critical; however, it may not be rapid enough to prevent neurological sequelae [5, 6]. For those reasons, antiviral medications are critical when available and can be lifesaving treatments as is the case with acyclovir and herpes simplex encephalitis [7]. However, for viruses such as HIV, combined anti-retroviral (ARV) therapy can suppress viral replication but does not always clear the infection, especially in the CNS compartment, where varying CNS

penetration of medications allow viral replication in this compartment and can lead to varying neurological consequences [8, 21]. Given the lack of curative treatment for HIV and its potential for devastating effects, research efforts have focused on better understanding the problems caused by long-term HIV infection in the CNS along with possible treatment [8, 9].

Recent studies have utilized futuristic techniques including nanotechnology and forms of artificial intelligence to address these longstanding questions. Nanotechnology can modulate pre-existing treatment to improve pharmacodynamics and targeting within the body, which can be important tools when developing new medications is costly and inefficient. Artificial intelligence offers a rapid means of doing complex calculations and analysis that could take years if done through manpower alone. Herein, we focus on the cutting-edge research that seeks to understand the pathogenesis and find potential treatments for HIV infection in the CNS.

## 2 HIV and the Central Nervous System

HIV can be found in the CNS early on in infection, and in people who are neurologically intact [10, 11]. HIV gains access to the CNS via infected monocytes and macrophages, where it then resides in astrocytes and microglia [12]. In the early days of the AIDS epidemic, prior to widespread use of ARVs, severe cognitive impairment was commonly seen in patients living with AIDS and was largely due to astrocyte and microglia infection that leads to disruption and impairment of neuronal activity [13]. With the widespread use of ARVs, the more severe forms of dementia are not as commonly seen, but lesser degrees of cognitive impairment are still commonly seen within the spectrum of HIV-associated neurocognitive disorders (HAND) [13, 14]. The criteria for HAND were revised in 2006 to reflect these changes. The new criteria, developed in Frascati, Italy, describe three categories of HAND: asymptomatic neurocognitive impairment (ANI), mild neurocognitive disorder (MND), and HIV-associated dementia (HAD). The three criteria are distinguished by degree of impairment on neuropsychological testing and level of functional impairment [15]. Theories behind the shift from more severe cognitive impairment pre-ARVs to an equally common, but less severe form of HAND in the ARV-era ranges from low-grade infection of the HIV virus versus increased local inflammation and toxicity due to CNS compartmentalization [16].

CNS compartmentalization, also known as viral escape, occurs when the HIV viral load in the CSF is either (1) elevated compared to the plasma by  $>0.5 \log_{10}$  copies/mL or (2)  $>200$  copies/mL where plasma HIV RNA  $<50$  copies/mL [17]. The prevalence of CNS compartmentalization is thought to be anywhere between 4–21% of ARV-experienced people [16–22]. The causes of CNS compartmentalization are not fully known but are thought to be due to poor CNS penetration by different ARV regimens, ARV resistant HIV in the CNS, low level viremia or low CD4 nadir [22]. Poor CNS penetration by different ARV regimens has received much attention over the years as the widespread use of ARVs has led to declines in the prevalence of AIDS and HIV-associated complications while the prevalence of cognitive issues

associated with HIV has not decreased as much [23]. However, it is unclear whether poor CNS penetration of ARVs leads to CNS compartmentalization and HAND, or if inflammation and toxic effects of the ARVs, themselves, may play a role.

In 2008, the CHARTER group established a CNS penetration effectiveness (CPE) score to delineate which antiretrovirals had the best CNS penetration [23]. This was revised in 2010 [24]. Antiretrovirals such as Zidovudine, Nevirapine and Indinavir all have a high CPE score, while others such as Tenofovir, Ritonavir, Saquinavir show less degree of CNS penetration [24]. Studies have shown that combinations of ARVs with higher CPE scores may lead to improved performance on neuropsychological testing, indicative of the benefit of higher ARV concentration in the brain to control CSF viremia [25]. While the CPE score is a useful tool to indicate CNS concentrations of ARVs, the true measure depends on a balance of their ability to traverse the BBB, and degree to which it undergoes efflux. Nanotechnology offers a novel mechanism to take pre-existing medicines and alter their characteristics to improve CNS penetration.

It is for those reasons that much research has gone into developing novel nanotechnology formulations for the treatment of HIV in the CNS. In addition, techniques that utilize forms of artificial intelligence, such as machine learning, have sought to better understand the mechanism behind HAND- providing unique new means with which future treatment can be developed [26, 48].

These new techniques may prove to decrease rates of neurocognitive impairment by creating new formulations that cross the barrier more easily, changing the pharmacokinetics of the medications and reducing their efflux [27].

### **3 Nanotechnology and Antiretrovirals: Treatment of HIV in the CNS**

Nanotechnology provides a unique means of increasing antiretroviral permeability across the BBB by modifying the packaging of the drug to either cross the lipid membrane more easily or activate transcription factors to cross with transporters. Specific properties of nanoparticle formulations include reduced drug toxicity, reduced molecular size and prolonged circulation by avoiding metabolism through the reticulo-endothelial system. Studies over the past few years have examined liposomes, inorganic nanoparticles, and magnetic fields to assist in delivery of ARVs to the brain [28].

#### ***3.1 Solid Lipid Nanoparticles, Liposomes, and Nanoparticles***

Given the ability of lipids to readily traverse membranes, they have become the focus of much research in nanotechnology. Solid lipid nanoparticles (SLN) are one such technology that can be used to make drugs more stable and lipophilic and have

been recently used in combination with specific antiretrovirals. SLNs contain a mixture of solid lipids, emulsifiers (to stabilize lipid dispersion) and water, and can be administered in multiple ways. Once administered, lipases break down the lipid component into free fatty acids and release the drug [29]. As a result, they are non-toxic, biodegradable, and can reduce systemic toxicity by modulating lipid compounds to have different degradation rates (longer chain fatty-acids taking more time to degrade than shorter chain) [27, 30]. Efavirenz, a nonnucleoside reverse transcriptase inhibitor (NNRTI) commonly used as a first-line medication in HIV treatment, is a highly lipophilic drug with a sizable degree of first-pass metabolism and low bioavailability in the CNS. Recent work examined SLN formulations of Efavirenz that were intranasally injected into adult rats. The intranasal injections showed absorption rates 70 times greater than non-SLN formulated Efavirenz, and concentrations of drug in the brain that were 150 times higher, compared to the traditional oral route. The drug also showed a higher percentage of release for more sustained durations within the serum, when compared to the plain drug suspension [31].

While the SLN formulation of Efavirenz showed a high percentage of drug release and increased drug concentration *in vivo*, lipid formulation nanotechnology, such as liposomes, can have a short half-life when systemically administered, due to capture of the preparations by the reticulo-endothelial system [32]. As a result, a more targeted approach needs to be developed in order to increase delivery of the lipid formulations to the brain and avoid elimination in the blood during circulation. This has been achieved by using magnetic formulations of liposomal nanoparticles (containing iron oxide particles) that can target their delivery and increase permeability across the endothelial cells that compose the BBB. Application of an external magnetic field to the target site, allows delivery of the magnetic liposomes in an efficient manner and reducing deleterious side effects [33]. This approach was used to create a liposome of phosphatidyl choline-cholesterol and magnetite encapsulating azidothymidine 5-triphosphate (AZTTP), which is the active formulation of zidovudine (AZT), (a nucleoside reverse transcriptase inhibitor (NRTI)). The results showed a 3-fold higher permeability across an *in vitro* BBB model compared to free AZTTP and enhanced transendothelial migration of the monocytes, to which the magnetic field was applied [29].

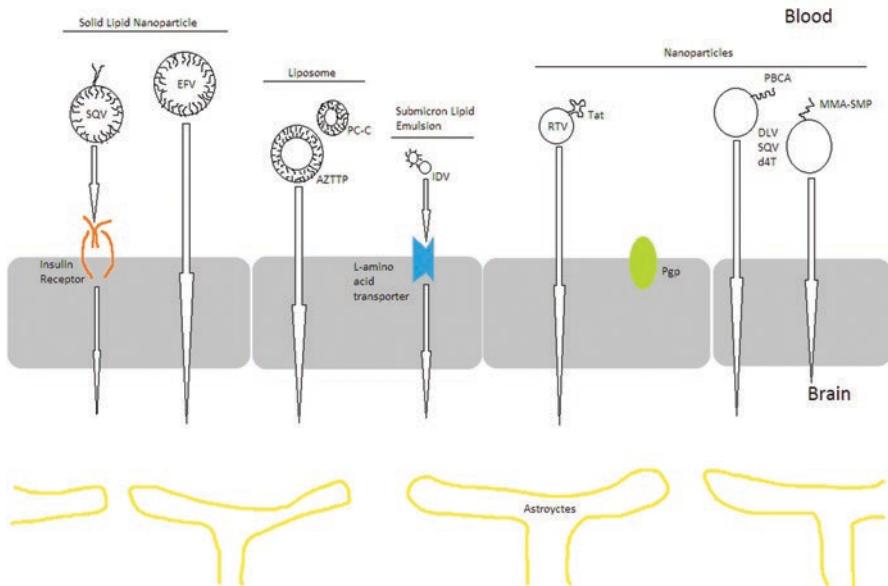
One potential factor that works against the utility of lipid formulations like liposomes and SLNs is the increased size of the formulations. Submicron lipid emulsions (SLEs) are lipid formulations with a size of 210-220 nm that can be attached to different molecules and pharmaceuticals to target their delivery and avoid increased molecular size. The L-amino acid transporter is a transporter on the surface of the BBB that brings branch-chain amino acids across the barrier, including lipo-amino acids that contain both the amino acid component needed for the transporter as well as improved lipophilicity. Bollam et al. used SLEs containing lipo-amino acids conjugated with Indinavir to improve uptake across the BBB and sustained drug concentrations. The Indinavir SLEs attached to lipo-amino acids showed higher brain concentrations at all time points compared to free Indinavir or Indinavir attached to just SLEs. However, there were no differences in concentrations among formulations measured in the heart, liver, or spleen. There was a slight

increase in drug concentration in the kidneys, but not to the same degree as in the brain (1.2–1.5 vs 2.5–3.3-fold, respectively), showing the precise targeting to the CNS that this system may offer [34].

In addition to the increased permeability that liposomes and SLNs can inherently provide, they can also be useful in formulations where protein ligands are attached to the surface and serve to activate receptors or block drug efflux. Studies looking at Saquinavir, a protease inhibitor, used SLNs with a peptide monoclonal antibody (83–14MAb) with a high affinity to the  $\alpha$ -subunit of the human insulin receptor grafted to its surface. The 83–14 MAb SLN stimulated endocytosis into human brain monocytes and increased brain permeability [35].

However, increased permeability still doesn't solve the problem of drug efflux certain transporters. The multidrug resistance gene transporter, p-glycoprotein (Pgp), is an ATP-dependent efflux pump located on the capillary endothelial cells of the BBB. Pgp plays a large role in the efflux of protease inhibitors and results in decreased CNS concentrations of that class of antiretrovirals [36, 37]. The HIV-1 trans-activating transcriptor (TAT) peptide can be used as a molecular beacon to enhance cellular delivery of drugs due to its ability to traverse membranes without transporters or receptor-mediated endocytosis. Rao et al. used Ritonavir (a protease inhibitor) loaded in nanoparticles and conjugated with TAT to determine if they could increase permeability across the BBB and decrease efflux by Pgp in both wild type and Pgp overexpressing multidrug resistant cell lines. In both models, there was a greater permeability of Ritonavir nanoparticles conjugated with TAT; however, there was a sustained concentration of the TAT conjugated nanoparticles seen up to 3 hours after infusion as compared to the non-TAT conjugated particles; and fluorescent microscopy of the rat brains showed the TAT-conjugated particles localized to the brain parenchyma throughout the sustained period [38]. They then proceeded to test whether the increased brain concentration of the Ritonavir, TAT-conjugated nanoparticles showed any neurotoxicity and if they were effective in suppressing HIV replication in the CNS. Results showed no toxicity of the Ritonavir conjugated nanoparticles compared to free Ritonavir. They also showed the Ritonavir conjugated nanoparticles had a 68% decrease in reverse transcriptase activity at 5 and 7 days compared to the controls, and a 70% decrease in HIV p24 levels. The decrease in p24 levels, a measure of HIV replication, was not significantly different compared to free Ritonavir. The results of the study showed that TAT-conjugated nanoparticles are not only safe, but efficacious [39].

While lipid formulations increase permeability of various ARVs by making them more lipophilic, other nanoparticles have been used to modify different properties of drug formulations to achieve similar results. Polybutyl-cyanoacrylate (PBCA) and methyl-methacrylate-sulfo-propyl-methacrylate (MMA-SMP) nanoparticles are charged colloids which modify the ionogenic properties of pharmaceuticals to increase absorption. Many drugs become highly ionogenic after metabolism due to dissociation of protons, which makes them more lipophobic. Charged colloids can reverse this. Kuo et al. developed formulations of Stavudine, Delavirdine and Saquinavir incorporated with either PBCA, MMA-SMP nanoparticles or SLNs and determined the differences in permeability across the BBB. The permeability of the



**Fig. 1** This figure shows the nanotechnology developed for different ARV formulations discussed in the text to improve permeability across the BBB. Abbreviations: EFV efavirenz, AZTTP azidothymidine, IDV indinavir, SQV saquinavir, RTV ritonavir, DLV delavirdine, D4T stavudine, PC-C phosphatidyl choline-cholesterol, Pgp p-glycoprotein, Tat HIV-1 trans-activating transcriptional activator, PBCA Poly-butyl-cyanoacrylate, MMA-SMP methyl-methacrylate-sulfo-propyl-methacrylate

different ARVs was enhanced 12–16 fold for PBCA formulations, 3–7 fold for MMA-SPM, and 4–11 fold for SLNs. Formulations with the PBCA and MMA-SPM nanoparticles were thought to increase permeability through several possible mechanisms, including increased uptake by the low-density lipoprotein (LDL) receptor as they were coated with polysorbate-80, which bears a similar structure to LDL particles [40]. Figure 1 shows many of the previously discussed ARV nanotechnology and their mechanisms for crossing the BBB.

Later work by Kuo combined the Saquinavir-PBCA, MMA-SMP, SLN formulations and exposed it to a magnetic field in an *in vitro* BBB model. The combined formulation under a magnetic field showed enhanced permeability; however, the magnetic field needed to be modulated as higher levels showed apoptosis of monocytes [41].

## 4 Future Directions in Treatment

While nanotechnology has provided a way of modulating pre-existing formulations of ARVs and improving CNS concentrations, it also offers a mechanism to develop entirely new therapeutic options. RNA interference (RNAi) is a system by which

double stranded RNA of pathogens are processed into short interfering RNAs (siRNA), associate with intracellular proteins, and thus cleave and degrade them. siRNA was first described using cells from *C.elegans*, and has been recognized as a major cellular defense mechanism in many organisms [42]. The introduction of siRNA into a system has been shown to stimulate the RNAi pathway and provides a new mechanism to disrupt pathogenic RNA [43, 44]. Chitosan nanoparticles conjugated to antibodies have been developed as a means of delivering siRNA across the BBB and target astrocytes to prevent HIV-1 replication. The chitosan nanoparticles were conjugated to a transferrin antibody and bradykinin B2 antibody that were theorized to bind to the transferrin receptor and bradykinin B2 receptor respectively. The ability to effectively target siRNA across the BBB is particularly important as siRNA tends to have limited stability and is quickly degraded by nucleases. The antibody conjugated chitosan particles then delivered siRNA to two targeted genes: SART3 and hCycT1 [47]. SART3 is involved in Tat transactivation and hCycT1 interacts with Tat and helps elongate RNA polymerase II at the HIV-1 pro-motor [45, 46] involved in TAT transactivation and transcription of HIV. Antibody conjugated chitosan nanoparticles showed increased uptake across the BBB and there was a significant reduction in the expression of the targeted genes [47]. While the potential effect on HIV replication was not quantified in the study, it provides a novel mechanism for future drug development.

## 5 Artificial Intelligence to Understand HIV in the CNS

CNS compartmentalization and low-level CNS viremia are thought to play a role in development of HAND; however, the mechanism is not fully known. Nanotechnology can increase the CNS concentrations of pharmaceuticals and theoretically help suppress CNS viremia, but that approach applies treatments to situations where the etiology is not fully understood. Recent research that uses machine learning, is seeking to better understand the pathogenesis of HAND [50–52]. While the studies are currently focusing on the underlying mechanisms, the results may provide new targets for therapeutics down the road [26, 48]. Machine learning is a branch of artificial intelligence, and a method of data analysis that uses computer algorithms that learn with experience. The program is able to predict outcomes based on the modified underlying algorithm. It typically involves multiple stages, beginning with the development of an algorithm that a researcher hopes will answer a question, then input of multiple data sets with increasingly unknown variables. The machine starts out by testing the algorithm against the known variables and learns from those, modifying the original code, to predict answers for unknown variables that can then be tested in a laboratory setting [49].

With this approach, several studies have been performed to predict biomarkers or genetic variations of HIV-associated proteins to determine if any specific genetic signatures predict HAND. A protein of specific interest is *gp120* glycoprotein as it is thought to mediate neuronal damage through induction of apoptosis.

*gp120* is encoded by the viral envelope gene, *env*. Initial work by Holman et al. in 2012, used machine learning on a meta-dataset of 860 *env* sequences to identify any that correlated with HAND and were able to identify 5 amino acid sequences that were associated with HAD ( $p < 0.05$ ) and 75% predictive accuracy [50]. Ogishi et al. furthered this work and created an algorithm based on 2494 *env* sequences. Sequences analyzed came from 9 studies and involved 85 patient specimens. From this model Ogishi et al. were able to identify only three genetic features that correlated with amino acid positions in the *gp120* glycoprotein. Furthermore, they were able to take the genetic features from their model and apply it to the Los Alamos HIV Sequence Database to predict the global burden of HAND based on how many specimens contained the genetic features. Based on this extrapolation, they estimated the global burden of HAND to be approximately 46% [51].

CNS strains of HIV are found to be genetically and phenotypically different from plasma isolates, which could also play a role in neurovirulence and development of HAND. Pillai et al. used machine learning to analyze the genetic features associated with HIV- neurotropism and neurovirulence. They enrolled 21 participants involved in studies at the HIV Neurobehavioral Research Center (HNRC) who had plasma and CSF HIV RNA over 500 copies/mL and no evidence of systemic or CNS illness (infection or malignancy), and then divided them into those that met criteria for CNS compartmentalization and those who did not. Eleven of the 21 participants were found to have CNS compartmentalization. They then analyzed samples from serum and CSF with a focus on the genetic signature of the C2-V3 region of the *env* protein to determine any differences between groups. They found that 4 positions in the V3 loop correlated with CSF compartmentalization, and a proline or histidine amino acid in position 13 of the V3 loop (*gp160* position 308) showed the greatest degree of compartmentalization ( $p < 0.044$ ) compared to the other amino acids found at that site- serine, threonine and asparagine. They then compared the *env* sequence results with neuropsychological testing and showed that the presence of a serine at position 5 of the V3 loop (*gp160* position 300) was the strongest predictor cognitive dysfunction [52].

## 6 Conclusions

Futuristic techniques such as nanotechnology, artificial intelligence, and machine learning are changing the way that research is done and providing insight into new therapeutic options for the treatment of diseases. This is particularly of importance with diseases and infections where the current treatment options are toxic, not completely curative, or show variability in pharmacokinetics and permeability. Within virology, these techniques are highly suitable for research on HIV, especially within the CNS, where infection leads to damage and disability, despite advances in treatment that suppress HIV in the serum.

The current approach is two-fold; modifying pre-existing therapies to improve their pharmacokinetics and targeting capability and using technology to better understand the physiological basis of HAND to potentially develop new pharmaceuticals. Nanotechnology, such as liposomes, SLNs and nanoparticles have shown to increase permeability across the BBB, sustain concentrations and avoid efflux of multiple ARVs. Notably, they have been used to improve CNS concentrations of multiple protease inhibitors- a class of ARVs with active efflux by the Pgp pump and a low CPE score [34–36, 39, 40]. Machine learning, and potentially other forms of artificial intelligence, are being utilized to elaborate the pathological mechanisms of HAND development. Results are showing that a finite number of regions on the C2-V3 region of the *env* protein may convey neurotropism and lead to cognitive dysfunction [50–52]. These sequences provide a target for future therapeutics and provide a means of minimizing CNS damage from cytotoxic compounds.

In addition, future studies with machine learning may predict the onset of HAND. Recent studies have utilized machine learning to develop multivariate models that consider either volumetric changes on MRI (e.g. in grey and white matter) or connectivity through functional MRI to predict HAND in small populations of patients. While the specific application to individuals has yet to show positive predictive results within the small sample sizes that were used, the models do show promise and with time may prove useful tools [53–55].

The greatest long-term utility within the field of HIV and the CNS, and virology in general, may lie in that of machine learning and similar tools with powerful computational abilities that can better understand pathogenesis and predict outcomes. The ability of the models to sort through thousands of variables and highlight potential targets for research, can save time and money. While nanotechnology has proven valuable at modifying ARVs and targeting them to the CNS, more research needs to be done on their overall safety. In addition, the precise mechanism of HAND development is not completely understood, and may ultimately not be due to low level CNS viremia or compartmentalization. There is research to show that changes in cognitive functioning in the ARV-era may have more to do with cardiovascular risk factors associated with ageing and the higher rates of hypertension, hyperlipidemia and diabetes associated with metabolic derangements from long-term ARV use [56]. In addition, ARVs themselves can be toxic to neurons, with greater toxicity at greater concentrations [57]. As a result, nanotechnology that targets ARVs to the CNS and increases their concentrations may not have a large impact on HAND development, or conversely, may increase neuronal damage.

Despite those concerns, these futuristic techniques provide a new means of assessing old problems. The combination of machine learning and nanotechnology offer a novel solution for understanding, treating and possibly preventing cognitive changes in HIV. Further work needs to be done to fully understand their utility in research and healthcare, but they show great promise.

**Acknowledgments** The authors would like to acknowledge the Neuro-AIDS Division at Icahn School of Medicine at Mount Sinai Hospital, including David Simpson MD and Susan Morgello MD for their support in this work.

## References

1. Nair S, Diamond MS. Innate immune interactions within the central nervous system modulate pathogenesis of viral infections. *Curr Opin Immunol.* 2015;36:47–53.
2. Ransohoff RM, Brown MA. Innate immunity in the CNS. *J Clin Invest.* 2012;122:1164–71. <https://doi.org/10.1172/JCI58644>.
3. Berlin LE, Rorabaugh ML, Heldrich F, et al. Aseptic meningitis in infants <2 years of age: diagnosis and etiology. *J Infect Dis.* 1993;168(4):888–92.
4. Teoh HL, Mohammad SS, Britton PN, et al. Clinical characteristics and functional motor outcomes of enterovirus 71 neurological disease in children. *JAMA Neurol.* 2016;73(3):300–7.
5. Frederickson BL. The neuroimmune response to West Nile virus. *J Neurovirol.* 2014;20(2):113–21.
6. Shih C, Liao CC, Chang YS, et al. Immunocompetent and immunodeficient mouse models for enterovirus 71 pathogenesis and therapy. *Viruses.* 2018;10(12):674. <https://doi.org/10.3390/v10120674>.
7. Whitley RJ, Alford CA, Hirsch MS, et al. Vidarabine versus acyclovir therapy in herpes simplex encephalitis. *N Engl J Med.* 1986;314(3):144–9.
8. Caniglia EC, Phillips A, Porter K, et al. Commonly prescribed antiretroviral therapy regimens and incidence of AIDS-defining neurological conditions. *J Acquir Immune Defic Syndr.* 2018;77(1):102–9.
9. Joseph J, Colosi DA, Rao VR. HIV-1 induced CNS dysfunction: current overview and research priorities. *Curr HIV Res.* 2016;14(5):389–99.
10. Gartner S, Liu Y. HIV neuroinvasion. In: Shapshak P, Levine AJ, Foley BT, Somboonwit C, Singer E, Chiappelli F, Sinnott JT, editors. *Global virology II-HIV and NeuroAIDS.* New York: Springer; 2017. p. 111–42.
11. Spudich S, Gisslen M, Hagberg L, et al. Central nervous system immune activation characterizes primary human immunodeficiency virus 1 infection even in participants with minimal cerebrospinal fluid viral burden. *J Infect Dis.* 2012;206(2):275–82.
12. Fischer-Smoth T, Bell C, Croul S, et al. Monocyte/macrophage trafficking in acquired immunodeficiency syndrome encephalitis: lessons from human and nonhuman primate studies. *J Neurovirol.* 2008;14(4):318–32.
13. Siminoi S, Cavassini M, Annoni JM, et al. Cognitive dysfunction in HIV patients despite long-standing suppression of viremia. *AIDS.* 2010;24(9):1243–50.
14. Heaton RK, Clifford DB, Franklin DR, et al. HIV-associated neurocognitive disorders persist in the era of potent antiretroviral therapy; a Charter Study. *Neurology.* 2010;75:2087–96.
15. Antinori A, Arendt G, Becker JT, et al. Updated research nosology for HIV-associated neurocognitive disorders. *Neurology.* 2007;69(18):1789–99.
16. de Almeida SM, Rotta I, Ribeiro CE, Smith D, HNRC Group, et al. Blood-CSF barrier and compartmentalization of CNS cellular immune response in HIV infection. *J Neuroimmunol.* 2016;301:41–8.
17. Rawson T, Muir D, Mackie NE, et al. Factors associated with cerebrospinal fluid HIV RNA in HIV infected subjects undergoing lumbar puncture examination in a clinical setting. *J Infect.* 2012;65:239–45.
18. Eden A, Fuchs D, Hagberg L, et al. HIV-1 viral escape in cerebrospinal fluid of subjects on suppressive antiretroviral treatment. *J Infect Dis.* 2010;202:1819–25.
19. Peluso MJ, Ferretti F, Peterson J, et al. Cerebrospinal fluid HIV escape associated with progressive neurologic dysfunction in patients on antiretroviral therapy with well controlled plasma viral load. *AIDS.* 2012;26:1765–74.
20. Nightingale S, Geretti AM, Beloukas A, et al. Discordant CSF/plasmaHIV-1 RNA in patients with unexplained low-level viraemia. *J Neurovirol.* 2016;22:852–60.
21. Anderson AM, Munoz-Moreno JA, McCleron D, et al. Prevalence andcorrelates of persistent HIV-1 RNA in cerebrospinalfluid duringantiretroviral therapy. *J Infect Dis.* 2016;215:105–13.

22. Mukerji SS, Misra V, Lorenz D, et al. Temporal patterns and drug resistance in CSF viral escape among ART experienced HIV-1 infected adults. *J Acquir Immune Defic Syndr.* 2017;75(2):246–55.
23. Letendre S, Marquie-Beck J, Capparelli E, CHARTER GROUP, et al. Validation of the CNS penetration-effectiveness rank for quantifying antiretroviral penetration into the central nervous system. *Arch Neurol.* 2008;65(1):65–70.
24. Antinori A, Lorenzini P, Giancola ML, et al. Antiretroviral CNS Penetration-Effectiveness (CPE) 2010 ranking predicts CSF Viral Suppression Only in Patients with an Undetectable HIV-1 RNA in Plasma. 2011. Conference on retroviruses and opportunistic infections, Boston MA. [http://www.natap.org/2011/CROI/croi\\_139.htm](http://www.natap.org/2011/CROI/croi_139.htm).
25. Smurzynski M, Wu K, Letendre S. Effects of central nervous system antiretroviral penetration on cognitive functioning in the ALLRT cohort. *AIDS.* 2011;25(3):357–65.
26. Bonet I. Machine learning for prediction of HIV drug resistance: a review. *Curr Bioinforma.* 2015;10(5):579–85. <https://doi.org/10.2174/1574893610666151008011731>.
27. Reynolds JL, Mahato RI. Nanomedicines for the treatment of CNS diseases. *J Neuroimmune Pharmacol.* 2017;12:1–5.
28. Teleanu DM, Chircov C, Grumezescu AM, et al. Blood-brain delivery methods using nanotechnology. *Pharmaceutics.* 2018;10(4):269.
29. Mehner W, Mader K. Solid lipid nanoparticles: production, characterization and applications. *Adv Drug Deliv Rev.* 2001;47(2–3):165–96.
30. Fiandra L, Capetti A, Sorrentino L, Corsi F. Nanoformulated antiretrovirals for penetration of the central nervous system: state of the art. *J Neuroimmune Pharmacol.* 2017;12:17–30.
31. Gupta S, Kesarla R, Chotai N, Misra A, Omri A. Systematic approach for the formulation and optimization of solid lipid nanoparticles of Efavirenz by high pressure homogenization using design of experiments for brain targeting and enhanced bioavailability. *Biomed Res Int.* 2017;2017:5984014.
32. Torchilin VP. Recent advances with liposomes as pharmaceutical carriers. *Nat Rev Drug Discov.* 2005;4(2):145–60.
33. Saiyed ZM, Gandhi NH, Nair MPN. Magnetic nanoformulation of azidothymidine 5'-triphosphate for targeted delivery across the blood-brain barrier. *Int J Nanomedicine.* 2010;5:157–66.
34. Bollam S, Kandadi P, Apte SS, Veerabrahma K. Development of indinavir submicron lipid emulsions loaded with lipoamino acids- in vivo pharmacokinetics and brain-specific delivery. *AAPS PharmSciTech.* 2011;12(1):422–30.
35. Kuo YC, Ko HF. Targeting delivery of saquinavir to the brain using 83-14 monoclonal antibody-grafted solid lipid nanoparticles. *Biomaterials.* 2013;34(20):4818–30.
36. Polli JW, Jarrett JL, Studenberg SD, et al. Role of P-glycoprotein on the CNS disposition of amprenavir (141W94), an HIV protease inhibitor. *Pharm Res.* 1999;16(8):1206–12.
37. Van der Sandt IC, Vos CM, Nabulsi L, et al. Assessment of active transport of HIV protease inhibitor in various cell lines and the in vitro blood-brain barrier. *AIDS.* 2001;15(4):483–91.
38. Rao KS, Reddy MK, Horning JL, Labhasetwar V. TAT-conjugated nanoparticles for the CNS delivery of anti-HIV drugs. *Biomaterials.* 2008;29(33):4429–38.
39. Borgmann K, Rao KS, Labhasetwar V, Ghorpade A. Efficacy of TAT-conjugated ritonavir-loaded nanoparticles in reducing HIV-1 replication in monocyte derived macrophages and cytocompatibility with macrophages and human neurons. *AIDS Res Hum Retrovir.* 2011;27(8):853–62.
40. Kuo YC, Su FL. Transport of stavudine, delavirdine, and saquinavir across the blood-brain barrier by polybutylcyanoacrylate, methylmethacrylate-sulfopropylmethacrylate, and solid lipid nanoparticles. *Int J Pharm.* 2007;340(1–2):143–52.
41. Kuo YC, Kuo CY. Electromagnetic interference in the permeability of saquinavir across the blood-brain barrier using nanoparticulate carriers. *Int J Pharm.* 2008;351(1–2):271–81.
42. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in *Caenorhabditis elegans*. *Nature.* 1998;391(6669):806–11.

43. Scarborough RJ, Gatignol A. RNA interference therapies for an HIV-1 functional cure. *Viruses*. 2018;10(1):8.
44. Bobbin M, Burnett JC, Rossi JJ. RNA interference approaches for treatment of HIV-1 infection. *Genome Med*. 2015;7(1):50.
45. Long L, Thelen JP, Furgason M, et al. The U4/U6 recycling factor SART3 has histone chaperone activity and associates with USP15 to regulate H2B deubiquitination. *J Biol Chem*. 2014;289:8916–30.
46. Chiu YL, Cao H, Jacque JM, Stevenson M, Rana TM. Inhibition of human immunodeficiency virus type 1 replication by RNA interference directed against human transcription elongation factor P-TEFb (CDK9/cyclinT1). *J Virol*. 2004;78:2517–29.
47. Gu J, Al-Bayati K, Ho EA. Development of antibody-modified chitosan nanoparticles for the targeted delivery of siRNA across the blood-brain barrier as a strategy for inhibiting HIV replication in astrocytes. *Drug Deliv Transl Res*. 2017;7(4):497–506.
48. Zhang L, Tan J, Han D, Zhu H. From machine learning to deep learning: progress in machine intelligence for rational drug discovery. *Drug Discov Today*. 2017;22(11):1680–5.
49. Libbrecht MW. Machine learning in genetics and genomics. *Nat Rev Genet*. 2015;16(6):321–32.
50. Holman AG, Gabuzda D. A machine learning approach for identifying amino acid signatures in the HIV env gene predictive of dementia. *PLoS One*. 2012;7(11):e49538.
51. Ogishi M, Yotsuyanagi H. Prediction of HIV-associated neurocognitive disorder (HAND) from three genetic features of envelope pg120 glycoprotein. *Retrovirology*. 2018;15:12.
52. Pillai SK, Kosakovsky Pond SL, Liu Y, et al. Genetic attributes of cerebrospinal fluid-derived HIV-1 env. *Brain*. 2006;129(7):1872–83.
53. Underwood J, Cole JH, Leech R, et al. Multivariate pattern analysis of volumetric neuroimaging data and its relationship with cognitive function in treated HIV disease. *J Acquir Immune Defic Syndr*. 2018;78(4):429–36.
54. Dsouza AM, Abidin AZ, Leistritz L, Wismuller A. Identifying HIV associated neurocognitive disorder using large-scale granger causality analysis on resting-state functional MRI. *Proc SPIE Int Opt Eng*. 2017;10133:101330M.
55. Dsouza AM, Abidin AZ, Wismuller A. Investigating changes in resting-state connectivity from functional MRI data in patients with HIV associated neurocognitive disorder using MCA and machine learning. *Proc SPIE Int Soc Opt Eng*. 2017;10137:101371C.
56. Wright EJ, Grund B, Robertson K, et al. Cardiovascular risk factors associated with lower baseline cognitive performance in HIV-positive persons. *Neurology*. 2010;75(10):864–73.
57. Robertson K, Liner J, Meeker RB. Antiretroviral neurotoxicity. *J Neurovirol*. 2012;18(5):388–99.

# Tuberculosis: Advances in Diagnostics and Treatment



Ju Hee Katzman, Mindy Sampson, and Beata Casañas

**Abstract** The World Health Organization’s “End TB Strategy” aims to achieve a 90% reduction in tuberculosis deaths and an 80% reduction in incidence by 2030; this calls for innovative approaches to counteract the high burden of disease through a two-prong strategy of early detection coupled with appropriate, timely treatment.

Artificial intelligence (AI) and machine learning are two remarkable technological breakthroughs that could revolutionize the management of tuberculosis. Examples of these technological advances include recognition of *tubercle bacilli* and abnormal chest radiograph recognition through the artificial neural networks. AI can learn to correlate massive amounts of data, which can be translated into clinically significant results. Another approach that has become promising is the development of virtual directly observed treatment (VDOT) via video, in areas with broadband internet and affordable devices that reach even remote places with tuberculosis patients. VDOT and drone drug delivery allow the treatment of tuberculosis in areas with minimal resources.

**Keywords** Tuberculosis · Artificial intelligence · Machine learning · Deep learning · Artificial neural network · Smear microscopy · Tuberculous pleural effusion · Chest x-ray · Computer aided programs · Host-directed therapy · Immunomodulation · Drug resistance · Virtually observed therapy

## Core Message

The eradication of tuberculosis (TB) remains a challenge worldwide. In this chapter, we discuss innovative strategies and technologic advancements in the diagnosis and treatment of tuberculosis.

J. H. Katzman · M. Sampson · B. Casañas (✉)  
Division of Infectious Diseases and International Health, Department of Medicine,  
Morsani College of Medicine, University of South Florida, Tampa, FL, USA  
e-mail: [juheekim@health.usf.edu](mailto:juheekim@health.usf.edu); [mindysampson@health.usf.edu](mailto:mindysampson@health.usf.edu); [beata@usf.edu](mailto:beata@usf.edu)

## 1 Introduction

The contagious nature of TB was elucidated by the French physician Jean Antoine Villemin in the 1860s with the discovery of the *tubercle bacilli* by Robert Koch in 1882 [1]. This discovery was pivotal in step in the fight against TB. Humans have been struggling with TB for thousands of years, and efforts to eradicate TB, globally, have led to the discovery of effective drug regimens since the 1940s, resulting in 60 million people swiftly treated and cured since 2000 [2]. However, TB remains one of the top ten global leading causes of death, as well as a leading cause of mortality from a single infectious agent before HIV/AIDS [2], calling for innovative approaches to counteract the high burden of disease. This is accomplished currently through early detection coupled with appropriate treatment. Among the notified incidence of TB cases in 2017, 60% had a documented HIV test result [2]. The battle to eradicate TB revolves around underdiagnosis, underreporting of detected cases, access to care, and gaps in detection and treatment of multidrug-resistant TB (MDR TB) and HIV-associated TB.

Delays in reaching prompt eradication relative to the developments of ground-breaking diagnostics and therapeutics warrant a synopsis of our current strategy towards TB. The use of current and future technology harbors hopes for both clinician and patient. The capabilities of artificial intelligence (AI) and its implications are vast – from massive data analysis to diagnostics and machine learning [3].

## 2 Artificial Intelligence in Diagnostics for Tuberculosis

### 2.1 *The Artificial Neural Network*

The term **AI** was introduced in 1955 as “the science and engineering of making intelligent machines” [4]. The use of AI has evolved from solving complex mathematical problems to finding novel solutions in complex challenges in the field of military, security, transport, manufacturing, and medicine. Machine learning or Deep Learning is an application of AI, where it uses mathematical algorithms to learn. Its learning ability and performance are improved through experience, i.e., the utilization of vast amounts of data. Experience is supplemented either through previous examples or the use of rewards and punishments [4]. An **Artificial Neural Network** (ANN) is a tool in AI that processes information through highly interconnected processing elements in response to external inputs [5]. The application of the tools in AI is being extensively studied in the biomedical field including medical informatics, diagnosis, statistics, and robotics. Concrete examples include, and are not limited to, discovering novel therapeutic targets (e.g. proteins) amongst thousands of molecules, identifying risks for chronic disease using specific clinical algorithms, detecting emotional disturbances among patients with psychiatric conditions, and using robots in surgery.

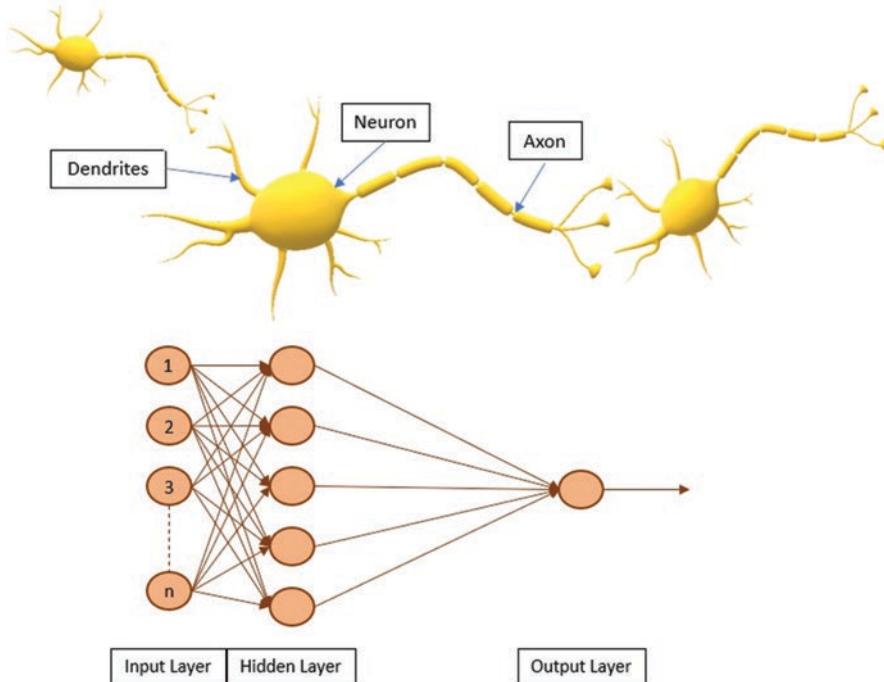
**Table 1** Terminology used in Artificial Neural Networks [5]

Term	Definition
Neuron	Information processing artificial nodes
Input Layer	A neuron that receives information outside the neural network
Output Layer	A neuron that contains classifications or interpretations of the neural network
Hidden Layer	Neurons that helps in processing information between input and output layers
Weights	Determines the strength of the connection that is activated with training and interpretation between neurons
Threshold	A parameter of the neuron expressed in the real number set to determine output from the sum of weight*input
Training an ANN	A procedure to assign values to weights and determining the combination to result in a minimum possible error
Learning	The adaptation of a neuron to alter its input/output due to changes in operating parameters

Artificial Neural Networks are modeled from the human brain and its complex networking system. Table 1 shows common terminologies used in ANN. The ANN functions similarly to a neuron complex of the human brain as seen in Fig. 1.

The key to the benefits of ANN includes its structural stratification. The first neurons on an input layer receive and transmit data to the hidden layer, where the data are mathematically multiplied according to the weights that are assigned to each input. Weight is a number that determines the strength of the association between two neurons. Following mathematical processing of the data, the results are forwarded to the following hidden layers of neurons then ultimately to the output layer. The output layer then reports the processed data and conclusions [5]. The ANN can be trained with a set of predetermined “correct” values to increase the accuracy of the output values [6]. An ANN has different structures based on the number of hidden layers (increases as the complexity between the input vs. output data) and propagation of the data whether it be forward towards the output layer or back-propagation (data returned to previously hidden layers) [6]. To explain this concept, consider a group of patients who were diagnosed with active TB with positive sputum smear or culture. This set of patients’ information can be used as a training set (supervised training) for an ANN designed to predict whether a patient has TB or not. Their symptoms such as cough, weight loss, fatigue, HIV status, etc. will be assigned a certain weight and will be used as inputs. The ANN’s output will then be compared to the “correct” output, i.e., patients who have active TB. During unsupervised training, if the ANN’s output is inaccurate, it can “learn” to alter the weights that have been previously assigned to calculate the output accurately. The advantage of ANN lays in its ability to tolerate vast amount of complex data and recognize new patterns [6].

Researchers have utilized different aspects of these concepts to aid the diagnosis of TB, such as direct identification of TB *bacilli* in smear microscopy, reading chest x-rays, and using clinical algorithms to classify patients with symptoms of TB. The following sections will further elaborate on these examples.



**Fig. 1** Human brain neuron complex contrasted to a single-layered forward artificial neural network

### 2.1.1 Detection of *Mycobacterium bacilli*

Microscopic sputum findings and culture tests are one of the most important diagnostic tests for the diagnosis of TB. Sputum obtained from patients are transferred on a slide applied with special stains to be microscopically examined. This microscopy process, which was developed more than 100 years ago, depends on the number of bacteria present in a given sample [2]. Microscopy is a time-consuming and arduous task for pathologists because of the minuscule size of the *bacilli*, requiring the use of high-power magnification X 1000 with limited visual areas [7]. TB bacilli are 2–4  $\mu$  in length and 0.2–0.5  $\mu$  in width [8]. Culture technology requires a fully equipped laboratory, and the TB culture process is very slow, i.e., up to 12 weeks. These techniques remain necessary for monitoring TB patients; however, many countries are now showing a preference for rapid molecular testing [2]. These newer techniques including polymerase chain reaction (PCR), transcription-mediated amplification, and strand displacement amplification. However, these advanced techniques are not feasible for extensive use at the desired level of sensitivity due to expense. An ideal screening test is one that would have a high sensitivity coupled with low cost and ease of use.

One basic approach to the use of ANN in the detection of *Mycobacterium tuberculosis bacilli* involves microscopy. The ANN is trained through multiple images of TB bacilli, identified as purple-red thin rods, with acid-fast staining. The pathologists determine the training set of images. One study resulted in a sensitivity and specificity of 97.9% and 84.7%, respectively, after processing more than 3 million samples. The reduced specificity stems from the AI's inability to recognize contaminant *bacilli* from pathogenic *bacilli*, especially when they are surrounded by histological reactions such as caseous necrosis, granulomas, and inflammation. Given the high sensitivity with moderate specificity, the authors have concluded that this approach and new technique could be used to help pathologists' workload [7].

In the twentieth century, scientists developed "electronic nose detection system" to detect and discriminate patterns and profiles of volatile organic compounds (VOC) in early diagnosis of cardiovascular disease, chronic obstructive lung disease, asthma, urinary tract infections, wound infections and breast and lung cancer [9–11]. An electronic detector (e-detector) detects complex volatile compounds from an odor through an array of partially specific chemical sensors coupled with a pattern-recognition system [12]. *In vitro* study has shown the ability to discriminate *M. tuberculosis* from control and unknown samples including *M. avium*, *M. scrofulaceum* and *Pseudomonas aeruginosa* [10]. A handheld, point-of-care device developed by researchers in the Netherlands is undergoing extensive clinical trials for the diagnosis of a variety of illness [13]. One study performed in Bangladesh uses this device that takes less than 10 minutes per subject for screening. This study revealed a sensitivity of 95.9% and a specificity of 98.5% in differentiating between healthy controls versus patients with TB. The authors postulate that this device if used to screen patients with TB, could potentially increase the detection of actual prevalence by 300–400%, likely due to the increased sensitivity and specificity of detecting patients with TB.

Diagnostic accuracy critically depends on the correct subdivision and selection of the samples in the initial training set for the ANN, because of the mathematical model used is derived from it. The limitation also springs from the problem of selecting the gold standard, which is to be used to diagnose TB via sputum culture. Once the accuracy of the reference or gold standard can be improved and hence more accurate training can occur for ANN, the output from the ANN will achieve higher sensitivity and specificity [11].

Another study performed in Egypt used the e-detector as a screening tool coupled with a prior qualitative estimation of TB diagnosis using ANN on blood, urine, breath, and sputum samples. These samples were collected in respective sealed containers, and the odors of volatile organic compounds (VOCs) present over these samples was carried by dry air into the sensor of the e-detector. Of the 260 TB patients examined, this study showed a sensitivity >95% for blood, breath, and urine samples with a specificity 100% for blood and breath samples in distinguishing patients with TB versus healthy controls [12].

### 2.1.2 Diagnosis of TB in Pleural Effusions

The diagnosis of pleural tuberculosis is especially a challenge since the current methods of diagnosis, i.e., cultures or nucleic acid amplification (NAA) have high specificity but low sensitivity [14]. Nevertheless, early diagnosis is imperative to reduce complications such as tuberculous empyema or fibrosis, and, ultimately, progression to pulmonary TB. An invasive pleural biopsy is often required to establish the diagnosis with 80% sensitivity [15]; however, the procedure poses a significant risk for both complications and expense. No gold-standard test exists for diagnosis of pulmonary TB. Biomarkers such as adenosine deaminase (ADA) and interferon-gamma are currently tested in the pleural effusion [16]. These markers are non-specific inflammatory markers that have high sensitivity in the inflammatory process. Moreover, these diagnostic tools are still not available in remote settings. Seixas et al. [14] proposed to aid in the diagnosis of pleural TB in a non-invasive way using the ANN. Their research showed an increase in sensitivity of the ANN when multiple biochemical studies, including ADA, NAA and IgA-enzyme-linked immunosorbent assay were combined. Their ANN model has shown to increase the pretest probability from 70% to over 90% among HIV patients coinfected with tuberculosis. The ideal clinical application of this study is the possibility of avoiding a pleural biopsy for diagnosis of pleural TB. This model could also be developed for a mobile phone or tablet without the need for internet connection. The study, however, was limited by a low number of controls, hence the assessment of specificity [14]. Another limitation of this model is the same as previously mentioned, in that the accuracy of the ANN method depends on the accuracy of the training set, i.e., having the available gold standard to diagnose pleural effusions. Other approaches have similarly used a combination of clinical and biochemical data (ADA, chemistry, cytokines) using a logistic regression model [17, 18] and have shown to improve diagnostic performance. Unfortunately, these studies have used multiple biomarkers, which may not be available worldwide.

Interestingly, Li et al. [19] used a support vector machine to prioritize and simplify clinical and biochemical data to be used to diagnose TB among patients without HIV infection. In machine learning, support vector machine improves generalization in the data to be classified. The researchers identified the top five indices out of 77 different variables: Pleural effusion ADA, pleural effusion lymphocyte percentage, age, temperature and color of the pleural effusion. The 147-person study population was any patient age >15 years old with pleural effusion of unknown etiology in China. The “gold standard” was pleural fluid cultures, pleural biopsy, thoracentesis and, thoracoscopy as a last resort. Thoracentesis and thoracoscopy are invasive procedures that require drainage and direct visualization of the pleural space, respectively. This rapid and cost-effective model achieved a sensitivity of 93.4% and specificity of 97.6%. Their model could easily be installed on mobile phones or tablets without the internet. The study was limited by sample size and have a high prevalence of TB among the study population [19]. Using only a few variables to achieve such high diagnostic value is very promising.

### 2.1.3 Imaging

The concept of AI in imaging predates that of the development of the technology itself. Its application commenced in 1967; a computer was used to highlight differences in the optical densities of mammogram films produced by a fax scanner, a telecommunication device that transmitted and reproduce documents by radio wave [20]. Currently, AI technology has greatly improved with the use of digital imaging coupled with the increased computing power of AIs [21]. Digital radiography (DR) has many advantages over conventional radiograph such as improved image quality, availability, lower operation costs, and safety [22]. DR uses phosphor plate to capture x rays which are later digitized into a picture archiving and communication systems (PACS) as opposed to the conventional film-screen, which is more expensive, requires storage space, and needs more staff to be handled. Using the more efficient phosphor plate also means less x-ray dose for patients. Once the image is in PACS, it can be intensified, i.e., brightened without the need for increased radiation dose [22]. Currently, artificial neural networks are used in radiology as the most advanced AI system [23]. As described above, ANN has the capability of being trained through supervised and unsupervised learning. Supervised learning compares the anticipated versus “correct” output. Unsupervised learning involves assigning weights derived from observation, correlation, and the input data through process of either feedforward or back-propagation [23]. ANN that has been trained by human observations can continue to learn through direct digitized images as inputs through extrapolating its bits of knowledge of more straightforward cases [23].

The critical application of AI to software and computer-aided programs (CAD) are already in use in mammography [24–27] and lung cancer screening [28–32]. The limitation of AI-based CAD lies in its high false-positive rates [23].

Scientists are exploring methods to incorporate CAD to screen for TB in chest x-rays. CAD technology would be beneficial in areas without available radiologists and would at least lessen the burden for the radiologists. CAD should not replace a radiologist but act as an adjunct to the radiologist [23]. A chest radiograph is one of the essential pieces of the puzzle to establish the diagnosis of TB. Chest radiographs are also used to rule out active TB disease for people living with human immunodeficiency virus (HIV) on antiretroviral therapy (ART) and their HIV-negative household contacts >5 years of age. The accuracy of chest radiographs depends on the experience of the radiologist or technician for interpretation and radiographic imaging equipment [2, 22]. Moreover, CAD achieves a more standardized interpretation compared to high intra- and inter-reader variability [21]. The existing technology achieves high sensitivity >85% in patients with symptoms of presumptive TB but with low, variable specificities ranging from 23% to 69% [21]. Several methodologies in these studies [33–36] that limit generalization and clinical application include patient populations with a high burden of TB and HIV, potential bias in excluding patients who were lost to follow-up and minimal data on patient characteristics such as age, HIV status, gender, history of active TB, co-morbidities and sputum smear status [21]. Further studies are needed to thoroughly assess performance and cost-effectiveness in clinical care [21]. Future CAD could be trained to detect general abnormalities in chest x rays, not only TB.

### 3 Treatments

Treatment for TB infection continues to be a challenging public health topic throughout the world. Currently, the first-line treatment consists of isoniazid, rifampin, ethambutol, and pyrazinamide for a minimum of 4 months [37]. Treatment durations are extended up to 20 months for extensively drug-resistant TB regimens [38]. Drug toxicities and prolonged duration of therapy make completing treatment for TB a challenge. Therefore, extensive research to consider extended treatment options is being carried out.

Personalization of therapy by optimizing drug regimens is being studied to assist in the treatment of drug-resistant TB. Currently, phenotypic minimum-inhibitory-concentration (MICs) are utilized to determine drug susceptibility. There is evidence that by increasing the dosage of some drugs, they will continue to be active against TB even with MICs, which would typically be considered resistant. Therefore, it may be possible to optimize dosing and continue to utilize drugs that were previously considered resistant. Also, TB gene sequencing has resulted in recognition of drug resistance mutations. However, before this becomes FDA-approved and clinically available, further databases for drugs and mutations need to be created, along with easier access to mutation testing [39].

Host-directed therapies, which include treatments that augment the patient's immune response to TB infections are being investigated [39]. The immune system plays an integral role in the control of TB [40]. In the HIV-infected population, the necessary balance of immune function is often observed, where lack of viral control places people at increased risk for development of TB infection [41]. Alternatively, a rapid increase in immune function can lead to deleterious clinical outcomes such as immune reconstitution inflammatory syndrome (IRIS) [42]. Investigations in host-directed specific therapies have focused both on supporting the ability of the immune system to clear TB infections and on limiting excessive inflammatory responses, with associated consequences [39].

Corticosteroids are an example of a broadly acting host-directed therapy. They have been shown to control the immune response in TB meningitis, leading to decreased mortality [43]. In patients with TB infection and HIV, with a CD4 count of less than 100, the risks of IRIS and associated morbidities were reduced when corticosteroids were utilized [42].

More specific immune system targets are also being investigated to avoid the consequences of broadly acting immunosuppressive therapies such as corticosteroids. For example, there is evidence that the metabolic system, which impacts the immune system is altered by infections like Mtb. Two metabolic energy sensors AMP-activated protein kinase (AMPK), and sirtuin 1 (SIRT1) can affect the host's response to infections [44]. Increased activity of AMPK and SIRT1 is associated with improved clearance of TB and a reduction of infection-related inflammation [45]. Pharmaceutical research in SIRT1-activating compounds such as resveratrol is ongoing [45].

Other novel areas of research for host-directed therapy include immune modulation with phosphodiesterase inhibitors, aspirin, nonsteroidal anti-inflammatory

drugs, vitamin D, tumor necrosis factor inhibitors, and statins. Currently, most of these novel therapies are still experimental, and further research is needed. It is also prudent to note that immune system modulation can increase the risk for other infections and malignancies, which would need to be balanced by the potential benefit of controlling TB infection [46, 47].

Drug resistance remains a significant problem or the treatment of TB around the world. New drug development is necessary to combat this problem. Machine learning was used for rapid screening for the molecules with activity against specific TB genome targets leading to prioritizing in-vivo testing [48]. This technology accelerated drug discovery by successfully identifying a compound. An example was Bedaquiline, which is approved for the treatment of multidrug-resistant TB [48, 49]. Machine learning continues to be utilized and improved to identify compounds that have activity against many other genome targets including topoisomerase I [48].

While there is much research focusing on new drug development and optimizing the effectiveness of our current drugs, there has also been an emphasis on increasing our ability to distribute medications adequately. This is particularly challenging for TB given the large number of remote locations with endemic TB. Directly observed therapy is the practice of observing patients take their TB medications [37]. This has been a standard of care as it has been correlated with increased sputum smear conversion and treatment success [37]. This approach, however, is not always feasible in places with limited resources. Video-assisted DOT is effective in several areas of the United States and abroad including California and Vietnam [50, 51]. Drones are also being investigated as a tool to deliver medications to difficult to access areas such as Papua, New Guinea, Madagascar, and Vanuatu [52–54]. As the technology for drones continues to progress, perhaps more remote areas in the world could be reached.

## 4 Conclusion

This chapter described the new innovative approaches to battle against TB. Promising technology uses artificial intelligence to aid in the diagnosis of tuberculosis, such as in the detection of tubercle bacilli in microscopy, reading abnormal radiographs, and clinical algorithms using available clinical information. The tool behind this ability is the artificial neural network, modeled after a human brain which processes information through highly interconnected processing elements. The clinical and global application of this technology is yet to be determined. Artificial intelligence and machine learning are also being used to identify target compounds for drug development. Personalization of therapy is being studied to optimize TB drug dosages per patient, especially when limited by choice due to drug-resistance. Host-directed therapies including the use of immune modulators are being investigated and are still at an experimental stage. The use of cellular phones and the internet is tapped into assist directly observed therapy into video-assisted directly observed therapy.

Current available diagnostic and therapies for tuberculosis have led to the cure of millions of people worldwide. Despite these efforts, tuberculosis remains a global threat. Understanding the multifactorial nature, social aspects of TB eradication, increase in funding is vital, as well, to continue efforts to eradicate TB.

**Conflict of Interest** The authors report no conflicts of interest.

## References

1. Murray JF, Schraufnagel DE, Hopewell PC. Treatment of tuberculosis. A historical perspective. *Ann Am Thorac Soc*. 2015;12(12):1749–59.
2. WHO. Global tuberculosis report 2018.
3. Doshi R, Falzon D, Thomas BV, Temesgen Z, Sadasivan L, Migliori GB, et al. Tuberculosis control, and the where and why of artificial intelligence. *ERJ Open Res*. 2017;3(2):00056-2017.
4. Hamet P, Tremblay J. Artificial intelligence in medicine. *Metabolism*. 2017;69S:S36–40.
5. Dande P, Samant P. Acquaintance to Artificial Neural Networks and use of artificial intelligence as a diagnostic tool for tuberculosis: a review. *Tuberculosis (Edinb)*. 2018;108:1–9.
6. FrontlineSolvers. Training an artificial neural network – intro 2019 [21 Jan 2019]. Available from: <https://www.solver.com/training-artificial-neural-network-intro>.
7. Xiong Y, Ba X, Hou A, Zhang K, Chen L, Li T. Automatic detection of mycobacterium tuberculosis using artificial intelligence. *J Thorac Dis*. 2018;10(3):1936–40.
8. Murray PR, Rosenthal KS, Pfaller MA. Medical microbiology. Elsevier Health Sciences; 2015, Philadelphia, PA
9. Turner AP, Magan N. Electronic noses and disease diagnostics. *Nat Rev Microbiol*. 2004;2(2):161–6.
10. Pavlou AK, Magan N, Jones JM, Brown J, Klatser P, Turner AP. Detection of Mycobacterium tuberculosis (TB) in vitro and in situ using an electronic nose in combination with a neural network system. *Biosens Bioelectron*. 2004;20(3):538–44.
11. Bruins M, Rahim Z, Bos A, van de Sande WW, Endtz HP, van Belkum A. Diagnosis of active tuberculosis by e-nose analysis of exhaled air. *Tuberculosis (Edinb)*. 2013;93(2):232–8.
12. Mohamed EI, Mohamed MA, Moustafa MH, Abdel-Mageed SM, Moro AM, Baess AI, et al. Qualitative analysis of biological tuberculosis samples by an electronic nose-based artificial neural network. *Int J Tuberc Lung Dis*. 2017;21(7):810–7.
13. Aeonase, Clinical Results [Internet]. 2013 [cited 28 Oct 2018]. Available from: <http://www.enose.nl/clinical-results/tuberculosis/>.
14. Seixas JM, Faria J, Souza Filho JB, Vieira AF, Kritski A, Trajman A. Artificial neural network models to support the diagnosis of pleural tuberculosis in adult patients. *Int J Tuberc Lung Dis*. 2013;17(5):682–6.
15. Valdes L, Alvarez D, San Jose E, Penela P, Valle JM, Garcia-Pazos JM, et al. Tuberculous pleurisy: a study of 254 patients. *Arch Intern Med*. 1998;158(18):2017–21.
16. Trajman A, Kaiserermann C, Luiz RR, Sperhake RD, Rossetti ML, Feres Saad MH, et al. Pleural fluid ADA, IgA-ELISA and PCR sensitivities for the diagnosis of pleural tuberculosis. *Scand J Clin Lab Invest*. 2007;67(8):877–84.
17. Klimiuk J, Safianowska A, Chazan R, Korczynski P, Krenke R. Development and evaluation of the new predictive models in tuberculous pleuritis. *Adv Exp Med Biol*. 2015;873:53–63.
18. Shu CC, Wang JY, Hsu CL, Keng LT, Tsui K, Lin JF, et al. Diagnostic role of inflammatory and anti-inflammatory cytokines and effector molecules of cytotoxic T lymphocytes in tuberculous pleural effusion. *Respirology*. 2015;20(1):147–54.
19. Li C, Hou L, Sharma BY, Li H, Chen C, Li Y, et al. Developing a new intelligent system for the diagnosis of tuberculous pleural effusion. *Comput Methods Prog Biomed*. 2018;153:211–25.

20. Winsberg F, Elkin M, Josiah Macy J, Bordaz V, Weymouth W. Detection of radiographic abnormalities in mammograms by means of optical scanning and computer analysis. *Radiology*. 1967;89(2):211–5.
21. Ahmad Khan F, Pande T, Tessema B, Song R, Benedetti A, Pai M, et al. Computer-aided reading of tuberculosis chest radiography: moving the research agenda forward to inform policy. *Eur Respir J*. 2017;50(1):1700953. <https://doi.org/10.1183/13993003.00953-2017>.
22. Bansal GJ. Digital radiography. A comparison with modern conventional imaging. *Postgrad Med J*. 2006;82(969):425–8.
23. Fazal MI, Patel ME, Tye J, Gupta Y. The past, present and future role of artificial intelligence in imaging. *Eur J Radiol*. 2018;105:246–50.
24. Georgian-Smith D, Moore RH, Halpern E, Yeh ED, Rafferty EA, D'Alessandro HA, et al. Blinded comparison of computer-aided detection with human second reading in screening mammography. *Am J Roentgenol*. 2007;189(5):1135–41.
25. Nishikawa RM. Current status and future directions of computer-aided diagnosis in mammography. *Comput Med Imaging Graph*. 2007;31(4–5):224–35.
26. Dean JC, Ilvento CC. Improved cancer detection using computer-aided detection with diagnostic and screening mammography: prospective study of 104 cancers. *Am J Roentgenol*. 2006;187(1):20–8.
27. Ko JM, Nicholas MJ, Mendel JB, Slanetz PJ. Prospective assessment of computer-aided detection in interpretation of screening mammography. *Am J Roentgenol*. 2006;187(6):1483–91.
28. Yuan R, Vos PM, Cooperberg PL. Computer-aided detection in screening CT for pulmonary nodules. *AJR Am J Roentgenol*. 2006;186(5):1280–7.
29. Kligerman S, Cai L, White CS. The effect of computer-aided detection on radiologist performance in the detection of lung cancers previously missed on a chest radiograph. *J Thorac Imaging*. 2013;28(4):244–52.
30. Liang M, Tang W, Xu DM, Jirapatnakul AC, Reeves AP, Henschke CI, et al. Low-dose CT screening for lung cancer: computer-aided detection of missed lung cancers. *Radiology*. 2016;281(1):279–88.
31. Kobayashi H, Ohkubo M, Narita A, Marasinghe JC, Murao K, Matsumoto T, et al. A method for evaluating the performance of computer-aided detection of pulmonary nodules in lung cancer CT screening: detection limit for nodule size and density. *Br J Radiol*. 2017;90(1070):20160313.
32. Das M, Muhlenbruch G, Mahnken AH, Flohr TG, Gundel L, Stanzel S, et al. Small pulmonary nodules: effect of two computer-aided detection systems on radiologist performance. *Radiology*. 2006;241(2):564–71.
33. Pande T, Pai M, Khan FA, Denkinger CM. Use of chest radiography in the 22 highest tuberculosis burden countries. *Eur Respir J*. 2015;46(6):1816–9.
34. Breuninger M, van Ginneken B, Philipsen RH, Mhimbira F, Hella JJ, Lwillia F, et al. Diagnostic accuracy of computer-aided detection of pulmonary tuberculosis in chest radiographs: a validation study from sub-Saharan Africa. *PLoS One*. 2014;9(9):e106381.
35. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*. 2017;284(2):574–82.
36. Pande T, Cohen C, Pai M, Ahmad Khan F. Computer-aided detection of pulmonary tuberculosis on digital chest radiographs: a systematic review. *Int J Tuberc Lung Dis*. 2016;20(9):1226–30.
37. Nahid P, Dorman SE, Alipanah N, Barry PM, Brozek JL, Cattamanchi A, et al. Official American Thoracic Society/Centers for Disease Control and Prevention/Infectious Diseases Society of America Clinical Practice Guidelines: Treatment of Drug-Susceptible Tuberculosis. *Clin Infect Dis*. 2016;63(7):e147–e95.
38. Falzon D, Schünemann HJ, Harausz E, González-Angulo L, Lienhardt C, Jaramillo E, et al. World Health Organization treatment guidelines for drug-resistant tuberculosis, 2016 update. *Eur Respir J*. 2017;49(3):1602308.
39. Lange C, Alghamdi WA, Al-Shaer MH, Brighenti S, Diacon AH, DiNardo AR, et al. Perspectives for personalized therapy for patients with multidrug-resistant tuberculosis. *J Intern Med*. 2018; <https://doi.org/10.1111/joim.12780>.

40. Sia JK, Georgieva M, Rengarajan J. Innate immune defenses in human tuberculosis: an overview of the interactions between *Mycobacterium tuberculosis* and innate immune cells. *J Immunol Res.* 2015;2015:747543.
41. Low A, Gavriilidis G, Larke N, MR BL, Drouin O, Stover J, et al. Incidence of opportunistic infections and the impact of antiretroviral therapy among HIV-infected adults in low- and middle-income countries: a systematic review and meta-analysis. *Clin Infect Dis.* 2016;62(12):1595–603.
42. Meintjes G, Stek C, Blumenthal L, Thienemann F, Schutz C, Buyze J, et al. Prednisone for the prevention of paradoxical tuberculosis-associated IRIS. *N Engl J Med.* 2018;379(20):1915–25.
43. Prasad K, Singh MB, Ryan H. Corticosteroids for managing tuberculous meningitis. *Cochrane Database Syst Rev.* 2016;4:CD002244.
44. Canto C, Auwerx J. PGC-1alpha, SIRT1 and AMPK, an energy sensing network that controls energy expenditure. *Curr Opin Lipidol.* 2009;20(2):98–105.
45. Cheng CY, Bohme J, Singhal A. Metabolic energy sensors as targets for designing host-directed therapies for tuberculosis. *J Leukoc Biol.* 2018;103(2):215–23.
46. Tobin DM. Host-directed therapies for tuberculosis. *Cold Spring Harb Perspect Med.* 2015;5(10):a021196.
47. Palucci I, Delogu G. Host directed therapies for tuberculosis: futures strategies for an ancient disease. *Chemotherapy.* 2018;63(3):172–80.
48. Ekins S, Godbole AA, Keri G, Orfi L, Pato J, Bhat RS, et al. Machine learning and docking models for *Mycobacterium tuberculosis* topoisomerase I. *Tuberculosis (Edinb).* 2017;103:52–60.
49. Andries K, Verhasselt P, Guillemont J, Gohlmann HW, Neefs JM, Winkler H, et al. A diarylquinoline drug active on the ATP synthase of *Mycobacterium tuberculosis*. *Science.* 2005;307(5707):223–7.
50. Garfein RS, Liu L, Cuevas-Mota J, Collins K, Munoz F, Catanzaro DG, et al. Tuberculosis treatment monitoring by video directly observed therapy in 5 health districts, California, USA. *Emerg Infect Dis.* 2018;24(10):1806–15.
51. Nguyen TA, Pham MT, Nguyen TL, Nguyen VN, Pham DC, Nguyen BH, et al. Video Directly Observed Therapy to support adherence with treatment for tuberculosis in Vietnam: a prospective cohort study. *Int J Infect Dis.* 2017;65:85–9.
52. Small P. DrOTS: Drone Observed Therapy System: Stony Brook University. Available from: <https://www.stonybrook.edu/commcms/ghi/projects/drots.php>.
53. Frontieres MS. Innovating to reach remote TB patients and improve access to treatment 2014 [18 Dec 2018]. Available from: <https://www.msf.org/papua-new-guinea-innovating-reach-remote-tb-patients-and-improve-access-treatment>.
54. Unicef. Child given world's first drone-delivered vaccine in Vanuatu – UNICEF 2018 [18 Dec 2018]. Available from: <https://www.unicef.org/press-releases/child-given-worlds-first-drone-delivered-vaccine-vanuatu-unicef>.

# Astrovirology, Astrobiology, Artificial Intelligence: Extra-Solar System Investigations



Paul Shapshak

**Abstract** This chapter attempts to encompass and tackle a large problem in Astrovirology and Astrobiology. There is a huge anthropomorphic prejudice that although life is unlikely, the just-right Goldilocks terrestrial conditions mean that the just-right balance of minerals and basic small molecules inevitably result in life as we know it throughout our solar system, galaxy, and the rest of the universe. Moreover, when such conditions on planets such as ours may not be quite right for the origin of life, it is popularly opined that asteroids and comets magically produce life or at the very least, the important, if not crucial components of terrestrial life so that life then blooms, when their fragments cruise the solar system, stars, and galaxies, and plummet onto appropriately bedecked planets and moons.

It is no longer extraordinary to detect extraterrestrial solar systems. Moreover, since extra-solar system space exploration has commenced, this provides the problem of detecting life with enhanced achievability. Small organisms, which replicate outside of a living cell or host, would not be catalogued as viruses. How about viruses that cohabit with life? On the Earth, viruses are a major, if underestimated, condition of life – will that be the case elsewhere? Detection of extra-solar system viruses, if they exist, requires finding life, since viruses necessitate life to replicate. (It should be noted, though, that viruses could be detected through various types of portable ultra-microscopes, including Electron Microscopes (EM) (scanning and transmission) as well as Atomic Force Microscopes (AFM).) However, extra-solar system detection of life does not oblige that viruses exist ubiquitously. Viruses are important potential components of biospheres because of their multiple interactions and influence on evolution, although viruses are small and obligatory parasitic. In addition, nanotechnology – living or replicating nano-synthetic machine organisms might also be present out there, and require consideration as well. An imposing caveat is that, if found, could some extraterrestrial viruses and synthetic nanotechnological microorganisms infect humans?

---

This chapter is dedicated to Ettore Majorana, who changed the view of the universe.

---

P. Shapshak (✉)

Division of Infectious Diseases and International Health, Department of Internal Medicine,  
University of South Florida, Morsani College of Medicine, Tampa, FL, USA  
e-mail: [pshapsha@usf.edu](mailto:pshapsha@usf.edu)

Possibly, intelligence and cognition may at times be contemporaneous with life. Concomitantly, life and viruses that may be detected, could well be impacted upon by intelligences existing on such exoplanets (and *vice versa*). Coming to an understanding of the plurality of extraterrestrial intelligence is an optimal objective, in order to avoid causing harm on exoplanets, as well as avoiding conflict and possible human devastation. This is especially the case if we encounter greatly advanced galactic-level civilizations, compared to terrestrial civilizations. Their machine and bionic technologies on the Dyson engineering civilization scale may be prominently superior to ours; their biological expertise may be similarly critically radical. For example, they may use viruses for purposes for which we are barely aware, and which could be utterly deadly for humans.

A series of steps is being taken in space exploration. Scientists hypothesize and claim that types of life may be near the Earth, in the solar system, and outside the solar system, similar to ours in the sense that only such conditions, Goldilocks conditions, are key *sine qua non* requirements, based on our terrestrial chemistry and biochemistry. If detected within the solar system, will life or its remnants resemble terrestrial life? Outside the solar system a similar chauvinism exists, although the likelihood for life, in any event, remains probably low, according to more cautious approaches to the problem. The study of our solar system includes planets, asteroids, comets, and other planetesimals that have been in overall contiguity during several billion years; anthropomorphisms claims life consequently has been developing along terrestrial-type mechanisms. However, a non-anthropomorphic view would surmise, probably not, especially for extra-solar system locales. The prime warning and admonition in all these deliberations is the contamination and damage, which current and past practice and procedures has caused and continues, due to insufficient biocontainment concepts and technology to date.

Advances in the development of robotics, artificial intelligence (AI), and high capacity ultrafast quantum computers (QC) greatly enhance the sophisticated control and logical development of extra-solar system studies. Consequently, future long-range manned space exploration seems unwarranted. Clearly, reduced dangers to human health and safety, will result from the use of intelligent machine-based investigations and besides, with increased cost-effectiveness. Space exploration comes at great cost to humanity as a whole and utilizes global resources. Consequently, appropriate organizational measures and planning/cooperation need to be in place. Moreover, the bottom line is that despite all the slogans and claims, there have been next to no financial benefits to our planet as a whole. Such financial and heedless difficulties need to be addressed, the sooner the better. In addition, prior to exposure to exoplanetary life, deep understanding of the problems of infectious diseases and immune dysfunction risks are needed. In addition, global efforts should avoid serendipity and stochasticity as this work should be directed with long-term organization, commitment, scientific, and technological methodology. This chapter briefly reviews such questions assuming a new paradigm for oversight of extrasolar system viral investigations including intelligence and life. Finances are included as an essential adjunct.

**Keywords** Virus · Life · Intelligence · Cognition · Astrovirology · Astrobiology · Exobiology · Extreme environments and caves · Infectious disease · Isotope-effect · Isotope-radioisotope quantification and ratio · Chemical composition · Detection · Carbon · Nitrogen · Oxygen · Sulfur · Selenium · Silicon · Fractal · Thermodynamics · Entropy · Enthalpy · Sagan · Anaxagoras · Arrhenius · Schrodinger · Gibbs · Maxwell · Boltzmann · Fermi · Feynman · von Neumann · Majorana · Margulies · Extra-solar system life · Goldilocks · Through the looking glass paradigm · Enceladus · Europa · Mars · Contamination · Feed-back contamination · BSL-4 · Genetics · Inheritance · Biological hybrid · Propagatory system · DNA sequencing · Genome · Evolution · Robots · Nanotechnology · Atomic force microscopes (AFM) · Laser communications · Gravitational lens · General relativity · Neutrino · Artificial intelligence (AI) · Quantum computers (QC) · NASA · CDC · NIH

### Key Concepts

Virus detection in extra-solar system locales depends on the presence of life. Intelligence detection in extra-solar system locales depends on the presence of life or on the presence of machines. Can these be detected? What is the extent of contamination with biological and junked material of the prior pristine extra-terrestrial environments in our Solar System? Have all the planets, asteroids, and planetesimals to which spacecraft were sent already damaged and contaminated the entire program of life detection in our solar system with biological, toxic, and junk contamination? In case terrestrial and extra-terrestrial biologies were similar, can biological hybrids form that are more pathogenic or toxic compared to either parental progenitor? Is this a danger even if they are dissimilar? Will this admonition endure with spacecraft having departed from our solar system? The highest level of biosafety biocontainment level (BSL-4) is used for the deadliest of known terrestrial viruses, such as Ebola virus. The severe problem of extraterrestrial contamination by terrestrial microbes and *vice versa*, the reverse possibility, requires careful and detailed examination, analysis, and technological improvement. Assessment of many constraints requires that intelligent machines sent to investigate life beyond the solar system, should have preeminent capabilities including Artificial Intelligence (AI), Quantum computers (QC), robotics, self-replication, and repair, nanotechnology, and von Neumann's "Universal Constructor". It is counter-productive to foster human exploration of life beyond the solar system, current human, financial, and technological conditions all considered. This chapter briefly provides synopses of diverse background and significant advances for virology/life detection elsewhere, within and beyond the solar system.<sup>1</sup>

---

<sup>1</sup>This includes viruses that can grow in microbes as well as in larger living systems.

## 1 Introduction

The Goldilocks or through the looking glass paradigm assumes, usually that universally, all life embodies the same or similar terrestrial processes. This terrestrial-centric view of life is limited in scope. The complete chemistry of all the natural elements, metals and non-metals, is unknown, and the predictive abilities of quantum chemistry are thus far unable to solve such salient issues that face us, while extraterrestrial exploration continues.

Humans have been anthropomorphic, ethnocentric, and Earth-centric for millennia, as far back in time that we know of. This led to widespread provincial limited unscientific deficiencies to grasp intelligence and life in the universe. However, many ancient Greek and modern scientists and technologists opposed such views. As investigations commence of locales outside our terrestrial and solar system environments, we need to re-examine fundamental assumptions, hypotheses, and directions. Key concerns in such explorations include unnecessary human life endangerment, global wastefulness, improvidence, duplicative financial cost, and huge irresponsible toxic accumulation of terrestrial and space junk. This chapter contends that unrelenting development of AI, QC, and intelligent robotics for extra-terrestrial exploration will remove human risk and endangerment from the equation, as well as reduce global costs.) [1–4]

## 2 Extra-Terrestrial Life

Could galactic-level civilizations exist that are greatly advanced? Is the Fermi-paradox of detection still extant? Astronomers are inventorying the solar system including planets, planetoids, asteroids, comets, dust, particles, etc. Nonetheless, although objects circulate within the solar system, some arrive from extra-solar systems and interstellar space, can life be detected anywhere among these? If not detected directly, which objects may have putative signatures of life? [5–13].

### 2.1 *Opposing View – The Hidden Civilization Survival Hypothesis*

Wandel epitomizes the view that there is an inimitable abundance of life in the universe; this is based on threadbare evidence, if at all. The Drake equation is applied, the biosignature paradigm for the presence of life is applied, and the conclusion drawn. The assumption is, as usual, that Goldilocks planets are the harbingers of life, and percentages calculated to purportedly demonstrate the anticipation as to just how life must be so abundant [14].

This returns to a chapter theme that if there were any life elsewhere in the universe the problem of contact must be faced in regard to human health and safety based on virology, biology, and intelligence estimates therefrom. Moreover, there have already been proposals as to human measures of extraterrestrial intelligence base on feats of engineering accomplishments. Dyson proposed that sufficiently advanced civilizations could build spheres around their suns to provide greater energy capture needed for advanced technological progress – a stage of civilization technological development. Kardashev further proposed stages of development that proceed from the Dyson sphere capability, to the ability to move planets, and then finally to civilizations that could modify the structure of space-time itself to suit their needs. However, nothing of the sort has been detected as yet [6, 7, 11, 14].

The absence of life detection based on the Goldilocks paradigm could be the result of no existence of life in the first place out there, or it may be rarer than thought. In addition, intelligent life may survive some time prior to extinguishing itself. Possibly, intelligent life that does survive, may then rapidly hide itself. In this regard, several thousands of years of human history demonstrates the folly of displayed ostentatious resources readily discerned by others compared to less displayed wealth and existence.

Thucydides, in his history of the Peloponnesian war<sup>2</sup> describes how Athens was able to initially grow and flourish, developing civilization and resources, because it was founded in an arid area that most other civilizations and cultures avoided [15]. Similarly, civilizations that hope to survive the unidentified dangers of the universe may know how to hide, strategically, from potential dangers and risks, and thus, to actually avoid Goldilocks zones.

## 2.2 *Life Cycles*

Various forms of life may have wide range cyclic durations, from transitory evanescence to long-term endurance. The time scales and examination methods used are critical and fundamental. Investigators could miss important clues as to the existence of such living systems, depending on the intervals, flexibility, and adaptability of how and when analyses were to be accomplished. A few demonstrative exemplars follow.

Possibly, investigators could miss cicada-like organisms that hide away for lengthy interludes. In Northeastern America, for example, distinctive cicada species have 13- and 17-year synchronous life-cycles. These two separate periodicities, generally, do not overlap, geographically. However, where there is geographical overlap, 221-year cycles occur [16, 17].

At the other end of the life cycle temporal spectrum, extra-solar system investigators could miss organisms that appear randomly (temporally and geographically),

---

<sup>2</sup>Peloponnesian War occurred 431–404 B.C.

and very briefly. For example, species of mayfly adults live near fresh watersheds in England and have brief lifespans that vary from 5 minutes to 14 days. Their eggs are found in river beds and the next stage of development, nymphs, are found in proximity on river water plants. Subsequently, two separate additional adult stages develop on other riverside plants [18]. Consequently, if the investigators happened to look at night, or with inappropriate durations, or miss any of the precise locales, such explorers might not detect such mayflies. Moreover, even if they did detect various separate components of, as yet unknown (to them) life cycles, would they be able link the various mayfly developmental and maturational stages to then amalgamate the information into mayfly life cycles?

Undoubtedly, sequenced DNA genome phylogenies would divulge a lot; however, we cannot dubiously assume, per chance, that extra-solar system explorers will find living systems that utilize DNA. Some other types of genetics, inheritance, or propagatory systems may be used that have other means of genetics or ‘memory’ required as to how organisms reproduce or propagate. Panoplies of temporal range, location, and multiplicity of stages are thus important considerations investigators must address. Intelligent robotics, AI, and QC’s are exceedingly applicable to conceive, plan, and address what needs to be constructed and assembled for successful and efficient extra-solar system investigations. Much work is currently being accomplished and the paradigm shift from utilization of humans to utilization of AI will accelerate the understanding and implementation of what needs to be done in preparation for such explorations. Additionally, large intelligent mechanical inventive systems necessitate development, which will implement obligatory changes upon arrival at extra-solar system target locales. Clearly AI will outperform humans and human intelligence [19, 20].

## 2.3 *Microfossils and Isotope-Radioisotope Quantification*

The detection of microfossils is important in confirming the early appearance of life on our planet and the use of isotope/radioisotope ratios is vital in enhanced dating methodologies [21, 22]. By the same token, in unmanned exploration of exoplanets, sophisticated apparatus using robots, AI, and QC are needed as well that can design and carry out measurement for isotope-radioisotope quantification *in situ*, in microfossils. This would have to be done across vast expanses of regolith, basalt, and sedimentary layers on such planets, carrying out in briefer times, what has taken more than a century of scientific and technological development and progress to accomplish, on our planet.

In addition, the conundrum of possible microfossils in foreign bodies that have plummeted onto planetary surfaces, fall into this domain. However, artefacts produced by severe heating when such objects fall through planetary atmospheres, and survive impact without complete annihilation, become questions of credibility and require additional support studies. Subsequently, how convincing is the claim for

microfossils in such circumstances? Moreover, when falling onto planets or moons that are devoid of any atmospheres, such as the Moon, it is anticipated that such fossils will be more likely to survive under those conditions. To wit, if Martian microfossils have fallen to Earth, then as a control, should we not detect a plethora of such fossils on our Moon, which, after all, has weaker gravitation, negligible atmosphere, and a softer powdery regolith. An in depth analysis of the degree of flaw of claimed microfossil findings in terrestrial Martian rocks, is provided by Grady et al., who ask the question: "Martian meteorites are ancient microfossils. Do the data demand faith, hope, or charity?" Based on questions of chemical analysis techniques, generated artefacts in these fallen objects, and methods of evaluation, Grady et al. oppose the microfossil conclusion purported findings. It should be noted that as of January 2019, 224 of 61,000 (0.4%) meteorites that fell to Earth were ejected from Mars [23, 24]. This does not support the occurrence of exobiology in our solar system.

### 2.3.1 Viral and Particle Contamination – Biosafety

Space exploration requires analysis of the conditions of physical and biological contamination including inanimate, such as dust, grains, and particles, and animate such as organisms of various sizes. Exploration of the Moon commenced some 50 years ago. In 1969, 1971, and 1972, six Apollo missions 11, 12, 14, 15, 16, and 17 landed on the moon and exhibited problems with lunar dust contamination of personnel clothing, space suits, and space vehicle cabins. Ubiquitous dangerous and deleterious effects of lunar dust were included in ten categories: inhalation and irritation, vision obscuration, false instrument readings, dust coating and contamination, astronaut space suit surfaces, loss of traction, thermal control problems, abrasion, clogging of mechanisms, and ball-bearing and seal failures. There were no effective procedures available to diminish the most serious problems of abrasion, clogging, and weakened heat rejection. The authors concluded that prior studies of these problems had been insufficient, so that additional studies will be needed prior to returning to the Moon and similarly prior to exploring Mars, if that is done. 21st Century ultra-sterile futuristic BSL-4 methods to prevent material and biological contamination should be used [25–29].

The complexity of detecting extra-solar system viruses and associated life is further heightened in consideration of incursion and clash with life that is potentially deleterious or hostile towards humans and their vehicles, AI, and other derivatives of our machines and technologies. Tribulations arising may be intentional or unintentional. Such warnings are exemplified using the following example of an unanticipated grim and dangerous lethal terrestrial virus, for which there is, as yet, no fully efficacious cure, Ebola.

This example is a huge international public health problem. Ebola epidemics are unanticipated, sporadic, and have high mortality. Based on phylogenetic studies, it appears that Ebola evolved in Africa for at least 1200 years, prior to its discovery.

Ebola virus disease was discovered in 1976 in the Sudan and also occurred in the same year in Democratic Republic of the Congo (DRC).<sup>3</sup> Ebola outbreaks and epidemics occurred in DRC in 1977, Sudan 1979, Philippines in 1989, Gabon 1994, Cote d'Ivoire in 1994, Gabon, 1995; DRC 1995, Gabon 1996 and 1997, Uganda 2000, Gabon 2001 and 2002, DRC 2001, 2002, and 2003, Sudan 2004, DRC 2005 and 2007, Uganda 2007, DRC 2008, Spain 2010, Uganda 2011 and 2012, DRC 2012, and a huge outbreak in West Africa (Liberia, Guinea) in 2013. Human Ebola virus infections were also detected in Lagos, Ghana, and Sierra Leone. Fruit bats are the reservoir for Ebola virus. Virus infected bats have been found in the DRC, Gabon, Lagos, Nigeria, and Ghana. Moreover, satellite telemetry is used to track bat migrations. Direct exposure to fruit bats can result in Ebola virus infection. In addition, Ebola virus has been identified in several animals including great apes such as chimpanzees. Ebola virus laboratory, clinical, and epidemiological studies are continuing due the devastating impact of this largely unanticipated disease [30–32].

At the start of the last few Ebola outbreaks and epidemics, however, international organizations, the NIH, and CDC were well prepared and promptly mustered the indispensable organizational, professional skill, complex technology, and rapid response required for the situations that unfolded. Biohazard suits were available for emergency teams and BSL-4 laboratories were already equipped and prepared to receive specimens for optical and electron microscopy as well as for various virologic, immunological, biochemical, and molecular analyses. Procedures for the proper transport of BSL-4-level specimens via commercial public airlines were in place and immediately utilized. In fact, there are 50 known BSL-4 laboratories world-wide, so the planet has a high degree of improved preparedness for various types of biological contamination problems. Indeed, the current 2019 Ebola epidemic in the Democratic Republic of the Congo is considered now at its worst and still spreading [29, 33–38].

National and international space programs should coordinate with US and International biological, microbiological, and virologic organizations, including NIH, CDC, and WHO. These Biomedical organizations are highly skilled expert and accomplished. They are well-rehearsed and prepared for preventing and dealing with viral including biological and material contamination and are expert in carefully controlled, statistical analysis of experimental-design limitations, as well as augmented predictive capabilities. Such cooperative planning will reduce unnecessary duplication and financial waste [39].

Such considerations, including the panoply and plethora of emerging virus infections in recent decades, cast great doubts on the dubious optimism shown by many astronomers and exobiologists that exoplanetary microorganisms are unlikely to cause damage to humans. This is counter to extensive evidence that viruses cross species barriers [40]. Moreover, whether microorganisms and viruses

---

<sup>3</sup>There was a Marburg virus outbreak in 1967 in Uganda; Marburg virus is a Filovirus related to Ebola virus.

have arisen on Goldilocks<sup>4</sup> or on non-Goldilocks exoplanetary environments, it is counter-productive and damaging to credibility, when avowed that the potential infectious pathogenicity and lethality of exoplanetary microorganisms and viruses, if they exist, are shrugged aside.

### 2.3.2 Viruses Cross Species Barriers

There is a plethora of viruses that have their origin in species other than where they are initially discovered. The following examples of such viruses, their original host, and original species jump-time are as follows: Measles virus (cattle and monkeys, since the origin of *Homo sapiens*), Smallpox virus (other primates and camels, more than 10,000 years ago), Influenza virus (water birds, pigs, horses, more than 5000 years ago), canine parvovirus (CPV) (cats and feline carnivores, since 1970's), HIV-1 and HIV-2 (old world primates, chimpanzees, early 20th century), SARS Corona virus (bats, since 1970's), Dengue virus (old world primates, less than 500 years ago), Nipah virus (fruit bats, continuously), Marburg and Ebola viruses (bats, continuously), Myxoma virus (rabbits, since 1950's), Hendra virus (fruit bats, continuously), and canine influenza virus (horses, 21st century) [41].

Remarkably, more recent studies among 19 virus families demonstrate that host-switching is universal among the viruses studied including *Hepadnaviridae*, *Polyomaviridae*, *Poxviridae*, *Papillomaviridae*, *Adenoviridae*, *Caliciviridae*, *Coronaviridae*, *Potyviridae*, *Herpesviridae*, *Paramyxoviridae*, *Parvoviridae*, *Togaviridae*, *Retroviridae*, *Flaviviridae*, *Bunyaviridae*, *Orthomyxoviridae*, *Reoviridae*, *Picornaviridae*, and *Rhaboviridae* [42].

## 2.4 Origin of Life

Where and when were organic compounds first synthesized, how did they accumulate, and how were they disbursed? Possibly, there are multiple developmental phases resulting in the origin of life from abiotic biomolecular synthesis and organelle formation required for living systems. In addition, was there a slow or rapid process during which living systems arose and assembled? Be that as it may, the key issue raised is a contrast between the problem of how life arose on the Earth, vs. its arising on asteroids and comets, and in these cases infalling onto the Earth. Temperatures of ejection are important in understanding the temperatures rocks were exposed to prior to reaching Earth. For example, rocks on the Earth from Mars could have been exposed to ejection temperatures of 400 °C. Additional studies

---

<sup>4</sup>Frequently, a paradigm, termed Goldilocks, or through the looking glass, is used that life found in the panoply of terrestrial environments is indicative of what should be anticipated for life in the rest of the solar system and universe.

indicate a higher temperature when asteroids collide expelling chondrites. The ranges of temperature maximum of 1850–1900 °K and 17–20 GPa for pressure peak were attained [43–46].

However, infall temperatures for rocks reaching the Earth could reach more than a thousand degrees K. For example, Jenniskens et al. found that cometary debris and asteroids that traverse the Earth’s atmosphere are exposed to atmospheric temperature range of 2900–6000 °K (traveling at 19–61 km/sec) and 7600–14,000 °K (traveling at 71–85 km/sec). However, depending on whether such events occur in a predominantly CO<sub>2</sub> or O<sub>2</sub> atmosphere influences the results. As the ejecta are expelled from the infall rocks, they cool the further away they travel. Organic molecules could then be synthesized in a predominantly CO<sub>2</sub> atmosphere compared to an O<sub>2</sub> atmosphere where C would react to produce CO and CO<sub>2</sub> [47].

## 2.5 *Early Sources of Organic Molecules*

Chyba and Sagan hypothesized [48] that the early Earth, prior to 3.5 Gyr ago, had several equivalent sources of organic molecules: impact shock-, UV radiation-, electrical discharge-propelled synthesis and infall from extraterrestrial objects. However, questions can be raised as to the actual quantification of each putative source, the accuracy of measurements, the degree of solar nebula opacity to UV radiation, influence and types of radioactivity, the strength and destructiveness of the solar wind during the course of these events, the reaction-rate and chemistry of isotope effects of deuterium in the aqueous environment, and the relative distributions of infall from chondrites, comets, and other solar system debris.

## 2.6 *Organic Compound Survival Under Extraterrestrial-Interplanetary Conditions: A Paradox*

Immanuel Kant, in the 18th century (in 1755), first proposed that the solar system condensed from gases and particles. In this theory, he relies on the concept of the atom derived from Lucretius in combination with the application of Newton’s laws of gravitation, leading to processes of condensation [49]. The work of Immanuel Kant set the stage for subsequent inferences to be based on planetary evolutionary phenomena, leading to the origin of life.

The evolution of carbon-containing molecules is of subsequent relevance in terms of possibly detecting Goldilocks-zone hypothesized Earth-like life, as advanced in the 20th century. The analyses of the Murchison and many other carbonaceous chondrites have been typical study foci. However, the question as to the survival of organic prebiological compounds continues to be a prime question in

studies related to the origin of life in the solar system, as well. For example, carbonaceous chondrites have been sources of amino acids, purines, and perhaps some pyrimidines. Do they reflect conditions related to the origin of life or synthesis during transit to the earth's surface? [50–53]. First, were compounds continually produced during the early history of the solar system (e.g. during the first Gyr) and then, second, various conditions would degrade and destroy these compounds (e.g. heat, impacts, radiation, radioactivity, ultraviolet light, cosmic rays, and solar wind.) Olivine basalt was returned from the lunar regolith to the Cornell University Space Center Planetary Laboratory (Ithaca, NY) and was pulverized and then mixed with amino acids. These mixtures were subjected to lunar surface proton irradiation, temperature, and high vacuum, and resulted in complete amino acid destruction, in entirety, in 5-cm mixture columns. The half-life of the amino acids was calculated at 4000 years [54].

Conditions would be different at various times and durations on asteroids, chondrites, proto-planetoids, planetesimals, Earth, other planets, satellites, and comets. Thus, one may ask the crucial question, did pre-biotic biochemistry occur and evolve on planets such as Earth, Mars, and Venus at various times in their histories or did this occur on a solar system-wide panorama, early in the solar system history. However, if it were solar system-wide, then did this occur in two phases for the Venus, Earth, and Mars? First as the planetary material precursors and compounds accreted, then when some planetary critical mass was attained resulting in volcanism and melting, consequent concomitant destruction of biological compounds ensued. Next, was it only after some level of cooling that there was *in situ* synthesis of protobiologicals on the Earth (and Venus and Mars) as well as raining down of such molecules, as described by Chyba and Sagan in 1974 [48].

Is life then a product of individually rare events such as on planets alone, or as a result of solar system-wide events that contributed increased probabilities of biomolecular assembly and biogenesis. As time goes by, with additional well-controlled studies, this paradox should be resolved. Be that as it may, Loeb proposed that because at a redshift of 100–137, because the cosmic background temperature was 273–373 °K, (thus, a widespread Goldilocks universe) organic molecules could have formed and perhaps life originated, though at that early stage of development of the expansion of the universe (only 10–17 million years of age) [55]. Using this slant, there could have been very early widespread origins of life over several million years, prior to the further cooling and evolution of the universe, towards what we observe today. This overall vantage is also proposed by Gibson [56].

## 2.7 Terrestrial Ribosomes and the Origin of Life

The thermodynamics involved in ribosome evolution and multiple linkages among proteins and nucleic acids may be involved in the energetics imposing evolutionary constraints in selection at the molecular intracellular level, especially during the epoch when the terrestrial atmosphere changed from a UV-penetrating atmosphere

with no oxygen to one with little UV and an abundance of oxygen.<sup>5</sup> The complexity of the origin of ribosomes involving proteins and RNA's may have taken place at an early stage of life, when RNA genetics predominated prior to the proliferation of DNA genetics, termed the RNA world [57–62].

Briefly, subsequent evolution of life including ribosomes and cell organelles, point to endosymbiont evolutionary processes. Eukaryotes (with mitochondria, plastids (chloroplasts), and mitotic apparatus) appeared due to the fusion of several separate endosymbiont prokaryote groups. Accordingly, note then, that at some stages of evolution, some prokaryote nucleated cells interacted with mitochondrial symbionts and others with both mitochondrial and plastid endosymbionts. The endosymbiont origin hypotheses were originated by Mereschkowsky in 1905 and championed by Margulis since 1967 [63–67].

Extra-solar system life searches utilize panels of sophisticated experiments available to establish the presence of many types of chemical reactions, pre- and post-origin of life, depending on the states of development, where exploratory vehicles land. Of under-rated importance, there should be improved biocontainment procedures in place to avoid upsetting and perturbing the evolution of life in such environments. As technology improves and further exploratory vehicles are sent, the danger of biocontamination becomes even greater. Each exploratory vehicle should have a probability value (p-value) attached - which should be lower than some crucial value that should be determined with appropriate scientific investigation. Many control experiments are required to fully investigate such potential contamination outcomes. The degrees of complexity involved additionally require AI and QC to accomplish such difficult and multifaceted tasks without destroying such environments. Such issues are obviously important in extra-terrestrial Goldilocks zone environments.

## 2.8 *Thermodynamics and Life*

Besides the many apparatuses and characteristics of life mentioned, it is also recognized that analysis of organization, symmetries, energetics, and thermodynamics are crucial components in the analysis of life. This way early recognized by Schrodinger. Life as we know, demonstrates several biochemical cycles including Nitrogen, Carbon, Oxygen, and Sulfur cycles as well as electron flow cycles in organelles including mitochondria and chloroplasts. For energy production, many electron-chain oxidation cycles are based on sulfur or oxygen. Molecular biosynthesis- and degradation-coupled cycles are additionally often cited including nitrogen fixation, urea cycle, glycolysis, Krebs cycle, and photosynthesis. Demonstrating extreme plasticity and adaptability, terrestrial life has infiltrated wide varieties of inhospitable extreme environments, nonetheless, of temperature, earth, atmosphere, and water (however, all considered in the Goldilocks zone) [68–74].

---

<sup>5</sup>Such considerations may be obviated for extra-terrestrial life, where other conditions may become imposed, in the event of the absence of UV and/or oxygen in the first place.

Life demonstrates increased order and lowered entropy. The Gibbs equation is fundamental in such calculations.

$$\Delta G = \Delta H - T\Delta S$$

where G is Gibbs free energy, H is enthalpy, T is temperature, and S is entropy [75–77].

Since the external environment is an open system, it supplies energy for this to occur as well as an increase in disorder and increased entropy to offset life's demands on energetics. Indeed, according to some cosmologies, entropy is increasing in the universe (second law of thermodynamics) [68, 78, 79]. Thermodynamics and kinetics of chemistry in space exploration is productive as exemplified for Enceladus [80]. An additional example of living thermodynamics is the Gibbs-Donnan equilibrium effect [81]. Technology needs to be developed to detect such thermodynamic effects due to the presence of life, prior to taking samples for analysis, thereby potentially harming and destroying what is under examination. The application of difficult and sophisticated methods including fractal, multifractal, and thermodynamic approaches would require some time to develop and optimize [82–89]. Anthropomorphism not intended, in terms of intelligence, scanning for the extraterrestrial equivalent of neurons (from an external vantage) is a daunting task. Robotics, AI, and QC would surely require development, in order to do such difficult and complex tasks.

Unanticipated and counter-intuitive properties of replication occur in terrestrial life, which are being comprehensively assessed, as indicated, for example, by studies of prions and prion-like proteins. On the one hand, prion proteins are produced from coding DNA exons; on the other hand, infectious prions can induce malformation of correctly folded proteins into misfolded neuropathological prions. This transformation is associated with a variety of human and animal diseases that have been carefully and scientifically studied since the early 20th century. Moreover, many proteins have prion-like sequences and may additionally participate in various disease processes. Unanticipated, amyloid fibril fatal associated involvement in Alzheimer's disease, may involve synergistic action with prions. Further relevant to Astrovirology, there is negligible immune response to prion disease itself (transmissible encephalopathies) in the natural setting. If anything, incipient immune reactions promote replication and spread of prion disease [28, 90–94]. Thus, exobiology portends unanticipated elevated health risks for human exploration, should life and viruses be discovered elsewhere.

Biomedical scientists recognize and widely discuss our incomplete understanding of the human genome. The salient feature of the incomplete information is that, remarkably, 98% of the human genome does not code for proteins. Therefore, the human genome is appropriately under close scrutiny using cutting edge contemporary molecular analytic technologies; these studies lead towards individualized medicine and are being done through extensive institutional organizational skills coupled with critical peer review [87, 88, 90, 95–102].

## 2.9 Abiogenesis

Abiogenesis may depend on how water interacts with other environmental variables. It should be noted, however, that although there is little evidence that such conditions intrinsically produce life (besides the single example of the Earth), such zones are certainly environments to where terrestrial life could relocate. As an additional example in the field, along such lines, using the anthropomorphic insular approach, a probability equation is postulated, to calculate probabilities for life on Earth [103].

However, an earlier Bayesian approach came to the conclusion that despite the existence of life on the Earth, it is an extremely unlikely event [104]. From this, it may be inferred that the focus on the Goldilocks zone in this restricted search for life may not be the most productive approach to take; detection may fail for many other life-forms.

The possibility of material exchange among planets is used to enhance the probability of abiogenesis [105]. There is a range of histories and types of planetary systems under which such exchanges could occur. Histories of migrating planets and their roving *excreta* further tone the clarification of planetary surface properties to be included in the framework of exploration of extra-solar systems [106–108].

## 2.10 Extinction Events

There have been five commonly acknowledged terrestrial extinctions of life to date on the Earth. Across the last 560 million years (MY), the mass extinction names and approximate dates are late-Ordovician, 440 MY ago; late-Devonian, 370–350 MY ago; end-Permian, 250 MY ago; end-Triassic, 199 MY ago; and end-Cretaceous, 65 MY ago. How such cataclysmic events terminate life is under study. For loss of life, a ‘trigger’ mechanism is separated conceptually from a ‘kill’ mechanism. Trigger mechanisms are whatever events (e.g. asteroids) cause or initiate kill mechanisms. An interesting example, though subtle, is a kill mechanism proposed when atmospheric partial pressure pCO<sub>2</sub> levels of 560 ppmv are associated with an oceanic toxic pH of 7.9. This level can be reached both from below and from above [109].

Subsequent to each extinction event, the fundamental building blocks and biochemistry of life persisted, as some organisms survived and continued to evolve each time. A question posed is to what extent intelligent life evolved prior to or after each extinction event. However, the consensus appears to be that highest intelligence was attained after the fifth extinction, with the rise of mammals, primates, and humans. What about extinction events in our scientific exploration of extra-solar system life? Our explorations necessitate including methods to ascertain whether extinction events have taken place, if our arrival is in the midst of such an event, and whether our arrival itself may cause such events. Dealing with questions of contamination is crucial. The footprint impacts of our exploration and arrival elsewhere should be explored scientifically.

### 3 Goldilocks Zone – Through the Looking Glass

Many biochemists and biologists have concluded that although several possible lineages of life may have got started at the dawn of life, there was one lineage originated that gave rise to all life on Earth, though exceedingly complex and diversified. This approach is not as straight forward as it once appeared because all life on the planet is divided among three groups (Archaeabacteria, Eubacteria, and Eukaryotes), which occupy ranges of several extreme environmental niches, and utilize many different and disparate energy sources. The unity of life derived from a single lineage is a charismatic hypothesis but not yet really scientifically fully proven [68, 72]. Be that as it may, it is not an optimally scientific tactic in the exploration of the universe, like looking in a mirror, to pursue evidence only for what we already presuppose to know. Chemistry, biochemistry, physics, and thermodynamics need to be examined in concert, since that will improve the chances of detecting life of any sort, without prior prejudice. Fractal forms and symmetries should be compared among various environments. Possibly fractal forms in differentiated environments may be able to discriminate the presence or absence of life. Various principles have been formulated [82, 102]. Statistical analysis and dynamics has become much more complex, including chaos theory and fractals vs. the methods of the last century, such as Gibbs, Maxwell, and Boltzmann theory. The need to utilize new methods looms over future life studies and exploration of exobiology/virology [83–85].

Goldilocks conditions have been located on Enceladus, one of Saturn’s moons and the presence of methane has been ascribed as possibly due to microorganisms that are similar to terrestrial deep-sea microorganisms. The terrestrial microorganisms referred to include a methanogenic *archaeon*, *Methanothermococcus okinawensis*, that lives in non-aerobic environments in the Pacific Ocean. It is reported that these unique methanogens rely on natural hydrogen production and that they can produce methane. Furthermore, the hydrogen production rate on Enceladus is sufficient to support methanogenic life that is hydrogenotrophic and autotrophic. That an detection of methanol are also discussed in terms of possible life biomarkers [80, 110]. The fundamental question is when a chemical profile is a biosignature.

#### 3.1 Goldilocks Elements

The carbon-, nitrogen-, oxygen-, and sulfur-based life positioned on DNA, RNA, proteins, and the rules of terrestrial molecular biology, may not comprise life elsewhere. Detecting organic molecules in exo-environments may relate to other forms of life that we simply have not yet imagined. Even among some terrestrial life forms, for example, silicon enters into metabolism: radiolarians, silici-sponges, and diatoms secrete silica [109]. Life detection elsewhere should include diverse unanticipated scenarios.

### 3.2 *Goldilocks Abiogenesis*

That life as we know it arises on Goldilocks habitable zones, relies on the assumption that terrestrial-like environments are the primary harbingers of life [103]. It should be noted that although there is little evidence that such conditions intrinsically produce life (besides the single example of the Earth), though such zones are certainly environments to where terrestrial life could move. An equation was postulated to calculate probabilities for life on Earth. The possibility of material exchange among planets was used to enhance the probability of abiogenesis. However, a Bayesian approach came to the conclusion that despite the existence of life on the Earth, it remains an extremely unlikely event. Moreover, it should be noted that earlier speculations considered that life resembling known terrestrial biochemistry was unlikely [9, 11, 68, 104, 105]. From this one may infer that the focus on the Goldilocks zone conditions in the search for life is an excessively restrictive approach.

## 4 Technology Development

Since the advent of the Industrial Revolution, science, medicine, and technology intensely developed and progressed. Continued development as it points to extra-solar system investigations requires appropriate additional improvements and advances.

### 4.1 *von Neumann Universal Constructor Machines*

von Neumann established in the last century, the feasibility of machine replication and error detection and correction. Central is the concept of the “Universal Constructor” of von Neumann’s [111, 112]. Taking this to the next stage of AI, QC, and intelligent robotics, it appears we are embarking on a new age of life made of various metals, semi-conductors, plastics, and other synthetics.

### 4.2 *AI Robotics in Space Exploration*

Robotics and AI are burgeoning fields in extra-solar system exploration. Such work is proceeding at an accelerated pace. For example, robotic self-supervised learning was accomplished in the laboratory environment to assist independent robot performance in the space exploration environment [113]. This is indicative of progress to fully function and independent robotics that can carry out all functions required ultimately for extra-solar system exploration.

### ***4.3 AI and Machine Learning Computation Approaches Related to Medicine and Infectious Diseases (ID)***

AI and machine learning computation approaches related to Medicine and ID were used to develop a more advanced search engine-decision support system of biomedical ontologies than hitherto accomplished. This system utilized infectious disease and antibiotics information for diagnosis classification. This system is for use by ID physicians and care-givers. This is especially useful, when extrapolation of exceedingly complex situations is required, when there is missing data, and when nonetheless, decisions require being made [114].

### ***4.4 Extra-Solar System Interplanetary Lasers***

Lasers have numerous uses, including communications, optical transponders, precision clocks, pathfinding, orbital dynamics, altimetry, laser propulsion in space, navigation, attitude control, construction, precision alignment, structural control in space, power generation and distribution, resource location including 3D-sensing, materials science, planetary, satellite, asteroid, cometary interior structure and geology, regolith and ice surface mapping, and of course, fulfilling measurement requirements due to Special and General Relativity. Further development will be essential for exo-solar system interplanetary exploration [115, 116]. Such development is perfect for robotics, AI, and QCs.

Very recently, a closer realization of the use of lasers in communications and life-detection has been proposed. Stellar gravitational lensing as described in Albert Einstein's theory of General Relativity was used to recommend an inter-solar system means of communication. This is shown to be feasible using an equivalent to a 1 Watt 1 nm channel laser light in juxtaposition between stars such as the Sun and alpha-Centauri. Possibly, more advanced civilizations may communicate with each other, using even more advanced related technology [117]. This is hardly unreasonable considering how recently we commenced developing our current technologies.

### ***4.5 Complexity: AI and QC***

Given the numbers of calculations and complexity of interactions that would be required for life to start, let alone the detection of intelligent life, the use of artificial intelligence (AI) as well as Quantum computers (QC) are anticipated to make the calculations and carry out the modeling required to estimate various possibilities, alternatives, and exigencies [87, 88, 118–121]. Scientific development will be greatly accelerated by the development of independent robotics, AI, and QC. Moreover, the sophistication of roving laboratories in planetary and other

environments will be greatly enhanced with the use of AI and QC improvements. QC's will have to be enhanced so that they operate at several ambient temperatures, utilizing materials that do not require cooling to close to zero degrees Kelvin. These are monumental tasks.

#### **4.6 A Recent Solar System Event: 1I/2017 U1 (*Oumuamua*) Transit**

To date, 750,000 asteroids and comets have been detected. However, an unanticipated novel object, 1I/2017 U1, was observed during several nights in October 2017, which twice transited the plane of the solar system, traveling at anomalously high velocity, in an unusual orbit. 1I/2017 U1 is the first object recognized as an extra-solar system visitor. It was suggested that due to various complex calculations, such interstellar visitors have been missed previously. However, it could not be traveled to by direct means. 1I/2017 U1 was initially identified as probably an asteroid, but not a comet. It traversed our solar system with a jolt from the Sun that assisted its hyperbolic exit. It was oblong shaped, with axes that were 200 by 20 meters. Spectroscopic analysis indicated it had a red color, coated with damaged or degraded organic molecules of some kind, and possibly hollow (perhaps with or without trapped internal ice). It was also reported that, unfortunately, we did not have any rockets prepared, which could reach the object and place a probe on it to plot its path as it exited the Solar system beyond visual range. Could laser technology have been used to extend the tracking range [122–124].

Currently, analysis continues for the 4 days of observational data produced from this extra-solar system object. It is most recently considered to have a spotty red color or graded red color scheme. The analysis of this color issue also involves concluding that it has a tumble rate that influences the apparent color analysis. Moreover, 1I/2017 U1 may have been involved in a collision in the distant past, perhaps the cause of its expulsion from the solar system from which it originated [125]. Analysis of the object for signs or signatures of life remain indefinite. Loeb, in 2018, [126] proposed an artificial extraterrestrial interstellar origin for 1I/2017 U1. Of further interest is the proposal of a few stars, originally, calculated to within 2 pc of 1I/2017 U1 as possible prior sources with ranges of distance and travel times of 0.6 pc at 1 Mya and 1.6 pc at 3.8 Mya [127]. However, it is as yet not anticipated which star, previously to that or originally, may have given rise to 1I/2017 U1. (How old is it? Could it be an intergalactic as well as an interstellar messenger?)

Supporting the notion that interstellar traveling objects have entered our solar system, Siraj and Loeb, in 2019, proposed a 0.45 m size meteorite (2014-01-08 17:05:34 UTC), with an unbounded hyperbolic orbit and asymptotic extra-solar system velocity of 43.8 km/sec (60 km/sec away from the velocity of the Local Standard of Rest [LSR]), is of interstellar origin. In addition, they estimate that there are eight such interstellar objects that have fallen on the Earth [128].

## 5 Launching Extra-Solar System Exploration: Possible Protocol

As time goes by, the planning and implementation of extra-solar system missions will be increasingly effective and accelerate with augmented utilization of AI and QC to produce the robotics, universal construction, and analytic and self-sufficient intelligent equipment required. The paradigm suggested for exobiological exploration is to: (1) first produce the robotics (including universal construction), vehicles, communication skills, AI, and QC necessary to analyze, select, and focus on extra-solar planets for detecting intelligence, life, and viruses; (2) This should be done at the feasible distant limits of our solar system; (3) Next, send all these and all necessary probes and landers to various targets and groups of targets (which by then may have different priorities compared to what we know currently); (4) Interstellar space is cold so that with long range communications, QCs that may require cold environments to function, could be stationed at suitable distances from stars. Calculations will be performed at central hubs during extra-solar system investigations; (5) Orbit the selected targets (stars, planets, or other objects); (6) Assess the presence of intelligence<sup>6</sup> and decide whether to exit or proceed with communications and/or direct exploration; (7) Then, map the targets and analyze them across several wavelengths (X-ray, IR, and visible spectra) and sound, and detect atmospheric chemistry, biochemistry, and particulates (including odors); (8) Send the lander vehicles that will contain the apparatus required to carry out the experiments; (9) All the while, communication links will be maintained; (9) Include detection of migrant planets and planetoids within and outside of solar systems; (10) Along the way, consider the prospect that other such investigations may be taking place. (11) Last and not least, definitely, the appropriate parallel controls for all components of such missions will be continued in suitable places within our solar system, evolving here while the missions evolve distantly during the investigations.

### 5.1 Europa Lander Mission Report of 2016

The development of the Europa Lander Mission report of 2016 demonstrates evolution in approach towards such ends with a wide scope in scientific planning and organization [129, 130]. The 264-page technical report and 734-page planned budget report are very wide-ranging, convey detailed planning, expectations when probes arrive, as well as the sophistication required to get there. These are early steps towards exploration within the solar system; however, the \$19 billion requested for work 2018–2022 may not be ratified and realistically, the deadline may be shifted towards the late 2020's.

---

<sup>6</sup>Models of extraterrestrial intelligence include both ‘natural’ and ‘artificial’ [89]. Both require evaluation prior to interaction with any exobiology or exorobotics.

## 6 NIH and NASA Budgets

What has been the cost for the work done by the US space exploration agency, NASA? According to Wikipedia, the NASA budget from 1958 to 2020 is estimated \$1,188,919,000,000. In comparison, the NIH budget from 1938 to 2018 is estimated \$737,016,000,000. The NIH budget expended across 80 years is approximately 62% of the NASA budget consumed across 62 years. (NIH-supported researchers received 90 Nobel prizes across 78 years, 1939–2017.) [131–133].

## 7 What Are the Caveats, if there Is Life Elsewhere?

NASA scientists, most recently, produced a body of work, related to the possibility of life on Mars.

### 7.1 *Space Exploration and Potential Contamination of Extraterrestrial Worlds, Planets, and Planetesimals*

Criteria and standards of material/particle contamination have been published for example, related to the European Space Agency (ESA) [134, 135]. However, the implementation, enforcement, and monitoring quality control, followed by peer review publications is not yet implemented globally. Moreover, this entails manufacturing, assembly, and launch stages and there is no systematic program globally in this regard. In addition, since there is a variety of components and procedures internationally, without international monitoring, publication, and quality control, more needs to be done. Clearly, where only materials and particles are monitored, biological (including potential microbial contamination) are not addressed. Although some procedures used could be toxic to microorganisms, it does not mean that living microorganisms are not present as well; furthermore, this cannot be taken to mean that the biological fragments and components of life are absent. Experiments, controls, and simulations are continually required, followed by public reports and expert peer review.

There are many sources of contamination by biological and microbiological organisms, materials, by-products, waste, and detritus. The International Space station, currently and in its prior stages of development, are prime examples. NASA points out that such issues need discussion and practice [136].

## 7.2 Point-Counter-Point Paradigms: Interplanetary and Inter-Stellar Spread of Life – Svante Arrhenius and Goldilocks

Sidestepping the all-inclusive question of the probability of more than a single origin of life event within and outside the solar system, is the possibility that microscopic life, once it has originated, could spread throughout any galaxy including our Milky Way. For example, upto 1 cm size meteorites, originating on a planet such as the Earth, which has microbial life, could be ejected and more than one could reach exoplanets at least 20 light years (ly) distance. Upon terrestrial ejection, such meteorites could reach comets. If the microorganisms remained alive on such ejecta, coupled with comets and other interstellar travelers, then the entire galaxy could be accessible. Along similar lines, the authors state that if life originated, for example,  $10^{10}$  (ten billion) years ago, anywhere in the Milky Way, then, by 4.6 billion years ago, life forms could have reached the Earth [137].

There have been heated arguments in contemporary times, relating to the panspermia hypothesis of Svante Arrhenius by many scientist including Wickramasinghe and Hoyle. It should be noted that in ancient times, Anaxagoras (born 510 BCE) first proposed that life came from elsewhere and Arrhenius, in more recent times, proposed this in 1903 [138–141].

Ginsburg and colleagues succinctly summarized ideas and models related to panspermia, with the pervasive presumption of the Goldilocks paradigm. In addition, they assert the possibility of terrestrial ejecta traversing the solar system as well as Milky Way, a central canonical issue for panspermia. Furthermore, they include viruses in their discussions. Suttle and others had previously pointed out that there are approximately  $10^{30}$  viruses, currently, in the Earth's oceans and that viruses have a huge impact globally – that they exhibit the widest genetic diversity on Earth, cause mortality, affect phytoplankton (one of the bases of the ecosystem and food chain on the planet), and drive geochemical cycles. Gonzalez recently pointed out, in a more restricted sense, that although Arrhenius' original canonical panspermia hypothesis is unlikely, what he terms as lithopanspermia of microorganisms that are endolithic and evolve in the same solar system, have increased positive consideration. In addition, terrestrial endolithic life shows unanticipated propensities to survive in caves and other hostile environments [106, 107, 142–146].

However, as inviting as these proposals are in regards to panspermia, for the single known origin of life on the Earth, no evidence except for organic compounds and Goldilocks environments are provided in support. Moreover and most damaging, is the hard experimental evidence that amino acids will not survive the solar wind to a depth of 5 cm of regolith, let alone the high temperature susceptibilities of biological components and organelles found in any Biochemistry and Biology textbooks [54].

### ***7.3 Reverse Interplanetary and Inter-Stellar Spread of Pathogen Paradigms: Big Bad Wolves Visiting Little Red Riding Hood Habitats***

The dangers of contamination as a direct result of human exploration has been pointed out and discussed for some time – especially in regards to contamination of the Moon and Mars. In addition, it is keenly indicated that samples returned to Earth then inevitably, will contain such contamination and thus distort any subsequent analyses [147, 148] Moreover, depending upon the ability of microbial life to replicate, which it has shown even on the Earth to have great capabilities to survive extreme environments, such replication could result in further evolution of such life to then become more toxic and pathogenic. This is an additional stipulation unexplored as yet.

The obverse issue of contamination is whether terrestrial environments could be harmed by toxic and disease-causing reverse contamination – i.e. derived from outside the Earth [149]. These authors conclude that there is very little reason to be concerned and provide a list of reasons why this is of very low probability to occur in the future. The authors thus ignore the lessons taught by all the hundreds of millions of human and animal mortalities and morbidities that have occurred due to microorganisms, including viruses, throughout the last few thousand years of known history on the Earth. (Cf. other chapters in this book as well as references [40, 49]. Most of the epidemics and pandemics were unanticipated, occurred due to lack of understanding of the biology and epidemiology of infectious diseases, as well as lack of understanding of basic molecular biology and biochemistry. This is of course, an ongoing process, and emphasizes the need for greater caution than is the norm.

Before the grave issue of whether there is life on Mars or not is decided, proposals are being made to terraform Mars and for example, increase the Martial surface temperature to bring it more into the Goldilocks zone, [150] i.e. more fit for human habitation. Clearly, this will degrade the problem from being feasible with scientific difficultly, to impossible, and a great opportunity for such study will be lost. That is to say, if it has not already been lost, due to the various objects that have been propelled to Mars. It is important to note that methods for prevention of microbial contamination were in a lower lack of expertise when the first landers were placed on Mars. Even upto contemporary times, as mentioned, although there are many cogent proposals for improved microbial disinfection, there is a lack of organization and standardization that demonstrates lack of clarity and purpose in this regard – since the inception of Martian exploration - to prevent terrestrial microbial infection and biological waste on Mars.

### ***7.4 Deliberate Panspermia – Astrobiology Stem Cells***

There is a concern, in addition to the conundrum as to whether panspermia has occurred among stars and solar systems within galaxies: synthetic panspermia. In addition to the accidental spread of viral and bacterial contamination of Goldilocks

potential habitats, as previously discussed, there is the possibility of intentional seeding of life with synthetic organisms, synthetic biology. The first such steps have been taken by synthesis of microorganism genomes and their use in the assembly of bacteria.

Synthetic biology has appeared as a promising field for research into the nature of what is alive, health, and financial dynamics. A few key stepping stones are the complete synthesis of several microorganism genomes: poliovirus cDNA by Cello et al., in 2002; phiX174 bacteriophage DNA by Smith et al., 2003; SARS Coronavirus genome by Becker et al. in 2008; and the complete synthesis of the Mycoplasma genitalium genome in 2008. The assembly of a bacterial cell with synthesis of its genome by Venter and colleagues was completed in 2010. Hutchison synthesized a minimal bacterial genome in 2016. Improvements in DNA synthesis techniques assist continued work – difficulties encountered include production of longer than 1 kb synthons, their assembly into larger structures, the presence of DNA sequences that are toxic to the host organism, sequences that have increased secondary structures, and repetitive sequences [151–153].

As part of studies in evolutionary biology, as synthetic biology advances, organisms may be produced that have evolutionary potential. Adult stem cells and pluripotent cells are used to assist in organ repair in various medical situations. A major hypothesis being researched is whether there are universal stem cells in the adult that exhibit a sufficiently elevated degree of plasticity, circulate throughout the blood stream, able to enter various organs, and then perform their functions [154, 155].

Additionally, it should be noted that possibly, the ability to produce stem cells must have occurred early in evolution prior to the divergence within the eukaryote kingdom, since both plants and animals have stem cells. However, some plants tested can be produced from adult isolated single cells. This is of interest because both plants and animals were exposed to Darwinian evolution [154–156].

In application to panspermia and the origin of life, as genetic and biological engineering progress, a point may be reached whereby cells (Astrobiology stem cells) are produced, which have the capacity for evolution into various unknown and unanticipated life-forms under controlled laboratory conditions, or when released into Astrobiological Goldilocks environmental conditions. Such astrobiology stem cells could be included as von Neumann universal astrobiological constructor living machines [111, 112].

## 7.5 Astrobiology Ethics

Ethical considerations are needed for all the topics covered by this chapter as in any scientific field. This is essential in order to better understand the ethical imperatives of past, current, and future space exploration and its impact on the Earth, Solar system, and galactic Astrobiology. Ethics in science is a well-developed field today and is being applied in the above contexts as well. In addition, the question of sustainability in connection with ethical analysis and understanding are also proposed [12, 13, 157–159].

## 7.6 Concluding Postscript – Neutrinos and Astrovirology

Neutrinos are fundamental particles (lepton Fermions) that are distinguished in what are termed, actually, three flavors, electron, muon, and tau neutrinos. They are produced in stars, nuclear reactors, particle accelerators, radioactivity, and in nuclear (bomb) explosions. Once created, through what is termed the Weak force ( $W^\pm$  and  $Z^0$  Bosons), although much lighter, compared to all the other known particles, neutrinos crisscross the universe, rarely interacting with matter. However, they can metamorphose into each other, in transit, called oscillations, and have internal ‘clocks’, governing them when to do so. Be that as it may, at the Big Bang, apparently though, a special fourth neutrino was produced that is different from the electron, muon, and tau neutrinos, and has been navigating the universe for 14 billion years since the Big Bang [160–168]. All in all, the universe is bathed in several generations and epochs of neutrinos and they are currently being mapped and characterized with a high level of interest.<sup>7</sup>

The fundamental unitary matrix equation describing neutrino oscillations in terms of their flavors, generations, and masses is:

$$\nu_i = U_{ij} \nu_j$$

where  $\nu_i$  represent the three neutrino flavors,  $U_{ij}$  is the unitary matrix, and  $\nu_j$  is the three putative neutrino masses, which are currently under investigation.

It has been proposed, originally by Pasachoff and colleagues in 1979, that neutrino production by advanced civilizations may someday be detected when our own technologies may advance sufficiently. Strikingly, Stancil et al. were able to demonstrate communication using detectors aimed at terrestrial-produced neutrinos [117, 169–174].

*In fine*, Ettore Majorana proposed new types of fundamental particles that revolutionized unanticipated concepts of what matter composes the universe, and thereby influenced how we may approach the problems of Astrobiology and Astrovirology in the known universe<sup>8</sup> [175–178].

## 8 Conclusions

This chapter attempts to encompass and tackle a large problem in Astrovirology and Astrobiology. There is a huge anthropomorphic prejudice that although life is unlikely, the just-right Goldilocks terrestrial conditions mean that the just-right balance of minerals and basic small molecules inevitably result in life as we know it throughout our solar system, galaxy, and the rest of the universe. Moreover, when

---

<sup>7</sup>Trillions of neutrinos traverse each person per second, to provide an idea of their ubiquity.

<sup>8</sup>E.g. are neutrinos Dirac or Majorana fermions?

such conditions on planets such as ours may not be quite right for the origin of life, it is popularly opined that asteroids and comets magically produce life or at the very least, the important, if not crucial components of terrestrial life, so that life then blooms, when their fragments cruise the solar system, stars, and galaxies, and plummet onto appropriately bedecked planets and moons. Be that as it may, it is agreed that life shapes the *milieu* in which viruses evolve (and *vice versa.*)

On the one hand, we need to understand how viruses and life originated and evolved on the Earth and indeed great progress has been made. On the other hand, we should not assume it is safe to collide with similar or different forms of life as we and/or our machines travel elsewhere. The identification of Goldilocks zone worlds does not self-evidently support the assumption of life, but is indicative, rather, of our possible ability to live on such worlds, hopefully without damaging them. Space agencies and their governments have been well aware of the problems of biological contamination from Earth of extraterrestrial places to be visited and explored. Improved priorities and regulation were set [179, 180].

Although such priorities were promulgated repeatedly, the responsible officials apparently over-all ignored the forewarnings and difficulties pointed out, and contamination has resulted, within a few decades of sloppy and premature exploration, severely annihilating our current and future ability to explore and possibly detect life under the correct pristine scientific conditions that our solar system had provided during some 5 billion years of its prior evolution. Examples of problems already in progress include innumerable artificial satellites orbit the Earth, many have been sent to orbit Mars, the Moon, and other planets, asteroids, and comets, etc. and many probes have crashed or landed on Mars, the Moon, as well as other objects, and several have been propelled out of the solar system. Consequently, these need to be fully catalogued and inventoried to ascertain more fully the extent of the contamination by terrestrial microorganisms and materials that contaminated these probes. This microbial contamination problem is being addressed from a peer review and public perspective and should be corrected before further exacerbations of the problems are continued [181, 182].

Clearly, if life were to exist on Mars, it could be deleterious and pathogenic for any terrestrial life, especially if the Goldilocks approach to life in our solar system were correct. In addition, some Martian and terrestrial organisms could produce new life forms (by synergistic (symbiotic) interactions as well as inter-breeding, depending on their ‘biochemistries’), which could then be pathogenic and have unanticipated stark effects. Surely, as one surveys the history of terrestrial catastrophic epidemics and pandemics, most were unanticipated and should be lessons not to under-rate the aggressiveness of many life-forms. Thus, in summary, pathogenicity to any terrestrial life could result from Martian organisms, terrestrial organisms that were conveyed to Mars, and pathogenicity could result in the Martian and terrestrial contexts, including new organisms that interbred. In the obverse, terrestrial life could be inimical to Martian life. This applies to all such exoplanetary explorations.

Finally, to obtain an enhanced perspective of the magnitude and difficulty of the biocontainment problem, please refer to the chapter in this book by Logue et al., which addresses the issue of terrestrial biocontainment at the highest level, BSL-4, for the most pathogenic terrestrial known viruses [29].

**Acknowledgments** Conversations and personal communications are acknowledged: Gilbert Baumslag (Institute for Advanced Study, Princeton, New Jersey); Charles Smith (Princeton University, Princeton, New Jersey); Bishun Khare, Thomas Gold, Frank Drake, and Carl Sagan (Center for Radiophysics and Space Research, Cornell University, Ithaca, NY); Andras Pellionisz (Mountain View, CA); G Rajasekaran, (Institute of Mathematical Sciences, Chennai, India); Andre de Gouvea (Northwestern University, Evanston, IL); Martin Pohl (ESA, Zurich, Switzerland); and Robert Wagoner (Stanford University, Palo Alto, CA).

**Conflicts of Interest** The author reports no conflicts of interest and also that no robots, AI's, nor QC's were harmed during writing this chapter.

## References

1. Waste and duplication in NASA Programs. [https://docs.lib.psu.edu/cgi/viewcontent.cgi?article=1096&context=lib\\_fsdocs](https://docs.lib.psu.edu/cgi/viewcontent.cgi?article=1096&context=lib_fsdocs).
2. Space Debris Remediation: an International Relations Approach [https://www.researchgate.net/publication/293481473\\_Space\\_Debris\\_Remediation\\_An\\_International\\_Relations\\_Approach](https://www.researchgate.net/publication/293481473_Space_Debris_Remediation_An_International_Relations_Approach) Spiegel DS, Turner EL. Bayesian analysis of the astrobiological implications of life's early emergence on Earth. Proc Natl Acad Sci U S A. 2012;109:395–400.
3. NASA Office of Inspector General Annual report, 2018. <https://oig.nasa.gov/docs/MC-2018.pdf>.
4. Bohlmann UM, Burger MJF. Anthropomorphism in the search for extra-terrestrial intelligence – the limits of cognition? Acta Astronaut. 2018;143:163–8. ISSN 0094-5765. <https://doi.org/10.1016/j.actaastro.2017.11.033>.
5. Armstrong S, Sandberg A. Eternity in six hours: intergalactic spreading of intelligent life and sharpening the Fermi paradox. Acta Astronaut. 2013;89:1–13. <https://doi.org/10.1016/j.actaastro.2013.04.002>. <https://flightfromperfection.com>.
6. Dyson FJ. Search for artificial stellar sources of infra-red radiation. Science. 1960;131(3414):1667–8.
7. Dyson FJ. The search for extraterrestrial technologies. In: Marshak RE, editor. Perspect mod phys. New York: Wiley; 1966.
8. Newman WI, Sagan C. Galactic civilizations: populations dynamics and interstellar diffusion. Icarus. 1981;46:293–327.
9. Sagan C. Direct contact among galactic civilizations by relativistic interstellar spaceflight. Planet Space Sci. 1963;11:485–98.
10. Kardashev N. On the inevitability and the possible structures of supercivilizations. Search for Extraterrestrial Life. Proc. Symp. Boston, Massachusetts. June, 1984. Dordrecht Publ. Co.; 1985. p. 497–504.
11. Sagan C, Dyson FJ, Morrison D. Cosmic connection: an extraterrestrial retrospective. 1973. ISBN 978-0-521-78303-3.
12. Cockell CS. Astrobiology and the ethics of new science. Interdisc Sci Rev. 2001;26 <https://doi.org/10.1179/0308018012772533>.
13. Cockell CS. Using exoplanets to test the universality of biology. Nat Astronomy. 2018;2:758–9.
14. Wandel A. On the abundance of extraterrestrial life after the Kepler mission. Int J Astrobiol. 2015;14(3):511–6. <https://doi.org/10.1017/S1473550414000767>.
15. Thucydides. The History of the Peloponnesian War. 431–404 B.C. (Translated by R. Crawley). The Internet Classics Archive. <http://classics.mit.edu/Thucydides/pelopwar.html>; <http://classics.mit.edu/Thucydides/pelopwar.mb.txt>.
16. Marshall DC. Periodical cicada (Homoptera: Cicadidae) life-cycle variations, the historical emergence record, and the geographic stability of brood distributions. Ann Entomol Soc Am. 2001;94:386–99.

17. Grant PR. The priming of periodical cicada life cycles. *Trends Ecol Evol.* 2005;20:169–74.
18. Mayfly. [http://www.wildtrout.org/sites/default/files/projects/teachers\\_introduction\\_to\\_mayfly\\_in\\_the\\_classroom.pdf](http://www.wildtrout.org/sites/default/files/projects/teachers_introduction_to_mayfly_in_the_classroom.pdf).
19. Girimonte D, Izzo D. AI for space applications. 2007. doi:[https://doi.org/10.1007/978-1-84628-943-9\\_12](https://doi.org/10.1007/978-1-84628-943-9_12).
20. Shabbir J, Anwer T. AI and its role in near future. April 1, 2018. arXiv:1804.01396v1 [cs.AI].
21. Javaux EJ, Lepot K. The Paleoproterozoic fossil record: implications for the evolution of the biosphere during Earth's middle-age. *Earth Sci Rev.* 2017; <https://doi.org/10.1016/j.earscirev.2017.10.001>.
22. Schopf JW. The fossil record of cyanobacteria. In: Whitton B, editor. *Ecology of cyanobacteria II*. Dordrecht: Springer; 2012. [https://doi.org/10.1007/978-94-007-3855-3\\_2](https://doi.org/10.1007/978-94-007-3855-3_2).
23. Grady MM, Wright IP, Pillinger CT. Microfossils from Mars: a question of faith. *Astron Geophys.* 1997;38:26–9. <https://academic.oup.com/astrogeo/article-abstract/38/1/26/224846>.
24. Meteorite society 2019. <http://meteoritical.org>.
25. Christoffersen R, Lindsay JF, Noble SK, Meador MA, Kosmo JJ, Lawrence JA, Brostoff L, Young A, McCue T. Lunar dust effects on spacesuit systems: insights from the apollo spacesuits. NASA/TP-2008-000000. <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20090015239.pdf>.
26. Gaier JR. The effects of lunar dust on EVA systems during the apollo missions. NASA/TM-2005-213610. <https://www.hq.nasa.gov/alsj/TM-2005-213610.pdf>.
27. O'Brien BJ, Gaier JR. Indicative basic issues about lunar dust in the lunar environment. A white paper for the National Academies Planetary Sciences Decadal Survey. 2009. <https://www3.nd.edu/~cneal/Lunar-L/LunarDustBasics.pdf>.
28. Fernandez F, Minagar A, Alekseeva N, Shapshak P. Neuropsychiatric aspects of prion disease. In: Sadock BJ, Sadock VA, Ruiz P, editors. *Comprehensive textbook of psychiatry*. New York: Kluwer and Lippincott Publ; 2017. p. 601–18.
29. Logue J, Solomon J, Niemeyer BF, Benam KH, Lin AE, Bjornson Z, Jiang S, McIlwain DR, Nolan GP, Palacios G, Kuhn JH. Innovative technologies for advancement of WHO risk group 4 pathogens research. New York: Springer; 2019. Chapter in this volume.
30. Chippaux J-P. Outbreaks of Ebola virus disease in Africa: the beginning of a tragic saga. *J Venom Anim Toxins Incl Trop Dis.* 2014;20:1–14.
31. Ladner JT, Wiley MR, Mate S, Dudas G, Prieto K, Lovett S, Nagle ER, Beitzel B, Gilbert ML, Fakoli L, DiClaro JW II, Schoepp RJ, Fair J, Kuhn JH, Hensley LE, Park DJ, Sabeti PC, Rambaut A, Sanchez-Lockhart M, Bolay FK, Kugelman JR, Palacios G. Evolution and spread of Ebola virus in Liberia, 2014–2015. *Cell Host Microbe.* 2015;18:659–69.
32. Kuhn JH. In: Kuhn JH, Calisher CH, editors. *Filoviruses. A compendium of 40 years of epidemiological, clinical, and laboratory studies*. New York: Springer; 2015.
33. BSL-4.: [https://en.wikipedia.org/wiki/Biosafety\\_level](https://en.wikipedia.org/wiki/Biosafety_level).
34. Bradfute SB, Jahrling PB, Kuhn JH. Chapter 20. Ebola virus disease. *Global virology I – identifying ad investigating viral diseases*. New York: Springer; 2015. p. 543–59.
35. Jahrling PB, Keith L, St. Claire M, Johnson RF, Bollinger L, Matthew G, Lackemeyer MG, Lisa E, Hensley LE, Jason Kindrachuk J, Kuhn JH. The NIAID Integrated Research Facility at Frederick, Maryland: a unique international resource to facilitate medical countermeasure development for BSL-4 pathogens. *Pathog Dis.* 2014;71:213–8.
36. Janosko K, Holbrook MR, Adams R, Barr J, Bollinger L, Newton JT, Ntiforo C, Coe L, Wada J, Pusl D, Jahrling PB, Kuhn JH, Lackemeyer MG. Safety precautions and operating procedures in an (A)BSL-4 laboratory: 1. Biosafety level 4, suit laboratory, suite entry, and exit procedures. *J Vis Exp.* 2016;116:e52317. <https://doi.org/10.3791/52317>.
37. Dyer O. Congo's Ebola epidemic is now at its worst ever and still spreading. *BMJ.* 2019;362:1433–41. <https://doi.org/10.1136/bmj.1433>.
38. DRC Health Ministry. 2019. <https://us13.campaign-archive.com/?u=89e5755d2cca4840b1af93176&id=21512e200b>.
39. Christaki E. New technologies in predicting, preventing and controlling emerging infectious diseases. *Virulence.* 2015;6:558–65.

40. Shapshak P, Sinnott JT, Somboonwit C, Kuhn JH. Global virology I – identifying and investigating viral diseases. New York: Springer; 2015b.
41. Parrish CR, Holmes EC, Morens DM, Park EC, Burke DS, Calisher CH, Laughlin CA, Saif LJ, Dazak P. Cross-species virus transmission and the emergence of new epidemic diseases. *Microbiol Mol Biol Rev*. 2008;72:457–70.
42. Geoghegan JL, Duchenne S, Holmes EC. Comparative analysis estimates of the relative frequencies of co-divergence and cross-species transmission within viral families. *PLoS Pathol*. 2017;13:1–17. <https://doi.org/10.1371/journal.ppat.1006215>.
43. NASA meteor report. [https://www.nasa.gov/sites/default/files/files/Meteors\\_Meteorites\\_Lithograph.pdf](https://www.nasa.gov/sites/default/files/files/Meteors_Meteorites_Lithograph.pdf).
44. Min K, Reiners PW. High-temperature Mars-to-Earth transfer of meteorite ALH84001. *Earth Planet Sci Lett*. 2007;260:72–85.
45. Tayro EAM, Scott ERD, Sharma SK, Misra AK. The pressures and temperatures of meteorite impact: evidence from micro-Raman mapping of mineral phases in the strongly shocked Taiban ordinary chondrite. *Am Mineral*. 2013;98:859–69. <https://doi.org/10.2138/am.2013.4300>.
46. Acosta-Maeda TE, Scott ERD, Sharma SK, Misra AK. The pressures and temperatures of meteorite impact: evidence from micro-Raman mapping of mineral phases in the strongly shocked Taiban ordinary chondrite. *Am Mineral*. 2013;98:859–69.
47. Jenniskens P, Laux CO, Wilson MA, Schaller EL. The mass and speed dependence of meteor air plasma temperatures. *Astrobiology*. 2004;4:1–14.
48. Chyba C, Sagan C. Endogenous production, exogenous delivery and impact-shock synthesis of organic molecules: an inventory for the origins of life. *Nature*. 1992;355:125–32. <https://doi.org/10.1038/355125a0>.
49. Kant I. Allgemeine Naturgeschichte und Theorie des Himmels. Konigsberg, Leipzig: Petersen Publications; 1755.
50. Callahan MP, Smith KE, HJ Cleaves II, Ruzick J, Stern JC, Glavin DP, House CH, Dworkin JP. Carbonaceous meteorites contain a wide range of extraterrestrial nucleobases. *Proc Natl Acad Sci U S A*. 2011;108:13995–8. <https://doi.org/10.1073/pnas.1106493108>.
51. Schmitt-Kopplin P, et al. High molecular diversity of extraterrestrial organic matter in Murchison meteorite revealed 40 years after its fall. *Proc Natl Acad Sci U S A*. 2010;107:2763–8.
52. Sephton MA. Organic compounds in carbonaceous meteorites. *Nat Prod Rep*. 2002;19:292–311.
53. Engel MH, Macko SA. Isotopic evidence for extraterrestrial non-racemic amino acids in the Murchison meteorite. *Nature*. 1997;389:265–8.
54. Sagan C, Bilson E, Raulin F, Shapshak P. Amino acid destruction under simulated lunar conditions. Center for Radiophysics and Space Research, Cornell University, Ithaca. Report number 488 1971. <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/19790024970.pdf>.
55. Loeb A. The habitable epoch of the early universe. 2014. arXiv:12.0613v3 [astro-ph.CO].
56. Gibson CH. The biological big bang: the first oceans of primordial planets at 2–8 million years explain Hoyle/Wickramasinghe cometary panspermia. *Proc SPIE* 8152-37. 2011. p. 1–19.
57. Sagan C. Ultraviolet selection pressure on the earliest organisms. Ithaca. Report number 445: Center for Radiophysics and Space Research, Cornell University; 1971.
58. Sagan C, Shapshak P. On ultraviolet light and the origin of ribosomes. Ithaca). Report number 446: Center for Radiophysics and Space Research, Cornell University; 1971.
59. Fox GE. Origin and evolution of the ribosome. *Cold Spring Harb Perspect Biol*. 2010;2:a003483. <https://doi.org/10.1101/cshperspect.a003483>.
60. Benner SA, Kim HJ, Yang Z. Setting the stage: the history, chemistry, and geobiology behind RNA. *Cold Spring Harb Perspect Biol*. 2012;4:a003541. <https://doi.org/10.1101/cshperspect.a003541>.
61. Robertson MP, Joyce GF. The origins of the RNA world. *Cold Spring Harb Perspect Biol*. 2012;4:a003608. <https://doi.org/10.1101/cshperspect.a003608>.
62. Blaustein R. Advances in astrobiology. *Bioscience*. 2015;65:460–5. <https://doi.org/10.1093/biosci/bivs05>.

63. Gray MW. Lynn Margulis and the endosymbiont hypothesis: 50 years later. *Mol Biol Cell.* 2017;28:1285–7. <https://doi.org/10.1091/mbc.E16-07-0509>. PMID: 28495966. PMCID: PMC5426843.
64. Lopez-Garcia P, Emeb L, Moreira D. Symbiosis in eukaryotic evolution. *J Theor Biol.* 2017; <https://doi.org/10.1016/j.jtbi.2017.02.031>.
65. Mereschkowsky C. Über natur und usprung der chromatophoren im pflanzenreiche. *Biol Cent.* 1905;25:593–604.
66. Mereschkowsky K. Theorie der zwei Plasmaarten als Grundlage der Symbiogenesis, einer neuen Lehre von der Entstehung der Organismen. *Biol Cent.* 1910;30:353–67.
67. Sagan L. On the origin of mitosing cells. *J Theor Biol.* 1967;14:255–74.
68. Sagan C. Definitions of life. This chapter originally appeared as the first section in the entry for “Life” in Encyclopedia Britannica, pp. 1083–1083A, Chicago: Encyclopedia Britannica Incorporated, 1970. <http://dnapunctuation.org/~poptsova/course2012/Sagan%20Definitions%20of%20life.pdf>.
69. Schrodinger E. What is life? Mind and matter. Cambridge: Cambridge University Press; 1967. (First published in 1944).
70. Rabinowitch E, Govindjee. Photosynthesis. New York: Wiley; 1969. ISBN 471 704245.
71. Haaker H. Biochemistry and physiology of nitrogen fixation. *BioEssays.* 1988;9:112. <https://doi.org/10.1002/bies.950090403>.
72. Leningher A. Principles of biochemistry. 7th ed. Accessed 4 Sept 2018. [http://www.esalq.usp.br/lepe/imgs/conteudo\\_thumb/mini/Principles-of-Biochemistry-by-ALbert-Leningher.pdf](http://www.esalq.usp.br/lepe/imgs/conteudo_thumb/mini/Principles-of-Biochemistry-by-ALbert-Leningher.pdf).
73. Wilson EK, Walker J. Principles and techniques of biochemistry and molecular biology. In: Wilson K, Walker J, editors. Cambridge, UK: Cambridge University Press; 2010. <http://www.kau.edu.sa/Files/0017514/Subjects/principals%20and%20techniques%20of%20biochemistry%20and%20molecular%20biology%207th%20ed%20wilson%20walker.pdf>.
74. Vasudevan DM, Sreekumari S, Vaidyanathan K. Textbook of biochemistry for medical students. New Delhi: Jaypee Brothers Medical Publ. (P) Ltd; 2011. <https://ia802205.us.archive.org/1/items/pdfy-5vClyqSbVzIGpuT2/DM%20Vasudevan%20-%20Textbook%20of%20Biochemistry%20For%20Medical%20Students%2C%206th%20Edition.pdf>.
75. Prigogine I. Thermodynamics of irreversible processes. New York: Wiley; 1967.
76. Klotz IM, Rosenberg RM. Chemical thermodynamics: basic concepts and methods. Hoboken: Wiley; 2008.
77. Klotz IM. Energetics in biochemical reactions. New York: Academic Press Inc; 1957.
78. Tolman RC. Relativity, thermodynamics, and cosmology. Mineola: Dover Publications; 1987.
79. de Waele ATAM. The first, second, and third laws of thermodynamics (ThLaws05.tex). 2009. <http://cryocourse2011.grenoble.cnrs.fr/IMG/file/Lectures/2011-deWaele-ThLaws05.pdf>.
80. Taubner RS, Pappenreiter P, Zwicker J, Smrzka D, Pruckner C, Kolar P, Bernacchi S, Seifert AH, Krajete A, Bach W, Peckmann J, Paulik C, Firneis MG, Schleper C, Rittmann SKMR. Biological methane production under putative Enceladus-like conditions. *Nat Commun.* 2018;9:748. <https://doi.org/10.1038/s41467-018-02876-y>.
81. Masuda T, Dobson GP, Veech RL. The Gibbs-Donnan near-equilibrium system of heart. *J Biol Chem.* 1990;265:20321–34.
82. Tel T. Fractals, multifractals, and thermodynamics. *Z. Naturforsch.* 1988;43a:1154–74.
83. Denisov S. Fractal binary sequences: tsallis thermodynamics and Zipf's law. *Electromagnet Stud.* 1998;I:64–8.
84. Gaspard P. Chaos, fractals, and thermodynamics. *Bull Cl Sci Acad R Belg.* 2000;6e-XI:9–48.
85. Deppman D. Thermodynamics with fractal structure, Tsallis statistics, and hadrons. 2016. arXiv:1601.02400v1 [hep-ph] 11 Jan 2016.
86. Weberszilp J, Chen W. Generalized maxwell relations in thermodynamics with metric derivatives. *Entropy.* 2017;19:407–19. <https://doi.org/10.3390/e19080407>.
87. Shapshak P, Somboonwit C, Sinnott JT. Artificial Intelligence and Virology – *quo vadis*. *Bioinformation.* 2017a;13(12):410–1.
88. Shapshak P. Artificial intelligence and brain. *Bioinformation.* 2018;14(1):038–41.

89. Zak M. A model of emerging intelligence in Universe. *Int J Astrobiol.* 2019;18(3):251–8. <https://doi.org/10.1017/S1473550417000489>.
90. Balaji S, Akash R, Krittika N, Shapshak P. Sequence accuracy in primary databases: a case study on HIV-1B. In: Shapshak P, Levine AJ, Somboonwit C, Foley BT, Singer E, Chiappelli F, Sinnott JT, editors. *Global virology II. HIV and NeuroAIDS.* New York: Springer; 2017.
91. Sneha P, Balaji S, Shapshak P. Amyloidogenic pattern prediction of HIV-1 proteins. In: Shapshak P, et al., editors. Chapter 33 in *Global virology II – HIV and NeuroAIDS.* New York: Springer; 2017. p. 823–95. [https://doi.org/10.1007/978-1-4939-7290-6\\_33](https://doi.org/10.1007/978-1-4939-7290-6_33).
92. Geschwind MD. Prion diseases. *Continuum (Minneapolis Minn).* 2015;21(6 Neuroinfectious disease):1612–38. <https://doi.org/10.1212/CON.0000000000000251>.
93. Furr A, Young AJ, Richt J. The immune system in the pathogenesis and prevention of prion diseases. *J Biotech Biodef.* 2012;S1:012. <https://doi.org/10.4172/2157-2526.S1-012>.
94. Gianluigi F, Balducci C. Beta-amyloid oligomers and prion protein – fatal attraction? *Prion.* 2011;5:10–5.
95. Boland CR. Non-coding RNA: it's not junk. *Dig Dis Sci.* 2017;62:1107–9. <https://doi.org/10.1007/s10620-017-4506-1>.
96. Antonarakis SE. Human genome sequence variation. In: Speicher M, Antonarakis SE, Motulsky AG, editors. *Human genetics: problems and approaches.* New York: Springer; 2009. p. 981.
97. NIH Roadmap and personalized medicine. <https://commonfund.nih.gov/sites/default/files/ADecadeofDiscoveryNIHRoadmapCF.pdf> <https://newsinhealth.nih.gov/2013/12/personalized-medicine>.
98. Shapshak P, Chiappelli F, Commins D, Singer E, Levine AJ, Somboonwit C, Minagar A, Pellionisz A. Molecular epigenetics, chromatin, and NeuroAIDS/HIV: translational implications. *Bioinformation.* 2008;3:53–7.
99. Shapshak P, Levine AJ, Somboonwit C, Foley BT, Singer E, Chiappelli F, Sinnott JT. *Global virology II. HIV and NeuroAIDS.* New York: Springer; 2017b.
100. Shapshak P. Challenges in health research funding: an opinion. *Bioinformation.* 2015c;11(2):55–6.
101. Steward CA, Parker APJ, Minassian BA, Sisodiya SM, Frankish A, Harrow J. Genome annotation for clinical genomic diagnostics: strengths and weaknesses. *Genome Med.* 2017;9:49. <https://doi.org/10.1186/s13073-017-0441-1>.
102. Pellionisz AJ. The principle of recursive genome function. *Cerebellum.* 2008;7:348–59. <https://doi.org/10.1007/s12311-008-0035-y>.
103. Forget F. On the probability of habitable planets. *Int J Astrobiol.* 2013; <https://doi.org/10.1017/S1473550413000128>.
104. Spiegel DS, Turner EL. Bayesian analysis of the astrobiological implications of life's early emergence on Earth. *Proc Natl Acad Sci U S A.* 2012;109:395–400.
105. Scharf C, Cronin L. Quantifying the origins of life on a planetary scale. *Proc Natl Acad Sci U S A.* 2016; <https://doi.org/10.1073/pnas.1523233113>.
106. Gonzalez G. Review: setting the stage for habitable planets. *Life.* 2014;4:1–27. <https://doi.org/10.3390/life40x000x>.
107. Gonzalez Oreja JA. *Quo vadis, panspermia?* Del origen de la vida en la Tierra a una ecología interplanetaria. *eVOLUCION.* 2016;11(1):71–88.
108. Woolfson M. Planet formation and the evolution of the Solar System. 2017. <https://arxiv.org/pdf/1709.07294.pdf>.
109. Wooldridge SA. Mass extinctions past and present: a unifying hypothesis. *Biogeosci Discuss.* 2008;5:2401–23.
110. Drabek-Maunder E, Greaves J, Fraser H, Clements D, Alconcel L. Ground-based detection of a cloud of methanol from Enceladus: when is a biomarker not a biomarker? *Int J Astrobiol.* 2017;1–8. <https://doi.org/10.1017/S1473550417000428>.
111. von Neumann J. Theory of self-replicating automata. In: Burks AW, editor. Urbana: University of Illinois Press; 1966.

112. Pesavento U. An implementation of von Neumann's self-reproducing machine. *Artif Life.* 1995;2:337–54. (Princeton University and Massachusetts Institute of Technology).
113. van Hecke K, de Croon GCHE, Hennes D, Setterfield TP, Saenz-Otero A. Self-supervised learning as an enabling technology for future space exploration robots: ISS experiments on monocular distance learning. *Acta Astronaut.* 2017;140:1–9. <https://doi.org/10.1016/j.actaastro.2017.07.038>.
114. Shen T, Yuan K, Chen D, Colloc J, Yang M, Li Y, Lei K. An ontology-driven clinical decision support system (IDDAP) for infectious disease diagnosis and antibiotic prescription. *Artificial Intelligence in Medicine.* 2018. <https://www.sciencedirect.com/science/article/pii/S0933365717302348>.
115. Dirkx D, Noomen R, Visser PNAM, Gurvits L, Vermeersen LLA. Space-time dynamics estimation from space mission tracking data. *Astron Astrophys.* 2015;. <https://arxiv.org/pdf/1512.06685.pdf>
116. Turyshev SG, Williams JG, Shao M, Anderson JD. Laser ranging to the Moon, Mars, and beyond. The 2004 NASA/JPL Workshop on Physics for Planetary Exploration. April 20–22, 2004, Solvang, CA. <https://arxiv.org/abs/gr-qc/0411082v1>.
117. Hippke M. Interstellar communication. II. Application to the solar gravitational lens. *Acta Astronaut.* 2018;142:64–74.
118. Feynman R. New Directions in Physics: the Los Alamos 40th anniversary volume. In: Metropolis N, Kerr DM, Rota G-C, editors. Orlando: Academic Press, Inc.; 1987.
119. Feynman R. Quantum mechanical computers. *Optic News.* 1985;11:11–20. <https://doi.org/10.1364/ON.11.2.000011>.
120. Gudder SP. Stochastic methods in quantum mechanics. Mineola: Dover Publications, Inc; 1979.
121. Shapshak P. Challenges in Health Research Funding: an opinion. *Bioinformation.* 2015a;11(2):55–6. PMCID: PMC4369678.
122. Meech KJ, Weryk R, Micheli M, Kleyna JT, Hainaut OR, Jedicke R, Wainscoat RJ, Chambers KC, Keane JV, Petric A, Denneau L, Magnier E, Berger T, Huber ME, Flewelling H, Waters C, Schunova-Lilly E, Chastel S. A brief visit from a red and extremely elongated interstellar asteroid. *Nature.* 2017;552:378–81. <https://doi.org/10.1038/nature25020>.
123. Schneider J. Is I/2017 U1 really of interstellar origin. 2017. arXiv:1711.05735v1.
124. Fitzsimmons A, Snodgrass C, Rozitis B, Yang B, Hyland M, Seccull T, Bannister MT, Fraser WC, Jedicke R, Lacerda P. Spectroscopy and thermal modelling of the first interstellar object I/2017 U1 ‘Oumuamua. ArXiv:1712.06552v1.
125. Fraser WC, Pravec P, Fitzsimmons A, Lacerda P, Bannister MT, Snodgrass C, Smolic I. The tumbling rotational state of I/‘Oumuamua. *Nat Astronomy.* 2018; <https://doi.org/10.1038/s41550-018-0398-z>.
126. Loeb A. Six strange facts about our first interstellar guest, ‘Oumuamua’. 2018. arXiv:1811.08832.
127. Bailer-Jones CAL, Farnocchia D, Meech KJ, Brasser R, Micheli M, Chakrabarti S, Buie MW, Hainaut OR. Plausible home stars of the interstellar object ‘Oumuamua’ found in Gaia DR2. 2018. arXiv: 1809.09009v1.
128. Siraj A, Loeb A. Identifying interstellar objects trapped in the solar system through their orbital parameters. 2019. arXiv:1811.09632v5 [astro-ph.EP] 4 Feb 2019.
129. Europa Lander Mission, Lander Study – 2016 report – JPL D-97667. NASA. Hand KP and the project engineering team. 2017.
130. Europa mission. <https://www.space.com/36993-nasa-europa-mission-launch-date-2018-budget.html>.
131. NASA budget. [https://en.m.wikipedia.org/wiki/Budget\\_of\\_NASA](https://en.m.wikipedia.org/wiki/Budget_of_NASA).
132. NIH Budget. [https://en.wikipedia.org/wiki/National\\_Institutes\\_of\\_Health](https://en.wikipedia.org/wiki/National_Institutes_of_Health).
133. NIH Nobel Laureates. <https://www.nih.gov/about-nih/what-we-do/nih-almanac/nobel-laureates>.
134. Faye D, Rampini R. Contamination control policy: last publication standards in standardizations. [http://esmat.esa.int/Materials\\_News/ISME09/pdf/6-Contamination/S8%20-%20Faye.pdf](http://esmat.esa.int/Materials_News/ISME09/pdf/6-Contamination/S8%20-%20Faye.pdf).

135. ECSS Secretariat, ESA-ESTEC, Requirements & Standards Division, Noordwijk, The Netherlands. Space product assurance Cleanliness and contamination control. 2008. <https://ecssies.org/download/webDocumentFile?id=62823>.
136. NASA. International Space Station, 2009. [https://www.nasa.gov/pdf/393789main\\_iss\\_utilization\\_brochure.pdf](https://www.nasa.gov/pdf/393789main_iss_utilization_brochure.pdf).
137. Hara T, Takagi K, Kajiura D. Transfer of life-bearing meteorites from earth to other planets. *J Astrobiol Space Sci Rev*. 2019;1:299–310.
138. Arrhenius S. Die Verbreitung des Lebens im Weltenraum. Die Umschau, Frankfurt a.M. 1903;7:481–6.
139. O'Leary MR. Anaxagoras and the origin of Panspermia theory. In iUniverse. ISBN 978-0-595-49596-2. OCLC 757322661. 2008.
140. Wickramasinghe C. The astrobiological case for our cosmic ancestry. *Int J Astrobiol*. 2010;9:119–25.
141. Wickramasinghe MK, Wickramasinghe C. Interstellar transfer of planetary microbiota. *Mon Not R Astron Soc*. 2004;348:52–7. <https://doi.org/10.1111/j.1365-2966.2004.07355.x>.
142. Boston PJ, Spilde MN, Northup DE, Melim LA, Soroka DA, Kleina LG, Lavoie KH, Hose LD, Mallory LM, Dahm CN, Crossey LJ, Scheble RT. Cave biosignature suites: microbes, minerals and Mars. *Astrobiology*. 2001;1:25–55.
143. Melim LA, Liescheidt R, Northup DE, Spilde MN, Boston PJ, Queen JM. A biosignature suite from cave pool precipitates, Cottonwood cave, New Mexico. *Astrobiology*. 2009;9:907–17.
144. Ginsburg I, Lingam M, Loeb A. Galactic panspermia. 2018. arXiv:1810.04307v2 [astro-ph.EP].
145. Suttle C. Viruses in the sea. *Nature*. 2005;437:356–9.
146. Weinberg KD. Viruses in marine ecosystems: from open waters to coral reefs. *Adv Virus Res*. 2018;101:1–38. <https://doi.org/10.1016/bs.avir.2018.02.001>.
147. Glavin DP, Dworkin JP, Lupisella M, Kminek G, Rumme JD. Biological contamination studies of lunar landing sites: implications for future planetary protection and life detection on the Moon and Mars. *Int J Astrobiol*. 2005;265–71. <https://doi.org/10.1017/S1473550404001958>. <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20040084463.pdf>.
148. Glavin DP, Dworkin JP, Lupisella M, Kminek G, and Rumme JD. *In situ* biological contamination studies of the moon: implications for future planetary protection and life detection missions. 2010. <https://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20100036597.pdf>.
149. Netea MG, van de Veerdonk FL, Strous M, van der Meer JWM. Infection risk of a human mission to Mars. *J Astrobiol Space Sci Rev*. 2019;1:144–55.
150. Ridder NN, Maan DC, Summerer L. Terraforming Mars: generating greenhouse gases to increase martian surface temperatures. *J Astrobiol Space Sci Rev*. 2019;1:338–52.
151. Hughes RA, Ellington AD. Synthetic DNA synthesis and assembly: putting the synthetic in synthetic biology. *Cold Spring Harb Perspect Biol*. 2017;9:a023812.
152. Hutchison CA 3rd, Chuang RY, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, Gill J, Kannan K, Karas BJ, Ma L. Design and synthesis of a minimal bacterial genome. *Science*. 2016;351:aad6253.
153. Zhang LY, Chang SH, Wang J. Synthetic biology: from the first synthetic cell to see its current situation and future development. *Chin Sci Bull*. 2011;56:229–37. <https://doi.org/10.1007/s11434-010-4304-z>.
154. Chakraborty C, Agoramoorthy G. Stem cells in the light of evolution. Review article. *Indian J Med Res*. 2012;135:813–9.
155. The stem cell book-NIH stem cell information. <https://stemcells.nih.gov> Cited 4-15-2019.
156. Hartenstein V. Stem cells in the context of evolution and development. *Dev Genes Evol*. 2013;223 <https://doi.org/10.1007/s00427-012-0430-8>.
157. Arnould J. Astrobiology, sustainability and ethical perspectives. *Sustainability*. 2009;1:1323–30. <https://doi.org/10.3390/su1041323>.
158. Losch A. The need of an ethics of planetary sustainability. *Int J Astrobiol*. 2017; <https://doi.org/10.1017/S1473550417000490>.

159. Rodriguez HE, Lakshmi S, Somboonwit C, Oxner A, Guerra L, Addisu A, Gutierrez L, Sinnott JT, Nilofer C, Kangueane P, Shapshak P. Gene therapy blueprints for NeuroAIDS. In: Shapshak P, Levine AJ, Foley BT, Somboonwit C, Singer E, Chiappelli F, Sinnott JT, editors. *Global virology II – HIV and NeuroAIDS*. New York: Springer; 2017. p. 953–93.
160. de Gouveia A. Neutrino Mass Models. *Ann Rev Nucl Part Sci*. 2016;66:197–215.
161. de Gouveia A. Neutrino Anomalies & CEvNS. PIRE Workshop. COFI February 6–7, 2017a.
162. de Gouveia A. Neutrino physics. Evanston). Lectures at Institute for Advanced Study (Princeton, NJ): Northwestern University; 2017b.
163. de Gouveia A. [https://cpb-us-e1.wpmucdn.com/sites.northwestern.edu/dist/8/307/files/2017/08/Andre-de-Gouveia-PIRE\\_PR\\_2017-16t79ju.pdf](https://cpb-us-e1.wpmucdn.com/sites.northwestern.edu/dist/8/307/files/2017/08/Andre-de-Gouveia-PIRE_PR_2017-16t79ju.pdf); 2017c.
164. Hannestad S. Aspects of neutrino physics in the early universe. PhD Thesis. Institute of Physics and Astronomy, University of Aarhus, Copenhagen, Denmark. 1997. [http://phys.au.dk/fileadmin/site\\_files/publikationer/phd/Steen\\_Hannestad.pdf](http://phys.au.dk/fileadmin/site_files/publikationer/phd/Steen_Hannestad.pdf).
165. Freedman WL. The Hubble constant and the expansion age of the Universe. *Phys Rep*. 2000;334:13–31.
166. Rajasekaran G. Phenomenology of neutrino oscillations. *Pramana*. 2000;55:19–5.
167. Indumathi D, Murthy MVN, Rajasekaran G. Perspectives in Neutrino Physics. *Proc India Natl Sci Acad–Part A*. 2004;70:1–15.
168. INO web-site: <http://www.imsc.res.in/~ino>.
169. Pasachoff JM, Kutner ML. Neutrinos for interstellar communication. *Cosmic Search*. 1979;2:21.
170. Hippke M. Interstellar communication. IV. Benchmarking information carriers. *Acta Astronaut*. 2018b;151:53–62.
171. Learned JG, Pakvasa S, Zee A. Galactic neutrino communication. *Phys Lett B*. 2009;671:15–9.
172. Stancil DD, Brooks W, Alania M, 110 additional authors. Demonstration of communication using neutrinos. *Mod Phys Lett A*. 2012;1–10.
173. Silagadze ZK. SETI and muon collider. 2008. arXiv:0803.0409v1.
174. Cavanna F, Costantinia ML, Palamarab O, Vissani F. Neutrinos as astrophysical probes. 2003. arXiv:astro-ph/0311256v1 11 Nov 2003.
175. Recami E. Ettore Majorana: the scientist and the man. *Int J Mod Phys D*. 2014;16:1–23. [https://www.researchgate.net/publication/269762476\\_ETTORE\\_MAJORANA\\_HIS\\_WORK\\_AND\\_HIS\\_LIFE\\_in\\_English\\_-\\_and\\_with\\_some\\_up-dated\\_Bibliography](https://www.researchgate.net/publication/269762476_ETTORE_MAJORANA_HIS_WORK_AND_HIS_LIFE_in_English_-_and_with_some_up-dated_Bibliography).
176. Esposito S. Majorana solution of the Thomas-Fermi equation. *Am J Phys*. 2002a;70:852–63.
177. Esposito S. Majorana transformation for differential equations. *Int J Theor Phys*. 2002b;41:2417–31.
178. Di Grezia E, Esposito S. Fermi, Majorana and the statistical model of atoms. *Found Phys*. 2004;34:1431–52.
179. Hall LB, Miles JR, Bruch CW, Tarver P. The objectives and technology of spacecraft sterilization. Washington, DC: NASA History Division, NASA Headquarters; 1995. 542pp.
180. Meltzer M. When biospheres collide. A history of NASA's planetary protection program. 2011. NASA SP-2011-4234.
181. Rhawn J. Sterilization failure and fungal contamination of Mars and NASA's Mars rovers. *J Cosmol*. 2018;30:51–97. <https://www.researchgate.net/publication/322702459>.
182. Faire AG, Parro V, Schulze-Makuch D, Whyte L. Searching for life on Mars before it is too late. *Astrobiology*. 2018;17:962–70. <https://doi.org/10.1089/ast.2017.1703>.

# Climate Crisis Impact on AIDS, IRIS and Neuro-AIDS



Francesco Chiappelli, Emma Reyes, and Ruth Toruño

**Abstract** Our awareness of the climate crisis/catastrophe is a convoluted story, which began in chemistry laboratories several decades ago. In that context, scientists first described the greenhouse effect: the potential CO<sub>2</sub> accumulation allows solar heat to penetrate the atmosphere, but prevents radiated warmth to escape from it. A theoretical proposition at first, scientists first described the greenhouse effect as the potential accumulation of CO<sub>2</sub>, allowing solar heat to penetrate the atmosphere while simultaneously preventing the radiated warmth from escaping it. This increase in atmospheric CO<sub>2</sub> would hypothetically then raise the planet's temperature, and if left unchecked, this phenomenon could lead to a multitude of issues including the melting of polar ice caps, rising of oceanic waters, alterations in the acidity and temperature of global water reservoirs (i.e., rivers, lakes, and seas), and changes in the patterns and strength of the gulf streams and consequently the jet stream.

Global warming is now a reality that is substantiated by an abundance of facts and evidence. The best available evidence further confirms that the climate crisis we have engendered with wanton human activity since the industrial revolution and with renewed vigor following WWII, has exacerbated within the last five decades, proffering serious threats to the health of children, adults and the elderly alike.

We discuss the implications of the climate crisis to susceptibility of HIV disease, AIDS, Neuro-AIDS and IRIS, and proffer the current measles outbreak in the Philippines as proof-of-concept of the proposition that global warming can affect morbidity and mortality to viral infections. We also propose a general Artificial Intelligence-driven statistical space-time mixture algorithm as a Bayesian predictive model for climate change-associated medical emergencies.

**Keywords** Human Immunodeficiency virus (HIV) · Acquired immune deficiency syndrome (AIDS) · Neuro-AIDS · Immune reconstitution inflammatory syndrome (IRIS) · Macroenvironment · Microenvironment · Best evidence base · Systematic review · Translational Environmental Restoration (TER) · Climate crisis · Artificial Intelligence (AI) · Bayesian semi-parametric multiple regression · Space-time model

---

F. Chiappelli (✉) · E. Reyes · R. Toruño

Department of the Health Sciences (Biostatistics), California State University, Northridge, CA, USA; <http://francescochiappelli.com/>

## 1 Introduction

Our global climate is changing, and the situation is growing worse by the day [1, 2]. It matters not what the weather is in any given site on our planet on any given day: climate patterns have changed widely worldwide. Opponents argue that global patterns of climate change are not sound scientific data. On the contrary, observations of dramatic changes in climatic patterns around the globe rest on rigorous best evidence criteria.

Skeptics defend that perhaps we are indeed undergoing radical climate change, but such cyclic patterns are naturally occurring, and thus independent of human activity. The scientific data clearly show otherwise – every analysis shows that the changes in climate registered in the last 75–100 years derive primarily from devastating human activities, such as massive and irresponsible deforestation, garbage and plastic pollution, and increased air pollutant emission, among many others. Human activities detrimental to the global climate have been so intense and persistent over the past century that, even if their effects have been moderately curbed within the past decade, the changes in climate they have caused and continue to engender have become an emergency climate crisis, which is rapidly evolving into a worldwide catastrophe.

Other misinformed commentators who agree on one hand that the recorded significant rise in global temperature associated with climate change is directly consequential to human interference in the macro-environmental balance have suggested on the other hand, that warmer temperatures are better and therefore the alarmist stance of the scientific community must be contested. Quite on the contrary, every bit of scientific data demonstrates that all living organisms are immediately experiencing the many consequences of the climate changing such as heat stress, increased acidity of air and water due to elevated carbon dioxide ( $\text{CO}_2$ ) levels, clean water deprivation and drought, and release of toxic gases: all of which pose a serious threat to public health [1, 2].

Recent evidence indicates that global heat is best measured in the oceans because they constitute the principal source of thermal inertia in the climate system. Data show that oceans have gained  $1.33 \pm 0.20 \times 10^{22}$  joules of heat/year on average between 1991 and 2016. This is equivalent to a planetary energy imbalance of  $0.83 \pm 0.11$  watts/ $\text{m}^2$  of the planet's surface, very much in the upper quartile end of previous estimates. These data are interpreted to mean that there is more  $\text{CO}_2$  trapped by the oceans than we originally feared, and the oceans – the main thermal regulators of our planet – are warming at a rate greater than previously anticipated [1].

Experts no longer warn of a simple climate change, but rather of a climate crisis. In fact, the UN Intergovernmental Panel on Climate Change (IPCC) generated a report earlier this year (October, 2018) alerting scientists and policy makers alike that our world is heading toward a climate catastrophe unless we revise our goals and practices with respect to climate warming. Specifically, the IPCC-2018 special report [3] analyzes the impacts of global warming of  $1.5^\circ\text{C}$  above pre-industrial

levels and related global greenhouse gas emission pathways. It proposes interventions and solutions for strengthening the global response to the threat of climate change, implementing methods of sustainable development. Short of those, the report predicts in no uncertain terms that global warming will reach the status of an irreversible global catastrophe sometime between 2030 and 2052 if detrimental human activities continue uncurbed at the current rate. Long-term changes in the climate system of the planet, such as sea level rise with its associated impacts have in fact already begun, the report asserts [3]. Evidence of this encompasses serious climate-related risks of natural and human systems, from marine biodiversity and fisheries, to coral reef ecosystems, and their functions and services to human survival [3]. The report confirms the best available evidence that outlines the direct nefarious effects of the climate crisis to animal and vegetal livelihood on the planet, and consequentially food security, water supply, human security, and more generally the consequential risks to human health [1–3]. Case in point, global warming can generate heat stress-mediated impairment of immunity. Expectations are, therefore, that HIV/AIDS and other infectious diseases (e.g., Ebola, Zika) will threaten public health more aggressively in the proximal future [2].

## 2 HIV-Disease, AIDS, IRIS and Neuro-AIDS

### 2.1 *Reprogramming of the Metabolic Status and Immunity*

Over one third of the patients seropositive for the human immunodeficiency virus (HIV) with signs of the acquired immune deficiency syndrome (AIDS) or neurological manifestations of AIDS (Neuro-AIDS), and under treatment with anti-retroviral (ART) interventions, develop the immune reconstitution inflammatory syndrome (IRIS) [4–7]. The variables that determine the incidence of ART-related IRIS are not fully characterized, and are suspect to include immunological as well as psycho-biological factors [4]. The events that determine and control cellular immune surveillance are also dependent on psychoneuroendocrine modulation, as well as physiological homeostasis and allostasis [2, 4, 8–10].

The state of the cellular immune system, and particularly the number and function of its primary cellular component the T lymphocytes that express cluster of differentiation #4 (CD4), are key determinants of cellular immune surveillance. The various subpopulations of CD+ lymphocytes work in concert to initiate and to regulate cellular immunity for the production of antibodies, as well as for the regulation of CD8+ positive cells and other cytotoxic subpopulations of white blood cells involved in viral and cancer immune surveillance. Indeed, reprogramming of the metabolic status of immune cells in general and particularly of CD4+ T lymphocytes, such as that engendered by the allostatic process [2], can and does have determinant regulatory outcomes on cellular immune surveillance, and immunity in general [11].

## 2.2 AIDS, Neuro-AIDS and Related Pathologies

Depending on the metabolic status of the HIV-seropositive patient, it may take many months to several years for AIDS to develop. AIDS is a syndrome consisting of a myriad of primarily immune disorders, and manifests in the most advanced stages of HIV infection. AIDS may also manifest in the nervous system and can be associated with a wide range of severe neurological disorders (Neuro-AIDS) [7, 12].

AIDS-related disorders of the nervous system may be caused directly by HIV or by conditions related to the state of increased immune deficiency that follows HIV infections, which makes HIV-seropositive patients more prone to certain cancers and opportunistic infections. These and other Neuro-AIDS disorders of unknown etiology are influenced by the inflammatory responses that result from HIV infection, but are not caused directly by the virus [7, 12].

Patients with Neuro-AIDS show inflammation of the central nervous system (CNS) as well as the peripheral nervous system (PNS) associated with HIV infection. AIDS-related *sequelae* may damage the brain and the spinal cord. Milder cognitive complaints in Neuro-AIDS are relatively common, and subsumed under the generic terminology of HIV-associated neurocognitive disorder (HAND). Neuro-psychologic testing can reveal these subtle deficits even in asymptomatic HIV-seropositive patients [12].

The AIDS dementia complex (ADC) and HIV-associated dementia (HAD) occur primarily in patients with advanced HIV infection. ADC and HAD can manifest as encephalitis with behavioral changes and a gradual decline in cognitive functions, including loss in concentration, memory, and attention. Patients with ADC, but not HAD, show progressive loss of motor function, dexterity and coordination. ADC is lethal, if left untreated [12].

Neuro-AIDS is a serious complication of AIDS that afflicts close to 50% of AIDS patients in the US and its initial symptoms are observed as progressive and intense peripheral neuropathy - PNS involvement in conjunction with CNS involvement (mental confusion, and cognitive impairments from forgetfulness, to behavioral and personality changes, and chronic headaches). Inflammatory processes consequential to HIV infection contribute to the patient's deteriorating condition by altering the size of certain brain structures involved in learning and information processing significantly [12].

Nervous system damage in AIDS patients can also derive from the drugs used to prevent or block HIV infection and propagation, to dampen or hinder the advancement of AIDS, and to blunt the progression of neuro-AIDS. Nervous system complications that result from therapy side-effects are usually mild at onset, but progressively intensify: they involve both the CNS and the PNS, and may include more acute neuropathy and pain perception, seizures, shingles, spinal cord problems, lack of coordination, difficult or painful swallowing, anxiety disorders, depression, fever, vision loss, gait disorders, destruction of brain tissue, sleep disorders, temporomandibular disorders, movement disorders, and coma [12].

In HIV-seropositive children, neuro-AIDS presents significant neurological complications and developmental delays in motor, cognitive and psychological functions. In addition, loss of previously achieved milestones are often observed in pediatric neuro-AIDS in association with brain lesions, neuropathic pain, smaller than normal skull size and brain volume, slow growth, eye coordination problems, recurring bacterial infections [12].

### **2.3 HIV/AIDS-Related CNS Pathologies**

A wide spectrum of CNS pathologies that are generally not considered neuro-AIDS *per se*, but that may show increased relative prevalence among HIV-seropositive patients include CNS lymphomas which are life-threatening cancerous tumors of the brain that either begin in the brain or result from peripheral metastases. CNS lymphomas are almost always associated with infection with the Epstein-Barr virus (human herpesvirus 4, HHV-4), which tends to be more prevalent among immune depressed patients [13]. Patients with CNS lymphomas typically complain of headaches, seizures, vision problems, dizziness, speech disturbances, paralysis, and mental deterioration. Prognosis is poor due to advanced and increasing immunodeficiency, and to the fact that this condition most often manifests as multiple CNS lymphoma foci [12].

Cytomegalovirus (CMV; human herpesvirus 5, HHV-5) infections can occur concurrently with HIV and other viral infections because of the generalized state of immune depression. CMV encephalitis manifests as weakness in the arms and legs, problems with hearing and balance, altered mental states, dementia, peripheral neuropathy, coma, and retinal disease that may lead to blindness. CMV infection of the spinal cord and nerves results in weakness in the lower limbs and some paralysis, severe lower back pain, and loss of bladder function. CMV can also cause pneumonia and gastrointestinal disease [12, 14].

The herpes zoster virus (Varicella-zoster virus, human herpesvirus type 3, HHV-3), which causes chickenpox and shingles, can also infect the brain and produce encephalitis and myelitis; that is, inflammation of the spinal cord. Shingles manifests as an eruption of intensely painful blisters along an area of skin supplied by an infected nerve. The virus may lay dormant in the nerve tissue for years until certain conditions, including immune depression, allostatic [2] and metabolic reprogramming [11], reactivate the virus' expression. HHV-3 reactivation and neuropathic shingles is common in immune suppressed patients, including HIV/AIDS [12, 15].

HIV-seropositive patients have a plethora of signs of immune suppression, which often include a greater susceptibility to fungal infections. The fungus *Cryptococcus neoformans*, an encapsulated yeast fungal organism that causes more fulminant disease in immunocompromised individuals, is commonly found in dirt and bird and rodent droppings. The fungus first invades the lungs and rapidly spreads to the

covering of the brain and spinal cord, causing meningeal inflammation. Cryptococcal meningitis is seen in about 10% of untreated individuals with HIV/AIDS, and manifests as increased fatigue, fever, headaches, nausea, memory loss, confusion, drowsiness, and vomiting. If left untreated, patients with cryptococcal meningitis may lapse into a coma and die [12, 16].

Other opportunistic conditions of the CNS and of the PNS further complicate the clinical outlook of HIV-seropositive patients, with clinically relevant signs of the immune suppression. They can include the following:

- neurosyphilis, the neurological manifestation of the sexually transmitted infection by the bacterium *Treponema pallidum* bacterium, subspecies *pallidum*, which causes syphilis,
- progressive multifocal leukoencephalopathy (PML), caused by the John Cunningham (JC) virus (human polyomavirus, papovavirus),
- toxoplasma encephalitis, resulting from infection with *Toxoplasma gondii*, the obligate intracellular, parasitic alveolate that causes toxoplasmosis,
- as well as several related infectious conditions [12].

Progressive multifocal leukoencephalopathy, as well as the other opportunistic infections of the CNS in HIV/AIDS, and neuro-AIDS proper can be relentlessly progressive in certain HIV-seropositive patients, or rarely affect other patients. Certain HIV/AIDS patients treated for immune reconstitution with highly active antiretroviral therapy (HAART) can experience a significant slowing down of the aggressive progression of HIV-associated CNS and PNS pathologies [12].

The variables that determine which patients will most benefit from ART/HAART-mediated immune reconstitution remain to be fully clarified. Data indicate that HIV/AIDS patients with low CD4 cell count, and HIV/AIDS patients whose CD4 count recovery show a sharp slope (suggesting a particularly fast immune reconstitution), are at greater risk of developing IRIS, which presents as a significant generalized syndrome characterized by a spectrum of acute clinical inflammation.

Certain macro-environmental variables, including nutrition, sleep, and heat stress can act in concordance to alter the patient's microenvironmental physiological balance (i.e., allostasis, [2]), epigenetic regulation including chromatin assembly, repair and remodeling (CARR), RNA interfering signaling complex (RISC) and molecular cartography [7], and significantly alter both the patient's metabolic profile in general [11] and immune allostasiome [2]. Together, these cellular and nuclear events and processes reprogram the metabolic status [11] of the white blood cells responsible for immune surveillance to these pathogens.

Hence, the processes that the patient's physiological systems undergo to regain homeostasis (i.e., allostasis) following external stressors – including heat stress and other climatic environmental factors – may, in many cases and in certain patients, hamper cellular immune surveillance to HIV and the HIV/AIDS opportunistic pathogens, and contribute to the increased risk of IRIS [4, 6], neuro-AIDS and other opportunistic conditions of the CNS and of the PNS in HIV-seropositive patients.

### 3 From Climate Change to Climate Crisis to an Anticipated Climate Catastrophe: Associated Medical Emergencies

#### 3.1 *From Climate Change to Climate Crisis to Climate Catastrophe*

Climate patterns describe cyclical variation in several meteorological variables, including temperature, humidity, atmospheric pressure, sea currents, polar ice sheet, eternal snow packs, wind and precipitation, atmospheric particle count ( $O_2$  content,  $CO_2$  concentrations, methane and other toxic gases and particulates) in various regions of the globe over long periods of time. Climate patterns are distinct from the weather, which simply describes the short-term conditions of these variables in any given region.

The climate of any specific region of the globe is determined by that region's climate system, which consists of three principal domains: the air, the earth, and the waters. Scientists describe these three determinants of climate as being composed of five components:

- atmosphere (i.e., layers of  $O_2$ ,  $CO_2$  and other breathable gases),
- hydrosphere (i.e., the surrounding mass of water – sea or fresh, drinkable or toxic),
- cryosphere (i.e., the proportion of surrounding water that is in ice form),
- lithosphere (i.e., the solid mass in that region: ground, plains, mountains, sand, rocks), and
- biosphere (i.e., the extent to which that region can sustain life).

Significant variations in one or more of these components have been recorded across several regions of the planet (e.g., increased global temperature: global warming) in the last decades. Taken together, these alterations constitute an important change in the earth's climate, which are above and beyond expectations due to causes such as processes internal to the earth (e.g., volcanic activity), or external forces (e.g., variations in sunlight intensity). Climate variability subsumed the entire set of those variations, proving them beyond doubt to be of non-human origins.

The atmosphere is a thin layer of gases that surrounds the earth. Gravity keeps the gases from drifting away from the planet into outer space. The lower layer of the atmosphere is rich enough in oxygen ( $O_2$ ) for plants and animals to survive. A substantial portion of the  $CO_2$  released by the planet – in greater proportions from human activities beginning during the industrial revolution in the mid-1800s and increasing substantially after WWII, with levels continuing to rise in the third quarter of the twentieth century – is trapped by the atmosphere, and creates the greenhouse effect that is responsible for increasing the planet's temperature.

There are two types of earth's crust:

- the continental crust contains the continents and rocks of lighter density, and
- the oceanic crust contains dense rocks from the upper mantle that are rich in iron minerals.

Both crusts are in constant motion, as a complex function of both the inner and outer temperature of the planet. The top soil that covers the continental crust absorbs a substantial proportion of the atmospheric gases, including CO<sub>2</sub> and methane. Top soil gases become entrapped in polar ice caps, which are then released into the atmosphere as the polar ice and permafrost melt consequential to the rise in global temperature. Oceans are more complex and diverse systems, compared to the earth's crust and its atmosphere. There are four major oceanic zones where plants and animals live, which together contain the largest ecosystem on earth.

- Intertidal zone: the area of the seafloor between high tide and low tide, which signifies the bridge between land and water. Tide pools, estuaries, mangrove swamps and rocky coastal areas are examples of the intertidal zone. Much of the superficial water pollution (i.e., oil, floating debris, plastics) is observable in the intertidal zone.
- Neritic zone: the waters found above the continental, which include coral reefs, underwater forests of kelp, and meadows of sea grass that house tiny fish, green turtles, sea cows, seahorses and shrimp. Excessive gaseous pollutants, such as CO<sub>2</sub> and methane, are found in elevated concentrations in this oceanic zone and, together with increased temperature of these waters, are responsible for the destruction of much of the water's flora and fauna.
- Open ocean zone: the water body that lies beyond the continental slope and contains close to 70–75% of the oceanic waters. This zone is further divided into four subzones.
- The sunlit subzone is where photosynthesis takes place. Plankton, jellyfish and most animals living in the open ocean inhabit the sunlit zone. Giant oil spills, immense plastic and garbage patches, and increased CO<sub>2</sub> and temperature directly threaten this habitat.
- The twilight subzone is the layer of oceanic waters at a depth of 3000–4000 feet, where some light can still penetrate. Viperfish, firefly squid, deep water fish, bioluminescent jellyfish, and the chambered nautilus live in these waters.
- The midnight subzone is the deepest layer of oceanic waters, and extends into darkness to the seafloor. The luminosity of the waters, which ought to be minimal, is altered by temperature and gaseous rise which endangers this unique habitat.
- The benthic subzone or “the seafloor” is not a layer of oceanic waters *per se*, but its properties are nevertheless key to the overall health of the ocean, and it is threatened by the CO<sub>2</sub> absorbed in the oceanic waters. The seafloor houses close to a quarter of a million species of plants and animals, which do not need sunlight to exist and which survive at relatively cold temperature and strong pressure. Hydrothermal vents provide an exquisite variety of flora and fauna in this abyss. A healthy seafloor is chalky white, largely made up of calcite (CaCO<sub>3</sub>), one among the most stable polymorphs of calcium carbonate. It formed from the skeletons and shells of planktonic organisms and corals. CaCO<sub>3</sub> has an important role in stabilizing the seas because it neutralizes CO<sub>2</sub> acidity of the seawaters. Large unhealthy murky brown patches of calcite in the North Atlantic and the

southern oceans have been observed, as benthic  $\text{CaCO}_3$  is disappearing. At those sites,  $\text{CO}_2$  levels in the waters are dangerously elevated, rendering them more acidic and warmer. In those regions of benthic  $\text{CaCO}_3$  depletion, the oceanic flora and fauna die at an accelerated pace [22].

In brief, as  $\text{CO}_2$  is absorbed into the oceans, the acidity of the waters increase, which is toxic to corals and to most marine life [23]. As the planet's temperature increases, the ocean water temperature increases as well, which is deadly to most fish species, plankton and other marine life. As the planetary temperatures increase, not only do oceanic waters warm up, but their levels rise consequential to polar ice melting. It follows that atmospheric wind patterns (e.g., jet stream) change, and oceanic currents (e.g., gulf stream) are disrupted. The combined effect of increased water, land and air temperature, and increased  $\text{CO}_2$  acidity kills crops and fish, and negatively impacts human health [1–3].

What biologists and physiologists call the macro-environment, earth scientists call the climate: that is, the set of data and evidence that defines and characterizes global patterns of weather beyond decades, into centuries and millennia. It is timely and critical to develop, test and evaluate effective solutions to the climate crisis, because it is a serious existential threat to all flora and fauna of the planet, including humankind. These solutions must go beyond the current interventions, some of which we have briefly outlined above, and must be stringently evaluated for effectiveness in the same manner as we stringently evaluate clinical interventions for patients: our planet is the patient, and it deserves the same attentive deployment of research activities. Novel interventions must be directed not only at preventing further disruption of the planet's climate by blunting or blocking global warming, but more importantly restoring the overall balance of our environmental system [2].

### ***3.2 Multiplicity of Climate Solutions: Toward Translational Environmental Restoration (TER)***

Concerted research on the systemic complexity of climate change must inform policies directed at countering, salvaging and restoring healthy survival on our planet. There are three principal classes of policies that can be designed to counter, and possibly reverse the human-caused climate crisis.

- Firstly, solutions to the climate crisis may encompass determinate individual actions. Individual personal lifestyle changes and choices can make an important contribution in reducing society's overall carbon impact, and therefore help lower greenhouse gas emissions to safer levels. A voluntary drive to eliminate the burning of coal, oil and, eventually, natural gas, for example, or to use recyclable plastics can go a long way toward that end. Individuals may be advised or required to move closer to their place of work, or alternatively to engage in cycling, commuting or utilizing shared or public transportation. Additionally,

individuals should be encouraged to avoid using plastic straws and utensils. Unfortunately, these are indeed the trends observed in cities such as Tokyo, Manila, Los Angeles, New York, Paris and Rome, all notorious for people preferentially choosing to drive alone rather than to carpool, and use plastics rather than washable and reusable containers and the like. Nonetheless, many citizens of richer nations rely for their comfort on products from such fossilized materials, and the energy stored in such fuels, which are fundamental to the global economy. Many citizens of developing nations want and arguably deserve the same comforts.

- Secondly, solutions to the climate crisis depend on the community, local, national and international politics as well as institutions and organizations, international consortia, and cross-national treaties such as the ambitious Paris Agreement (signed by 195 of 196 nations globally, effective 4 November 2016). Concerted actions call for infrastructure upgrades in almost every city and in every country: buildings worldwide contribute over 33% of greenhouse gas emissions, and poor road conditions lower the fuel economy. Visionary politicians must understand that investing in infrastructure – buildings, roads, bridges, and the like – help cut greenhouse gas emissions and drive economic growth by generating new and profitable jobs. One drawback to this calculus is that the cement required for infrastructure rebuilding produces large volumes of CO<sub>2</sub>: in the US alone production of cement in 2005 liberated close to 51 million metric tons of CO<sub>2</sub>. Mining copper and other elements needed for electrical wiring also causes climate warming pollution. Therefore, it is timely and critical to invest in the development of new and improved cement production techniques and mining. Nuclear power could replace fossil power for this and other demands, but it would certainly not be as safe and environmentally sound and clean as wind, sea current, or solar energy. Coal still supplies nearly 50% of the electricity demands in the US on average, and no alternative can reliably reduce the dependence on fossil fuels in the US as of current. Fossil fuels accounted for 81% of the world's energy consumption in 1987: 30 years later the mark is still 81%. Some states such as California have made the laudable commitment to be 100% fossil fuel-independent by 2045 (cf., SB-100), and certain countries, such as Sweden have met their 2030 green energy target as early as 2018. In other words: it is possible! There is hope! As arduous as the process will be, the tipping point is in sight although we are still far from the goal in large part because we have not yet been able to establish systematically which of the solutions now at hand is the most effective.
- Thirdly, besides energy requirements, our planet is afflicted by serious pollution. Avoiding or minimizing pollution in the first place must be an individual as well as a societal concern, because the clean-up is expensive and energy-consuming as well. Currently some solutions are operational and others are still in the discussion, planning or development stages. To cite only one example, plastic pollution is a most serious threat to our land masses, as it is to our oceans and fresh waters. Polystyrene and expanded polystyrene foam (i.e., Styrofoam) are plastics made from styrene and benzene, two petroleum based chemicals that are

classified as ‘probably carcinogenic’ for humans. Both chemicals have a very long half-life of biodegradability. Across the globe, human populations are surrounded by plastics, and plastics are in particularly heavy use in facilitating medical care. Despite their usefulness, plastics are serious environmental hazards that contribute both to the medical emergencies consequential to the climate crisis and to the pollution and thus endangerment of our water supplies. One of the more arduous problems of fighting plastic pollution is its slow biodegradability. Nonetheless, plastics can be collected, ground, heated, processed chemically to link the diverse organic substrata, and molded into second-use plastics. At present, this process is still rather inefficient and expensive but it is one that greatly exceeds the production of plastics from fracking byproducts. It is timely and urgent that societies around the world invest in scientific endeavors to develop new, improved and more economical means to collect, process and recycle plastics. Critical situations call for immediate intervention, like the Great Pacific Plastic Garbage Patch, a gyre of close to 2 trillion pieces of plastic refuse floating between Hawaii and California in the central North Pacific Ocean that is estimated to weigh between 75 and 100 thousand metric tons. However, presently, current solutions still meet serious feasibility challenges, including wanton individual, societal and political obstructionism and denial.

- Individual citizens could elect to avoid plastic: polystyrene products are presently found everywhere, from coffee cups and lids to straws, cutlery, plastic plates, Ziploc and plastic bags, buckets and plastic containers for food and plants. Nonetheless, individuals can abide by the very simple mindful resolution of minimizing their use of first-use plastic or avoiding it altogether; or, at the very least limiting use of plastics to reusable and recycled forms. At the local, national and international levels, interventions can be multi-faceted. Non-profit organizations and advocacy go a long way in fighting the scourge of plastics in our times. Case in point, the work of the ‘5Gyres’ ([5gyres.org](http://5gyres.org)): its *Asia Pacific Action Against Plastic Pollution* project aims to reduce plastic pollution in the Philippines, Indonesia and across Southeast Asia. At last count, it had reached close to 10,000 people in about 150 villages with the purpose of scaling up the ground efforts to implement zero waste strategies by preventing 14,000 tons of plastic waste from entering waterways and oceans annually. The project proffers a multi-dimensional model with strong emphasis on communications, training, and exchanges to share best practices, as well as including the creation of financial instruments and the establishment of programs that facilitate the development of new and lasting zero-waste materials recovery infrastructure to ensure sustainability and ongoing impact. To be clear, intervention models such as these will meet their maximal efficiency only when they are empowered by local, national and international political will, determination, policies and budget.

In brief, the natural gas boom of the last decades has made plastic feedstocks cheap and readily available. An estimated \$50 billion will be invested into new and expanded plastic production facilities in the US alone, tripling the amount of new plastic exports by 2030! That project will require 400 new plastic processing

facilities, in addition to new plastic manufacturing facilities and new plastic additive processing facilities. The project will also generate new gargantuan volumes of some of the most significantly harmful chemicals to human health, and to the safety of our planet, including phthalates and brominated flame retardants. Polyethylene production alone is expected to increase by as much as 75% by 2022.

In other words, should individual demand for plastic cups, bags, straws, cutlery, and other items not continue to increase so dramatically, and should the plastic recycling industry be contemporaneously improved, pollution of the planet by plastics could be curbed, and the surge in non-biodegradable plastic solid waste may be arrested and perhaps even reversed in the next decade.

The solutions that are now at hand to our current plastics crisis – a significant contributor to the climate crisis we experience – need to urgently be more effectively coordinated at the level of the individual consumer, community awareness, and national and international interests. Concerted scientific effort must be directed at obtaining evidence-based interventions to the climate crisis.

One model of this approach is provided by the extensive study of climate change in Tasmania supported by the Australian Government (Department of Climate Change and Energy Efficiency) and the National Climate Change Adaptation Research Facility in 2013. Several domains of the system of climate change in that region of the continent were examined by means of stringent research synthesis design. Meta-analysis of the systematic review findings established that significant gaps of knowledge remain in our understanding of adaptation needs and successes across all areas of the system under study, including marine life, land use, infrastructure, business, stakeholders and policy makers. Several research priorities were identified as potential solutions, which, based on the best available evidence, must address a need for better climate system understanding and improved evaluation of adaptation options. Stakeholder feedback analysis evinced a myriad of societal issues around adaptation to the climate crisis, including threats to public health and emerging climate change-related health emergencies, including cancers, infectious vector-borne diseases (e.g., Zika, Ebola), and worsening conditions in the chronically ill, including HIV-seropositive, AIDS, Neuro-AIDS and IRIS patients. Recommendations for future evidence-based research were proffered in several domains, particularly with respect to engaging stakeholders and end-users at the earliest stages of research design, including the refinement of research objectives and questions [24–27].

Taken together, the salient elements of the current climate crisis and the plethora of solutions that often seem uncoordinated and under-researched, and consequently fail to gather sufficient political support and budget, led us to propose a Translational Environmental Restoration (TER) paradigm that integrates the model proposed by McDonald and collaborators [28], and follows the framework of translational health care [19, 20]. In brief, the novel science of comparative effectiveness health care emerged as a response to the need to find the best available research evidence from fundamental basic biomedical research, to incorporate the best evidence base into novel clinical interventions, and to evaluate them for effectiveness. Two sides of the same coin soon emerged to ensure optimal patient-centered care [19, 20].

- translational research, for basic research on clinical biopsies obtained from the patient to suggest to the clinician treatments that are targeted to that patient, and
- translational effectiveness, for direct head-to-head comparison of clinical protocols for efficacy and effectiveness.

In translational research, new and improved physiologic, histologic, cellular, biochemical and molecular methodologies were tested and deployed, which now provide valid and reliable data to characterize the pathologic processes in patient biopsies. Based on these characteristics, a patient and pathology profile can be drawn by the clinician, who can, at this stage, entertain alternative treatment modalities. Based on these criteria, a clinical research question is generated that defines and characterizes the patient (P), the preferred intervention (I) and alternative comparator (C) interventions, the clinical outcome (O) sought, within the projected timeline (T) and in the available clinical setting (S). The resulting PICOTS question, and the analytical framework that is derived from it in consultation with stakeholders, informs the hypothesis-driven process of the systematic review for translational effectiveness. This research component of translational healthcare utilizes the research synthesis design, psychometrically validated instruments, acceptable sampling analysis and meta-analysis to establish the best evidence base in support of that, among the projected intervention, which is preferable to the others on criteria of effectiveness. Translational health care culminates in the translation of the systematic review into a critical review, a statement of the best evidence base in clinician-friendly and in patient-friendly language, which can be disseminated by tele-care and other means, so that the best evidence base information becomes available to all clinicians globally. In brief, translational health care is, by its very nature and definition, patient-centered care, effectiveness-focused care, and evidence-based care [19, 20].

Here, we propose that specific public health issues and medical problems and emergencies consequential to the climate crisis must be investigated by a similar two-prong process to the study of clinical interventions. Our TER paradigm is designed to maximize effectiveness of treatment modalities for climate crisis-related medical emergencies, and, as importantly, to provide a structure to uncover the best available evidence in support of efficacious and effective solutions to curb and reverse the causes of global warming.

### ***3.3 Health Outcomes and Medical Emergencies Associated with the Climate Crisis***

The survival of all species is threatened by climate change. The global climate crisis we are experiencing is an existential risk for all prokaryotes and eukaryotes, including mammals and human beings. All organisms on this planet live and survive by adapting to the ever-changing demands of a complex system of interactions among atmospheric, earth crust, and oceanic variables. As noted above (cf., Sect. 1), a

substantial portion of the CO<sub>2</sub> released by the planet is trapped by the atmosphere and creates the greenhouse effect that is responsible for increasing the planet's temperature. Increased temperatures alter wind and oceanic current patterns, promote melting of polar ice, and release of more CO<sub>2</sub> and other gases from the frozen permafrost. The earth's crust and oceanic waters trap CO<sub>2</sub>, and become more acidic, which in turn affects crops and kills coral and fish. Petroleum byproducts, such as plastics, constitute a large proportion of the human activities responsible for climate change. Plastics biodegrade slowly, and become one of the predominant pollutants of the oceans.

The climate crisis brings forth gargantuan macro-environmental alterations that Earth's organisms are ill-equipped to face. It engenders medical emergencies that threatens our survival because it causes serious threats to the biological system. Together, the macro-environmental stressors of heat challenge, de-oxygenation and CO<sub>2</sub> pollution of potable and sea water sources with consequential poisoning of planktons and fish, air particulates and their impact on depressing cellular immune surveillance to a variety of blood and solid tumors across vertebrate species, worsening lung disease, emphysema and asthmatic conditions, and psycho-cognitive and sleep disturbances are major threats to health in our society. Taken together, climate data best available evidence base convincingly show the precipitating downward pattern of life on earth as a direct consequence of climate change [29–32].

Increasing temperatures alter ecosystem dynamics, making it easier for mosquitoes and other organisms to contact human populations and spread infectious disease. The most recent Intergovernmental Panel on Climate Change report (IPCC-2018), global climate change has established direct health impacts tied to changes in the frequency of extreme weather events including heat, drought and intense rain [3].

Just as serious is the outcome of the climate crisis on ocean and air current (e.g., jet stream, gulf stream), which pushes temperate climatic zone towards the pole. Consequently, the ice caps show dangerous melting patterns, and torrid humidity and heat spreading wider from the equatorial band. With the expansion of the tropical zones, mosquito-borne disease has spread wider and faster into heavily inhabited cities. Novel infectious diseases, including Ebola, Zika, Dengue and others pose new challenges to public health and health care in Western societies [2, 8, 17, 18]. When making the case for patient-centered care [19, 20], it behooves us to take into serious consideration the important role of the individual responses of each individual patient to the climatic macro-environmental stressors. To be clear, patient-centered translational health care will only be achieved when the specific physiologic profile of each patient as it responds to the demands of the climate crisis - the allostasiome [2], the reprogramming of the metabolic status [11] - will be systematically defined and characterized, and taken in account in the clinical treatment planning and delivery.

In brief, it behooves us to entertain the validity of climate solutions, such as those discussed above, from the perspective of systemic psychobiology [10, 21]. Our survival, and the survival of all prokaryotes and eukaryotes, including mammals, and ultimately our species, depends upon their ability to adapt to changes in their

microenvironmental milieu and to the challenges of their surrounding macro-environment. Our microenvironment is our physiology: the context within which our organs, tissues, and the cells that compose our bodies survive, thrive, grow and divide. It is a biological system made up of complex and finely controlled pathways, regulatory feed-back loops and delicate biochemical check-and-balances, which together modify and modulate the expression of our genes to the ultimate end of improving our adaptability to the demands of the surroundings. These epigenetic alterations, which can be short-lived (i.e., one cell division), or sustained for several cell multiplication cycles, are concerted sub-cellular changes, intended to ensure our survival, although they may precipitate cell death either by necrosis or by the programmed process of apoptosis. Epigenetic changes are fundamental alterations in the organism's molecular, biochemical, cellular and physiological balance, which we call homeostasis, in response to, and for adapting to disturbances in the complex systems of biologic processes and responses that constitute its microenvironment. The microenvironment system, which preserves us into health and prevents us from falling into 'dis-ease', is constantly challenged by alterations in the macro-environment system multidimensional set of factors (e.g., temperature, humidity, altitude) that surrounds us [2].

The climate crisis causes serious threats to biological system, above and beyond the aforementioned macro-environmental stressors of heat challenge, de-oxygenation and carbon dioxide pollution of potable and sea water sources with consequential poisoning of planktons and fish, air particulates and their impact on depressing cellular immune surveillance to a variety of blood and solid tumors across vertebrate species, worsening lung disease, emphysema and asthmatic conditions, and psycho-cognitive and sleep disturbances. As serious is the outcome of the climate crisis on ocean and air current (e.g., jet stream, gulf stream), which pushes temperate climatic zone towards the pole. Consequently, the ice caps show dangerous melting patterns, and torrid humidity and heat spreading wider from the equatorial band. With that expansion of the tropical zones, mosquito-borne disease spread wider and faster into heavily inhabited cities.

Case in point, The Democratic Republic of Congo is now facing the worst Ebola outbreak in the country's history. More than 200 people have died from the disease since August 2018 and almost 330 confirmed or probable cases have been reported. Two health workers died in one attack, according to the minister, while last month 11 civilians and one soldier were killed in Beni, a city of 800,000 people and the epicenter of the outbreak. This outbreak, the second of calendar year 2018, began in North Kivu province before spreading to Ituri province in the east of the country. More than one million refugees and internally displaced people are in North Kivu and Ituri, according to World Health Organization (WHO), and their movement through and out of the provinces is a potential risk factor for the spread of Ebola. This current outbreak is the tenth since 1976 that Ebola has struck that African country.

The current crisis in climate change causes diverse public health threats [30]. The incidence of certain infectious diseases has sharply increased because of recent natural disasters:

- Cholera and enterotoxigenic *Escherichia coli* in Bangladesh 1998 and 2004;
- *Salmonella enterica*, *Cryptosporidium parvum*, and hepatitis A and E in Indonesia 1992, 2002, and 2004;
- Norovirus, salmonella, V cholera following Hurricane Katrina in US 2003;
- Leptospirosis in Taiwan 2001, Brazil 1996, and Puerto Rico 2017 and 2018;
- Measles in the Philippines following the Pinatubo eruption, and in Pakistan following the 2005 earthquake;
- Meningitis in Pakistan following the 2005 earthquake;
- Vector-borne diseases: Malaria in numerous locations after flooding and earthquakes;
- Tetanus and mucormycosis following natural disasters and associated trauma in numerous locations;
- *Coccidioidomycosis* in California following dust storms triggered by earthquake-driven landslides

The Zika virus outbreak in Brazil helps to illustrate how Outbreaks of disease in the setting of climate change are multifactorial. The El Niño drove a drying and warming condition in northeastern Brazil, which is where Zika first appeared. The drying of rivers leaves standing pools of water where mosquitoes can more readily replicate, allowing them to flourish. As rivers dried, people began to bring water to their homes in buckets, another perfect breeding ground for the vector mosquitoes. As Zika virus has disproportionately affected people of lower economic status who live in poorer conditions, the public health awareness of outbreaks has been low. Immunocompromised persons and pregnant women were further disproportionately affected by the disease [27].

Air pollution across the globe has killed an estimated 600,000 children in 2016. People living in low- and middle-income countries disproportionately experience the burden of outdoor air pollution with 91% (of the 4.2 million premature deaths) occurring in low- and middle-income countries, and the greatest burden in the WHO South-East Asia and Western Pacific regions. The latest burden estimates reflect the very significant role air pollution plays in cardiovascular illness and death. More and more, evidence demonstrating the linkages between ambient air pollution and the cardiovascular disease risk is becoming available, including studies from highly polluted areas [31].

Our increased understanding of the human microbiome has brought insight into the role it plays in health and disease, including HIV infection. Studies have shown that the gut microbiome is less diverse in individuals with HIV infection than in non-infected control subjects. Efforts to modify the microbiome to bolster immune reconstitution in people with HIV infection have so far been unsuccessful. The vaginal microbiome affects risk of HIV acquisition, with *Lactobacillus* dominance being protective compared with vaginosis characterized by larger populations of *Gardnerella*. The vaginal microbiome might also affect efficacy of topical tenofovir disoproxil fumarate pre-exposure prophylaxis [32].

Table 1 summarizes certain among the most imminent threats to public health consequential to the current climate crisis. Shortfalls in world preparedness to

**Table 1** Medical emergencies associated with the current climate crisis

Pathology	Summary of Main Finding(s)
<b>Viral Diseases</b>	
Hepatitis	Heavy precipitation and sewage overflow cause local flooding, which promotes the spread of <b>dysentery</b> and <b>hepatitis</b> [33].
Zika	Drying of lakes and rivers leave stagnant pools of water that serve as a breeding ground for mosquitoes carrying <b>malaria</b> and <b>Zika</b> .
Dengue fever	The relationship between <b>malaria</b> outbreaks and the El Niño-Southern Oscillation cycle has been documented in South America [36].
Chikungunya	<b>Chikungunya</b> and <b>dengue</b> are now being reported within the southern United States, with Zika on the horizon [40].
Ebola	Malnutrition and crowded living conditions along with a particularly dry season due to rising temperatures provided a favorable environment for <b>Ebola</b> transmission resulting in a sizable outbreak in West Africa in 2014 [58].
Malaria	Biodiversity loss leads to greater vector-borne pathogen transmission, including those that cause <b>Lyme disease</b> and <b>West Nile</b> [33].
West Nile Virus	Changing temperatures affect the migration timing of wild fowls, which are raising a threat to North America and Europe by bringing <b>avian influenza</b> to the regions [48].
Tick-borne encephalitis	The incidence of <b>malaria</b> , <b>Zika</b> , and other mosquito-borne infections rises with warming temperatures and changing rainfall patterns [36].
Lyme Disease	
Avian influenza	
<b>Airborne diseases</b>	Low humidity, dry winds, substantial rainfall, and high levels of dust and particulate matter facilitate the development of <b>meningitis</b> when they penetrate the upper respiratory mucosae to enter the bloodstream and spinal meninges [50].
	>25% of HIV-TB dual infections occur in South Africa, where the current synergistic epidemic of HIV-TB is triggering the proliferation of new and increasingly drug-resistant strains of <b>tuberculosis</b> .
	Accelerated migration compounds pollution levels, decreasing air quality and making the urban environment more conducive to efficient <i>Myobacterium tuberculosis</i> aerosol transmission [43].
<b>G-I diseases</b>	
Cholera	Apparent links exist between temperature and rainfall events with higher amount of reports of gastrointestinal illnesses (G-I).
Salmonella	Increased rainfall and subsequent sewage overflow of contaminated water facilitates the spread of <b>cholera</b> and <b>typhoid</b> [33].
Norovirus	<b>Salmonella</b> and cholera bacteria proliferate rapidly at warmer temperatures [46].
	During an El Niño year when temperatures in Lima, Peru were 5 °C above normal, <b>diarrheal disease</b> -related hospitalization rates among children doubled [47].
	Since floods are associated with <b>Norovirus</b> outbreaks, the predicted increase of heavy rainfall events ascribed to climate change may lead to more reported cases [51].

(continued)

**Table 1** (continued)

Pathology	Summary of Main Finding(s)
<b>Fungal diseases</b> <i>Coccidioidomycosis</i> (Valley Fever) Mucormycosis	Incidence of valley fever has climbed in recent years due to longer dry seasons and more frequent windstorms that aerosolize the fungal spores responsible for the lung disease [40]. Mucormycosis is a <b>fungal infection</b> caused by the inhalation of or absorption of <i>mucormycetes</i> into the lungs and skin The mold spores are disseminated during climate-related natural disasters such as volcanic eruptions, tornadoes, and tsunamis Development of mucormycosis mainly affects the immunocompromised [54].
<b>Skin diseases</b> Atopic dermatitis (eczema) Cutaneous leishmaniasis Hand-foot-and-mouth disease	Pollen-induced allergic diseases including atopic dermatitis, various types of <b>skin cancer</b> , and <b>cutaneous leishmaniasis</b> are all significant health problems consequential to the climate crisis. Increasingly-infectious cycles of leishmaniasis can be attributed to drier weather, drought, elevated temperatures, and habitat fragmentation due to human activity. Movement of the <i>Leishmania</i> parasite rodent reservoir and sand fly vector are expected to bring the skin disease northward to the US-Canadian border by 2080. <b>Hand-foot-and-mouth disease</b> is a seasonal viral infection whose incidence correlates with increased temperatures and humidity [8].
<b>Pulmonary/ Respiratory and Lung Diseases</b> Asthma Acute rhinitis (Hay fever)	Global warming is correlated with longer pollen seasons and increased aero-allergen production [36]. Higher frequencies of pollen-induced <b>respiratory and allergic diseases</b> like asthma and hay fever can accordingly be expected. The prevalence of allergic and respiratory diseases and cancers will continue to rise in Europe (e.g., Mediterranean regions), America, and South-Asia due to hotter and more humid conditions [39, 49].
<b>Cardiovascular disease and related disorders</b>	Heatwave-related deaths have been increasingly reported in people living with pre-existing <b>cardiovascular diseases</b> and complications. At greatest risk are the elderly, who have a diminished capacity for thermal regulation [36]. Extreme heat and prolonged drought are instrumental to the initiation of wildfires, which have dramatically risen in frequency in the US, Russia, and Mediterranean [52]. Exposure to toxic wildfire smoke emissions (composed of particulate matter, ozone gas, and other harmful substances) is associated with escalations in respiratory and cardiovascular-related hospital admissions [53].
<b>Neoplasms</b>	Chlorofluorocarbon (CFC) use has contributed to the depletion of the ozone layer, a phenomenon that can have significant impacts on human health Exposure to solar UVR is associated with growth of <b>skin cancers</b> including basal and squamous cell carcinomas and cutaneous melanomas. Evidence shows that smog contributes to respiratory carcinogenesis, thus the projected increase in air pollution levels will likely lead to the development of more <b>lung cancer</b> cases [49].

(continued)

**Table 1** (continued)

Pathology	Summary of Main Finding(s)
<b>Mental, psychological, developmental and neurocognitive disorders</b>	Warmer North Atlantic Ocean temperatures facilitate the uptake of pollutant mercury by fish, and consumption of the fish impairs <b>fetal neurocognitive development</b> , resulting in lifelong developmental issues [37]. Emotions including <b>anxiety</b> , despair, and uncertainty magnify within vulnerable individuals in response to climate-related natural disasters. The consequences from growing frequency and intensity of extreme weather events may aggravate the conditions of those living with <b>post-traumatic stress, anxiety, and depressive disorders</b> . High heat and humidity have been noted to increase hospital admissions for mood and <b>behavioral disorders</b> including <b>schizophrenia, mania</b> , neurotic and other <b>personality disorders</b> [59].
<b>Sleep disorders</b>	Poorer air quality is associated with <b>sleep-related breathing problems</b> . Higher daytime temperatures lead to persistently higher temperatures at night, making adequate amounts of sleep harder to attain for those who already suffer from sleep disorders such as insomnia. Diagnosis of insomnia occurs more frequently in those who have previously experienced wildfires or other disaster-related stressors [57].

respond to the stressors imposed by elevated temperatures leave some regions more vulnerable than others to the emergence and spread of infectious diseases. Excessive bouts of precipitation and the subsequent overflowing of sewage systems may promote the spread of cholera, typhoid, dysentery, and hepatitis [33].

Cholera is primarily a water-borne disease that is found in close association with algae and crustacea; increased sunlight and temperature promotes the blooming of phytoplankton, which both raises the water pH and provides a food source to favor growth of *V. cholera*. Epidemic periods of the disease in fact have been reported to occur during warmer seasons. Although improvements in sanitation and more abundant reservoirs of safe drinking water have made the disease rare in the modern world, some poorer urban areas have supplied an ideal environment for cholera due to overcrowding and lack of properly functioning sewage disposal systems [34]. In 2015, most cholera deaths were reported in Africa, but an increasing amount of cases have been reported in the United Kingdom and United States, as well as in Haiti following the catastrophic 2010 earthquake. It is likely that higher frequencies of reported cholera cases have been due to human activity, which has introduced strains of the *V. cholerae* bacteria to new regions from distant geographic sources [8]. Excessive rainfall and flooding then facilitate the entry of human and animal wastes into waterways, contaminating drinking water and potentiating related water-borne diseases [36]. One significant instance regarding contamination of potable water with *Cryptosporidium* due to heavy rains occurred in Milwaukee in 1993, resulting in an outbreak of over 403,000 cases of diarrheal disease [41].

While excessive flooding and rainfall is problematic, so is the disappearance of water caused by climate change. The drying of rivers and lakes resulting from hotter world temperatures has left behind stagnant pools of water, which serve as an ideal breeding ground for mosquitoes carrying malaria, Zika, and other viral diseases. For instance, changing weather patterns brought on by the 2015 El Niño fueled the outbreak of Zika virus in Brazil. El Niño is a naturally-occurring phenomenon that occurs periodically with varying intensity; when combined with pre-existing climate changes, a conducive breeding ground for the *Aedes aegyptii* mosquito vector is generated. The 2015 El Niño was in fact one of the strongest on record, instigating severe droughts in some areas and heavy rainfall in other areas, along with global temperature spikes. The El Niño therefore played a vital role in igniting the recent Zika outbreaks, exacerbated by factors including the vulnerability of the unexposed South American population, international travel to the region, the virulence of the strain itself, and co-infections with other viruses such as dengue [36].

Tropical species of *Anopheles* mosquitoes carrying malaria are more active at higher temperatures. These vectors require warmer climates to complete their life cycles, and the *Plasmodium* parasites that cause them to develop more rapidly at temperatures above 20 °C. If temperature conditions are ideal, one mosquito can infect up to 200 humans. Epidemics of malaria occur during rainy seasons in the tropics, as well as following annual weather events like the El Niño-Southern Oscillation. There is other evidence of the expansion of malaria in the African highlands in association with local warming, although many environmental factors play a role such as local geographic disturbances, short-term variability in climate, and prolonged climate trends [36]. On the other hand, epidemics of West Nile Virus occur during droughts, when mosquitoes are brought into closer contact with birds and humans. Natural predators of the mosquitoes are also reduced when wet areas dry out, contributing to spikes in cases of the virus [41].

It is predicted that climate change will compound the burden of other vector-borne diseases such as meningococcal meningitis, viral and tick-borne encephalitis, Lyme disease, and dengue and yellow fever. Cases of tick-borne (viral) encephalitis have reportedly already risen in Sweden in response to a succession of warmer winters occurring within the past two decades. The geographic range of ticks that transmit Lyme disease and viral encephalitis has extended northwards in the area, accompanying recent climate trends. The transmission of dengue fever specifically is affected by warmer temperatures, which shorten the time that mosquitoes carrying the virus become infectious, therefore increasing the probability of transmission of the disease to human hosts [36].

Studies in Europe and North America have shown a positive association between the occurrence of heatwaves and mortality in elderly people, especially older women, who have a reduced physiological capacity for thermoregulation. Other research has indicated that susceptible populations including the mentally ill, the financially disadvantaged, and young children who live in thermally-stressful environments or who have pre-existing illnesses are particularly vulnerable to the health effects associated with climate change. Most heatwave-related deaths occur in those individuals living with pre-existing respiratory or cardiovascular diseases, often

with a history of strokes and/or heart attacks. Those individuals living in urban environments with thermally-inefficient housing are at the greatest risk [36].

Extreme weather events related to the climate crisis such as heavy rainfall, flash flooding, heat waves, and cold spells give little time for human response and preparation, posing significant threats to the HIV/AIDS population [45]. Flooding can be detrimental to the health of HIV/AIDS patients in that flash floods facilitate the spread of water-borne diseases that can be perilous to the immunocompromised. Other climate change patterns including depleting air quality caused by high ozone, particulate matter and levels of various sulphur oxides ( $\text{SO}_x$ ) have a direct negative influence on human health, increasing the morbidity of respiratory and cardiovascular diseases [44]. Extended temperature inversions during the winter seasons may trap and concentrate pollutants closer to the earth's surface; rising temperatures and continued carbon dioxide emissions coupled with periods of drought will intensify this phenomenon. Inhalation of air pollutants and subsequent absorption into the bloodstream is associated with depressed immune functioning, which compromises lung operation and increases HIV/AIDs patients' vulnerability to lung infections like tuberculosis and pneumonia. Furthermore, predictions of receding water quality and availability consequently to climate change threatens food security, leading to increased rates of hunger and malnutrition in already-impoverished and water-scarce conditions. Soil drying and irregular rainfall caused by rising temperatures further limits soil quality and therefore the arability of lands used for agricultural production, crippling poorer farming communities. Those living in poverty are hence among the most vulnerable to the HIV epidemic; malnutrition compromises the immune system further [42].

Research on the impacts of climate change on allergic diseases has been somewhat neglected, but it is important to recognize that global warming indeed alters the timing and duration of the pollen and spore seasons, as well as the geographic range of such aeroallergens. The pollen amount, allergenicity, and distribution have all been greatly affected by the climate emergency at hand, aggravating patients who suffer from allergic disorders like hay fever and asthma. Atmospheric variables that have impacts on the dispersion of aeroallergens like pollen and mold spores include elevated carbon dioxide levels and temperatures, which have together been associated with higher pollen concentrations. Significantly stronger allergenicity has been found in pollen released by trees that grow exclusively at warmer temperatures. The spatial distribution of allergens is influenced by atmospheric conditions such as wind speed and direction, amount of rainfall, and humidity. Furthermore, it is estimated that many areas will have pollen seasons that begin earlier and last longer. *Gramineae* pollens are a major cause of seasonal allergic rhinitis (hay fever) in the Asia Minor, a Mediterranean region connecting Asia to Europe. The issue is mediated by worsening air pollution and altered local and regional pollen production [39]. Air pollution overcomes the mucosal barriers that then trigger an allergy-induced response, eliciting inflammation of the airways. Since allergic diseases are already significant health issues in developed countries including the UK, Australia, the US, and New Zealand, the potential for adverse impacts in this field of health is a serious and pressing concern [38].

The thermal environment is an essential determinant of sleep quality; even minuscule changes in ambient temperature can disturb the thermoregulatory processes involved in human sleep. New studies have found that higher temperatures are associated with overall lessened sleep quality in the United States. Most commonly, reported sleep-related complaints are among the elderly and low-income populations following climate-related disasters such as hurricanes, flash floods, and wildfires [57]. Other ramifications of the climate catastrophe to consider are the mental health consequences corresponding to stressors such as property and asset losses, displacement from flooded areas, and conflicts over dwindling natural resources. Unpredictable climatic conditions alter natural landscapes and disrupt agricultural and working conditions as well as habitable neighborhoods [59]. Social exclusion resulting from physical relocation following increased disasters is likely to fracture social connections while simultaneously cutting access off to services that are already deficient in impoverished communities [55]. Children are at elevated risk of climate-related health issues due to their immature physiology and sensitivity and their lifelong exposure to its impacts; disruptions have negative repercussions including developmental delays and other psychological effects [56]. In this regard, climate change essentially amplifies existing health threats by heightening financial and relationship tensions and inciting further sentiments of anguish, hopelessness, and distress. Taken together, these stressors threaten the most marginalized groups, including immune depressed patients with HIV/AIDS and Neuro-AIDS. Heat-related mental health morbidity tends to occur most often in persons with impaired thermoregulation, often those struggling with pre-existing mental health illnesses and those who take prescription medications. In that same vein, people who have substance abuse issues may also more vulnerable to the implications of the climate crisis [28].

In brief, the precarious and dangerous state of our climate yields a plethora of secondary effects, the magnitude of which remains to be fully evaluated. Thus, ongoing changes in our climate, including air and water pollution, rising temperatures, increased rate and intensity of natural disasters, bring along significant psycho-emotional and physiological stress to the population, which in turn alter individual responses [2, 11], and contribute to increasing both the relative prevalence and the incidence of certain pathologies. Cardiovascular diseases, pulmonary diseases, autoimmune diseases, metabolic diseases, skin diseases and cancers are but a few of the medical threats associated with, and directly consequential to our current climate crisis. Infectious diseases, including vector-borne (e.g., ticks, mosquitoes) viral diseases, are migrating from the tropical regions and rising sharply in more temperate latitudes. Rising temperatures also increasing pathogen and vector efficiency in spreading disease. The natural disasters associated with climate change lead to massive population displacement and living conditions conducive to the spread of infectious diseases [27].

## 4 Conclusion: Toward Solutions for More Efficiently Controlling Medical Emergencies Consequential to Climate Change

### 4.1 *Translational Health Care*

When making the case for patient-centered care [1] in the context of HIV-seropositive patients with AIDS, neuro-AIDS or IRIS as they experience the heat stress, air pollution stress and other conditions imposed by our climate crisis, it behooves us to take into serious consideration the important role of the individual allostatic responses of each individual patient to the climatic macro-environmental stressors discussed above. To be clear, patient-centered translational health care [19, 20] can only be achieved when the allostasiomic profile of each patient is successfully systematically defined and characterized [2], and taken in account of the clinical treatment planning and delivery, and individual patient data research and analysis [9, 10].

Climate change, as noted above, is a complex multi-dimensional systemic issue. Several solutions have emerged, but the research findings are scant with respect to which among those are the most effective. This paper proposed to tackle this problem in a similar manner to how complex systemic and multi-dimensional pathologies are addressed in health care: an equally complex and systemic societal problem. Translational health care, the interface between translational research and translational effectiveness, obtains the best evidence base for clinical intervention.

The same protocol can be engaged to uncover the best evidence base in support of the solutions designed and deployed to blunt, block or reverse the damages consequential to the climate crisis. The TER paradigm leads to the translation of best evidence base in climate change research into effective and efficacious policies for restorative renewal of our planet.

Conceptually, TER outcomes will contribute to a fuller interpretation and explanation of the allostasiomic profile [2] of each HIV-seropositive patient with AIDS, neuro-AIDS patient or IRIS patient, and yield a new and improved predictive model for immune resilience for these patients in the model we recently described [8].

### 4.2 *Translational Healthcare for AIDS and Neuro-AIDS in the Age of the Climate Catastrophe*

Climate change also undermines improvements in the management of existing disease outbreaks. This is the case in South Africa, a country where the government has become more aggressive in responding to the HIV/AIDS epidemic while establishing itself as a leader in the testing and treatment of HIV-positive individuals living through the climate crisis.

An estimated 36.7 million people worldwide were living with HIV/AIDS at the end of 2015, and about 2.1 million individuals became newly infected with HIV in the same year. Roughly two million patients receive treatment in South Africa each day, which means that more people are living with HIV there than in any other country. Thus, the government of South Africa has reduced the likelihood of patients progressing from HIV disease to AIDS, thereby extending the lives of many for years or decades [25].

Despite being plagued by plastics and other sources of global warming, South Africa has set a model for an effective approach of HIV management, as it addresses new and old challenges that arise from global climate change, not the least of which being ensuring that HIV/AIDS have access to fresh drinking water and nutritional food supplies [25].

Access to medications to control neuropathic pain, lethargy, depression, neural inflammation, psychotic disorders, dementia and other symptoms of neurological damage and cognitive decline in Neuro-AIDS patients is also as critical for the patients, as it is arduous for the clinical caregivers because of the increased intensity of flooding, drought, and other natural disasters consequential to the climate crisis South Africa [25]. Similarly, access to counseling and psychotherapy for HIV/AIDS patients and their relatives is rendered difficult, if not impossible. Last but not least, access to hospital-monitored delivery of aggressive antiretroviral therapy (ART, HAART) to prevent or treat AIDS and its dementia complex, vacuolar myopathy, progressive multifocal leukoencephalopathy, and cytomegalovirus encephalitis, and other pharmacological interventions, including antifungal or anti-malarial drugs to combat certain bacterial infections associated with the disorder, and penicillin to treat neurosyphilis, is often delayed or impeded by climate-related emergencies [25].

One major concern for people in this part of the world is less a shortage of antiretroviral drugs than a shortage of food and drinkable water. Climate crisis-related obstacles to getting care for HIV/AIDS patients can include:

- Difficulty in reaching critical areas of care due to climate-related variability and unpredictability;
- need to access fresh water and nutritional foods to maintain good health, not all of which are readily available to the patient populations;
- difficulty to ensure compliance in taking the prescribed drugs; and
- diverse diet among patients, which may help or hinder immunity, and the effectiveness of certain medications, thus precipitate secondary infections, mutations of HIV within the same patient, and exacerbation of the progression to AIDS and Neuro-AIDS.

In brief, managed HIV is survival, and this survival depends not just on access to antiretroviral drugs but also on a large spectrum of allied social and environmental resources that are necessary to meet the HIV/AIDS patients' health needs, and which are all coming under threat in our era of global climate crisis [27].

### **4.3 Conclusion: Toward A Bayesian Space-Time Modeling AI-Driven Algorithm for Predicting Climate Crisis Outcomes on Morbidity**

In conclusion, climate change is quickly progressing on our planet. The world is warming, and this warming reflects not just the usual planetary cycles of warming and cooling that have been present since the earth formed, but changes caused specifically by human activity. Left essentially unchecked for decades, climate change has now become a climate crisis. The October 2018 UN states in no uncertain terms that, unless drastic interventions are engaged presently, our global climate crisis is inexorably bound to become a climate catastrophe within the next 10–15 years [3, 27].

The changing global climate is producing increasingly unusual threats to global health relative to preindustrial conditions. In an absolute sense, changing climatic conditions constitute direct medical emergencies, which implicate mental health, immune suppression, increased susceptibility to vector-borne infectious diseases, and a wide variety of other serious patho-physiologies.

Proof-of-concept of the proposition that rise in global temperature affects morbidity and mortality to viral infection is proffered by the current measles outbreak in the Philippines. UNICEF-WHO (26 February 2019) indicates that 12,736 new measles cases and 203 deaths were officially reported through the routine surveillance system from the Philippines Department of Health (DoH) between 1 January and 23 February 2019, a rate 4–8 times higher than the corresponding period last year. The significant rise in morbidity mortality is attributed in part to the fact that, as of 23 February 2019, 63% of cases were not vaccinated, and to the putative synergistic effect of climate change (i.e., global warming, increased rains this Spring due to more extreme typhoons) on the infectivity of the measles virus (MeV, morbillivirus, single-stranded, negative-sense, enveloped, non-segmented RNA virus of the *Paramyxoviridae* family), and depressed immune surveillance of heat-stressed patients at risk.

Progress toward solutions to the impending dangers associated with climate change are generally hindered by a range of factors including expectations, memory limitations, and cognitive biases, which may be surmised as the ‘boiling frog’ denialism syndrome. The declining noteworthiness of historically extreme temperatures and associated medical emergencies, including a greater threat of HIV infection, AIDS, Neuro-AIDS and IRIS, is often not accompanied by a decline in the negative sentiment and psycho-social fear that they induce. In brief, social normalization of extreme conditions of global warming and related threats to human health globally, such as those that have been recently reported or that are anticipated (see discussion above) rather than adaptation to this global environmental, ecological and health crisis is setting a ‘new normal’, and consequently dangerously blunting best-practices targeted solution research and development [60].

Novel Artificial Intelligence (AI) algorithms could improve our ability to predict the effect of climate change on global ecological survival of animal and vegetal species, as well as, specifically on health and disease of human populations. A Bayesian statistical model was proposed [61], which, could be expanded and refined by integrating space-time latent models using mixture structures to open new AI avenues of health-targeted solutions to the current cataclysmic climate crisis.

The original model was successfully applied to the investigation of the effects of air pollution on asthma [61]. It is possible and even probable that further development of the algorithm by integrating a multiple regression multi-stage space-time mixture model Bayesian paradigm to study the effect of global climate crisis on infectious diseases, and other medical emergencies consequential to our present critical climate catastrophe, while including as well ‘dummy variables’ such as social normalization and deniability [60], will lay the foundations for AI-driven outcome predictions and interventions.

AI may initially find an important role in deploying measures to counter inherent analytical biases, such as, the temporal risk effects that can vary within the space-time domain. A subset of spatial areas can have a homogeneous temporal profile in risk, rendering global modeling inappropriate because of the restrictive assumption of common risk effects across all areas.

Adjustment of the Bayesian predictive model will adapt the space-time latent models with mixed structures to capture the heterogeneous temporal profiles of relative risks in space-time health data, as was done, on a much smaller scale, with ambulatory asthma county-level data previously [62].

## References

1. Resplandy L, Keeling RF, Eddebar Y, Brooks MK, Wang R, Bopp L, Long MC, Dunne JP, Koeve W, Oschlies A. Quantification of ocean heat uptake from changes in atmospheric O<sub>2</sub> and CO<sub>2</sub> composition. *Nature*. 2018;563:105–8. On Line Pub. 31 October 2018 [PMID: 30382201].
2. Chiappelli F. Bioinformation Informs the Allostasiome: Translational Environmental Restoration (TER) for the Climate Crisis Medical Emergency. *Bioinformation*. 2018;14:446–8. [PMID:30310252].
3. IPCC-2018. Global Warming ff 1.5 °C. 48th Session of the IPCC, Incheon, Republic of Korea, 6 October 2018. Accessed 1 Nov 2018. [http://report.ipcc.ch/sr15/pdf/sr15\\_spm\\_final.pdf](http://report.ipcc.ch/sr15/pdf/sr15_spm_final.pdf)
4. Khakshooy A, Chiappelli F. Hypothalamus-Pituitary-Adrenal cell-mediated immunity regulation in the Immune Restoration Inflammatory Syndrome. *Bioinformation*. 2016;12:28–31. [PMID: 27212842].
5. Dybul M, Fauci AS, Bartlett JG, Kaplan JE, Pau AK. Guidelines for using antiretroviral agents among HIV-infected adults and adolescents. *Ann Intern Med*. 2002;137:381–433. [PMID: 12617573].
6. Espinosa E, Ormsby CE, Vega-Barrientos RS, Ruiz-Cruz M, Moreno-Coutiño G, Peña-Jiménez A, Peralta-Prado AB, Cantoral-Díaz M, Romero-Rodríguez DP, Reyes-Terán G. Risk factors for immune reconstitution inflammatory syndrome under combination antiretroviral therapy can be aetiology-specific. *Int J STD AIDS*. 2010;21:573–9. [PMID: 20975091].
7. Chiappelli F, Shapshak P, Commins D, Singer E, Minagar A, Oluwadara O, Prolo P, Pellionisz AJ. Molecular epigenetics, chromatin, and NeuroAIDS/HIV: immunopathological implications. *Bioinformation*. 2008;3:47–52. [PMID:19052666].

8. Chiappelli F, Balenton N, Khakshooy A. Future Innovations in Viral Immune Surveillance: A Novel Place for Bioinformation and Artificial Intelligence in the Administration of Health Care. *Bioinformation*. 2018;14(5):201–5. [PMID:30108416].
9. Solomon GF. Psychoneuroimmunology: interactions between central nervous system and immune system. *J Neurosci Res*. 1987;18:1–9. [PMID: 3316677].
10. Chiappelli F, Abanomy A, Hodgson D, Mazey KA, Messadi DV, Mito RS, Nishimura I, Spigleman I. Chapter 64, Clinical, experimental and translational psychoneuroimmunology research models in oral biology and medicine. In: Ader R, et al., editors. *Psychoneuroimmunology, III*. San Diego: Academic Press; 2001. p. 645–70.
11. Kim J. Regulation of Immune Cell Functions by Metabolic Reprogramming. *J Immunol Res*. 2018;. On Line Pub. 13 Feb 2018: 8605471. [PMID: 29651445].
12. Neurological Complications of AIDS Fact Sheet: National Institute of Neurological Disorders and Stroke. Accessed 3 Oct 2018. [www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Neurological-Complications-AIDS-Fact-Sheet2/9](http://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Neurological-Complications-AIDS-Fact-Sheet2/9).
13. Brandsma D, Bromberg JEC. Primary CNS lymphoma in HIV infection. *Handb Clin Neurol*. 2018;152:177–86. [PMID:29604975].
14. Grønborg HL, Jespersen S, Hønqe BL, Jensen-Fangel S, Wejse C. Review of cytomegalovirus coinfection in HIV-infected individuals in Africa. *Rev Med Virol*. 2017. On Line Pub. 7 Oct. 2016. <https://doi.org/10.1002/rmv.1907.2016>. [PMID: 27714898].
15. Hung CH, Chang KH, Kuo HC, Huang CC, Liao MF, Tsai YT, Ro LS. Features of varicella zoster virus myelitis and dependence on immune status. *J Neurol Sci*. 2012;318:19–24. [PMID: 22564884].
16. Limper AH, Adenis A, Le T, Harrison TS. Fungal infections in HIV/AIDS. *Lancet Infect Dis*. 2017;17:e334–43. [PMID: 2877e4701].
17. Chiappelli F, Bakhordarian A, Thamess A, Du AM, Jan AL, Nahcivan M, Nguyen MT, Sama N, Manfrini E, Piva F, Rocha RM, Maida CA. Ebola: translational science considerations. *Translational Med*. 2015;13:11. <https://doi.org/10.1186/s12967-014-0362-3>. [PMID:25592846].
18. Chiappelli F, Santos SM, Caldeira Brant XM, Bakhordarian A, Thamess AD, Maida CA, Du AM, Jan AL, Nahcivan M, Nguyen MT, Sama N. Viral immune evasion in dengue: toward evidence-based revisions of clinical practice guidelines. *Bioinformation*. 2014;10:726–33. [PMID:25670874].
19. Chiappelli F. Fundamentals of Evidence-based Health Care and Translational Science. Heidelberg: Springer–Verlag; 2014.
20. Chiappelli F. Comparative Effectiveness Research (CER): New Methods, Challenges and Health Implications. Hauppauge: NovaScience Publisher, Inc.; 2016.
21. Chiappelli F. Advances in Psychophysiology Research. Hauppauge: NovaScience Publisher, Inc.; 2018.
22. Sulpis O, Boudreau BP, Mucci A, Jenkins C, Trossman DS, Arbic BK, Key RM. Current  $\text{CaCO}_3$  dissolution at the seafloor caused by anthropogenic  $\text{CO}_2$ . *Proc Natl Acad Sci USA*. 115:11700–5. On Line Pub. 29 Oct. 2018 [PMID:30373837].
23. National Ocean Service. What is Coral bleaching? 25 June 2018. Accessed 13 Nov 2018. [https://oceanservice.noaa.gov/facts/coral\\_bleach.html](https://oceanservice.noaa.gov/facts/coral_bleach.html)
24. Levins R, Richard Lewontin R. The Dialectical Biologist. Cambridge, MA: Harvard University Press; 1985.
25. National Institute of Neurological Disorders and Stroke Neurological Complications of AIDS Fact Sheet. Report 10/3/2018. Accessed 12 Nov 2018. <https://www.ninds.nih.gov/Disorders/Patient-Caregiver-Education/Fact-Sheets/Neurological-Complications-AIDS-Fact-Sheet5/9>.
26. Nogueira M, Da Silva Marinho RV, Harumi Narumiya I, Bach Q, Kasar V, Khorshad D, Chiappelli F. Chapter 20, Comparative Effectiveness Research in the Pharmacological Treatment of HIV/AIDS – The Immune Reconstitution Inflammatory Syndrome (IRIS). In: Chiappelli F, editor. Comparative Effectiveness Research (CER): New Methods, Challenges and Health Implications. Hauppauge: NovaScience Publisher, Inc.; 2016.
27. Schooley RT. Our Warming Planet: Is the HIV-1-Infected Population in the Crosshairs. *Top Antivir Med*. 2016;26:67–70. [PMID:29906791].

28. McDonald J, Harkin J, Harwood A, Hobday A, Lyth A, Meinke H. Supporting evidence-based adaptation decision-making in Tasmania: A synthesis of climate change adaptation research. Gold Coast: National Climate Change Adaptation Research Facility; 2013. p. 169.
29. Smith DL, Dushoff J, McKenzie FE. The risk of a mosquito-borne infection in a heterogeneous environment. *PLoS Biol.* 2004;2:e368. OnLine Pub. 26 Oct. 2004 [PMID:15510228].
30. McMichael A. *Climate Change and the Health of Nations: Famines, Fevers, and the Fate of Populations*. 1st ed. New York, NY: Oxford University Press; 2017.
31. World Health Organization. Ambient (outdoor) air quality and health. 2 May 2018. Accessed 13 Nov 2018. [http://www.who.int/news-room/fact-sheets/detail/ambient-\(outdoor\)-air-quality-and-health](http://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health).
32. Schooley RT. The human microbiome: implications for health and disease, including HIV infection. *Top Antivir Med.* 2018;26:75–8. [PMID:30384329].
33. Liang L, Gong P. Climate change and human infectious diseases: A synthesis of research findings from global and spatio-temporal perspectives. *Ann Rev Virol.* 2016;3(1):125–45. [PMID:27482902].
34. Chowdhury FR, Nur Z, Hassan N, von Seidlein L, Dunachie S. Pandemics, pathogenicity and changing molecular epidemiology of cholera in the era of global warming. *Ann Clin Microbiol Antimicrob.* 2017;16(1):10. [PMID:28270154].
35. Chin CS, Sorenson J, Harris JB, Robins WP, Charles RC, Jean-Charles RR, Bullard J, Webster DR, Kasarskis A, Peluso P, Paxinos EE, Yamaichi Y, Calderwood SB, Mekalanos JJ, Schadt EE, Waldor MK. The origin of the Haitian cholera outbreak strain. *N Engl J Med.* 2011;364(1):33–42. [PMID:21142692].
36. McMichael AJ, Woodruff RE, Hales S. Climate change and human health: present and future risks. *Lancet.* 2006;367(9513):859–69. [PMID:16530580].
37. Booth S, Zeller D. Mercury, Food Webs, and Marine Mammals: Implications of Diet and Climate Change for Human Health. *Environ Health Perspect.* 2005;113(5):521–6. [PMID:15866757].
38. Beggs PJ. Impacts of climate change on Aeroallergens: past and future. *Clinical and experimental allergy.* 2004(10):1507–13. [PMID:15479264].
39. Bajin M, Cingi C, Oghan F, Gurbuz M. Global warming and allergy in Asia Minor. *Eur Arch Otorhinolaryngol.* 2013;270(1):27–31. [PMID:22695877].
40. Kaffenberger B, Shetlar D, Norton S, Rosenbach M. The effect of climate change on skin disease in North America. *J Am Acad Dermatol.* 2017;76(1):140–7. [PMID:27742170].
41. Shuman EK. Global climate change and infectious diseases. *Int J Occup Environ Med.* 2011;2(1):11–9. [PMID:23022814].
42. Abayomi A, Cowan MN. The HIV/AIDS epidemic in South Africa: Convergence with tuberculosis, socioecological vulnerability, and climate change patterns. *S Afr Med J.* 2014;104(8):583. [PMID:26307805].
43. Cohen J. Reversals of misfortunes. *Science.* 2013;339(6122):898–903. [PMID:23430629].
44. Pope CA, Burnett RT, Thun MJ, Calle EE, Krewski D, Ito K, Thurston GD. Lung cancer, cardiopulmonary mortality, and long-term exposure to fine particulate air pollution. *JAMA.* 2002;287(9):1132–41. [PMID:11879110].
45. Abayomi A. HIV/AIDS disease burden complex in South Africa: Impact on health and environmental resources, and vulnerability to climate. *Climate Vulnerability.* 2013;1:125–43.
46. Akil L, Ahmad H, Reddy R. Effects of climate change on *Salmonella* infections. *Foodborne Pathog Dis.* 2014;11(12):974–80. [PMID:25496072].
47. Checkley W, Epstein LD, Gilman RH, Figueroa D, Cama RI, Patz JA, Black RE. Effect of El Niño and ambient temperature on hospital admissions for diarrhoeal diseases in Peruvian children. *Lancet.* 2000;355(9202):442–50. [PMID:10841124].
48. Gilbert M, Slingenbergh J, Xiao X. Climate change and avian influenza. *Rev Sci Tech.* 2008;27(2):459–66. [PMID:18819672].
49. Longstreth J. Anticipated public health consequences of global climate change. *Environ Health Perspect.* 1991;96:139–44. [PMID:1820256].

50. Palmgren H. Meningococcal disease and climate. *Glob Health Action.* 2009;2 <https://doi.org/10.3402/gha.v2i0.2061>. [PMID:20052424].
51. Semenza JC, Herbst S, Rechenburg A, Suk JE, Höser C, Schreiber C, Kistemann T. Climate Change Impact Assessment of Food- and Waterborne Diseases. *Crit Rev Environ Sci Technol.* 2011;42(8):857–90. [PMID:24808720].
52. McMichael AJ, Lindgren E. Climate change: present and future risks to health, and necessary responses. *J Intern Med.* 2011;270(5):401–13. [PMID:21682780].
53. Delfino RJ, Brummel S, Wu J, Stern H, Ostro B, Lipsett M, Winer A, Street DH, Zhang L, Tjoa T, Gillen DL. The relationship of respiratory and cardiovascular hospital admissions to the Southern California wildfires of 2003. *Occup Environ Med.* 2009;66(3):189–97. [PMID:19017694].
54. Benedict K, Park BJ. Invasive fungal infections after natural disasters. *Emerg Infect Dis.* 2014;20(3):349–55. [PMID:24565446].
55. Fritze J, Blashki G, Burke S, Wiseman J. Hope, despair and transformation: Climate change and the promotion of mental health and wellbeing. *Int J Ment Health Syst.* 2008;2(1):13. <https://doi.org/10.1186/1752-4458-2-13>. [PMID:18799005].
56. Burke S, Sanson A, Van Hoorn J. The Psychological Effects of Climate Change on Children. *Curr Psychiatry Rep.* 2018;20(5):35. <https://doi.org/10.1007/s11920-018-0896-9>. [PMID:29637319].
57. Rifkin D, Long M, Perry M. Climate change and sleep: A systematic review of the literature and conceptual framework. *Sleep Med Rev.* 2018;42:3–9. [PMID:30177247].
58. Gislason M. Climate change, health and infectious disease. *Virulence.* 2015;6(6):539–42. [PMID:26132053].
59. Hayes K, Blashki G, Wiseman J, Burke S, Reifels L. Climate change and mental health: risks, impacts and priority actions. *Int J Ment Health Syst.* 2018;12:28. <https://doi.org/10.1186/s13033-018-0210-6>. [PMID:29881451].
60. Moore FC, Obradovich N, Lehner F, Baylis P. Rapidly declining remarkableability of temperature anomalies may obscure public perception of climate change. *Proc Natl Acad Sci USA.* 2019;116:4905–10.. 2019 Epub Feb 25. [PMID:30804179].
61. Lawson AB, Choi J, Cai B, Hossain M, Kirby RS, Liu J. Bayesian 2-Stage Space-Time Mixture Modeling with Spatial Misalignment of the Exposure in Small Area Health Data. *J Agric Biol Environ Stat.* 2012;17:417–41. [PMID: 28943751].
62. Lawson AB, Song HR, Cai B, Hossain MM, Huang K. Space-time latent component modeling of geo-referenced health data. *Stat Med.* 2010;29:2012–27. [PMID: 20683893].

# 21st Century Virology: Critical Steps



Paul Shapshak

**Abstract** Modernization and refurbishing virology are rapidly advancing as we embark the third decade of the 21st Century. This is needed so as to deepen the impact of the global public health establishment on disease reduction and improvement in well-being. One of the worst global scenarios has occurred, despite alerts and caveats from scientist, namely, global warming, with severe consequences, which promote lower levels of health, lapses in care, vector spread, and provides complementary alternative evolutionary pathways for disease proliferation and progression.

This chapter approaches the application of equations and computer-related intelligence to the study of biological viruses and summarizes certain theoretical advances in our understanding of energy and order/disorder (entropy), which are essential for advances in virology. This chapter further promotes advances in virology that are essential for fundamental attacks on viral infectious disease, therapy, and vaccines. (Computer as well as biological viruses are mentioned, because of their syzygy of yoked integral understanding.)

**Keywords** Virus · Virus detection · Virology · Biological viruses · Electronic viruses · Molecular biology · DNA sequencing · Genome · Evolution · Vaccines · Thermodynamics · Entropy · Enthalpy · Equations · Fractals · Computer-related intelligence · Self-replication and repair · Artificial Intelligence (AI) · Quantum computers (QC) · Clausius · Feynman · Gibbs · Maxwell · Boltzmann · Einstein · von Neumann

---

P. Shapshak (✉)

Division of Infectious Diseases and International Medicine, Department of Internal Medicine,  
Tampa General Hospital, USF Morsani College of Medicine, Tampa, FL, USA  
e-mail: [pshapsha@usf.edu](mailto:pshapsha@usf.edu)

### Key Concepts

Second decade of the 21st Century. Remodernize, refurbish, and deepen understanding of virology, molecular biology, immunology, and infectious diseases. Increase and re-optimize impact on global human health. Strive for disease reduction. This chapter promotes advances in concepts related to virology and molecular biology, which are essential for fundamental attacks on viral infectious diseases. The increased onslaught of viral diseases since the 20th century requires new paradigm development and widening horizons in our understanding of viral diseases, therapy, and of course vaccines.

## 1 Introduction. Brief History of Biological and Computer Viruses

Viruses have a complex history with escalating consequences. Throughout thousands of years there were slight inklings of invisible noxious diseases that spread havoc and panic. Finally, in the nineteenth century, Koch's work on infectious disease causation, criteria leading from hypothesis to actual detection, were developed and applied in general and for virology. Contemporary work in the twenty-first century, yields great technological advances on many fronts, as well as the application of various computer technologies including, AI, neural networks, etc. Moreover, although frequently styled in opposition, whereas they are really in apposition, the study of biological and computer viruses actually assist one another in terms of understanding, modeling, propagation, evolution, as well as anti-viral therapies, procedures, and programming [1, 2].

In 1892, Dmitry Ivanovsky and Martinus Beijerinck discovered the first biological viruses, identified as Tobacco Mosaic Virus (TMV). In 1931, the electron microscope was invented by Ernst Ruska and Max Knoll. In 1941, the electron microscope was then used by William M. Stanley and Thomas F. Anderson to detect and visualize TMV. Since then, the study of viruses has expanded massively, including, for example, Zika, influenza, bunya, hemorrhagic fevers (such as epominous Ebola and Marburg), dengue, papilloma,<sup>1</sup> RSV, WNV, CMV, HIV, hepatitis viruses (including HCV, HBV, Hepatitis A virus), pico-, mega-, filo-, flavi-, henipa-, parvo-, lenti-(HIV-1, HIV-2), retro- (HTLV-1, HTLV-2) and oncornavirus. Last, but not least, are prions, enigmatic proteinaceous non-canonical infectious agents [3–15].

In addition, a cast of characters involving the human genome was discovered in the twentieth century. Namely, during the late 1940s, Barbara McClintock discovered jumping genes, which migrate in maize genomes; jumping genes, mobile elements, were then also demonstrated to have important roles in evolution. In addition, many segments of the human genome were identified as mobile elements, as well. Furthermore, vestiges of prior virus epidemics and pandemics left their marks, hidden in human genomes [16–21].

<sup>1</sup> Respiratory syncytial virus (RSV), West Nile virus (WNV), cytomegalovirus (CMV), human immunodeficiency virus (HIV), hepatitis B virus (HBV), Human T-cell leukemia virus (HTLV).

The first computer virus, called ‘Brain’, was created with respectable intentions by Basit Farooq Alvi and Amjad Farooq Alvi, in 1985. Brain was initially invented to secure medical programs from piracy and infected unauthorized PCs via floppy disks; be that as it may, Brain, in effect, became the first computer virus. Actually, earlier, in 1971, ‘Creeper virus’ was the first internet virus and spread on ARPANET, which was precursor to the internet. ‘Creeper’ was a self-replicating program, invented by Bob Thomas of BBN Technologies, reason unknown (originally Bolt, Beranek, and Newman Technology, Cambridge, MA) [22–26].

Spread globally since then, there have been explosive proliferations of malicious codes, malware, including botnet, worms, viruses, Trojan horses, backdoor, logic bombs, rabbits, and spyware. These have become denizens of the ‘Dark Web’. The Dark Web is huge, has not been characterized much, and is home and refuge for these codes and host to many more, increasing daily, ensconced in crime and terrorism [24–26].

Global electronic ‘webs’ are complex multilayered constructs. The tip of the iceberg that is accessed by public search engines such as Google, Chrome, Firefox, Safari, and Yahoo, include the internet and worldwide web – though, their content is indexed. The internet and worldwide web are minuscule (1/400 to 1/500 size) compared to the unknown unfathomed sizes of successive layers, the Deep Web and then the Dark Web. The Deep Web is partially unindexed and indexed. As yet, the deepest layer, the Dark Web, is intentionally hidden and clandestine. However, passwords and network permissions are required for its use. Codes in the Deep and Dark Webs can replicate, others not, and they show diverse degrees of independence, dependence, and inter-dependence. Moreover, the Dark Web is relatively unscathed by law enforcement organizations around the globe [24–26].

It should be noted that a converse comparison could be made between public health service institutions (including the NIH, CDC, and WHO) that attack human and animal viral infectious diseases and reservoirs vs. efforts by other governmental to investigate, control, and abolish the Dark Web. If NIH and other governmental health-related agencies conducted their attacks on viral infectious diseases as do the branches of governments that attack the nefarious criminal Dark Web, they would cut their spending by surely 90%, have only a few programs operating, and simply ignore the spread of infectious diseases that kill millions of people globally. Indeed, for example, Google searches indicate that there are stolen confidential health care databases for sale on the Dark web. Consequently, public healthcare faces threats of crime and bioterrorism – legal worldwide web information faces abuse and exploitation by the Dark web [24–26].

## 2 Simple Classic Virus Infectivity

One of the fundamental classic canonical simple approaches to virus propagation has been the calculation of overall (net) reproductive ratios,  $R$ . When cells or people are infected with a virus, this ratio is calculated by dividing the infection rate by the recovery rate, i.e. the reservoir inflow/outflow. This yields the following formula where  $S$  is the number of susceptible uninfected people, which is 1 at time zero,  $b$  is

the infection rate constant, and A is the reservoir mean lifetime inverse rate constant.  $R_0 = bS_0/A$ . When  $R = 1, 0$ , or  $< 1$ , then the virus dies out. Indubitably, infection proceeds when  $R > 1$ . Additional differential equations are further described in the literature, based on this simple model [27].

### 3 Databases and Published Literature

The plethora of data published during the last 70–80 years is available inefficiently for the most part, due to paper-archives. However, since the accelerated use of computers, scanning, and the internet, data has become more accessible. Peer reviewed publications and databases have significantly fostered growth and progress. As exemplified by the utility of, for example, systems biology, biochemistry, and molecular pathway databases. The problem of the accuracy of the information published in the literature and in databases, even after peer review, is complex and although not discussed much here, has been analyzed previously, in part, for viral sequence databases as an example, and found wanting. Indeed, there are many decades of spurious data accumulated in the databases since their inception [28]. Clearly improvements in data quality as well as improved curation are needed; although, that said, over decades, curation criteria have shown compliance, improvement, and willingness to progress.

Overall, pathway database literature includes multiple components, such as signaling, metabolic, and genetic diagrams. Due to questions of accuracy and means of comparisons among studies, curation is a major issue in progress, as generally, curation does not provide comparisons of various studies with meaningful associated p-values or significance level measures, or any other reliable levels of comparison. Some studies mention control experiments performed; however, there is generally no internationally agreed upon measures or protocols for such calculations and units of comparison. On the contrary, as a customary practice, several types of controls are generally mentioned, which for the most part are considered sufficient. Be it as it may, on the one hand, professional curators assemble information and provide their stamps of approval. On the other hand, text mining of abstracts using natural language processing contribute to the accumulating reviews and faulty curation processes, let alone the problems of laboratory criteria of fidelity, as mentioned above [28, 29].

Summaries and explanations of several databases are provided by Nagasaki et al. [29]. They report that diagrams are usually presented in graphic interchange format (GIF) files and thus are uneditable. File formats differ although Self Extensible Mark-up Language (XML) is common. DNA sequence databases include, for example, Los Alamos, Weizmann Institute, Santa Cruz, NIH, several in Europe, and the Human Genome Center of Kyoto. Several additional databases include: Kyoto Encyclopedia of Genes and Genomes (KEGG), Biocyc by SRI International, Ingenuity Pathway and Knowledgebase, Biobase Transpath, ResNet, Ariadne Genomics, Signal transduction knowledge environment, (STKE), Database of Cell

signaling, Cell Biology, GeneOntology, Reactome, Cold Spring Harbor and European Bioinformatics Institute, Metabolome.jp, BioCarta Signaling Pathways, INOH Signaling Pathways, iPath Signaling Pathways, and Pathway builder. These sites tend to lack simulation investigatory methods, experimental data support is missing, and there are problems of viewing and editing [29]. Moreover, global standards of analysis and practice are not in place.

Virus structure databases are available too, for example, which provide information related to specific capsid proteins, amino acids, etc. [30] Most databases appear to be stand-alone and non-interactive with other databases. However, common languages, ontologies, and shared semiotics have been devised. *Glossa* is such a language that promotes communications, semiotics, among computers [31]. Funding agencies need to be aware that further work needs to be done.

Intrinsic to the methods of databases include application of various types of algebras [32–35]. Named sets are one of the foundations of mathematics and various algebras utilize this and are key in their application to computer communications – languages. Semantics, combinatorics, and various logics are based on multisets – and these are named sets. In outline, relational databases use relational algebras and abstract states and automata are used for mathematical modeling and computer information processing [36, 37]. It should be emphasized that there appears to be absent international standards, codification, compilation, adjudication, or any scientific body, which stays abreast and provides any sort of systematization, regulation, let alone oversight for the venue of all these databases. Such lack of organization was not envisaged by founders of the computer age including Babbage, Feynman, Godel, Hopper, Lovelace, Nash, Turing, von Neumann, nor Zuse [36, 38].

Clearly, extrication from this conundrum requires technological progress equal to the task including especially artificial intelligence (AI) and high speed quantum computers (QC). These are global massive tasks with huge information volumes and are required to deal with task immensity and complexity.

## 4 AI Legacy

The application of AI to virology is a recent twenty-first century endeavor; however, AI has been utilized and is continuing rapid acceleration and development in other scientific fields, since the mid twentieth century [39]. For some perspective, for example, it is germane to remark upon AI advances in space programs and the aerospace industry, as described by D. Girimonte and D. Izzo of the European Space Agency (ESA). Areas of progress and application of AI includes space technology, space engineering, spacecraft system design, distributed AI, and enhanced situation self-awareness. Girimonte and Izzo explain that ‘emerging’ intelligence is required due to mutable and unpredictable environments. Along these lines, intelligence does not reside in one singular system, cogently, it effects noteworthy measures due to environmental interfaces. This approach is termed distributed AI (DAI). Examples of DAI include distributed computing and swarm intelligence [40].

Swarm intelligence is actually a sub-category of distributed AI and has a variety of definitions. Salient among them is that intelligent systems can be constituted by multiple components, which have limited sensing or cognitive abilities compared to their sum. Each of the components interacts with its environment and also interacts among each other, locally [40]. It should be apparent that possibly, a swarm intelligence approach could be used by a large central computer that simulates the automata of swarm intelligence to arrive at subsequent action modes. This approach could be used in virology, considering the complex innumerable multiple variables involved in characterizing and understanding the many components of virus structure, function, infection, replication, inhibition, and interaction with host environments. (Cf. other chapters in this book.)

Distributed computing involves shared computer resources and memory capabilities of interacting dispersed separate computers, without interdependencies. Using this mechanism, hundreds of terabytes of space program data have been analyzed per year. (400 terabytes equals about 60,000 CDs, which amounts to about 1.1 terabyte per day, or about 6.9 CDs per hour). Larger tasks involve distributed global optimization, and this compels even greater complexity as well as massive impediments and necessitates big data proliferation analytic techniques. This further promulgates interdependency among computers for such calculations [40].

As a result, the use of the many virus databases that exist globally for virus structure, function, genetics, receptor, metabolic interactions, growth and inhibitory factors, requirements in various cell and host environments, would necessitate advanced computer power of the types, as implied. Towards such goals, the many forms of AI and QCs need further development. Moreover, space exploration provides situations in which docking maneuvers are required, when contact is needed in numerous situations.

Correspondingly, by analogy, a consideration of molecular docking that requires various programmed gene expression requirements, maturation stages and conformations of proteins and possibly organelles involved, movement through nuclear pores, across Golgi apparatus, and conceivably within axons in either direction. As the various components agglomerate, the docking then takes place at a cell membrane or mitochondrial membrane, all the while taking into account the exigencies of cellular physiology [41]. Then, envisioning a specific desired molecular or reactive outcome, clearly the multidimensionality of the various independent variables involved are greatly exceed what any group of individuals with their personal computers can accomplish in any timely fashion and towards any reasonable goal.

The computer systems, codes, and calculations that could solve such complex problems, which go beyond the scope of big data, would have to have some degree of creativity and/or self-awareness. This is especially the case for unanticipated modes of reaction, equilibria, barriers, and thermodynamics that could occur. The degrees of stochastic modeling required, are overwhelming. AI creative coding production and facilitation are required towards these goals. Moreover, cloud-use by the various codes and computers involved is an additional step towards autonomous conduct and verisimilitude. The design and management of experiments as well as data analysis will be greatly enhanced by AI with advanced computer power and codes.

## 5 Entropy

Since Clausius first invented entropy, what entropy is, has been hotly debated for at least 150 years, by physicists including Gibbs, Helmholtz, Schrodinger, Klotz, Prigogine, Shannon, etc. On the one hand, Albert Einstein in 1949 is often quoted saying “Thermodynamics is the only physical theory of universal content which, within the framework of the applicability of its basic concepts, I am convinced will never be overthrown.” On the other hand, however John von Neumann is quoted as stating that “no one knows what entropy really is.” [42–49]

Be that as it may, entropy and its associated concepts, equations, and calculations are utilized throughout chemistry, physics, and cosmology. Analysis, research, and development utilize entropy across a wide range. For example, entropy considerations assist in the development of new materials through understanding the stability and melting of e.g. various ceramics and metal alloys [50, 51]. Entropy deliberations assist in understanding viruses as will be described below (*vide infra*).

Moving forward then, the following is a brief introductory review of entropy basics used in AI and virology. Based on the 2017 entropy review by Popovic [42], thermodynamic entropy, S, is described in Eq. 1 as follows.

$$S = S_0 + \int_{T=0}^T \frac{C_p}{T} dT \quad (1)$$

Note that  $S_0$  is residual or zero-point entropy at zero degrees Kelvin (0degK) and is due to a minimalist arrangement of particles in crystal lattices. T is temperature and  $C_p$  is heat capacity at constant pressure.  $\int(C_p/T) dT = S_{\text{Therm}}$  is thermal entropy due to particle motion that is consequent to heat. Next, Gibbs free energy  $\Delta G$  (a measure of non-expansive reversible work, available energy) and enthalpy  $\Delta H$  (a measure of internal energy for a given pressure and volume), then  $\Delta S = (\Delta H - \Delta G)/T$  (Gibbs Equation.)

Reversible processes are characterized by heat exchanged,  $Q_{\text{rev}}$ , where  $S_{\text{Therm}}$  is thermal entropy.  $dS_{\text{Therm}} = dQ_{\text{rev}}/T$ . Heat and entropy are proportional in reversible reactions.  $\Delta S$  is thus net unavailable energy. Entropy is not an obvious concept, still debated, used in various considerations and contexts, and is not a form of energy that can itself be used for work. In economics, for example, entropy would be a product of labor, which had been used and was no longer re-useable – a measure of uselessness or nonproductivity. Such thermodynamic systems are isolated and closed [42]. For detailed discussions and a global view, also refer to [43, 52].

Thus, irreversible processes generate entropy. Entropy, in non-equilibrium reactions, is a measure of irreversible disorder. In the following Eq. 2, external entropy exchange is designated by  $dS_e$  and production of entropy by irreversible actions including heat transport, diffusion, chemical reactions (reaction potential) are designated by  $dS_i$

$$dS = dS_e - dS_i \quad (2)$$

Consequently, for each thermodynamic force, so to say, there is a conjugate flow and corresponding differential equations that can be integrated to quantify the entropy flows and forces in each part of a system [42].

## 6 Statistical Mechanics

A more statistical approach is taken to describe the thermodynamics of systems with multiple particles. The problem is that given, say, a mole of a gas,  $6 \times 10^{23}$  particles, then the number of degrees of freedom in the system is  $18 \times 10^{23}$  spatially, and  $24 \times 10^{23}$  including time. Hence, the Ergodic theorem to the rescue – that microstate values, weighted and averaged, provide the probability of a system's macroscopic properties. The Ergodic theorem is also known as the Gibbs postulate. It is also statistical thermodynamics' second postulate. The internal energy equation is Eq. 3. Where  $\epsilon_i$  is the energy of the microstate and  $p_i$  is the probability of the microstate.

$$U = \sum_i p_i \epsilon_i \quad (3)$$

For  $T$ , temperature and  $k_b$ , Boltzmann constant, Eq. 4 is the Boltzmann distribution, the probability,  $p_i$ , of a microstate as a function of its energy.

$$p_i = e^{-\epsilon_i/kbT} / \sum_i e^{-\epsilon_i/kbT} \quad (4)$$

The Gibbs entropy Eq. 5 allows calculation of the entropy from the microstate probabilities.

$$S = -k_b \sum_i p_i \ln p_i \quad (5)$$

Consequently, the Boltzmann entropy Eq. 6 is at equilibrium, or summed over all microstates:

$$S = k_b \ln W_{eq} \sim k_b \ln W \quad (6)$$

In regard to living systems, Schrodinger defined ‘negentropy’ as a measure of order, as the opposite or negation of entropy. If  $D$  is the measure of disorder, entropy, then  $\ln(1/D) = -\ln D$  is a measure of order. Although there is much discussion and disagreement about this, possibly, it is applicable, at least to some degree, to living systems, including viruses.

Moving further, Shannon produced entropy measures of information. This placed a new feature as to what is entropy and impacts information itself about objects, translated into cornerstones of entropy. There are two parts to this as described in detail by Popovic [42].

First, Eq. 7.

$$I = K \ln(M) \quad (7)$$

K is a constant used for conversion among types of logarithms, M is the finite total of possible messages, and I is the system's total information. For example, then, if there are only two states for a message, then possible messages,  $M=2^N$ . Consequently, total information  $I = \log_2(M) = N$  bits.

Second, Shannon information entropy.  $S_{sh}$  is Shannon entropy and is defined as the message's average information per symbol,  $p_i$  is probability of the i symbol.  $N \times S_{sh} = I$ , number of message symbols, N [42, 43, 48, 52, 53].

$$S_{sh} = -K \sum p_i \ln p_i \quad (8)$$

Nonetheless, if a given material, which is uniform is divided equally into different pieces, then if the amount of heat required to change temperature is different for each, then this would indicate differences in Shannon entropy – i.e. – in whatever way the molecules are aligned, independent of any observer.

## 7 Viruses and Entropy

Using entropy in the study of biology is implicit – examining reactions, limits, and envelopes of what is alive. Viruses are the simplest forms of life on the planet, though obligatory intracellular parasites. For example, although influenza virus has had a tremendously detrimental impact on human health over the millennia and acutely during the 20th and 21st centuries, entropy considerations have been applied productively and predictively at the molecular level [46, 49, 54, 55].

Influenza is one of several viruses that demonstrate accelerated rates of molecular evolution. This ability allows such viruses to escape immune responses, attack the host, and broadens their ability to breach herd and individual immunity, thereby causing epidemics and pandemics. For Influenza, at the molecular level, this distills down to the evolution and immune escape of the H3N2 influenza A haemagglutinin (HA) glycoprotein. In a study by Pan and Deem [54], H3N2 molecular sequence comparisons were made across two epidemic seasons 1992–3 vs. 2009–10. Antibody-associated selection pressure and diversity across each amino acid were studied. Relative and Shannon entropy, state variables, were used. The authors state that studying an initial season allows predictions for the subsequent season's parameters. Comparisons of sequences among the USA, Europe, Japan, and China were made. As anticipated, virus diversity and rate of evolution were synchronized. The entropy methods used, confirmed 54 amino acid sites that previously evolved in evading immune responses. In addition, antibody binding was abrogated on HA epitopes A and B. Between seasons, Relative and Shannon entropy correlated with higher evolutionary rates. In addition, relative entropies showed migration of virus

from the epicenter, China, to Japan, and to the USA, followed by a new path to Europe. Moreover, selection pressure was confirmed predominantly on a 54 amino acid subset on the HA trimer (epitopes A and B) [54].

Several equations were used by Pan and Deem [54], where amino acid identity, k runs 1, ..., 20; j is position; and season is i.

Equation 9 Shannon entropy for the HA influenza A protein

$$D_{i,j} = -\sum_{k=1}^{20} f(k,i,j) \log f(k,i,j) \quad (9)$$

However, it should be noted that there are differing reports as to the amino composition of the HA protein. One report states that there are only 18 naturally occurring amino acids in Influenza proteins. These are ala, arg, asp, cys, glu, gly, his, isoleu, leu, lys, met, phe, pro, ser, tyr, threo, trypt, and val [56]. Another report states that the amino acid list additionally includes, asparagine and glutamine as well as glycosylated asparagine residues [57]. These should be included in such entropy analyses. In addition, the cysteine residue may be in the form of an SH/thio side-chain or could be part of a cys-cys, S-S (disulfide bond) dimer. The disulfide bond should have additional effects on protein entropy. Such additional annotations need to be included as well, in such calculations.

Furthermore, another problem should be dealt with in such molecular epidemiological studies – that the bird vectors involved, do not obey the arbitrariness of political boundaries and maps. *Au contraire*, the vector demes and habitats should be investigated and defined to ascertain and identify the impact of their biogeographic distributions, not their geopolitical distributions. This is advisable to address that particular contributory component of the epidemiology of vector-borne infectious diseases.

## 8 Noise in Entropy and Virology

Measurements made in virology specifically, and in biology in general, have unavoidable errors. Consequently, there are innumerable means of error measurements, classification, and analysis to stratify measurement reliability [58, 59]. The use of error reports has progressed and evolved to the degree that measurements, results, or reports from research laboratories or from clinics, without error measurements or estimates, tend to be ignored by the scientific community. The sophistication of error analysis stems from fundamental concepts of signal and noise.

Shannon, in 1948, pioneered the impact of noise (disorder and error) on communication (information) and towards that understanding that entropy is involved. Indeed, Shannon's perspective has had a profound and revolutionary impact on subsequent communication theory [48, 60].

Signals, like any other phenomenon, are not perfectly reproducible. Thus, Shannon states that a function describing a received signal E, and transmitting signal, S, must then include noise N, where N is a stochastic variable.

$$E = f(S, N) \quad (10)$$

Assume probabilities and numbers of states are finite, where  $\alpha$  is initial channel state,  $\beta$  is final channel state, i ranges over transmitted signals, and j ranges over received signals. p is the probability that channel  $\alpha$  in state i results in channel  $\beta$  in state j. Then we have  $p_{\alpha i}(\beta j)$ .

However, Shannon specifies that if successive channels are independently perturbed by noise, then there is only one state with the transition probability from i to j:  $p_{ij}$ . Nonetheless, he goes on to state that there are two processes occurring even if one source feeds a single channel. From this, one calculates a few entropies.  $H(x)$  is the input source entropy and  $H(y)$  is the output received channel entropy. In the absence of noise,  $H(x) = H(y)$ .  $H(xy)$  is the joint entropy of input/output. Conditional entropies are  $H_x(y)$  and  $H_y(x)$  where  $H_x(y)$  is the entropy of the output if the input is known and  $H_y(x)$  is the entropy of the input if the output is known.

$$H(x,y) = H(x) + H_x(y) = H(y) + H_y(x) \quad (11)$$

$H_y(x)$  is an ambiguity or equivocation in the received signal and thus the actual signal transmission is R, which it should be noted, tends to be a Bayesian approach. By implication, Shannon points out a theorem in informational entropy that the correction channel capacity must be at least  $H_y(x)$  and this is the information (that we may say, counters entropy), and which corrects and improves the received signal. An independent machine or channel, corrects the information flow enhancing the signal. It must observe the source and receiver and then proffer its corrections. This is not a perfect process, as errors exist in the actual performance of real machines, though perhaps not for virtual machines [48, 60].

$$R = H(x) - H_y(x) \quad (12)$$

## 9 Shannon Entropy Concepts and Viruses

At this point, skeptical readers will ask whether any of this really relates in the slightest to viruses, let alone cells and people, which viruses infect, parasitize, make ill, and often kill.

Indeed, Shannon's work, methods, and discoveries are relevant. Entropy of DNA, RNA, mRNA, siRNA, transcription, translation, protein synthesis, enzymes, metabolism, heat shock proteins, protein folding, maturation, etc. globally include much of what keeps cells alive; these include processes that are perturbed when viruses invade cells.

Finite fractal dimension and Shannon entropy can be analyzed concomitantly in order to determine their possible relationships, since both focus on information present in the various systems involved. Measurements of chemistry, biochemistry, biophysics, etc. provide the information flow being analyzed. As an example of recent sophisticated preliminary-type analyses, Holden and colleagues in 2013 [61] compared Shannon entropy and finite fractal dimension. Indeed, a wide array of such analyses produced evidence that supports the utility of such methods.

A very interesting approach was used. Two-dimensional (2D) maps of several genes from humans and mice were analyzed using DNA bioinformatics methods: fractal dimension and Shannon entropy. Shannon entropy provides analysis of informational organization vs. chaos and fractal analysis entails calculating a dimension that may be non-integer and is often based on a non-rectifiable curve. Fractals are generally non-commensurate with the spaces in which they are embedded; in addition, their self-similarity is maintained with changes in magnification. These would appear to be independent variables and thus amenable to statistical comparison. However, since they both derive from information content of the molecules being analyzed, they could be related. Indeed, the following equation indicates a relationship between topological entropy  $S$ ,  $D$  is dimension calculated – e.g. from Shannon, Renyi (Kolmogorov), or Hausdorff,  $r$  is the reciprocal of the reduced size of the covering element (termed  $\varepsilon$ ) [ $r = 1/\varepsilon$ ], and  $q$  is the order of the entropy (e.g. Renyi entropy). In summary, using various types of approaches, the authors conclude that fractals and entropy are directly related [62].

$$S_q(r) = D_q \ln(r) \quad (13)$$

Nevertheless, possibly, the particular choices of assumptions in deriving entropy and fractal *formulae* will influence the results.

Studies by Holden and colleagues in 2012–2013 [61, 63] compared mammalian molecular sequences including RNF4, Myc, LMNA, ESR-1, ID1, PLCZ1, PKM2, p53, and DYS14. They found that mice and humans showed relationships for RNF4, Myc, LMNA, and ESR-1 and a larger separation for ID1 and PLCZ1. They found a mouse-chimpanzee-human relationship for ID1 and tumor suppressor, p53. They found that PKM2 and glycolysis pathways could be targeted related to cancer. They proposed that DTS14 may offer protection from Alzheimer's disease. (DTS14 is a Y-chromosome gene that is involved in fetal micro-chimerism). In summary then, they proposed that there is a measure of non-coding sequence processes that exhibited a pressure of some kind related to the linear sequence or ponderal distance between mRNA and the processed protein-coding CDS [61, 63].

Methods used for analysis of fractal dimension and entropy by Holden et al. were applied to DNA sequences and based on a time-series model by Higuchi [61, 63–65]. Gene DNA sequence data were obtained from GenBank. AT, TA, GC, and CG from base-pairs across DNA strands. However, entropies at single nucleotide placements are made counting four ( $=2^2$ ). However, many calculations are based on dinucleotides, in which there are 16 possible occurrences:  $2^2 \times 2^2 = 2^4 = 16$ . Based on the

basic Shannon entropy equation above (Eq. 8), nucleotide sequence entropies are proportional to their nucleotide variations. Moreover, there are two binary bits per nucleotide place and four binary bits per contiguous nucleotide pair [61, 63].

The authors use a variation of the Higuchi methods for fractals. Difference series of spatial intensity, Int, are produced, in uniform breaks, in spatial ( $j-i$ ) lags:  $\text{Int}(j) - \text{Int}(i)$ .

Here,  $(j-i)$  are pairs that = k and the series-curve unnormalized length  $L(k)$ , is shown in Eq. 12.

$$L(k) = \Sigma |Int(j) - Int(i)|$$

Normalization is used to obtain the series length and the authors provide additional details, including the Weierstrass function, in the calculation plan, resulting in describing the slope of the plot of  $\log(L(k))$  vs  $\log(1/k)$ . This slope is linear and is the fractal dimension. (The error rate is  $</= 1\%$ .) Various plots are shown of dinucleotide entropy (bits/symbol) vs. fractal dimension; however, fractal dimension is stated as not an independent variable. Plots for several genes demonstrate various differences between and within species, mouse and human as well as wolf/dog, bovine, rat, Zebra fish, and chimpanzee. Additional studies analyzed Neanderthal DNA as well. The authors hypothesize that elevated fractal dimension may be associated with transcription factors [61, 63, 64].

Frequently, for entropy calculations involving codons, dinucleotide sequences are used instead of full trinucleotide codon sequences, since the third nucleotide is often redundant [66]. However, in this regard, it should be pointed out that the trinucleotide code will still require a spacer in the third place where such sequences are used for entropy calculations.<sup>2</sup> Analogously, it must be noted that for some genes, parts of exons are read by codon codes in opposite direction, on different strands, and sometimes overlapping on the same strand. Moreover, introns and exons are known for large numbers of genes, multiples splicing subtypes, as well as exons for genes within introns from other genes. This information should assist in these analyses and contribute to specificity. In addition, the genetic code for organelles including mitochondria and chloroplasts should be considered, compiled, and analyzed as well. Such approaches are applicable to viral as well as host DNA sequences.

## 10 Molecular Virology and Program Code Probity

Moving on, there have been remarkable advances in information theory. However, on the one hand, molecular biologists and mathematicians, analyzing biological molecules including DNA, RNA, and proteins, for the most part appear to assume

---

<sup>2</sup>An overlapping code would wreak havoc on such calculations!

that the informational approaches they utilize are universal and thus readily applicable to molecular biology, though originated from other engineering, physical, mathematical, and computer sciences. On the other hand, the mathematician, Kolmogorov, raised questions (1930s – 1980s) as to whether there are such universal codes or programs at all. Kolmogorov's school of thought has been extensively reviewed including components such as assumptions, intuition, various approaches, and methodologies. Further, Kolmogorov utilized the intuitive approach (discussed extensively in the literature) and additionally addressed the non-obvious issue, as to the commensurability of program codes coupled with other program codes, in the presence of varying degrees of channel noise [67–72].

This is a very important issue that needs further deliberation, because, as currently practiced, contemporary molecular biology, public health, and biomedical/molecular research and medical sciences appear to assume that any codes/programs, which are used, are universal and can be harnessed together and concatenated routinely and automatically, without much further consideration – in other words, that one glove fits all.

A straight forward example of this would be the abundance of meta-analyses used by multiple articles on a topic, reviewed and compared from the literature, with conclusions being made, where diverse study designs, statistical algorithms, and programs were used among the various compared studies. PubMed searches show that the literature is replete with such reviews of serendipitously combined medical and biomedical findings and conclusions. Additional examples include the use of results from various concatenated programs with final concluded results. Furthermore, generally, there are no propagated error analyses from each program to the next, to establish the true reliability of the inputs vs. intermediate vs. final outputs. Indubitably, greater use of informational entropy analyses in such studies, as well as the use of AI, are needed.

It should be highlighted that in 1918, Brower was the first to opine that intutional logic has a role and basis in mathematics. Great strides were made and the complexity that developed, is considered overwhelming by many subsequent mathematicians, during the 100 years since then [73, 74]. As logic and mathematics intution theory further developed, Kleene made many advances and conjectured Church's Rule in a weak form for a formula, which is closed and provable, using intutional number theory. He stated that if for all  $x$  there exists  $\phi f(x,y)$ , then the following is true –  $\phi(\bar{N}, F(\bar{n}))$  [75–77]. All this impacts harshly on approaches towards information theory. It cannot be sufficiently emphasized that this is no simplistic purely scholastic or heuristic matter, but impacts on the development of, for example, gaining deeper and more sophisticated understanding of virus evolution and virus vaccines, which are crucial clinical and public health issues: i.e. whose criteria have been or should be stated as these fields continue to advance – by an assortment of individuals or AIs?

The various descriptions and stages of development of the two widely different approaches to information theory, call them the Shannon school and the Kolmogorov school, are discussed in great detail in the literature, e.g. by Grunwald and Vitanyi [78]. In a nutshell, Shannon information theory, as well as Kolmogorov algorithmic

or complexity information theory, includes probabilistic information measurement calculations, using the bit as the information unit. Moreover, it is agreed that the length of description of an object is its amount of information.

In particular, Shannon information theory, encodes objects as outcomes of random (variable) sources that are known. The objects' features are not the outcome. Rather, the known random source descriptions determine the encoding and are the intent, via channels that verge on being error free. Minimizing bits in message communication is the objective. However, the desired outcome is one of many possible random outcomes.

Kolmogorov algorithmic or complexity information theory, treats individual isolated objects as encoded by brief program/codes. Once the code is enunciated, it comes to a halt – and that is the object form, compressed. Here, minimizing the bits needed to store or reconstruct the file is the objective, i.e. irrespective as to how the file originated [78, 79].

## 11 Conclusions

Twenty-First century virology is evolving beyond prior work, as virologists, molecular biologists, clinicians, and mathematicians are developing increased understanding and use of powerful tools involving such concepts and equations for entropy, fractals, and the application of AI. AI has become unavoidably pervasive in a plethora of fields including bioinformatics, brain-computer interfaces, neuroinformatics, DNA computing, autonomic computing, quantum cryptography, and QC. [80] It behooves virologists, globally, to further apply AI methodologies and technologies to advance virology and immunology and to enhance the attacks on virus diseases.

Since the late twentieth century, molecular biologists are becoming aware of the huge complexity and lack of consensus by mathematicians and logicians as to roles intuition might have on logic and mathematics, intuitional and extensional realizability, information theory, and ultimately, AI, which are included among the foundational pillars of molecular biology.

Notable caveats, beyond the scope of this chapter, are that mounting and enduring disasters, including global warming, poverty, violence, drug and human trafficking, and war promote declined health levels, lapses in healthcare, widespread disease vectors, and provide alternate additional evolutionary pathways. Their complexities require AI and QC, for viruses to be more accurately understood and more effectively combatted.

**Acknowledgements** Conversations with Dr. G. Baumslag (Institute for Advanced Study, Princeton, NJ), Dr. C. Smith (Princeton University, Princeton, NJ), and A. Pellionisz (Mountainview, CA) are acknowledged.

**Conflicts of Interest** The author reports no conflicts of interest.

## References

1. Kephart JO, Sorkin GB, Arnold WC, Chess DM, Tesauro GJ, White SR. Biologically inspired defenses against computer viruses. 1996;1:985–96. <https://www.ijcai.org/Proceedings/95-1/Papers/127.pdf>
2. Spafford EH. Computer viruses as artificial life. *Artif Life*. 1994;1:249–65. [www.scs.carleton.ca/~soma/biosec/readings/spafford-viruses.pdf](http://www.scs.carleton.ca/~soma/biosec/readings/spafford-viruses.pdf)
3. Cairns J, Stent GS, Watson JD. In: Watson JD, editor. *Phage and the origins of molecular biology*. Cold Spring Harbor: Cold Spring Harbor Lab; 1966.
4. Doerr R, Hallauer C. In: Hallauer C, editor. *Handbuch der Virusforschung – erste Halft*e. Vienna: Springer; 1938.
5. Fenner F. In: Gibbs A, editor. *Portraits of virology: a history of virology*. Basel: Karger; 1988.
6. Luria SE. *General virology*. New York: Wiley; 1953.
7. van Helvoort T. History of virus research in the 20th century: the problem of conceptual continuity. *Hist Sci*. 1994;32:185–235.
8. van Helvoort T. When did virology start? *Am Soc Microbiol News*. 1996;62:142–5.
9. Waterson AP, Wilkinson L. An introduction to the history of virology. Cambridge, MA: Cambridge University Press; 1978.
10. Emerman M, Malik HS. Paleovirology—modern consequences of ancient viruses. *PLoS Biol*. 2010;8:e1000301. <https://doi.org/10.1371/journal.pbio.1000301>.
11. Rybicki E. A short history of the discovery of viruses. 2018. <https://www.researchgate.net/publication/279758269>.
12. Flaviani F, Schroeder DC, Lebret K, Balestreri C, Highfield A, Schroeder JL, Thorpe SE, Moore K, Pasckiewicz K, Pfaff MC, Rybicki EP. Distinct oceanic microbiomes from viruses to protists located near the antarctic circumpolar current. *Front Microbiol*. 2018;9:1474. <https://doi.org/10.3389/fmicb.2018.01474>.
13. Fernandez F, Minagar A, Alekseeva N, Shapshak P. Neuropsychiatric aspects of prion disease. In: Sadock BJ, Sadock VA, Ruiz P, editors. *Comprehensive textbook of psychiatry*. Philadelphia: Kluwer and Lippincott Publ; 2017. p. 601–18.
14. Shapshak P, Somboonwit C, Kuhn J, Sinnott JT, editors. *Global virology I. Identifying and investigating viral diseases*. New York: Springer Publ; 2015.
15. Shapshak P, Levine AJ, Somboonwit C, Foley BT, Singer E, Chiappelli F, Sinnott JT. *Global virology II. HIV and NeuroAIDS*. New York: Springer Publ; 2017.
16. Ravindran S. Barbara McClintock and the discovery of jumping genes. *Proc Natl Acad Sci USA*. 2012;109:20198–9. [www.pnas.org/cgi/doi/10.1073/pnas.1219372109](https://www.pnas.org/cgi/doi/10.1073/pnas.1219372109).
17. Pandita D, Pandita A. Jumping genes- the other half of the human genome and the missing heritability conundrum of human genetic disorders. *Br Biotechnol J*. 2016;11:1–18. ISSN: 2231-2927. NLM ID: 101616695. [https://www.researchgate.net/publication/290210528\\_Jumping\\_Genes-The\\_Other\\_Half\\_of\\_the\\_Human\\_Genome\\_and\\_the\\_Missing\\_Heritability\\_Conundrum\\_of\\_Human\\_Genetic\\_Disorders](https://www.researchgate.net/publication/290210528_Jumping_Genes-The_Other_Half_of_the_Human_Genome_and_the_Missing_Heritability_Conundrum_of_Human_Genetic_Disorders)
18. Belshaw R, Pereira V, Katzourakis A, Talbot G, Paces J, Burt A, Tristem M. Long-term reinfection of the human genome by endogenous retroviruses. *Proc Natl Acad Sci U S A*. 2004;101:4894–9.
19. Katzourakis A, Gifford RJ. Endogenous viral elements in animal genomes. *PLoS Genet*. 2010;6:e1001191.
20. Koonin EV. Taming of the shrewd: novel eukaryotic genes from RNA viruses. *BMC Biol*. 2010;8:2–11.
21. Boeke JD, Stoye JP. Retrotransposons, endogenous retro-viruses, and the evolution of retro-elements. In: Coffin JM, Hughes SH, Varmus HE, editors. *Retroviruses*. Cold Spring Harbor: Cold Spring Harbor Press; 1997. p. 343–435.
22. Brain virus. [https://en.wikipedia.org/wiki/Brain\\_\(computer\\_virus\)](https://en.wikipedia.org/wiki/Brain_(computer_virus)).
23. Creeper virus. [https://en.wikipedia.org/wiki/Creeper\\_\(program\)](https://en.wikipedia.org/wiki/Creeper_(program)).

24. Zeidanloo HR, Tabatabaei SF, Amoli PV, Tajpour A. All about malwares (malicious codes). 2010. <https://pdfs.semanticscholar.org/a45e/50583a13e04b920f6ba04473612734967aa7.pdf>.
25. Finklea K. Dark Web. Security. Congressional Research Service. 2017. <https://fas.org/sgp/crs/misc/R44101.pdf>.
26. Sui D, Caverlee J, Rudesill D. The deep web and the darknet: a look inside the internet's massive black box. 2017. [www.wilsoncenter.org](http://www.wilsoncenter.org) [https://www.wilsoncenter.org/sites/default/files/stip\\_dark\\_web.pdf](https://www.wilsoncenter.org/sites/default/files/stip_dark_web.pdf).
27. MCB 137 Berkeley Virus Population Dynamics. 2016. [https://mcb.berkeley.edu/courses/mcb137/exercises/Virus\\_Dynamics.pdf](https://mcb.berkeley.edu/courses/mcb137/exercises/Virus_Dynamics.pdf).
28. Balaji S, Akash R, Krittika N, Shapshak P. Sequence accuracy in primary databases: a case study on HIV-1B. In: Shapshak P, Levine AJ, Somboonwit C, Foley BT, Singer E, Chiappelli F, Sinnott JT, editors. Global virology II. HIV and NeuroAIDS. New York: Springer Publ; 2017.
29. Nagasaki M, Saito A, Doi A, Matsuno H, Miyano S. Foundations of systems biology using Cell Illustrator and pathway databases. New York: Springer Publ; 2009. Chapter 2 Pathway databases. p. 5–18.
30. Virus structure database. [https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1347395/bin/nar\\_34\\_suppl-1\\_D386\\_index.html](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1347395/bin/nar_34_suppl-1_D386_index.html).
31. Kazic T. Semiotics: a semantics for sharing. Bioinformatics. 2000;16(12):1129–44. <https://doi.org/10.1093/bioinformatics/16.12.1129>.
32. Adamek J, Rosicky J, Vitale EM. Algebraic theories. Cambridge Tracts in Mathematics 184. Cambridge, UK: Cambridge University Press; 2011. isbn: 978-0-521-11922-1
33. Schultz P, Spivak DI, Vasilakopoulou V, Wisnesky R. Algebraic databases. 2016. <https://categoricaldata.net/fql/jfpslides.pdf>.
34. Abiteboul S, Hull R, Vianu V. Foundations of databases. Reading: Addison-Wesley; 1995. isbn: 0-201-53771-0
35. Adamek J, Rosicky J. Locally presentable and accessible categories. London Mathematical Society Lecture Note Series 189. Cambridge: Cambridge University Press; 1994. isbn: 0-521-42261-2
36. von Neumann J, Burks AW. Theory of Self-reproducing automata. Urbana, IL: University of Illinois Press; 1996.
37. Burgin M. Unified foundations for mathematics. Mathematics LO/0403186. 2004. 1–39. p.arXiv:math/0403186v1.
38. Pesavento U. An implementation of von Neumann's self-reproducing machine. Artif Life. 1995;2:337–54.
39. Berrar D, Sato N, Schuster A. *Quo Vadis*, Artificial Intelligence? Adv Artif Intell. 2010;2010:629869, 12. <https://doi.org/10.1155/2010/629869>.
40. Girimonte D, Izzo D. Artificial intelligence for space applications. In: Schuster A, editor. Intelligent computing everywhere. London: Springer; 2007. p. 235–43.
41. Good MC, Zalatan JG, Lim WA. Scaffold proteins: hubs for controlling the flow of cellular information. Science. 2011;332:680–6.
42. Popovic M. Researchers in an entropy wonderland: a review of the entropy concept. 2017. <https://arxiv.org/pdf/1711.07326>
43. Salamon P, Andresen B, Nulton J, and Konopka AK. The mathematical structure of thermodynamics. 1996. <http://www.sci.sdsu.edu/~salamon/MathThermoStates.pdf>
44. Tribus M, McIrving EC. Energy and information. Sci Am. 1971;225:179–88.
45. Clausius R. The mechanical theory of heat. London: John van Voorst; 1879. <https://www3.nd.edu/~powers/ame.20231/clausius1879.pdf>
46. Schrodinger E. What is life? The physical aspect of the living cell. Cambridge: Cambridge University press, X printing; 2003.
47. Boltzmann L. The second law of thermodynamics (Theoretical physics and philosophical problems). New York: Springer-Verlag New York, LLC; 1974. ISBN 978-90-277-0250-0
48. Shannon C. A mathematical theory of communication. Bell Syst Tech J. 1948;27:379–423.
49. Prigogine I, Wiame JM. Irreversible thermodynamics. Experientia. 1946;2:451–3.

50. Choi WM, Jung S, Jo YH, Lee S, Lee BJ. Design of new face-centered cubic high entropy alloys by thermodynamic calculation. *Met Mater Int.* 2017;23:839–47. <https://doi.org/10.1007/s12540-017-6701-1>.
51. Miracle DB, Senkov ON. A critical review of high entropy alloys and related concepts. *Acta Mater.* 2017;122:448–511.
52. Tolman RC. Relativity, thermodynamics, and cosmology. New York: Dover Publ., Inc.; 1987.
53. Santra SB. Thermodynamics and statistical mechanics, A brief overview. 2014. [http://www.iitg.ac.in/santra/course\\_files/ph443/hstm.pdf](http://www.iitg.ac.in/santra/course_files/ph443/hstm.pdf)
54. Pan K, Deem MW. Quantifying selection and diversity in viruses by entropy methods, with application to the haemagglutinin of H3N2 influenza. *J R Soc Interface.* 2011;8:1644–53. <https://doi.org/10.1098/rsif.2011.0105>.
55. Shapshak P, Chiappelli F, Somboonwit C, Sinnott JT. The influenza pandemic of 2009: Lessons and implications. *Mol Diag Ther.* 2011;15:63–81.
56. Hoyle L, Davies SP. Amino acid composition of the protein components of influenza virus A. *Virol.* 1961;13:53–7.
57. Ward CW, Dopheide TA. Amino acid sequence and oligosaccharide distribution of the HA from an early Hong Kong influenza virus variant A/Aichi/2/68(X-31). *Biochem J.* 1981;193:953–62.
58. Duncan RC, Knapp RG, Miller MC III. Introductory biostatistics for the health sciences. Albany: Delmar Publ. Inc.; 1983.
59. Rosner B. Fundamental of biostatistics. Pacific Grove: Duxbury Publ. Inc; 2000.
60. Reza FM. Information theory. New York: Dover Publ; 1994.
61. Holden T, Cheung E, Dehipawala S, Ye J, Tremberger G Jr, Lieberman D, Cheung T. Gene entropy-fractal dimension informatics with application to mouse-human translational medicine. *BioMed Res Intern.* 2013;2013:7. ID 582358. <https://doi.org/10.1155/2013/582358>
62. Zmeskal O, Dzik P, Vesely M. Entropy of fractal systems. *Comput Math Appl.* 2013;66:135–46.
63. Holden T, Tremberger G, Cheung B, Subramanian R, Sullivan R, Gadura N, Schneider P, Marchese P, Flamholz A, Cheung T, Lieberman D. Fractal analysis of 16S rRNA gene sequences in archaea thermophiles. *World Acad Sci Eng Technol Int J Bioeng Life Sci.* 2008;2:192–6.
64. Tremberger G Jr, Dehipawala S, Cheung E, Yao H, Gadura N, Schneider P, Lieberman D, Holden T, Cheung T. Fractal analysis of FOXP2 regulated accelerated conserved non-coding sequences in human fetal brain. *World Acad Sci Eng Technol.* 2012;67:881–6.
65. Higuchi T. Approach to an irregular time series on the basis of fractal theory. *Physica D.* 1988;31:277–83. Kolmogorov AN, Zur Deutung der Intuitionistischen Logik. Math; 35: 57–65. 1932.
66. Riyazuddin M. Information analysis of DNA. 2005. <https://arxiv.org/pdf/1010.4205>
67. Kolmogorov AN. Zur Deutung der Intuitionistischen Logik. *Math Z.* 1932;35:57–65.
68. Kolmogorov AN. Three approaches to the quantitative definition of information. *Probl Inf Transm.* 1965;1:1–7.
69. Kolmogorov AN. Complexity of algorithms and objective definition of randomness. *Uspekhi Mat Nauk.* 1974;29:155. Moscow Math Soc meeting 4/16/1974
70. Kolmogorov AN. Combinatorial foundations of information theory and the calculus of probabilities. *Russ Math Surv.* 1983;38:29–40.
71. Shen A, Vereshchagin N. Logical operations and Kolmogorov complexity. *Theor Comp Sci.* 2002;271:125–9.
72. Terwijn SA, Torenvliet L, Vitnyi PMB. Nonapproximability of the normalized information distance. *J Comp System Sci.* 2013;77:738–42.
73. Brouwer LEJ. Begründung der Mengenlehre unabh angig vom logischen Satz vom ausgeschlos-senen Dritten. Erster Teil, Allgemeine Mengenlehre, vol. 5. Kon Ned Ak Wet Verhandelingen; 1918. p. 1–43.
74. van Dalen D. Intuitionistic logic. In: Goble L, editor. The blackwell guide to philosophica logic. Oxford: Blackwell; 2001. p. 224–57.

75. Kleene SC. Realizability: a retrospective survey. In: Mathias ARD, Rogers H, editors. Cambridge Summer School in Mathematical Logic, volume 337, of Lecture Notes in Mathematics. Cambridge, UK: Springer-Verlag; 1973. p. 95–112.
76. van Oosten J. Axiomatizing higher-order Kleene realizability. Ann Pure Appl Logic. 1994;70:87–111.
77. van Oosten J. Extensional realizability. Ann Pure Appl Logic. 1997;84:317–49.
78. Grunwald P, Vitanyi P. Shannon information and Kolmogorov complexity. arXiv:cs/0410002v1 [cs.IT] 2004.
79. Gray RM. Entropy and information theory. Stanford, CA, Publ. 2013. <https://ee.stanford.edu/~gray/it.html>
80. Schuster AJ. In: Schuster AJ, editor. Intelligent computing everywhere. London: Springer Publ; 2007.

# Futuristic Methods for Determining HIV Co-receptor Use



Jacqueline K. Flynn, Matthew Gartner, Annamarie Laumaea,  
and Paul R. Gorry

**Abstract** HIV infection of two key immune cells, CD4<sup>+</sup> T cells and macrophages, plays a major role in the establishment of HIV infection and the seeding of the latent reservoir. There is a critical gap in our understanding of the determinants of viral tropism for these cell types and the molecular Env-receptor interactions involved. The development of novel futuristic methods including machine learning and neural network approaches, will allow the generation of sophisticated genotypic prediction algorithms which will greatly enhance accurate HIV coreceptor usage identification within the clinic. Furthermore, advancements in surface plasmon resonance, biolayer interferometry and glycol-FRET technologies will be fundamental for real time investigation of molecular mechanisms involvement in Env-receptor interactions, which will enhance our understanding of coreceptor usage and viral tropism determinants. These technologies will be important for the development and improvement of therapeutic efficacy, novel entry inhibitors and assist vaccine design.

**Keywords** HIV · CD4<sup>+</sup> T cells · Macrophages · HIV entry · Coreceptor usage · Cellular tropism · Machine learning · Surface plasmon resonance · BioLayer interferometry · Glyco-Fret

---

J. K. Flynn (✉)

School of Clinical Sciences at Monash Health, Monash University,  
Melbourne, VIC, Australia

School of Health and Biomedical Sciences, RMIT University, Melbourne, VIC, Australia

Centre for Biomedical Research, Burnet Institute, Melbourne, VIC, Australia  
e-mail: [Jacqueline.flynn@monash.edu](mailto:Jacqueline.flynn@monash.edu)

M. Gartner · P. R. Gorry

School of Health and Biomedical Sciences, RMIT University, Melbourne, VIC, Australia  
e-mail: [s3642483@student.rmit.edu.au](mailto:s3642483@student.rmit.edu.au); [paul.gorry@rmit.edu.au](mailto:paul.gorry@rmit.edu.au)

A. Laumaea

Centre for Biomedical Research, Burnet Institute, Melbourne, VIC, Australia  
e-mail: [annamarie.laumaea@burnet.edu.au](mailto:annamarie.laumaea@burnet.edu.au)

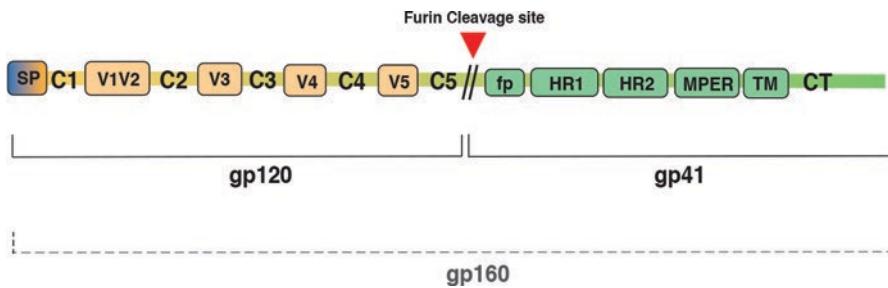
## 1 Introduction

Human immunodeficiency virus (HIV) is a lentivirus from the *retroviridae* family and leads to the onset of acquired immune deficiency syndrome (AIDS) through the infection and destruction of immune cells. HIV infects several immune cell types that express the CD4 receptor, including CD4<sup>+</sup> T cells, macrophages, dendritic cells and monocytes. Infection of CD4<sup>+</sup> T cells typically leads to death of infected cells as well as bystander cells [1], contributing to a severe reduction in CD4<sup>+</sup> T cells that culminates in disease progression to AIDS. Early studies in untreated patients suggested that a median period of 8 years was required for AIDS to develop following HIV infection [2–6]. Since the introduction of combination anti-retroviral therapy (cART) in the mid nineteen-nineties, patient morbidity and mortality have decreased considerably, with patients on continuous therapy with life spans similar to healthy individuals. cART successfully reduces patient viral load in the blood and assists to restore CD4<sup>+</sup> T cell counts within the blood through targeting multiple stages of the viral lifecycle to block viral replication. However, cART is not curative as it fails to clear HIV from the body. HIV persists through the establishment of viral reservoirs of latently infected cells [7]. Following cART interruption or the acquisition of viral-encoded resistance to cART drugs, these latently infected cells can re-establish a productive infection leading to a rapid rise in viral load and a drop in CD4<sup>+</sup> T cell count [8, 9]. This represents a major roadblock in achieving HIV cure using current therapeutic intervention. As such, it is imperative that the mechanisms determining viral tropism and those that allow the virus to infect and establish a latent infection are better understood.

This chapter outlines the HIV entry process and infection of two key immune cells CD4<sup>+</sup> T cells and macrophages and their role in HIV pathogenesis. It examines current and futuristic technologies for determining coreceptor usage and viral tropism, including the development of advanced machine learning and neural networks for increased accuracy and sensitivity of genotypic algorithms to verify HIV coreceptor usage in the clinic. This chapter also discusses novel technologies of surface plasmon resonance, biolayer interferometry and glycoFRET to determine key molecular interactions involved in HIV entry and transcytosis across mucosal barriers. Combined these technologies will enhance our understanding of Env-receptor interactions to improve therapeutic efficacy, future development of entry inhibitors and assist vaccine design.

## 2 HIV Entry

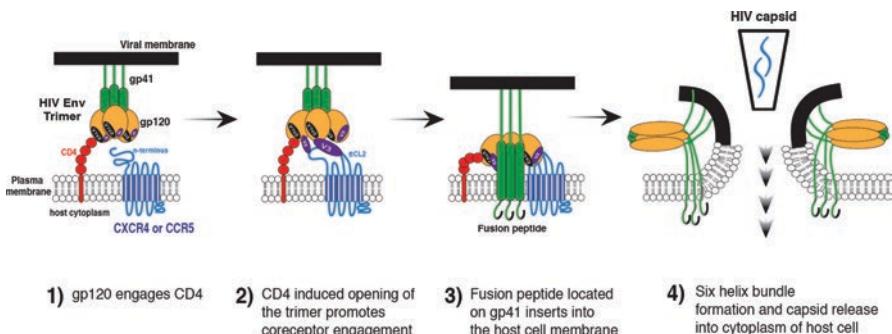
The infection of target cells with HIV is mediated by the envelope (Env) glycoprotein through interactions with cellular receptors. Env is a heterotrimeric protein that studs the outer surface of the virion and contains three gp120 and gp41 subunits [10],



**Fig. 1** The genetic organisation of the HIV envelope gene. Monomeric *Env* (gp160) is cleaved by furin to yield the surface exposed unit (gp120) and the transmembrane unit (gp41). Gp120 consists of a signal peptide (SP), five constant regions (C1–C5) and five variable loops (V1–V5), while gp41 contains a fusion peptide (FP), two heptad repeat regions (HR1 and HR2), the membrane proximal external region (MPER), a transmembrane domain (TM), and a cytoplasmic tail

11]. Gp120 is the surface exposed unit of Env and engages cellular receptors to initiate viral entry, while gp41 is buried within the viral membrane and is involved in mediating viral fusion. Five variable loops (V1–V5) and five conserved regions (C1–C5) comprise gp120 (Fig. 1). The variable regions are generally the sections of gp120 that are surface exposed for receptor engagement while the conserved regions make up the core of the monomeric protein. Gp41 contains a fusion peptide, responsible for mediating viral fusion between the plasma and viral membranes as well as two heptad repeat regions (HR1 and HR2), the membrane proximal external region, a transmembrane domain as well as the cytoplasmic tail, which is embedded within the virion (Fig. 1). Env is highly glycosylated with 26–30 N-linked glycan residues per monomer to protect the native protein from immune surveillance. These residues largely cover the surface exposed region of the trimer and are variable in their position within the trimer for immune evasion purposes.

Env mediates viral entry into host cells through high affinity interactions between gp120 and the CD4 host cell receptor [12]. Attachment of Env whilst in a closed conformation to host attachment factors on the cell surface such as  $\alpha 4\beta 7$  integrin brings Env in close proximity to CD4 [13–15]. Interaction of the CD4 binding site (CD4bs) within gp120 of Env with CD4 causes structural rearrangements of the V1/V2 loop followed by V3 loop rearrangement and formation of the bridging sheet, a four-stranded  $\beta$ -sheet formed between the C1, C2 and C4 domains of gp120 (Fig. 2) [16–19]. These conformational changes induce an open conformation, exposing key coreceptor binding sites (CoRbs) in gp120 that facilitate engagement with a chemokine receptor, either CCR5 or CXCR4 [11, 20]. Coreceptor engagement leads to subsequent exposure and insertion of the gp41-encoded fusion peptide into the cell membrane [10–12], allowing the viral membrane to tether to the host membrane (Fig. 2). During this process, the six-helix bundle is formed to facilitate the delivery of viral contents into the host cell via fusion [21–23].

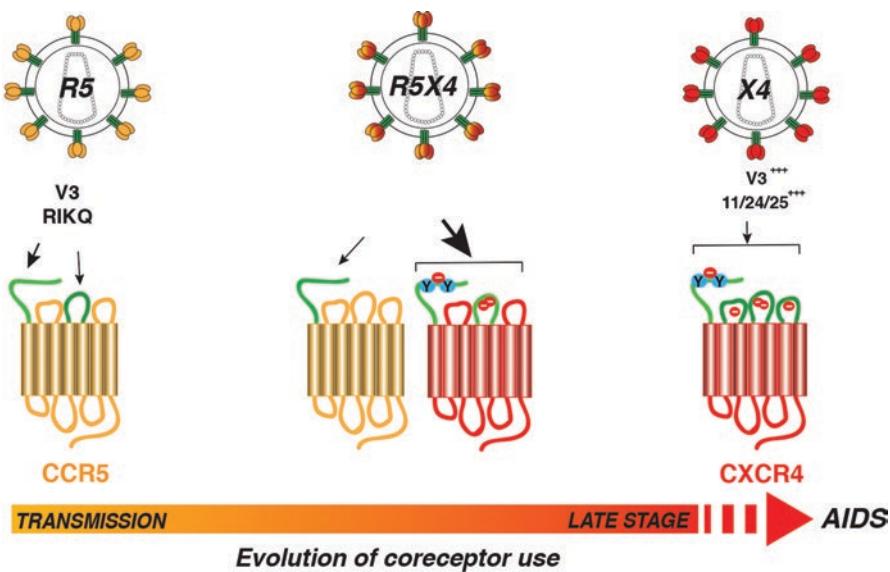


**Fig. 2** Schematic of the HIV entry process. (1) Following attachment to a cell via engagement with adhesion molecules, various sites within outer domain of gp120 engage the CD4 receptor. (2) CD4 engagement induces several conformational changes within the trimer that expose coreceptor binding residues, particularly within the V3 loop (green). (3) The V3 loop engages the coreceptor leading to insertion of the gp41-encoded fusion peptide into the host cell membrane. (4) Fusion peptide insertion leads refolding of gp41, with the HR1 and HR2 domains forming a six-helix coiled bundle that permits viral fusion; fusion occurs between the viral and host membranes to release the HIV capsid into the host cytoplasm

### 3 Coreceptor Usage

HIV Envs are phenotypically characterized by their ability to utilize CCR5 (R5), CXCR4 (X4) or both coreceptors (R5X4) for entry. Activated CD4<sup>+</sup> T cells and resting memory CD4<sup>+</sup> T cells express high levels of CD4 and varying levels of CCR5, while naïve CD4<sup>+</sup> T cells express little to no CCR5 [24–28]. In contrast, CXCR4 is expressed by all CD4<sup>+</sup> T cell subsets, although its expression is the highest on naïve CD4<sup>+</sup> T cells [24–28]. Immature and mature dendritic cells as well as macrophages express low levels of CD4 and varying levels of CCR5 and CXCR4 [25, 26, 29]. Historically CXCR4-using HIV strains were categorised as T cell tropic (T-tropic) and CCR5-using as macrophage tropic (M-tropic) based on efficient infection of each cell type, and dual-tropic CCR5 and CXCR4-using (R5X4) viruses were classed dual-tropic [30–33]. However, more recent studies have demonstrated that CCR5-using viruses that infected CD4<sup>+</sup> T cells do not always infect macrophages [34–37]. Moreover, several highly M-tropic HIV viruses have been shown to utilize CXCR4 for entry [34, 38]. As such, there are three main HIV-1 entry phenotypes; CCR5 T cell tropic, CXCR4 T cell tropic and CCR5 macrophage tropic, with X4 macrophage tropic viruses occurring rarely *in vivo*. Viruses, which use CCR5 as a coreceptor almost exclusively, facilitate HIV transmission [39–42] such that individuals harbouring a deletion mutation in the CCR5 gene ( $\Delta 32$  CCR5) are largely protected from infection [43, 44]. Dendritic cells (DCs), macrophages and memory CD4<sup>+</sup> T cells express both coreceptors, with the expression of CCR5 implicating these cells in playing a major role in the establishment of infection [45–48].

Transmitting viruses, termed transmitted/founder (T/F) viruses are typically CCR5-using, although CXCR4-using variants from recently infected individuals have been isolated [40, 49]. T/F viruses infect resting and activated CD4<sup>+</sup> T cells, and rarely macrophages [40, 50, 51]. Disease progression of from acute to chronic disease stages is associated with a switch in coreceptor usage from R5 to X4 in 40–50% in clade B infected individuals [52], whilst prevalence of coreceptor switching and X4-usage remains less clearly defined for other clades. It is thought that a switch in coreceptor usage would allow viruses to infect a broader range of CD4<sup>+</sup> T cells including naïve CD4<sup>+</sup> T cells, leading to a rapid depletion of these cells, supported by coreceptor switching being associated with accelerated disease progression to AIDS [53, 54] (Fig. 3).



**Fig. 3** The evolution of HIV coreceptor usage and switching during disease progression. Transmitter/founder viruses are characterised by CCR5 coreceptor usage. The HIV envelope sequence, including the presence/absence of the RIKQ sequence, and in particular the V3 loop sequence and charge play a role in determining coreceptor usage. During disease progression the evolution of a dual tropic CCR5 and CXCR4 virus can occur. This then can lead to the evolution of CXCR4-using viruses during chronic disease stages, especially in clade B HIV where 40–50% of viruses undergo a coreceptor switch from CCR5-using to CXCR4-using. The evolution of a coreceptor switch can occur from CCR5-using straight to CXCR4-using viruses or through an evolutionary intermediate dual tropic CCR5 and CXCR4-using HIV, termed a R5X4 virus. To determine coreceptor usage the V3 loop of gp120 is commonly analysed whereby a coreceptor switch is associated by a change in charge (indicated by the black plus symbols), with a net charge equal to or greater than +6 predicting X4-usage. Additionally, the presence of positive charged amino acids at positions 11 and 24/25 in V3 was consistent with X4 usage introduction of the V3 loops net charge into the rule. R5 indicates CCR5-using viruses, R5X4 indicates dual tropic viruses, X4 indicates CXCR4-using viruses and Y indicates tyrosines

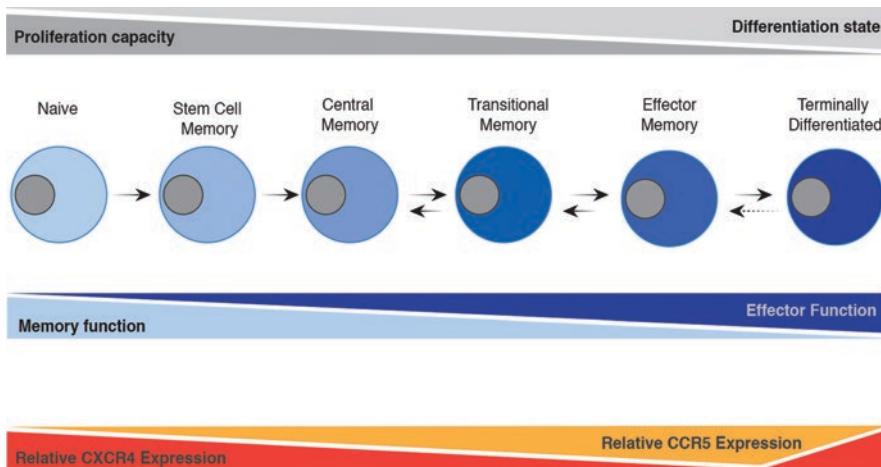
## 4 CD4<sup>+</sup> T Cell Tropism

### 4.1 Role of CD4<sup>+</sup> T Cells in HIV Disease Pathogenesis

CD4<sup>+</sup> T cells are the primary target cell for HIV-1 infection. Their contribution to the pathogenesis of infection is two-fold, with CD4<sup>+</sup> T cell depletion leading to AIDS [55, 56], and the latent infection of resting CD4<sup>+</sup> T cells leading to long-term HIV persistence despite therapy [57]. CD4<sup>+</sup> T cell depletion occurs through several mechanisms, including direct viral killing of infected cells, induction of syncytium formation followed by lysis, and pyroptosis of abortively infected T cells [58, 59]. Moreover, CD4<sup>+</sup> T cell depletion can also result from killing of infected CD4<sup>+</sup> T cells by CD8<sup>+</sup> cytotoxic T cells (CTLs) and expression of death ligands on activated macrophages and DCs (i.e. Fas Ligand and TRAIL) [59]. HIV can induce CTL-mediated killing of uninfected bystander CD4<sup>+</sup> T cells through gp120 shedding following engagement of a CD4 binding site-targeting antibody [60, 61]. Antibody engagement with this region leads to disassociation of the trimer, allowing monomeric gp120 to bind CD4 on uninfected cells, which is recognised and killed by CTLs [60, 62]. CD4<sup>+</sup> T cell depletion by HIV-1 infection was initially thought to be relatively slow throughout disease pathogenesis [3, 5]. However, more recent studies have revealed memory CCR5<sup>+</sup>CD4<sup>+</sup> T cells are preferentially and markedly depleted during acute infection within the gastrointestinal tract [63, 64]. As such, HIV CD4<sup>+</sup> T cell pathogenesis has two phases: the acute infection phase which is characterized by a rapid loss of resting memory CD4<sup>+</sup> T cells expressing CCR5 in mucosal tissue, and the chronic phase that is characterized by proinflammatory cytokine production, and slow peripheral CD4<sup>+</sup> T cell depletion [63–67].

### 4.2 Determinants of CD4<sup>+</sup> T Cell Tropic Viruses

CD4<sup>+</sup> T cells can be divided by their effector function and differentiation into naïve and memory T cell subsets. For example, following the clearance of a pathogen, the number of effector CD4<sup>+</sup> T cells contracts and a small subset of effector CD4<sup>+</sup> T cells transition into a resting or memory state, of which there are numerous subsets, including the stem cell memory (TSCM), central memory (CM), transitional memory (TM), effector memory (EM) and terminally differentiated (TEMRA) (Fig. 3). Of particular interest is the ability of HIV to infect the memory T cell subsets as these cells have long life spans that contribute to the persistence of the latent reservoir. These cells express differential levels of the HIV coreceptors CCR5 and CXCR4 [26–28]. CXCR4 expression is highest on naïve T cells and decreases as cells become more differentiated (Fig. 4). In contrast, CCR5 is rarely detected on most naïve T cells, with expression increasing as cells become more differentiated into effector subsets.



**Fig. 4** Properties of different memory CD4<sup>+</sup> T cell subsets. Memory CD4<sup>+</sup> T cell subsets have differing roles within the host immune response. Naïve CD4<sup>+</sup> T cells demonstrate the greatest proliferation capacity but the lowest effector function. After interaction with cognate antigen, naïve T cells progress through the differentiation pathway, losing proliferation capacity and gaining effector functions such as pro-inflammatory cytokine production. HIV is able to infect all of T cell types depicted in this figure, with the efficiency of infection influenced by the expression of CCR5 and CXCR4. CXCR4 expression is the greatest on naïve CD4<sup>+</sup> T cells and decreases as cells differentiate, while CCR5 expression is the lowest on naïve CD4<sup>+</sup> T cells and increases as cells progress through differentiation states. Terminally differentiated CD4<sup>+</sup> T cells are the exception to this, demonstrating decreased CCR5 and increased CXCR4 expression compared to effector memory cells

Studies describing the susceptibility of various CD4<sup>+</sup> T cell subsets with clinical R5 and X4 isolates are limited. Based on the expression patterns of CXCR4 and CCR5 on memory CD4<sup>+</sup> T cell subsets, it is conceivable that R5-using viruses will preferentially infect EM cells, while X4-using viruses will have preference for infecting naïve CD4<sup>+</sup> T cells. Tabler and colleagues analysed the efficiency of infection of different memory CD4<sup>+</sup> T cell subsets via infection with infectious molecular clones with a clade B X4-using and R5-using envelopes [24]. This study found that the efficiency of infection by R5-using viruses matched the level of CCR5 expression on the subsets. Furthermore, a trend was observed in X4-using viruses infection of memory cells decreasing as CXCR4 expression decreased. Consistent with this, Parrish et al. used clade C Env-pseudotyped infectious molecular clones from T/F viruses as well as chronic viruses to show EM cells were more susceptible to infection with CCR5-using Envs than CM, TEMRA or naïve CD4<sup>+</sup> T cells [68]. Furthermore, our laboratory has shown R5-using C-HIV Env-pseudotyped viruses preferentially infected memory CD4<sup>+</sup> T cell subsets, while X4 viruses preferentially infected naïve CD4<sup>+</sup> T cells [27, 69]. We also showed TSCM cells are preferentially infected by R5 isolates compared to X4 isolates, despite these cells expressing very low levels of CCR5 [27]. As such, coreceptor usage and the level of coreceptor

expression may be involved in determining cellular tropism, but is likely to not be the only determinant.

### 4.3 *The Role of CD4<sup>+</sup> T Cells in HIV Latency*

HIV is able to establish a latent infection and this is a major barrier to cure via current therapeutic interventions [7, 8]. A latent HIV infection is defined as a cell containing an integrated replication competent provirus that lack the transcriptional environment to produce viral proteins and infectious virus particles [70, 71]. Latently infected cells also have the potential to be reactivated, leading to a productive infection [70]. Latency is established predominantly in CD4<sup>+</sup> T cells with a resting phenotype through either the direct infection of a resting cell (termed pre-activation latency) or through the infection of an activated CD4<sup>+</sup> T cell that survives HIV-mediated cytopathic effects and transitions to a resting cell (termed post-activation latency) [72]. Recently, the Siliciano laboratory showed activated CD4<sup>+</sup> T cells that were in the process of transitioning to a resting state were more permissive to HIV infection than activated or resting cells [73]. These cells also demonstrated a favourable transcriptional environment to maintain a latent infection. Because of their central role in HIV pathogenesis, the study of how CD4<sup>+</sup> T cells become infected is of vital importance for future therapeutic intervention strategies.

There has been a large research effort to elucidate the contribution of different CD4<sup>+</sup> T cell subsets to the long-lived latent reservoir. A seminal study by Chomont et al. found CM and TM CD4<sup>+</sup> T cells were major long-lived reservoirs harboring CCR5-using HIV in cART-treated patients, with the relative contributions of these subsets to the latent reservoir varying between patients [74]. Consistent with this, Soriano-Sarabia et al. found replication-competent HIV DNA in CM cells, but rarely in TM cells of acute and chronic HIV infected patients receiving cART [75]. Recent studies have implicated TSCM CD4<sup>+</sup> T cells as a long-lived reservoir for HIV, with these cells exhibiting higher HIV DNA levels per cell following cART treatment than other CD4<sup>+</sup> T cell subsets [76, 77]. Consistent with this, a separate study revealed patients who maintained undetectable viral loads (termed elite controllers (ECs)) demonstrated a greater preservation of TSCM cell numbers and lower HIV DNA copies/per cell compared to cART-naïve B-HIV infected patients [78]. Together, these findings highlight the role of TSCM and CM CD4<sup>+</sup> T cells in the maintenance of a stable HIV-1 reservoir in cART-treated individuals and the importance of future intervention strategies targeting the infection of these long-lived cell types.

## 5 Macrophage Tropism

### 5.1 Role of Macrophages in HIV Disease Pathogenesis

Macrophages are an important innate immune cell, which are derived from the mononuclear phagocytic lineage [79–82]. Tissues can be reseeded and populated with circulating monocytes, which have migrated through blood vessel endothelium, where they differentiate into macrophages with differing functions; guided largely by the tissue environment [80, 83]. Macrophages are morphologically and functionally heterogeneous, with varying lifespans ranging from weeks to years largely dependent upon their tissue location [84, 85]. Macrophages are key antigen presenting cells and their main role is to survey their environment, phagocytose foreign antigens and present these to B and T cells, as well as to maintain tissue homeostasis by removing dead cells and/or respond to danger signals through their surface receptors [86–88].

Macrophages express CD4, albeit at lower levels than CD4<sup>+</sup> T cells [26], and express the HIV coreceptors CCR5 and CXCR4, and are important target cells for HIV [26, 89]. In comparison to CD4<sup>+</sup> T cells, macrophages are able to resist HIV infection largely due to the presence of the host cell restriction factor SamHD1, expressed in myeloid cells [89, 90] and macrophages have a reduced sensitivity to viral cytotoxicity compared to CD4<sup>+</sup> T cells [89, 91]. Despite a decline in the popularity of characterising macrophages by the M1-M2 paradigm [92], where M1 macrophages (classically activated) are associated with the secretion of pro-inflammatory cytokines (IL-1beta, IL-6, IL-12, TNF-alpha) and M2 macrophages (alternatively activated), which have an anti-inflammatory role and regulate wound healing and tissue repair, this classification has aided understanding of their role in HIV pathogenesis [93, 94]. In the acute stages of HIV infection it has been proposed that M1 macrophages are the more abundant macrophage phenotype [95], whereas later in more chronic stages of HIV infection where tissue injury and inflammation are common, the presence of elevated IL-4 and IL-13 is likely to contribute to a switch from M1 to M2 macrophages being the more abundant phenotype [95].

HIV infection, especially the chronic stages, is characterized by chronic immune activation and inflammation [96], and due to the macrophage's long lifespan they have been characterized as chronically and persistently infected cells [91, 97]. Studies in untreated AIDS patients, have demonstrated macrophages in the gut to be the main drivers of persistent inflammation [98]. Clayton et al. highlighted the contribution of macrophages to persistent immune activation and inflammation, particularly through poor clearance of HIV infected macrophages by CD8<sup>+</sup> T cells [99]. Thus, inflammatory macrophages are key contributors to both HIV disease pathogenesis and its co-morbidities including cardiovascular disease [100] and HIV associated neurocognitive disorders [101].

## 5.2 Determinants of Macrophage Tropic Viruses

Macrophages express CCR5 and CXCR4, with the expression of CXCR4 at a relatively lower level [102]. Expression of CD4 on macrophages is modest compared to CD4<sup>+</sup> T cells [103], which express relatively high levels of CD4 and varying levels of co-receptors CXCR4 and CCR5, largely depending upon state of differentiation [27, 104] (Fig. 4). This heterogeneity of receptor expression across different cell types appears to be an important determinant for cell susceptibility to HIV infection, although receptor expression alone does not seem to be the only determinant for cellular tropism. Studies by Gorry and colleagues [34, 36–38, 105] showed several T cell tropic R5 viruses were incapable of infecting CCR5-expressing macrophages, and X4 viruses infected macrophages in a relatively CXCR4<sup>low</sup> environment. The asymmetry of CCR5 use on macrophages by brain derived versus blood derived R5 viruses was further demonstrated by Peters et al. [106, 107], who showed brain Envs exhibited higher efficiency of receptor use. These studies highlighted an important distinction between cell and coreceptor tropism and emphasized the importance of factors such as immune pressure influencing coreceptor tropism evolution [106, 108, 109], as well as the need to further delineate their contributions to tropism.

Tissue location can also influence permissiveness to HIV-1 infection and the phenotype of macrophage present [110, 111]. The isolation and study of HIV infected macrophages from various anatomical locations has proven challenging and thus far macrophages from the alveolar space in the lung obtained via bronchoalveolar lavage [112, 113] and macrophages of the CNS post-mortem are the most well characterised subsets [114, 115]. However, the function of alveolar macrophages in HIV is less clear [112, 113], though the increased susceptibility of HIV infected individuals to respiratory infections suggests an impaired role of pulmonary immune cells, of which macrophages make a vital contribution.

In contrast, much work has been done with macrophages of the CNS. These are comprised of parenchymal microglia, perivascular, meningeal and choroid plexus macrophages; all of which are susceptible to direct HIV infection [116, 117], however perivascular macrophages and parenchymal microglia are suggested to be the main targets of HIV [114, 118, 119]. Macrophages of the CNS are crucial to HIV pathogenesis as they serve as an immune sanctuary for HIV infection due to poor bioavailability of cART coupled with poor immune responses. Furthermore, longevity of HIV infected CNS macrophages in the absence of severe cytopathic effects that often befall infected CD4<sup>+</sup> T cells [114, 120–125] suggest that these cells may be important long-term viral reservoirs, possibly constituting the major source of virus in later disease stages in the face of declining numbers of CD4<sup>+</sup> T cells [126, 127].

While CNS, rectal and lung macrophages can be directly infected by HIV [128–130], gut macrophages, though one of the most abundant cell population in the gastrointestinal tract, are largely refractory [131, 132]. This is thought to be due to lower expression of HIV's main receptors required for HIV entry [132]. Instead gut macrophages, as with those in other tissue locations, are hypothesised to ingest HIV

infected cells and harbour intact virus in what are known as virus containing compartments (VCC) or crypts inadvertently becoming productively infected [89, 133, 134]. Additionally, HIV can accumulate in these sanctuaries protected from neutralizing antibodies, thus becoming important viral reservoirs. HIV can subsequently be transferred via *trans*-infection through the virological synapse between macrophages and T cells to infect CD4<sup>+</sup> T cells [135–137] aided by the cell surface C-type lectin Siglec-1 [138]; a mode described as being the most efficient means for rapid virus dissemination. It is likely then that VCCs function principally to provide a continuous supply of virus for dissemination via *trans*-infection of CD4<sup>+</sup> T cells.

### 5.3 *The Role of Macrophages in HIV Latency*

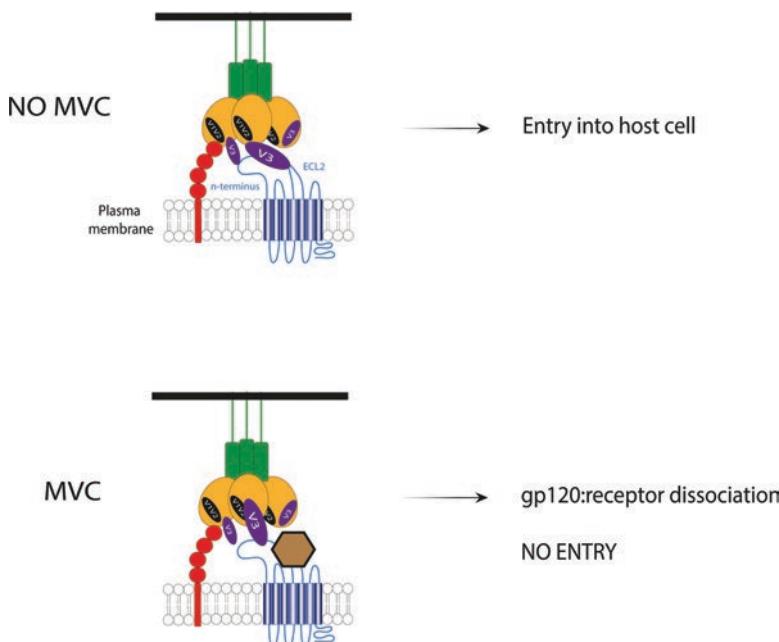
An important consideration for HIV cure efforts is the contribution of macrophages to HIV latency. The majority of studies have focused on the contribution of CD4<sup>+</sup> T cells to the latent reservoir, with few examining the contribution of macrophages. Recently however, Castellano et al. [139] using human fetal microglia, showed integration of HIV DNA into host DNA as early as 3 days. Importantly, their study was able to show that cells surviving after 21 days of infection had transitioned to a state of latency that was able to be reversed upon reactivation with latency reversing agents (LRAs). This further supports the notion that macrophages are important contributors to the latent reservoir, which may be exacerbated by their ability to secrete pro-inflammatory cytokines and chemokines to recruit T cells to the site of infection [91]; increasing the number of infected cells and contribution to the viral reservoir. Importantly, the long-lifespan of macrophages, the relative resistance to HIV-1 induced apoptosis and some cytotoxic effects of HIV, and their ability to release HIV-1 over an extended period, highlights the significant role macrophages play in HIV pathogenesis [140]. While significant technological advancements have been made in studying the latent reservoir in CD4<sup>+</sup> T cells, much work is to be done in the context of macrophages, most of which is reliant on characterizing coreceptor usage, determinants of cellular tropism, understanding the phenomenon of VCCs, and ascertaining the contributions each of these make to HIV disease pathogenesis.

## 6 HIV Entry Inhibitors and Developments in Coreceptor Screening

Since the discovery of involvement of the CD4 receptor and the coreceptors CCR5 and CXCR4 in mediating viral entry [30, 141, 142], the inhibition of the entry process has remained an attractive therapeutic strategy. Entry inhibitors are classed as either CD4-interaction, coreceptor-interaction or fusion inhibitors. CD4-interaction

inhibitors prevent gp120 from engaging with CD4 to block opening of the trimer, for example preventing the transition between envelope confirmation states (Fig. 2). Coreceptor-interaction inhibitors prevent gp120 from engaging with either CCR5 or CXCR4, while fusion inhibitors target gp41 to prevent fusion between the viral and host cellular membranes. To date, two entry inhibitors have been FDA approved for use in the clinic; Maraviroc (MVC), a CCR5 inhibitor [143] and Enfuvirtide (T-20), a fusion inhibitor [144].

The only FDA-approved CCR5 inhibitor, Maraviroc (MVC) binds within the hydrophobic pocket of the second extracellular loop (ECL2) of CCR5 and induces conformational changes within this region, preventing HIV from engaging CCR5 (Fig. 5) [145]. MVC has been shown to be an effective antiretroviral compound in treatment-experienced as well as naïve individuals that harbour CCR5-using viruses only [146–148]. However, resistance to MVC can occur and has been described *in vivo* during the MOTIVATE-1 and -2 phase III clinical trials. Of the 1049 patients, 228 patients experienced treatment failure at 48 weeks following MVC administration, 57% of treatment failure patients harboured dual or X4-using



**Fig. 5** The mechanism of action of maraviroc (MVC) in the inhibition of CCR5-using HIV. After CD4 engagement, several conformational changes within the trimer are induced. These expose the coreceptor binding residues, particularly within the V3 loop (green). The V3 loop engages the coreceptor CCR5, via the second extracellular loop (ECL2) and N-terminus of CCR5, leading to viral fusion between the host and viral membranes. Maraviroc (MVC) is a CCR5 antagonist that binds within the hydrophobic pocket of ECL2 within CCR5, causing conformational changes within this region. These conformational changes are sufficient to block engagement with CCR5 receptor and block HIV entry into the cell

viruses at week 48 of the study, while the remaining 43% of treatment failure patients maintained CCR5-using viruses that displayed MVC resistance at week 48 of the study [147]. A subsequent clinical study of the Maraviroc versus Efavirenz in treatment-naïve patients (MERIT) phase III clinical trial found that 31% of individuals with virological failure harboured CXCR4-using viral variants [149]. Studies have revealed that resistance to MVC in patients with CCR5-using viruses results from alterations in the CCR5-gp120 interaction. CCR5-using MVC-resistant viruses demonstrate increased reliance on interactions between gp120 and the N-terminus of CCR5 allowing these viruses to recognise the drug-bound form of CCR5 [150–154].

## 6.1 Phenotypic Coreceptor Screening Development

Due to MVC's mechanism of action, it is not prescribed to patients that harbour CXCR4-tropic viruses. As such, before MVC prescription, it is necessary to perform a coreceptor usage test of the viruses circulating within a patient. Methods utilised in the clinic to determine the coreceptor tropism of clinical strains include phenotypic and genotypic-based approaches. The Trofile assay is the 'Gold Standard' for coreceptor usage determination of clinical isolates [155–158]. This assay was developed by LabCorp and was first utilised in the MOTIVATE trials that led to the approval of MVC for use in the clinic [146]. Coreceptor usage is determined by measuring the efficiency of pseudoviruses containing patient-derived *Env* sequences to infect U87 cells expressing CD4 and either CCR5 or CXCR4. The sensitivity of the Trofile assay has improved overtime, however, these tests are expensive and extremely time-consuming in comparison to genotypic approaches, limiting the assay's use in the clinic particularly in resource-constrained countries.

## 6.2 Genotypic Coreceptor Screening Development

Genotypic coreceptor prediction algorithms provide an easy, affordable alternative to phenotypic coreceptor usage testing. The first determinants used to predict coreceptor usage in primary HIV isolates was the 11/25 rule [159], where early observations in clade B viruses revealed the presence of positive charged amino acids at positions 11 and 25 in V3 was consistent with X4 usage [159–162]. This simple rule was later bolstered by the introduction of the V3 loops net charge into the rule, whereby a net charge equal to or greater than +6 predicts X4-usage; this rule is termed Raymond's algorithm [163]. However, several studies have demonstrated this simple algorithm (11/25 as well as 11/25/charge) has low sensitivity compared to other more advanced genotypic approaches and the phenotypic Trofile assay [164–167].

More sophisticated computational approaches utilizing machine learning technology have been developed that have improved sensitivity and specificity that anal-

use the whole V3 loop region. Currently, the most common algorithms utilised for genetic coreceptor usage prediction are geno2pheno and WebPSSM [161, 168, 169]. The original WebPSSM likelihood matrices were generated using a set of clade B V3 sequences, and as such performed poorly when clade C sequences were analysed [170]. The authors later developed likelihood scoring matrices that predicted X4 usage in clade C sequences with improved sensitivity and specificity [170]. Geno2pheno utilises non-linear support vector machines (SVM), a method of supervised statistical machine learning that builds a model based on the set of training V3 loop sequences to predict the likelihood a sequence belongs to an X4-using virus [161]. However, while these two methods of coreceptor prediction remain popular, studies have shown they do not always predict the emergence of non-clade B X4-variants in clinical samples with exceptional sensitivity or specificity [165, 171–176], and there has been a considerable discordance shown between the two algorithms [165, 172, 173]; this necessitates the requirement for improved approaches to more accurately predict X4-variants.

Our laboratory developed coreceptor usage prediction algorithms specific for different clades of HIV. Phenoseq has separate models to determine coreceptor usage of clade A/AG, AE, B, C and D isolates with robust sensitivity and specificity. The different models were manually built through analysis of V3 loop characteristics from training sets of clade-specific V3 sequences and analysis of their influence on X4-tropism; properties analysed included the length and net charge of the V3 region, the number of N-linked glycosylation sites and the frequency of site-specific amino acid alterations [166, 177]. Clade-specific Phenoseq algorithms demonstrated similar or improved sensitivities and specificities for predicting X4-tropic variants in several cohorts of viruses representing different clades [177]. A more advanced coreceptor usage prediction platform that incorporates molecular dynamics simulations of structural and chemical variability within V3 loop:coreceptor interactions was recently described [178]. CoReceptor USage prediction for HIV-1 (CRUSH) utilises inter-residue interaction energies derived from computational models of V3:coreceptor complexes combined with four known rules for coreceptor usage; net charge, glycosylation motif, 11/24/25 rule, and length to predict CCR5 or CXCR4-usage [178]. Similar to geno2pheno, CRUSH utilised non-linear SVM training with a training set of sequences that included 235 X4-tropic and 2220 R5-tropic V3 sequences to generate the algorithm. CRUSH prediction of 876 V3 loops derived from diverse clades (A/AG, AE, B and C) demonstrated similar or improved performance when compared to the respective Phenoseq algorithms [178]. One limitation to this algorithm currently is that no studies have validated CRUSH in a clinical setting.

Genotypic-based coreceptor prediction algorithms are inexpensive, simple to use and generate results faster than phenotypic assays [164, 165, 179, 180]. However, there is a large discordance between different methods of coreceptor prediction, suggesting that more work is required to improve current algorithms. For example, Kalu et al. compared results obtained from 352 therapy-naïve clade C-infected Ethiopian individuals using geno2pheno (clinical and clonal models), clade C PSSM, Raymond's algorithm and Phenoseq-C; here, only 58.2% of predictions were concordant across the different algorithms [173]. Similar results were shown

by Trabaud et al., with 58% concordance between geno2pheno, PSSM and Raymond's algorithm using 50 individuals infected with diverse clades [172]. The studies discussed above demonstrate that further experimentation is required to further understand the viral determinants that influence coreceptor usage to improve current prediction algorithms.

### 6.3 Improving Current Approaches

Despite the large number of genotypic coreceptor usage prediction algorithms that have been developed, these algorithms do not predict the presence of X4-using variants with exceptional accuracy and precision. One caveat of current methods is that these algorithms were typically trained on V3 loop sequences only [161, 166, 170, 177, 178]. While the V3 loop is the centre-point of coreceptor interaction [159–162], numerous studies have demonstrated that regions outside the V3 region influence coreceptor usage [69, 181–188]. For instance, our laboratory and Coetzer et al. demonstrated that clade C HIV Envs required mutations within the V1/V2 loops as well as the V3 loops induce a coreceptor switch in clinical isolates [69, 188]. It is thought that these additional mutations within the V1 and V2 regions compensate for the subtle structural differences in Env conformation as a result of the V3 mutations [69]. Additional studies have revealed that several residues within gp41 are associated with coreceptor switching in clade B and C isolates [182, 183, 186]. As such, future prediction algorithms must take into account the full-length *envelope* sequence to predict coreceptor use.

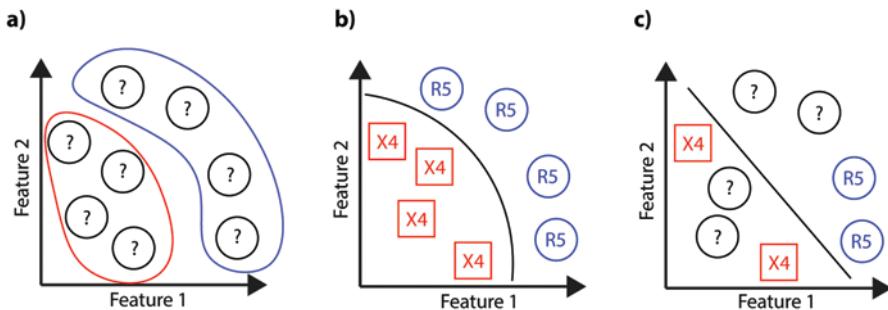
A recent study from Díez-Fuertes et al. devised an algorithm from a tree-augmented Naïve (TAN) Bayes classifier trained on full-length *Env* sequences from 429 isolates across diverse clades [189]. The authors used a wrapper method to select nucleotide sites that provided the most accurate model for coreceptor prediction. Of the 3648 nucleotide positions screened within the 429 sequences, 26 positions were selected by wrapper feature selection to generate the TAN algorithm. Of these, 16 positions were within gp120 and 10 were in gp41; 6 nucleotide positions were located within the V3 loop, further suggesting that features other than the V3 are required for prediction of coreceptor usage. The TAN classifier algorithm was compared against Geno2Pheno (false positive rate (FPR) of 2.5% and 10%) and WebPSSM using 177 clade B sequences and 252 non-clade B sequences and demonstrated improved or similar accuracy, specificity and sensitivity; although sensitivity for non-clade B sequences was slightly lower than geno2pheno with 10% FPR. Overall, this study highlights two important points that future coreceptor prediction strategies should consider. First, machine learning algorithms that train models on 'training' data sets are required; and second, that nucleotide features outside the V3 region are important to consider as these sites may affect the overall structure of an Env and thus ability to bind either CCR5 or CXCR4.

## 7 Futuristic Methods for Determining Coreceptor Usage – A Prediction Algorithm Approach

### 7.1 Machine Learning

Machine learning is a collection of computer-based analytical techniques that learn from a ‘training’ dataset to make predictions and decisions about unknown or unclassified data [190]. Machine learning algorithms find optimum separations to classify a set of data to generate a model for classification, based on mathematical rules and statistical assumptions to maximise correct classifications while minimising errors [191]. Machine learning models are typically generated from input data consisting of ‘labels’ and ‘features’. Labels are the output or the result that the model aims to predict, while the features refer to the measurements or data used to predict the labels [191]. Using coreceptor usage prediction as an example, the feature set may include a vector representing amino acid composition or the physiochemical properties of the amino acids within the V3 loop of a training set of sequences with phenotypically verified coreceptor tropism, while the labels are the prediction that an Env is R5- or X4-using. Features can be selected manually by the user from previous observations (ie V3 net charge, amino acid positions at 11 and 25), or they can be selected using a multitude of different feature selection techniques [192]. The ‘training’ or ‘learning’ process involves generating the optimum set of model parameters that optimise a given evaluation metric to assess model accuracy, precision or recall. Models are usually assessed using the area under curve (AUC) score if data sets are approximately equal for each label or F1 score if there is an imbalance in label numbers [193]. The ‘learning’ process involves adjustment of model parameters to optimise the chosen evaluation metric, and is halted once minimal improvements are being made or the model is over fitted [193].

Hundreds of machine learning algorithms have been published, with methods commonly split into three categories: supervised, unsupervised and semi-supervised learning [193]. Unsupervised methods are used when the labels or the output is unknown (Fig. 6) [193], and as such are unlikely to be used for coreceptor prediction as *Env* sequences used for training are usually phenotypically verified using the Trofile assay. Supervised learning methods are common in the generation of coreceptor usage prediction algorithms because the output labels are already known [194]. In supervised learning the output data labels are used to find common patterns within the features of the input data that are predictive of the output labels (Fig. 5). For example, discriminative features might include the amino acid residues at key positions that best predict X4-usage. In addition to supervised and unsupervised learning methods, semi-supervised learning is used when some but not all of the output labels are known (Fig. 6) [195]. Supervised learning methods can further be grouped into classifications or regressions, where classifications have categorical output values (eg. R5 or X4-using), and regression outputs are continuous variables [195]. Regression algorithms include linear regression, lasso regression, regression trees and multivariate regression, while classifiers include logistic regressions, support



**Fig. 6** Categories of machine learning methods include unsupervised, supervised and semi-supervised learning. **(a)** Unsupervised learning does not contain labelled output data, and thus aims to generate a model that explains the structure within the features of the training dataset. **(b)** Supervised learning methods are trained on known output data labels, aiming to devise a model using the input features to segregate the labels (eg. X4 vs R5-using Envs). **(c)** Semi-supervised learning utilises known and unknown labels to generate a model to categorise the dataset based on the features analysed

vector machines (SVM),  $\kappa$ -nearest neighbours, naïve Bayes and decision trees [191, 196]. Ensemble methods are an extension of supervised learning that combines multiple independent machine learning algorithms into a single predictive model to improve performance [195]. One example of an ensemble method is random decision forests, which combine multiple independent decisions trees to generate a model [197].

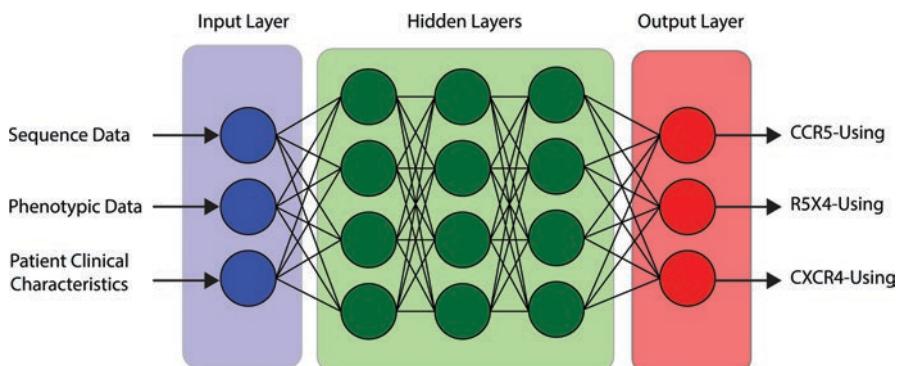
The performance of a model principally relies on the algorithm chosen and the input data to train the model. As such, the features selected play an important role in the overall performance of the model. Feature selection can be performed manually or can be determined by a plethora of different selection techniques [192]. Optimal feature selection allows for the reduction of input data to improve model performance and efficiency, and also prevents over fitting the model [195]. Feature selection techniques can be split into three categories; embedded, filter and wrapper methods (reviewed in [192]). Briefly, filter methods select features based on their correlation with predicting the right output, as such they select what they see as the most interesting features for the model [198]. While filter methods have the advantages of being time efficient, they can select features that are redundant as they do not consider relationships between different features. Wrapper methods detect interactions between different features and how those interactions change the ability to predict the correct outcome [199]. While these methods have a clear advantage over filter methods, they require greater computational time and have the risk of over fitting. Finally, embedded methods utilise the learning algorithm to perform feature selection and classification simultaneously [199]. Ideally, wrapper or embedded methods should be used for feature selection of genomic data sets evaluation the influence of different features individually and in combination on the accuracy of the model.

Machine learning approaches have previously been utilised to generate coreceptor prediction algorithms. The algorithms for CRUSH and geno2pheno were both generated using non-linear SVMs with V3 sequences as the training dataset [161, 178], while lesser known algorithms have also utilised this method for algorithm generation [180, 200–203]. SVMs are a supervised learning method that uses the training data for classification or regression analysis [191, 196]. Given a training dataset with known output labels, SVMs aim to construct a hyperplane between the known classes to achieve maximum separation between the classes (e.g. R5 or X4) [204]. Random forests are another method that has been utilised to generate coreceptor prediction algorithms [205, 206]. Random forests are a supervised learning method that consists of multiple decision trees that are combined to obtain a more accurate prediction [191]. Decision trees look similar to flow charts and consist of different nodes termed: root nodes, decision nodes or leafs and terminal nodes. The root node is where the tree begins with a question or criteria. The answers to each criteria lead to decision nodes where more criteria are assessed, while the terminal node represents the final outcome (R5-using or X4-using). For instance, Xu et al. assessed the influence of each amino acid residue within the V3 sequence to predict X4-usage using random forests [207], which demonstrated improved accuracy and precision when compared against WebPSSM for clade B, C and non-B non-C sequences. As such, methods of machine learning have been important for the creation of genotypic coreceptor prediction algorithms, allowing more accurate prediction of X4-using variants within the clinic.

## 7.2 Artificial Neural Networks and Deep Learning Approaches

Artificial neural networks were first proposed by Walter Pitts and Warren McCulloch in 1943 following the creation of a mathematic model to explain biological neural networks [208]. Training of artificial neural networks was made possible by the introduction of the concept of back propagation in 1985 by Paul Werbos [209]. These neural networks consist of several layers: the input layer, the hidden layer (middle layer) and the prediction or output layer. Here, the data in the input layer is passed through to the hidden layer where mathematical functions allow patterns in the data to be established to predict the output data (Fig. 7) [195, 210]. Training of the neural network involves multiple rounds of evaluation (termed ‘back propagation) and tuning of the hidden layer parameters until the accuracy of the network can no longer be improved [195].

Deep learning or next-generation machine learning is a subset of machine learning that has been developed in recent years [211–213]. Deep learning approaches utilise artificial neural networks that can contain thousands of hidden layers and are trained on highly complex datasets [195]. These deep learning approaches are possible due to the rapid rise in computational power. Deep learning utilises artificial neural networks to assess the input data for patterns that allow accurate prediction of output values for unknown data sets [210, 213]. This technique has demonstrated



**Fig. 7** Example of a neural network approach to HIV coreceptor usage prediction. The desired patient data including virus *Envelope* sequence data, clinical characteristics such as CD4 count and viral load at the time of sampling and the Trofile assay tested phenotype of sequences could all be processed and fed into the input later. Multiple hidden layers transform the input data into mathematical functions or ‘weights’ that allow the input data to be processed in a way that can predict the final outcome in the output later of X4-using, dual-tropic (R5X4-using) or CCR5-using

improved performance over other machine learning approaches in image and speech recognition as well as language translation [211, 212, 214].

Recent studies have utilised deep learning approaches for drug discovery approaches; for instance, studies from Ragoza et al. and Gomes et al. utilised a type of artificial neural network termed a convolutional neural network to predict binding affinity of different ligands to the protein of interest [215, 216]. Here three-dimensional crystal structure data was used as the input, allowing the network to learn the key protein-ligand interactions that correlated with binding and outperformed cheminformatics-based approaches. Deep learning algorithms work best when there is a very large data set for model training. As such, deep neural networks are a promising strategy for improving coreceptor usage prediction in HIV as they can incorporate larger training datasets to improve prediction performance.

Neural networks have previously been used in HIV research to generate coreceptor prediction algorithms [217, 218], however these studies were conducted on computers with drastically lower computational power than modern computers, and demonstrated modest accuracies of 89% and 75% [217, 218]. One advantage of the network developed by Lamers et al., is that it was the first method that could identify dual-tropic strains and separate these from X4 and R5-using strains, albeit with an accuracy of 75% [217]. Here, 149 V3 loop sequences that had known coreceptor usage data were used for feature selection. Features analysed within the amino acid sequences included charge, surface area, chemical property, mass (daltons) and the degree of hydrophobicity per amino acid position. The network generated in this study demonstrated a dual-tropic prediction accuracy of 77.4% and CCR5/CXCR4-using prediction accuracy of 73.7%. However, a disadvantage of this study is that V3 sequences instead of the whole *Env* sequence was used as the input for the network. The inclusion of full *Env* sequences is necessary to more accurately predict

coreceptor usage as mutations within regions outside of the V3 have been demonstrated to influence the ability of Envs to use CCR5 or CXCR4 [69, 181–188]. As such, the use of deep neural networks to generate a program of coreceptor usage (R5, X4 or dual-tropic) using full *Env* sequences is both plausible and recommended for improving current genotypic coreceptor usage evaluation in the clinic. Moreover, it is plausible to include structural information about Envs of interest including secondary structure prediction [219] and protein-protein interaction prediction [220] with coreceptors in addition to genetic sequences to obtain more accurate coreceptor usage prediction.

## 8 Futuristic Methods for Determining Molecular Env-Receptor Interactions

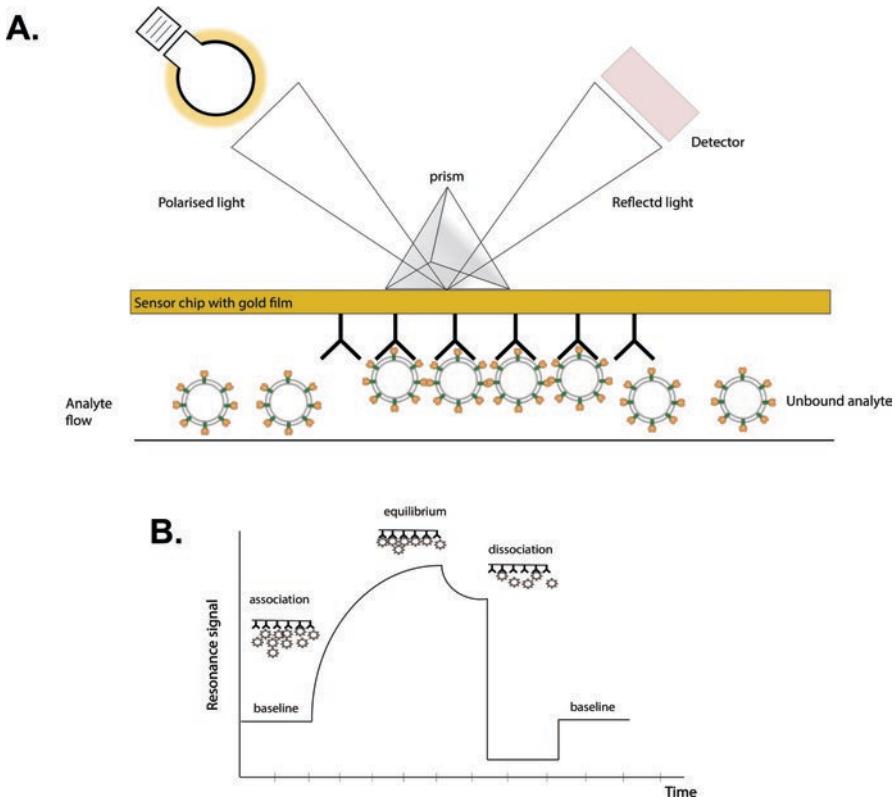
### 8.1 Surface Plasmon Resonance

The development of biosensor technology has significantly enhanced the investigation of biomolecular interactions in microbiological systems. The application of biosensor technology utilising electrochemical, electromechanical as well as optical and/or visual fluorescence biosensors are widespread; from food analysis [221] to protein engineering and drug discovery [222–225].

Surface plasmon resonance (SPR) in particular, an optical biosensor based technology, measures the resonance (oscillations) of surface electromagnetic waves known as surface plasmons [226–228]. SPR is based on the Kretschmann configuration [229] whereby incident light focussed through a glass prism (medium 1), upon striking plasmons on a non-magnetic gold film (sensor surface) at the interface of media with different refractive indices is then reflected and collected through a detector (Fig. 8). Changes in the refractive index of the second medium for instance an analyte with the addition/inclusion of foreign items such as antigens, is represented by a change in the intensity of the reflected light, which can then be monitored and analysed using a biosensor [228]. This real-time investigation of molecular interactions has been utilised extensively in biology.

In HIV research, SPR has been used to provide a comprehensive analysis of the structure and function of antibodies to various HIV antigens (reviewed in [230]). Importantly antibody-antigen interactions have been defined at different stages of the HIV life cycle advancing our understanding of the kinetics and affinity of antibody-antigen interactions as well as the thermodynamic parameters that control these interactions (Reviewed in [230]).

The early stages of the HIV lifecycle including entry, receptor engagement and receptor binding have been examined using SPR [231–234]. These studies demonstrated the utility of SPR in studying receptor-Env interactions by immobilisation of Env to the sensor surface [231–236] and were pivotal for identifying gp120 epitopes on CCR5. Limitations that were associated with receptor instability on the directly-



**Fig. 8** Schematic of SPR and sensorgram output for measurement of HIV pseudoparticles (analyte) to surface immobilised antibodies targeted to HIV Env. **(a)** Depiction of instrument set up showing the passing of polarised light from a light source through a prism onto the sensor chip coated with a gold-film where surface plasmons are generated at a critical angle to incident light. **(b)** Sensogram showing the phases during measurement. The association phase represents phase at which HIV pseudoparticles bind to immobilised Ab on the sensor surface. The equilibrium phase represents the steady state when all analyte molecules have occupied all binding sites/antibodies and the dissociation phase represents when binding sites become unoccupied, when analyte flow is disrupted. Each phase enables measurement of the rate of at which the molecules are interacting

bound sensor surface, resulted in the evolution to the antibody-capture systems that are now widespread [237]. Research by Gu et al.; [238] using SPR provided novel findings of the interaction of Env with various T cell immunoglobulins (TIM) and therefore together with earlier SPR analyses demonstrating the importance of DC-SIGN in HIV interaction with CD4 receptor in both CD4 dependent and independent systems [239], supporting a role for this tool in further characterising the role of other HIV entry factors such as the recently implicated, Siglec-1, particularly in macrophage infections [240].

More recently, the HIV vaccine field with its ever expanding broadly neutralising antibody panel [241–243], has employed the use of SPR for antibody characterisa-

tion such as that done to assess the therapeutic efficacy of the broadly neutralising antibody (bNAb) PGT121 [244]. SPR has also aided in demonstrating the importance of bNAbs in reducing transcytosis of HIV across mucosal barriers [245], a finding which has implications for HIV transmission and importantly therapy. Thus, highlighting a role for SPR in advancing HIV therapeutics.

However, while the benefits of SPR being a highly automated, non-invasive, label-free and highly sensitive affinity based technique, one minor drawback is its limited specificity [246]. Recently, Geuijen et al. [247] described a sensitive, specific and rapid multiplexed biotin-streptavidin capture SPR system for the study of Fc $\gamma$  receptor-IgG interactions that was able to measure the functionality of IgG prior and following various stress conditions, thus improving the specificity of SPR in measuring binding affinity and kinetics. Advancing from this, similar refinements using the streptavidin-avidin capture system can be adopted for investigating CD4 and coreceptor interactions with Env that may mitigate limitations surrounding the preparations of membrane/receptor complexes. These improvements in SPR will be important particularly when mimicking environments pertinent to that of the heterogenous receptor environment on the surface of macrophages in order for it to be applicable as a tool for coreceptor utilisation in the macrophage setting.

## 8.2 *BioLayer Interferometry*

BioLayer Interferometry (BLI) similar to SPR is built on the optical biosensor technology that utilizes fiber optic biosensors linked to probes to perform high throughput screening of targets in sample solutions. This technology, not dissimilar to SPR, works on the premise that interference or lack thereof during loading and offloading of samples creates an optical interference profile detected by spectrophotometers that can then be quantitated and translated to kinetics of binding [248]. However an important distinction is that BLI unlike SPR, is able to take measurements irrespective of differences in the refractive index of the solution and thus can be also be used for the assessment of crude samples such as patient serum and plasma and therefore has a more broader application.

Like SPR, BLI has also been used to characterise HIV bNAbs [249], as well as identify otherwise unknown receptors/factors that may contribute to HIV infection [250]. Importantly, the high specificity of BLI in comparison to SPR has been utilised to ascertain precise affinity interactions of Env trimers with antibodies [251] and particularly assess the impact of Env surface changes (glycan deletions/changes) on receptor interactions and their impact on macrophage tropism [252, 253]. Furthermore, BLI can also be used to ascertain receptor–Env stoichiometry assessments which are important in understanding molecular factors/predictors of tropism.

An important advantage of BLI as with SPR is their ability to be utilised to study systems controlled by allosteric mechanisms. Since Env is an allosteric ensemble whose conformational states are dictated by receptor binding, employing BLI to

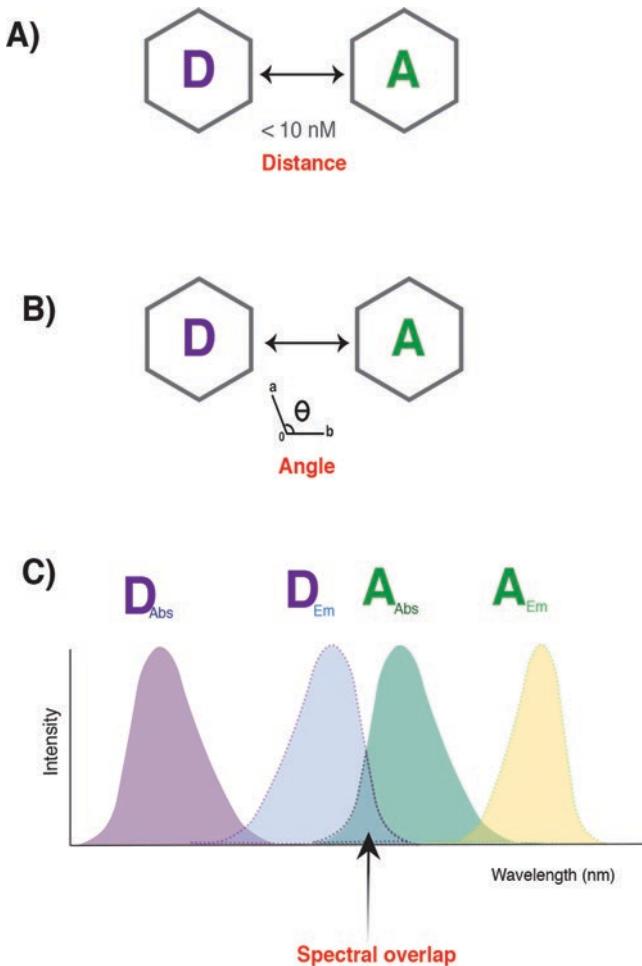
investigate not only the binding kinetics of CXCR4 and CCR5 to Env but also their influence on the important and precise conformational changes in Env brought on by these interactions would add to current understanding of the transitional intermediates primed for fusion and entry [254]. Further to determining the kinetics of antibody binding to gp120 [255, 256], BLI has more recently been used to [257] assess whether increased resistance to undergo CD4-induced transitions were due to low affinity of CD4, supporting a more expanded use of this tool in assessing resistance patterns and impact of affinity measurements. BLI and SPR are powerful tools that will enable the precise real time analysis of Env-receptor interactions to contribute to current understanding of coreceptor tropism and have recently been reviewed to be important technologies for the advancement of HIV vaccine development [258].

### 8.3 FRET and GlycoFRET

Förster (Fluorescence) resonance energy transfer (FRET) is a non-radiation dependent technique that is used for measuring nanometer scale distances between two molecules or points within a macromolecule, called a donor-acceptor pair [259]. The technique works on the phenomenon that the donor-fluorophore, upon excitation, is able to transfer energy to the acceptor group, given that both have complementary resonance frequencies or spectral profiles (Fig. 8). The result of this transfer manifests as a sequential decline of fluorescence of the donor and an increase in that of the acceptor group. This coupled effect can be measured using a fluorimeter or microscopically and displayed as ratio metric output.

The exclusive dependence on resonance means that FRET techniques are indifferent/insensitive to radiative energy transfer properties (ie photon emissions, optical properties) such as those required for SPR and BLI, however is highly dependent on: the distance between donor-acceptor ( $\leq 10$  nm); the orientation of donor-acceptor which is key when observing complex allosteric mechanisms such as Env-receptor interactions; and importantly the extent of spectral overlap between the donor-acceptor pairs (Fig. 9). FRET is a powerful tool with applications ranging from the study of protein-protein interactions [260], enzyme kinetics [261, 262] and importantly for Env-receptor interactions, can also be used for assessing the dynamics of protein conformations [259, 263, 264].

The study of molecular mechanisms of HIV has benefitted from the use of FRET [265–268], with single molecule FRET (smFRET) analysis used recently to demonstrate the now widely accepted assertion that unliganded Env transitions through three conformational states [269, 270]. This finding dramatically opened up the field to understanding the conformational dynamics associated with Env-receptor interactions [270], and lead to further results elucidating [269] the stoichiometry of Env-CD4 interactions pertaining/responding to these conformational states (ie.



**Fig. 9** Requirements for FRET. For FRET to be achieved, there are three main criteria that need to be met. (a) The distance between donor [purple D] and acceptor [green A] molecules must be no more than 10 nm. (b) The angle of orientation between the donor and acceptor must be compatible and (c) there must be spectral overlap between the fluorophores used

intermediate state corresponds to one-CD4 bound Env while the state 3 corresponds to Env-bound to 4 CD4s). Such findings highlight the rational design of futuristic assays to investigate Env-receptor interactions, demonstrating the strength of FRET as a tool to investigate the often complex and highly dynamic molecular interactions inherent with HIVs Env proteins.

Other developments in FRET technology include Stockmann and colleague's GlycoFRET assay [271], which exploits the inherent nature of cell surface receptors to undergo glycosylation by directly conjugating glycosylated receptors through inherent metabolic processing, with the highly spectrally efficient terbium-labelled

reporters-called metabolic glycan engineering. The process therefore has the benefits of avoiding the use of complex genetic engineering for the production of donor and acceptor molecules and furthermore exploits the inherent high signal-noise ratios of time resolved FRET analysis afforded by labels from the lanthanide series. Importantly this technique bypasses the need for the use of antibodies to aid analysis [271].

Because both CXCR4 and CCR5 are glycosylated, Glyco-FRET can be used to assess the degree of interaction of these coreceptors with Env and depending on the location of the label, and the presence and/or level of energy transfer, the magnitude of receptor-Env interaction can be inferred. Furthermore, the ease of these labels to penetrate cells and be incorporated into cell surface glycans during post-translational modification means that this assay could also be utilised to define whether different cell types/subsets express different receptor forms and their likely contribution to gp120-CKR interactions.

## 9 Conclusion

Identifying the mechanisms determining HIV receptor engagement and cellular tropism is an important and fast advancing field. Current techniques have proven vital in aiding our understanding and assisting the development of entry inhibitors, however further work is critical to investigate the role of Env-receptor interactions for improving therapeutic efficacy and the development of a vaccine. Advanced technologies encompassing sophisticated machine learning and neural network approaches would deliver enhanced coreceptor prediction algorithms, for more accurate prediction of R5, R5X4 and X4-using variants within the clinic. Additionally, rational design of assays to investigate env-receptor interactions using more advanced techniques such as SPR, BLI and glyco-FRET will be vital for the development of future entry inhibitor targeted therapeutics and assisting vaccine design.

**Acknowledgements** MG was supported by an RMIT PhD Scholarship. The authors thank Dr Andrew Guy, School of Science, RMIT University for his helpful input and discussions on the machine learning section of this chapter.

## References

1. Finkel TH, Tudor-Williams G, Banda NK, Cotton MF, Curiel T, Monks C, et al. Apoptosis occurs predominantly in bystander cells and not in productively infected cells of HIV- and SIV-infected lymph nodes. *Nat Med.* 1995;1(2):129–34.
2. Phair J, Jacobson L, Detels R, Rinaldo C, Saah A, Schrager L, et al. Acquired immune deficiency syndrome occurring within 5 years of infection with human immunodeficiency virus type-1: the Multicenter AIDS Cohort Study. *J Acquir Immune Defic Syndr.* 1992;5(5):490–6.

3. Munoz A, Wang MC, Bass S, Taylor JM, Kingsley LA, Chmiel JS, et al. Acquired immunodeficiency syndrome (AIDS)-free time after human immunodeficiency virus type 1 (HIV-1) seroconversion in homosexual men. Multicenter AIDS Cohort Study Group. *Am J Epidemiol.* 1989;130(3):530–9.
4. Hendriks JC, Medley GF, van Griensven GJ, Coutinho RA, Heisterkamp SH, van Druten HA. The treatment-free incubation period of AIDS in a cohort of homosexual men. *AIDS* (London, England). 1993;7(2):231–9.
5. Hendriks JC, Satten GA, van Ameijden EJ, van Druten HA, Coutinho RA, van Griensven GJ. The incubation period to AIDS in injecting drug users estimated from prevalent cohort data, accounting for death prior to an AIDS diagnosis. *AIDS* (London, England). 1998;12(12):1537–44.
6. Mellors JW, Rinaldo CR Jr, Gupta P, White RM, Todd JA, Kingsley LA. Prognosis in HIV-1 infection predicted by the quantity of virus in plasma. *Science.* 1996;272(5265):1167–70.
7. Finzi D, Hermankova M, Pierson T, Carruth LM, Buck C, Chaisson RE, et al. Identification of a reservoir for HIV-1 in patients on highly active antiretroviral therapy. *Science.* 1997;278(5341):1295–300.
8. Finzi D, Blankson J, Siliciano JD, Margolick JB, Chadwick K, Pierson T, et al. Latent infection of CD4+ T cells provides a mechanism for lifelong persistence of HIV-1, even in patients on effective combination therapy. *Nat Med.* 1999;5(5):512–7.
9. Wong JK, Hezareh M, Gunthard HF, Havlir DV, Ignacio CC, Spina CA, et al. Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science.* 1997;278(5341):1291–5.
10. Kwong PD, Wyatt R, Robinson J, Sweet RW, Sodroski J, Hendrickson WA. Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature.* 1998;393(6686):648–59.
11. Wyatt R, Kwong PD, Desjardins E, Sweet RW, Robinson J, Hendrickson WA, et al. The antigenic structure of the HIV gp120 envelope glycoprotein. *Nature.* 1998;393(6686):705–11.
12. Wyatt R, Sodroski J. The HIV-1 envelope glycoproteins: fusogens, antigens, and immunogens. *Science.* 1998;280(5371):1884–8.
13. Arthos J, Cicala C, Martinelli E, Macleod K, Van Ryk D, Wei D, et al. HIV-1 envelope protein binds to and signals through integrin alpha4beta7, the gut mucosal homing receptor for peripheral T cells. *Nat Immunol.* 2008;9(3):301–9.
14. Saphire AC, Bobardt MD, Zhang Z, David G, Gallay PA. Syndecans serve as attachment receptors for human immunodeficiency virus type 1 on macrophages. *J Virol.* 2001;75(19):9187–200.
15. Geijtenbeek TB, Kwon DS, Torensma R, van Vliet SJ, van Duijnoven GC, Middel J, et al. DC-SIGN, a dendritic cell-specific HIV-1-binding protein that enhances trans-infection of T cells. *Cell.* 2000;100(5):587–97.
16. Chen B, Vogan EM, Gong H, Skehel JJ, Wiley DC, Harrison SC. Structure of an unliganded simian immunodeficiency virus gp120 core. *Nature.* 2005;433(7028):834–41.
17. Huang CC, Tang M, Zhang MY, Majeed S, Montabana E, Stanfield RL, et al. Structure of a V3-containing HIV-1 gp120 core. *Science.* 2005;310(5750):1025–8.
18. Cormier EG, Dragic T. The crown and stem of the V3 loop play distinct roles in human immunodeficiency virus type 1 envelope glycoprotein interactions with the CCR5 coreceptor. *J Virol.* 2002;76(17):8953–7.
19. Kwon YD, Finzi A, Wu X, Dogo-Isonagie C, Lee LK, Moore LR, et al. Unliganded HIV-1 gp120 core structures assume the CD4-bound conformation with regulation by quaternary interactions and variable loops. *Proc Natl Acad Sci U S A.* 2012;109(15):5663–8.
20. Gallo SA, Finnegan CM, Viard M, Raviv Y, Dimitrov A, Rawat SS, et al. The HIV Env-mediated fusion reaction. *Biochim Biophys Acta.* 2003;1614(1):36–50.
21. Lyumkis D, Julien JP, de Val N, Cupo A, Potter CS, Klasse PJ, et al. Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer. *Science.* 2013;342(6165):1484–90.

22. Gallo SA, Puri A, Blumenthal R. HIV-1 gp41 six-helix bundle formation occurs rapidly after the engagement of gp120 by CXCR4 in the HIV-1 Env-mediated fusion process. *Biochemistry*. 2001;40(41):12231–6.
23. Markosyan RM, Leung MY, Cohen FS. The six-helix bundle of human immunodeficiency virus Env controls pore formation and enlargement and is initiated at residues proximal to the hairpin turn. *J Virol*. 2009;83(19):10048–57.
24. Tabler CO, Lucera MB, Haqqani AA, McDonald DJ, Migueles SA, Connors M, et al. CD4(+) memory stem cells are infected by HIV-1 in a manner regulated in part by SAMHD1 expression. *J Virol*. 2014;88(9):4976–86.
25. Bleul CC, Wu L, Hoxie JA, Springer TA, Mackay CR. The HIV coreceptors CXCR4 and CCR5 are differentially expressed and regulated on human T lymphocytes. *Proc Natl Acad Sci U S A*. 1997;94(5):1925–30.
26. Lee B, Sharron M, Montaner LJ, Weissman D, Doms RW. Quantification of CD4, CCR5, and CXCR4 levels on lymphocyte subsets, dendritic cells, and differentially conditioned monocyte-derived macrophages. *Proc Natl Acad Sci U S A*. 1999;96(9):5215–20.
27. Cashin K, Paukovics G, Jakobsen MR, Ostergaard L, Churchill MJ, Gorry PR, et al. Differences in coreceptor specificity contribute to alternative tropism of HIV-1 subtype C for CD4(+) T-cell subsets, including stem cell memory T-cells. *Retrovirology*. 2014;11:97.
28. Flynn JK, Paukovics G, Cashin K, Born K, Ellett A, Roche M, et al. Quantifying susceptibility of CD4+ stem memory T-cells to infection by laboratory adapted and clinical HIV-1 strains. *Viruses*. 2014;6(2):709–26.
29. Jobe O, Trinh HV, Kim J, Alsalmi W, Tovanabutra S, Ehrenberg PK, et al. Effect of cytokines on Siglec-1 and HIV-1 entry in monocyte-derived macrophages: the importance of HIV-1 envelope V1V2 region. *J Leukoc Biol*. 2016;99(6):1089–106.
30. Deng H, Liu R, Ellmeier W, Choe S, Unutmaz D, Burkhardt M, et al. Identification of a major co-receptor for primary isolates of HIV-1. *Nature*. 1996;381(6584):661–6.
31. Huang Y, Paxton WA, Wolinsky SM, Neumann AU, Zhang L, He T, et al. The role of a mutant CCR5 allele in HIV-1 transmission and disease progression. *Nat Med*. 1996;2(11):1240–3.
32. Yi Y, Isaacs SN, Williams DA, Frank I, Schols D, De Clercq E, et al. Role of CXCR4 in cell-cell fusion and infection of monocyte-derived macrophages by primary human immunodeficiency virus type 1 (HIV-1) strains: two distinct mechanisms of HIV-1 dual tropism. *J Virol*. 1999;73(9):7117–25.
33. Berger EA, Doms RW, Fenyo EM, Korber BT, Littman DR, Moore JP, et al. A new classification for HIV-1. *Nature*. 1998;391(6664):240.
34. Gorry PR, Bristol G, Zack JA, Ritola K, Swanstrom R, Birch CJ, et al. Macrophage tropism of human immunodeficiency virus type 1 isolates from brain and lymphoid tissues predicts neurotropism independent of coreceptor specificity. *J Virol*. 2001;75(21):10073–89.
35. Musich T, O'Connell O, Gonzalez-Perez MP, Derdeyn CA, Peters PJ, Clapham PR. HIV-1 non-macrophage-tropic R5 envelope glycoproteins are not more tropic for entry into primary CD4+ T-cells than envelopes highly adapted for macrophages. *Retrovirology*. 2015;12:25.
36. Flynn JK, Paukovics G, Moore MS, Ellett A, Gray LR, Duncan R, et al. The magnitude of HIV-1 resistance to the CCR5 antagonist maraviroc may impart a differential alteration in HIV-1 tropism for macrophages and T-cell subsets. *Virology*. 2013;442(1):51–8.
37. Gray L, Sterjovski J, Churchill M, Ellery P, Nasr N, Lewin SR, et al. Uncoupling coreceptor usage of human immunodeficiency virus type 1 (HIV-1) from macrophage tropism reveals biological properties of CCR5-restricted HIV-1 isolates from patients with acquired immunodeficiency syndrome. *Virology*. 2005;337(2):384–98.
38. Gray L, Roche M, Churchill MJ, Sterjovski J, Ellett A, Poumbourios P, et al. Tissue-specific sequence alterations in the human immunodeficiency virus type 1 envelope favoring CCR5 usage contribute to persistence of dual-tropic virus in the brain. *J Virol*. 2009;83(11):5430–41.
39. Parker ZF, Iyer SS, Wilen CB, Parrish NF, Chikere KC, Lee FH, et al. Transmitted/founder and chronic HIV-1 envelope proteins are distinguished by differential utilization of CCR5. *J Virol*. 2013;87(5):2401–11.

40. Keele BF, Giorgi EE, Salazar-Gonzalez JF, Decker JM, Pham KT, Salazar MG, et al. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. *Proc Natl Acad Sci U S A*. 2008;105(21):7552–7.
41. Margolis L, Shattock R. Selective transmission of CCR5-utilizing HIV-1: the ‘gatekeeper’ problem resolved? *Nat Rev Microbiol*. 2006;4:312–7.
42. van’t Wout AB, Kootstra NA, Mulder-Kampinga GA, Albrecht van Lent N. Macrophage-tropic variants initiate Human Immunodeficiency Virus type 1 infection after sexual, parenteral and vertical transmission. *J Clin Invest*. 1994;94:2060–7.
43. Shaw GM, Hunter E. HIV transmission. *Cold Spring Harb Perspect Med*. 2012;2(11):a006965.
44. Samson M, Libert F, Doranz BJ, Rucker J, Liesnard C, Farber CM, et al. Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. *Nature*. 1996;382(6593):722–5.
45. Schacker T, Little S, Connick E, Gebhard K, Zhang Z, Krieger J, et al. Productive infection of T cells in lymphoid tissues during primary and early Human Immunodeficiency Virus infection. *J Infect Dis*. 2001;183(4):555–62.
46. Pope M, Betjes MGH, Romani N, Hirmand H, Cameron PU, Hoffman L, et al. Conjugates of dendritic cells and memory T lymphocytes from skin facilitate productive infection with HIV-1. *Cell*. 1994;78:389–98.
47. Cameron PU, Freudenthal PS, Barker JM, Gezelter S, Inaba K, Steinman RM. Dendritic cells exposed to Human Immunodeficiency Virus type-1 transmit a vigorous cytopathic infection to CD4+ T cells. *Science*. 1992;257:383–7.
48. Cohen MS, Shaw GM, McMichael AJ, Haynes BF. Acute HIV-1 infection. *N Engl J Med*. 2011;364:1943–54.
49. Huang W, Toma J, Stawiski E, Fransen S, Wrin T, Parkin N, et al. Characterization of human immunodeficiency virus type 1 populations containing CXCR4-using variants from recently infected individuals. *AIDS Res Hum Retrovir*. 2009;25(8):795–802.
50. Ochsenbauer C, Edmonds TG, Ding H, Keele BF, Decker J, Salazar MG, et al. Generation of transmitted/founder HIV-1 infectious molecular clones and characterization of their replication capacity in CD4 T lymphocytes and monocyte-derived macrophages. *J Virol*. 2012;86(5):2715–28.
51. Wilen CB, Parrish NF, Pfaff JM, Decker JM, Henning EA, Haim H, et al. Phenotypic and immunologic comparison of clade B transmitted/founder and chronic HIV-1 envelope glycoproteins. *J Virol*. 2011;85(17):8514–27.
52. Jakobsen MR, Ellett A, Churchill MJ, Gorry PR. Viral tropism, fitness and pathogenicity of HIV-1 subtype C. *Futur Virol*. 2010;5(2):219–31.
53. Koot M, Keet IP, Vos AH, de Goede RE, Roos MT, Coutinho RA, et al. Prognostic value of HIV-1 syncytium-inducing phenotype for rate of CD4+ cell depletion and progression to AIDS. *Ann Intern Med*. 1993;118(9):681–8.
54. Connor RI, Sheridan KE, Ceradini D, Choe S, Landau NR. Change in coreceptor use correlates with disease progression in HIV-1--infected individuals. *J Exp Med*. 1997;185(4):621–8.
55. Klitzmann D, Barre-Sinoussi F, Nugeyre MT, Danquet C, Vilmer E, Griscelli C, et al. Selective tropism of lymphadenopathy associated virus (LAV) for helper-inducer T lymphocytes. *Science*. 1984;225(4657):59–63.
56. Masur H, Ognibene FP, Yarchoan R, Shelhamer JH, Baird BF, Travis W, et al. CD4 counts as predictors of opportunistic pneumonias in human immunodeficiency virus (HIV) infection. *Ann Intern Med*. 1989;111(3):223–31.
57. Siliciano JD, Kajdas J, Finzi D, Quinn TC, Chadwick K, Margolick JB, et al. Long-term follow-up studies confirm the stability of the latent reservoir for HIV-1 in resting CD4+ T cells. *Nat Med*. 2003;9(6):727–8.
58. Doitsh G, Galloway NLK, Geng X, Yang Z, Monroe KM, Zepeda O, et al. Cell death by pyroptosis drives CD4 T-cell depletion in HIV-1 infection. *Nature*. 2014;505(7484):509–14.
59. Cummins NW, Badley AD. Mechanisms of HIV-associated lymphocyte apoptosis: 2010. *Cell Death Dis*. 2010;1:e99.

60. Richard J, Prevost J, Baxter AE, von Bredow B, Ding S, Medjahed H, et al. Uninfected bystander cells impact the measurement of HIV-specific antibody-dependent cellular cytotoxicity responses. *MBio*. 2018;9(2):e00358–18.
61. Richard J, Veillette M, Ding S, Zoubchenok D, Alsahafi N, Couto M, et al. Small CD4 mimetics prevent HIV-1 uninfected bystander CD4 + T cell killing mediated by antibody-dependent cell-mediated cytotoxicity. *EBioMedicine*. 2016;3:122–34.
62. Moore JP, McKeating JA, Weiss RA, Sattentau QJ. Dissociation of gp120 from HIV-1 virions induced by soluble CD4. *Science*. 1990;250(4984):1139–42.
63. Mehandru S, Poles MA, Tenner-Racz K, Horowitz A, Hurley A, Hogan C, et al. Primary HIV-1 infection is associated with preferential depletion of CD4+ T lymphocytes from effector sites in the gastrointestinal tract. *J Exp Med*. 2004;200(6):761–70.
64. Guadalupe M, Reay E, Sankaran S, Prindiville T, Flamm J, McNeil A, et al. Severe CD4+ T-cell depletion in gut lymphoid tissue during primary human immunodeficiency virus type 1 infection and substantial delay in restoration following highly active antiretroviral therapy. *J Virol*. 2003;77(21):11708–17.
65. Fevrier M, Dorgham K, Rebollo A. CD4+ T cell depletion in human immunodeficiency virus (HIV) infection: role of apoptosis. *Viruses*. 2011;3(5):586–612.
66. Cossarizza A, Ortolani C, Mussini C, Borghi V, Guaraldi G, Mongiardo N, et al. Massive activation of immune cells with an intact T cell repertoire in acute human immunodeficiency virus syndrome. *J Infect Dis*. 1995;172(1):105–12.
67. Norris PJ, Pappalardo BL, Custer B, Spotts G, Hecht FM, Busch MP. Elevations in IL-10, TNF-alpha, and IFN-gamma from the earliest point of HIV Type 1 infection. *AIDS Res Hum Retrovir*. 2006;22(8):757–62.
68. Parrish NF, Wilen CB, Banks LB, Iyer SS, Pfaff JM, Salazar-Gonzalez JF, et al. Transmitted/founder and chronic subtype C HIV-1 use CD4 and CCR5 receptors with equal efficiency and are not inhibited by blocking the integrin alpha4beta7. *PLoS Pathog*. 2012;8(5):e1002686.
69. Jakobsen MR, Cashin K, Roche M, Sterjovski J, Ellett A, Borm K, et al. Longitudinal analysis of CCR5 and CXCR4 usage in a cohort of antiretroviral therapy-naïve subjects with progressive HIV-1 subtype C infection. *PLoS One*. 2013;8(6):e65950.
70. Gray LR, Roche M, Flynn JK, Wesselingh SL, Gorry PR, Churchill MJ. Is the central nervous system a reservoir of HIV-1? *Curr Opin HIV AIDS*. 2014;9(6):552–8.
71. Siliciano RF, Greene WC. HIV latency. *Cold Spring Harb Perspect Med*. 2011;1(1):a007096.
72. Chavez L, Calvanese V, Verdin E. HIV latency is established directly and early in both resting and activated primary CD4 T cells. *PLoS Pathog*. 2015;11(6):e1004955.
73. Shan L, Deng K, Gao H, Xing S, Capoferra AA, Durand CM, et al. Transcriptional reprogramming during effector-to-memory transition renders CD4(+) T cells permissive for latent HIV-1 infection. *Immunity*. 2017;47(4):766–75.e3.
74. Chomont N, El-Far M, Ancuta P, Trautmann L, Procopio FA, Yassine-Diab B, et al. HIV reservoir size and persistence are driven by T cell survival and homeostatic proliferation. *Nat Med*. 2009;15(8):893–900.
75. Soriano-Sarabia N, Bateson RE, Dahl NP, Crooks AM, Kuruc JD, Margolis DM, et al. Quantitation of replication-competent HIV-1 in populations of resting CD4+ T cells. *J Virol*. 2014;88(24):14070–7.
76. Buzon MJ, Sun H, Li C, Shaw A, Seiss K, Ouyang Z, et al. HIV-1 persistence in CD4+ T cells with stem cell-like properties. *Nat Med*. 2014;20(2):139–42.
77. Jaafoura S, de Goer de Herve MG, Hernandez-Vargas EA, Hendel-Chavez H, Abdoh M, Mateo MC, et al. Progressive contraction of the latent HIV reservoir around a core of less-differentiated CD4(+) memory T Cells. *Nat Commun*. 2014;5:5407.
78. Klatt NR, Bosinger SE, Peck M, Richert-Spuhler LE, Heigle A, Gile JP, et al. Limited HIV infection of central memory and stem cell memory CD4+ T cells is associated with lack of progression in viremic individuals. *PLoS Pathog*. 2014;10(8):e1004345.
79. Hoeffel G, Ginhoux F. Fetal monocytes and the origins of tissue-resident macrophages. *Cell Immunol*. 2018;330:5–15.

80. Epelman S, Levine Kory J, Randolph GJ. Origin and functions of tissue macrophages. *Immunity*. 2014;41(1):21–35.
81. Mass E. Delineating the origins, developmental programs and homeostatic functions of tissue-resident macrophages. *Int Immunol*. 2018;30(11):493–501.
82. Ginhoux F, Guilliams M. Tissue-resident macrophage ontogeny and homeostasis. *Immunity*. 2016;44(3):439–49.
83. Munro DAD, Hughes J. The origins and functions of tissue-resident macrophages in kidney development. *Front Physiol*. 2017;8:837.
84. Mosser DM, Edwards JP. Exploring the full spectrum of macrophage activation. *Nat Rev Immunol*. 2008;8:958.
85. Rojas J, Salazar J, Martinez MS, Palmar J, Bautista J, Chavez-Castillo M, et al. Macrophage heterogeneity and plasticity: impact of macrophage biomarkers on atherosclerosis. *Scientifica*. 2015;2015:851252.
86. Gordon S, Martinez-Pomares L. Physiological roles of macrophages. *Pflugers Arch Eur J Physiol*. 2017;469(3–4):365–74.
87. Grainger JR, Konkel JE, Zangerle-Murray T, Shaw TN. Macrophages in gastrointestinal homeostasis and inflammation. *Arch Eur J Physiol*. 2017;469(3):527–39.
88. Bain CC, Mowat AM. Macrophages in intestinal homeostasis and inflammation. *Immunol Rev*. 2014;260(1):102–17.
89. Koppensteiner H, Brack-Werner R, Schindler M. Macrophages and their relevance in Human Immunodeficiency Virus Type I infection. *Retrovirology*. 2012;9:82.
90. Laguette N, Sobhian B, Casartelli N, Ringeard M, Chable-Bessia C, Segéral E, et al. SAMHD1 is the dendritic- and myeloid-cell-specific HIV-1 restriction factor counteracted by Vpx. *Nature*. 2011;474(7353):654–7.
91. Flynn JK, Gorry PR. In: Shapshak PSJ, Somboonwit C, Kuhn JH, editors. Role of macrophages in the immunopathogenesis of HIV-1 infection. New York: Springer; 2015.
92. Martinez FO, Gordon S. The M1 and M2 paradigm of macrophage activation: time for reassessment. *F1000Prime Rep*. 2014;6:13.
93. Clayton KL, Garcia JV, Clements JE, Walker BD. HIV infection of macrophages: implications for pathogenesis and cure. *Pathog Immun*. 2017;2(2):179–92.
94. Porcheray F, Samah B, Leone C, Dereuddre-Bosquet N, Gras G. Macrophage activation and human immunodeficiency virus infection: HIV replication directs macrophages towards a pro-inflammatory phenotype while previous activation modulates macrophage susceptibility to infection and viral production. *Virology*. 2006;349(1):112–20.
95. Herbein G, Varin A. The macrophage in HIV-1 infection: from activation to deactivation? *Retrovirology*. 2010;7:33.
96. Paiardini M, Muller-Trutwin M. HIV-associated chronic immune activation. *Immunol Rev*. 2013;254(1):78–101.
97. Aquaro S, Bagnarelli P, Guenci T, De Luca A, Clementi M, Balestra E, et al. Long-term survival and virus production in human primary macrophages infected by human immunodeficiency virus. *J Med Virol*. 2002;68(4):479–88.
98. Cassol E, Rossouw T, Malfeld S, Mahasha P, Slavik T, Seebregts C, et al. CD14(+) macrophages that accumulate in the colon of African AIDS patients express pro-inflammatory cytokines and are responsive to lipopolysaccharide. *BMC Infect Dis*. 2015;15:430.
99. Clayton KL, Collins DR, Lengieza J, Ghebremichael M, Dotiwala F, Lieberman J, et al. Resistance of HIV-infected macrophages to CD8(+) T lymphocyte-mediated killing drives activation of the immune system. *Nat Immunol*. 2018;19(5):475–86.
100. Crowe SM, Westhorpe CL, Mukhamedova N, Jaworowski A, Sviridov D, Bukrinsky M. The macrophage: the intersection between HIV infection and atherosclerosis. *J Leukoc Biol*. 2010;87(4):589–98.
101. Saylor D, Dickens AM, Sacktor N, Haughey N, Slusher B, Pletnikov M, et al. HIV-associated neurocognitive disorder--pathogenesis and prospects for treatment. *Nat Rev Neurol*. 2016;12(4):234–48.

102. Tuttle DL, Harrison JK, Anders C, Slesman JW, Goodenow MM. Expression of CCR5 increases during monocyte differentiation and directly mediates macrophage susceptibility to infection by human immunodeficiency virus type 1. *J Virol.* 1998;72(6):4962–9.
103. Lewin SR, Sonza S, Irving LB, McDonald CF, Mills J, Crowe SM. Surface CD4 is critical to *in vitro* HIV infection of human alveolar macrophages. *AIDS Res Hum Retrovir.* 1996;12(10):877–83.
104. Flynn JK, Gorry PR. Stem memory T cells (TSCM)-their role in cancer and HIV immunotherapies. *Clin Transl Immunol.* 2014;3(7):e20.
105. Gorry PR, Francella N, Lewin SR, Collman RG. HIV-1 envelope-receptor interactions required for macrophage infection and implications for current HIV-1 cure strategies. *J Leukoc Biol.* 2014;95(1):71–81.
106. Peters PJ, Bhattacharya J, Hibbitts S, Dittmar MT, Simmons G, Bell J, et al. Biological analysis of human immunodeficiency virus type 1 R5 envelopes amplified from brain and lymph node tissues of AIDS patients with neuropathology reveals two distinct tropism phenotypes and identifies envelopes in the brain that confer an enhanced tropism and fusigenicity for macrophages. *J Virol.* 2004;78(13):6915–26.
107. Peters PJ, Sullivan WM, Duenas-Decamp MJ, Bhattacharya J, Ankghuambom C, Brown R, et al. Non-macrophage-tropic human immunodeficiency virus type 1 R5 envelopes predominate in blood, lymph nodes, and semen: implications for transmission and pathogenesis. *J Virol.* 2006;80(13):6324–32.
108. Gorry PR, Taylor J, Holm GH, Mehle A, Morgan T, Cayabyab M, et al. Increased CCR5 affinity and reduced CCR5/CD4 dependence of a neurovirulent primary human immunodeficiency virus type 1 isolate. *J Virol.* 2002;76(12):6277–92.
109. Dunfee RL, Thomas ER, Gorry PR, Wang J, Taylor J, Kunstman K, et al. The HIV Env variant N283 enhances macrophage tropism and is associated with brain infection and dementia. *Proc Natl Acad Sci U S A.* 2006;103(41):15160–5.
110. Murray PJ, Wynn TA. Protective and pathogenic functions of macrophage subsets. *Nat Rev Immunol.* 2011;11(11):723–37.
111. Gordon S, Taylor PR. Monocyte and macrophage heterogeneity. *Nat Rev Immunol.* 2005;5(12):953–64.
112. Jambo KC, Banda DH, Kankwatira AM, Sukumar N, Allain TJ, Heyderman RS, et al. Small alveolar macrophages are infected preferentially by HIV and exhibit impaired phagocytic function. *Mucosal Immunol.* 2014;7(5):1116–26.
113. Mitis E, Kamng’ona R, Rylance J, Solorzano C, Jesus Reine J, Mwandumba HC, et al. Human alveolar macrophages predominantly express combined classical M1 and M2 surface markers in steady state. *Respir Res.* 2018;19(1):66.
114. Burdo TH, Lackner A, Williams KC. Monocyte/macrophages and their role in HIV neuropathogenesis. *Immunol Rev.* 2013;254(1):102–13.
115. Joseph SB, Arrildt KT, Sturdevant CB, Swanstrom R. HIV-1 target cells in the CNS. *J Neurovirol.* 2015;21(3):276–89.
116. He J, Chen Y, Farzan M, Choe H, Ohagen A, Gartner S, et al. CCR3 and CCR5 are co-receptors for HIV-1 infection of microglia. *Nature.* 1997;385(6617):645–9.
117. Lavi E, Strizki JM, Ulrich AM, Zhang W, Fu L, Wang Q, et al. CXCR-4 (Fusin), a co-receptor for the type 1 human immunodeficiency virus (HIV-1), is expressed in the human brain in a variety of cell types, including microglia and neurons. *Am J Pathol.* 1997;151(4):1035–42.
118. Williams KC, Corey S, Westmoreland SV, Pauley D, Knight H, deBakker C, et al. Perivascular macrophages are the primary cell type productively infected by simian immunodeficiency virus in the brains of macaques: implications for the neuropathogenesis of AIDS. *J Exp Med.* 2001;193(8):905–15.
119. Bell JE. The neuropathology of adult HIV infection. *Rev Neurol (Paris).* 1998;154(12):816–29.
120. Best BM, Letendre SL, Koopmans P, Rossi SS, Clifford DB, Collier AC, et al. Low cerebrospinal fluid concentrations of the nucleotide HIV reverse transcriptase inhibitor, tenofovir. *J Acquir Immune Defic Syndr.* 2012;59(4):376–81.

121. Kumar A, Abbas W, Herbein G. HIV-1 latency in monocytes/macrophages. *Viruses*. 2014;6(4):1837–60.
122. Chun TW, Fauci AS. Latent reservoirs of HIV: obstacles to the eradication of virus. *Proc Natl Acad Sci U S A*. 1999;96(20):10958–61.
123. Gras G, Kaul M. Molecular mechanisms of neuroinvasion by monocytes-macrophages in HIV-1 infection. *Retrovirology*. 2010;7:30.
124. Nath A, Clements JE. Eradication of HIV from the brain: reasons for pause. *AIDS*. 2011;25(5):577–80.
125. Alexaki A, Liu Y, Wigdahl B. Cellular reservoirs of HIV-1 and their role in viral persistence. *Curr HIV Res*. 2008;6(5):388–400.
126. Igarashi T, Brown CR, Endo Y, Buckler-White A, Plishka R, Bischofberger N, et al. Macrophage are the principal reservoir and sustain high virus loads in rhesus macaques after the depletion of CD4+ T cells by a highly pathogenic simian immunodeficiency virus/HIV type 1 chimera (SHIV): implications for HIV-1 infections of humans. *Proc Natl Acad Sci U S A*. 2001;98(2):658–63.
127. Brown D, Mattapallil JJ. Gastrointestinal tract and the mucosal macrophage reservoir in HIV infection. *Clin Vaccine Immunol*. 2014;21(11):1469–73.
128. Costiniuk CT, Jenabian MA. The lungs as anatomical reservoirs of HIV infection. *Rev Med Virol*. 2014;24(1):35–54.
129. King DF, Siddiqui AA, Buffa V, Fischetti L, Gao Y, Stieh D, et al. Mucosal tissue tropism and dissemination of HIV-1 subtype B acute envelope-expressing chimeric virus. *J Virol*. 2013;87(2):890–9.
130. McElrath MJ, Smythe K, Randolph-Habecker J, Melton KR, Goodpaster TA, Hughes SM, et al. Comprehensive assessment of HIV target cells in the distal human gut suggests increasing HIV susceptibility toward the anus. *J Acquir Immune Defic Syndr*. 2013;63(3):263–71.
131. Li L, Meng G, Graham MF, Shaw GM, Smith PD. Intestinal macrophages display reduced permissiveness to human immunodeficiency virus 1 and decreased surface CCR5. *Gastroenterology*. 1999;116(5):1043–53.
132. Shen R, Meng G, Ochsenbauer C, Clapham PR, Grams J, Novak L, et al. Stromal down-regulation of macrophage CD4/CCR5 expression and NF- $\kappa$ B activation mediates HIV-1 non-permissiveness in intestinal macrophages. *PLoS Pathog*. 2011;7(5):e1002060.
133. Sharova N, Swingler C, Sharkey M, Stevenson M. Macrophages archive HIV-1 virions for dissemination in trans. *EMBO J*. 2005;24(13):2481–9.
134. Chu H, Wang JJ, Qi M, Yoon JJ, Wen X, Chen X, et al. The intracellular virus-containing compartments in primary human macrophages are largely inaccessible to antibodies and small molecules. *PLoS One*. 2012;7(5):e35297.
135. Groot F, Welsch S, Sattentau QJ. Efficient HIV-1 transmission from macrophages to T cells across transient virological synapses. *Blood*. 2008;111(9):4660–3.
136. Waki K, Freed EO. Macrophages and cell-cell spread of HIV-1. *Viruses*. 2010;2(8):1603–20.
137. Gousset K, Ablan SD, Coren LV, Ono A, Soheilian F, Nagashima K, et al. Real-time visualization of HIV-1 GAG trafficking in infected macrophages. *PLoS Pathog*. 2008;4(3):e1000015.
138. Hammonds JE, Beeman N, Ding L, Takushi S, Francis AC, Wang JJ, et al. Siglec-1 initiates formation of the virus-containing compartment and enhances macrophage-to-T cell transmission of HIV-1. *PLoS Pathog*. 2017;13(1):e1006181.
139. Castellano P, Prevedel L, Eugenin EA. HIV-infected macrophages and microglia that survive acute infection become viral reservoirs by a mechanism involving Bim. *Sci Rep*. 2017;7(1):12866.
140. Cribbs SK, Lennox J, Caliendo AM, Brown LA, Guidot DM. Healthy HIV-1-infected individuals on highly active antiretroviral therapy harbor HIV-1 in their alveolar macrophages. *AIDS Res Hum Retrovir*. 2015;31(1):64–70.
141. Dagleish AG, Beverley PC, Clapham PR, Crawford DH, Greaves MF, Weiss RA. The CD4 (T4) antigen is an essential component of the receptor for the AIDS retrovirus. *Nature*. 1984;312(5996):763–7.

142. Feng Y, Broder CC, Kennedy PE, Berger EA. HIV-1 entry cofactor: functional cDNA cloning of a seven-transmembrane, G protein-coupled receptor. *Science*. 1996;272(5263):872–7.
143. Dorr P, Westby M, Dobbs S, Griffin P, Irvine B, Macartney M, et al. Maraviroc (UK-427,857), a potent, orally bioavailable, and selective small-molecule inhibitor of chemokine receptor CCR5 with broad-spectrum anti-human immunodeficiency virus type 1 activity. *Antimicrob Agents Chemother*. 2005;49(11):4721–32.
144. Lalezari JP, Henry K, O’Hearn M, Montaner JS, Piliero PJ, Trottier B, et al. Enfuvirtide, an HIV-1 fusion inhibitor, for drug-resistant HIV infection in North and South America. *N Engl J Med*. 2003;348(22):2175–85.
145. Roche M, Born K, Flynn JK, Lewin SR, Churchill MJ, Gorry PR. Molecular gymnastics: mechanisms of HIV-1 resistance to CCR5 antagonists and impact on virus phenotypes. *Curr Top Med Chem*. 2016;16(10):1091–106.
146. Gulick RM, Lalezari J, Goodrich J, Clumeck N, DeJesus E, Horban A, et al. Maraviroc for previously treated patients with R5 HIV-1 infection. *N Engl J Med*. 2008;359(14):1429–41.
147. Fichtenheuer G, Nelson M, Lazzarin A, Konourina I, Hoepelman AI, Lampiris H, et al. Subgroup analyses of maraviroc in previously treated R5 HIV-1 infection. *N Engl J Med*. 2008;359(14):1442–55.
148. Cooper DA, Heera J, Ive P, Botes M, Dejesus E, Burnside R, et al. Efficacy and safety of maraviroc vs. efavirenz in treatment-naïve patients with HIV-1: 5-year findings. *AIDS* (London, England). 2014;28(5):717–25.
149. Cooper DA, Heera J, Goodrich J, Tawadrous M, Saag M, Dejesus E, et al. Maraviroc versus efavirenz, both in combination with zidovudine-lamivudine, for the treatment of antiretroviral-naïve subjects with CCR5-tropic HIV-1 infection. *J Infect Dis*. 2010;201(6):803–13.
150. Roche M, Jakobsen MR, Sterjovski J, Ellett A, Posta F, Lee B, et al. HIV-1 escape from the CCR5 antagonist maraviroc associated with an altered and less-efficient mechanism of gp120-CCR5 engagement that attenuates macrophage tropism. *J Virol*. 2011;85(9):4330–42.
151. Roche M, Salimi H, Duncan R, Wilkinson BL, Chikere K, Moore MS, et al. A common mechanism of clinical HIV-1 resistance to the CCR5 antagonist maraviroc despite divergent resistance levels and lack of common gp120 resistance mutations. *Retrovirology*. 2013;10:43.
152. Tilton JC, Wilen CB, Didigu CA, Sinha R, Harrison JE, Agrawal-Gamse C, et al. A maraviroc-resistant HIV-1 with narrow cross-resistance to other CCR5 antagonists depends on both N-terminal and extracellular loop domains of drug-bound CCR5. *J Virol*. 2010;84(20):10863–76.
153. Westby M, Smith-Burchnell C, Mori J, Lewis M, Mosley M, Stockdale M, et al. Reduced maximal inhibition in phenotypic susceptibility assays indicates that viral strains resistant to the CCR5 antagonist maraviroc utilize inhibitor-bound receptor for entry. *J Virol*. 2007;81(5):2359–71.
154. Flynn JK, Ellenberg P, Duncan R, Ellett A, Zhou J, Sterjovski J, et al. Analysis of clinical HIV-1 strains with resistance to maraviroc reveals strain-specific resistance mutations, variable degrees of resistance, and minimal cross-resistance to other CCR5 antagonists. *AIDS Res Hum Retrovir*. 2017;33(12):1220–35.
155. Reeves J, Coakley E, Petropoulos C, Whitcomb J. An enhanced sensitivity Trofile HIV coreceptor tropism assay for selecting patients for therapy with entry inhibitors targeting CCR5: a review of analytical and clinical studies. *J Viral Entry*. 2009;3(3):94–102.
156. Trinh L, Han D, Huang W, Wrin T, Larson J, Kiss L, et al. Validation of an enhanced sensitivity Trofile™ HIV-1 co-receptor tropism assay for selecting patients for therapy with entry inhibitors targeting CCR5. *J Int AIDS Soc*. 2008;11(1):P197.
157. Su Z, Gulick RM, Krambrink A, Coakley E, Hughes MD, Han D, et al. Response to vicriviroc in treatment-experienced subjects using an enhanced sensitivity co-receptor tropism assay: reanalysis of AIDS Clinical Trials Group A5211. *J Infect Dis*. 2009;200(11):1724–8.
158. Toma J, Frantzell A, Hoh R, Martin J, Deeks S, Petropoulos C, et al., editors. Determining HIV-1 co-receptor tropism using PBMC proviral DNA derived from aviremic blood samples. The 17th conference on retroviruses and opportunistic infections (CROI) San Francisco; 2010.

159. Fouchier RA, Groenink M, Kootstra NA, Tersmette M, Huisman HG, Miedema F, et al. Phenotype-associated sequence variation in the third variable domain of the human immunodeficiency virus type 1 gp120 molecule. *J Virol*. 1992;66(5):3183–7.
160. De Jong JJ, De Ronde A, Keulen W, Tersmette M, Goudsmit J. Minimal requirements for the human immunodeficiency virus type 1 V3 domain to support the syncytium-inducing phenotype: analysis by single amino acid substitution. *J Virol*. 1992;66(11):6777–80.
161. Lengauer T, Sander O, Sierra S, Thielen A, Kaiser R. Bioinformatics prediction of HIV coreceptor usage. *Nat Biotechnol*. 2007;25(12):1407–10.
162. Hwang SS, Boyle TJ, Lyerly HK, Cullen BR. Identification of the envelope V3 loop as the primary determinant of cell tropism in HIV-1. *Science*. 1991;253(5015):71–4.
163. Raymond S, Delobel P, Mavigner M, Ferradini L, Cazabat M, Souyris C, et al. Prediction of HIV type 1 subtype C tropism by genotypic algorithms built from subtype B viruses. *J Acquir Immune Defic Syndr*. 2010;53(2):167–75.
164. Sanchez V, Masia M, Robledano C, Padilla S, Ramos JM, Gutierrez F. Performance of genotypic algorithms for predicting HIV-1 tropism measured against the enhanced-sensitivity Trofile coreceptor tropism assay. *J Clin Microbiol*. 2010;48(11):4135–9.
165. Garrido C, Roulet V, Chueca N, Poveda E, Aguilera A, Skrabal K, et al. Evaluation of eight different bioinformatics tools to predict viral tropism in different human immunodeficiency virus type 1 subtypes. *J Clin Microbiol*. 2008;46(3):887–91.
166. Cashin K, Gray LR, Jakobsen MR, Sterjovski J, Churchill MJ, Gorry PR. CoRSeqV3-C: a novel HIV-1 subtype C specific V3 sequence based coreceptor usage prediction algorithm. *Retrovirology*. 2013;10:24.
167. Raymond S, Delobel P, Rogez S, Encinas S, Bruel P, Pasquier C, et al. Genotypic prediction of HIV-1 CRF01-AE tropism. *J Clin Microbiol*. 2013;51(2):564–70.
168. Jensen MA, Li FS, van 't Wout AB, Nickle DC, Shriner D, He XH, et al. Improved coreceptor usage prediction and genotypic monitoring of R5-to-X4 transition by motif analysis of human immunodeficiency virus type 1 env V3 loop sequences. *J Virol*. 2003;77(24):13376–88.
169. Soulie C, Derache A, Aime C, Marcellin AG, Carcelain G, Simon A, et al. Comparison of two genotypic algorithms to determine HIV-1 tropism. *HIV Med*. 2008;9(1):1–5.
170. Jensen MA, Coetzer M, van 't Wout AB, Morris L, Mullins JI. A reliable phenotype predictor for human immunodeficiency virus type 1 subtype C based on envelope V3 sequences. *J Virol*. 2006;80(10):4698–704.
171. Low AJ, Dong W, Chan D, Sing T, Swanstrom R, Jensen M, et al. Current V3 genotyping algorithms are inadequate for predicting X4 co-receptor usage in clinical isolates. *AIDS* (London, England). 2007;21(14):F17–24.
172. Trabaud MA, Icard V, Scholtes C, Perpoint T, Koffi J, Cotte L, et al. Discordance in HIV-1 coreceptor use prediction by different genotypic algorithms and phenotype assay: intermediate profile in relation to concordant predictions. *J Med Virol*. 2012;84(3):402–13.
173. Kalu AW, Telele NF, Gebreselasie S, Fekade D, Abdurahman S, Marrone G, et al. Prediction of coreceptor usage by five bioinformatics tools in a large Ethiopian HIV-1 subtype C cohort. *PLoS One*. 2017;12(8):e0182384.
174. Delgado E, Fernandez-Garcia A, Vega Y, Cuevas T, Pinilla M, Garcia V, et al. Evaluation of genotypic tropism prediction tests compared with in vitro co-receptor usage in HIV-1 primary isolates of diverse subtypes. *J Antimicrob Chemother*. 2012;67(1):25–31.
175. Recordon-Pinson P, Soulie C, Flandre P, Descamps D, Lazrek M, Charpentier C, et al. Evaluation of the genotypic prediction of HIV-1 coreceptor use versus a phenotypic assay and correlation with the virological response to maraviroc: the ANRS GenoTropism study. *Antimicrob Agents Chemother*. 2010;54(8):3335–40.
176. Seclen E, Garrido C, Gonzalez Mdel M, Gonzalez-Lahoz J, de Mendoza C, Soriano V, et al. High sensitivity of specific genotypic tools for detection of X4 variants in antiretroviral-experienced patients suitable to be treated with CCR5 antagonists. *J Antimicrob Chemother*. 2010;65(7):1486–92.

177. Cashin K, Gray LR, Harvey KL, Perez-Bercoff D, Lee GQ, Sterjovski J, et al. Reliable genotypic tropism tests for the major HIV-1 subtypes. *Sci Rep.* 2015;5:8543.
178. Kieslich CA, Tamamis P, Guzman YA, Onel M, Floudas CA. Highly accurate structure-based prediction of HIV-1 coreceptor usage suggests intermolecular interactions driving tropism. *PLoS One.* 2016;11(2):e0148974.
179. Poveda E, Briz V, Roulet V, Del Mar Gonzalez M, Faudon JL, Skrabal K, et al. Correlation between a phenotypic assay and three bioinformatic tools for determining HIV co-receptor use. *AIDS (London, England).* 2007;21(11):1487–90.
180. Skrabal K, Low AJ, Dong W, Sing T, Cheung PK, Mammano F, et al. Determining human immunodeficiency virus coreceptor use in a clinical setting: degree of correlation between two phenotypic assays and a bioinformatic model. *J Clin Microbiol.* 2007;45(2):279–84.
181. Hoffman NG, Seillier-Moiseiwitsch F, Ahn J, Walker JM, Swanstrom R. Variability in the human immunodeficiency virus type 1 gp120 Env protein linked to phenotype-associated changes in the V3 loop. *J Virol.* 2002;76(8):3852–64.
182. Dimonte S, Babakir-Mina M, Mercurio F, Di Pinto D, Ceccherini-Silberstein F, Svicher V, et al. Selected amino acid changes in HIV-1 subtype-C gp41 are associated with specific gp120(V3) signatures in the regulation of co-receptor usage. *Virus Res.* 2012;168(1–2):73–83.
183. Dimonte S, Mercurio F, Svicher V, D'Arrigo R, Perno CF, Ceccherini-Silberstein F. Selected amino acid mutations in HIV-1 B subtype gp41 are associated with specific gp120v(3) signatures in the regulation of co-receptor usage. *Retrovirology.* 2011;8:33.
184. Boyd MT, Simpson GR, Cann AJ, Johnson MA, Weiss RA. A single amino acid substitution in the V1 loop of human immunodeficiency virus type 1 gp120 alters cellular tropism. *J Virol.* 1993;67(6):3649–52.
185. Thielen A, Lengauer T, Swenson LC, Dong WW, McGovern RA, Lewis M, et al. Mutations in gp41 are correlated with coreceptor tropism but do not improve prediction methods substantially. *Antivir Ther.* 2011;16(3):319–28.
186. Huang W, Toma J, Fransen S, Stawiski E, Reeves JD, Whitcomb JM, et al. Coreceptor tropism can be influenced by amino acid substitutions in the gp41 transmembrane subunit of human immunodeficiency virus type 1 envelope protein. *J Virol.* 2008;82(11):5584–93.
187. Huang W, Eshleman SH, Toma J, Fransen S, Stawiski E, Paxinos EE, et al. Coreceptor tropism in human immunodeficiency virus type 1 subtype D: high prevalence of CXCR4 tropism and heterogeneous composition of viral populations. *J Virol.* 2007;81(15):7885–93.
188. Coetzer M, Nedellec R, Cilliers T, Meyers T, Morris L, Mosier DE. Extreme genetic divergence is required for coreceptor switching in HIV-1 subtype C. *J Acquir Immune Defic Syndr.* 2011;56(1):9–15.
189. Diez-Fuertes F, Delgado E, Vega Y, Fernandez-Garcia A, Cuevas MT, Pinilla M, et al. Improvement of HIV-1 coreceptor tropism prediction by employing selected nucleotide positions of the env gene in a Bayesian network classifier. *J Antimicrob Chemother.* 2013;68(7):1471–85.
190. Schriener DR, Kern AD. Supervised machine learning for population genetics: a new paradigm. *Trends Genet.* 2018;34(4):301–12.
191. Tarca AL, Carey VJ, Chen XW, Romero R, Draghici S. Machine learning and its applications to biology. *PLoS Comput Biol.* 2007;3(6):e116.
192. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England).* 2007;23(19):2507–17.
193. Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. *Nat Rev Genet.* 2015;16(6):321–32.
194. Kumari S, Chouhan U, Suryawanshi SK. Machine learning approaches to study HIV/AIDS infection: a review. *Biosci Biotechnol Res.* 2017;10(1):34–43.
195. Camacho DM, Collins KM, Powers RK, Costello JC, Collins JJ. Next-generation machine learning for biological networks. *Cell.* 2018;173(7):1581–92.
196. Yip KY, Cheng C, Gerstein M. Machine learning and genome annotation: a match meant to be? *Genome Biol.* 2013;14(5):205.

197. Ho TK, editor. Random decision forests. Document analysis and recognition, 1995, proceedings of the third international conference on; 1995: IEEE.
198. Torkkola K. Feature extraction by non parametric mutual information maximization. *J Mach Learn Res.* 2003;3:1415–38.
199. Boritz EA, Darko S, Swaszek L, Wolf G, Wells D, Wu X, et al. Multiple origins of virus persistence during natural control of HIV infection. *Cell.* 2016;166(4):1004–15.
200. Thielen A, Sichtig N, Kaiser R, Lam J, Harrigan PR, Lengauer T. Improved prediction of HIV-1 coreceptor usage with sequence information from the second hypervariable loop of gp120. *J Infect Dis.* 2010;202(9):1435–43.
201. Boisvert S, Marchand M, Laviolette F, Corbeil J. HIV-1 coreceptor usage prediction without multiple alignments: an application of string kernels. *Retrovirology.* 2008;5(1):110.
202. Pillai S, Good B, Richman D, Corbeil J. A new perspective on V3 phenotype prediction. *AIDS Res Hum Retrovir.* 2003;19(2):145–9.
203. Kumar R, Raghava GP. Hybrid approach for predicting coreceptor used by HIV-1 from its V3 loop amino acid sequence. *PLoS One.* 2013;8(4):e61437.
204. Bzdok D, Krzywinski M, Altman N. Machine learning: supervised methods. *Nat Methods.* 2018;15:5.
205. Dybowski JN, Heider D, Hoffmann D. Prediction of co-receptor usage of HIV-1 from genotype. *PLoS Comput Biol.* 2010;6(4):e1000743.
206. Bozek K, Lengauer T, Sierra S, Kaiser R, Domingues FS. Analysis of physicochemical and structural properties determining HIV-1 coreceptor usage. *PLoS Comput Biol.* 2013;9(3):e1002977.
207. Xu S, Huang X, Xu H, Zhang C. Improved prediction of coreceptor usage and phenotype of HIV-1 based on combined features of V3 loop sequence using random forest. *J Microbiol.* 2007;45(5):441–6.
208. McCulloch WS, Pitts W. A logical calculus of the ideas immanent in nervous activity. *Bull Math Biophys.* 1943;5(4):115–33.
209. Werbos PJ. Beyond regression: new tools for prediction and analysis in the behavioral science. Thesis (Ph. D.). Appl. Math. Harvard University, January 1974.
210. Angermueller C, Parnamaa T, Parts L, Stegle O. Deep learning for computational biology. *Mol Syst Biol.* 2016;12(7):878.
211. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Published in Advances in Neural Information Processing Systems 25 edited by F. Pereira and C.J.C. Burges and L. Bottou and K.Q. Weinberger. 2012. <https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks>
212. Graves A, Mohamed A-R, Hinton GE. Speech recognition with deep recurrent neural networks. Published in IEEE International Conference on Acoustics, Speech and Signal Processing, 26–31th May 2013, Vancouver, BC, Canada. <https://ieeexplore.ieee.org/document/6638947> <https://doi.org/10.1109/ICASSP>
213. Park Y, Kellis M. Deep learning for regulatory genomics. *Nat Biotechnol.* 2015;33(8):825–6.
214. Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. In arXiv. preprint arXiv:14090473. 2014. <https://arxiv.org/pdf/1409.0473.pdf>
215. Ragoza M, Hochuli J, Idrobo E, Sunseri J, Koes DR. Protein-ligand scoring with convolutional neural networks. *J Chem Inf Model.* 2017;57(4):942–57.
216. Gomes J, Ramsundar B, Feinberg NE, Pande V. Atomic convolutional networks for predicting protein-ligand binding affinity. 2017. In arXiv:1703.10603. <https://arxiv.org/pdf/1703.10603.pdf>
217. Lamers S, Salemi M, McGrath M, Fogel G. Prediction of R5, X4, and R5X4 HIV-1 coreceptor usage with evolved neural networks. *IEEE/ACM Trans Comput Biol Bioinform.* 2008;5(2):291–300.
218. Resch W, Hoffman N, Swanstrom R. Improved success of phenotype prediction of the human immunodeficiency virus type 1 from envelope variable loop 3 sequence using neural networks. *Virology.* 2001;288(1):51–62.

219. Sun T, Zhou B, Lai L, Pei J. Sequence-based prediction of protein protein interaction using a deep-learning algorithm. *BMC Bioinformatics.* 2017;18(1):277.
220. Wang S, Peng J, Ma J, Xu J. Protein secondary structure prediction using deep convolutional neural fields. *Sci Rep.* 2016;6:18962.
221. Luong JH, Bouvrette P, Male KB. Developments and applications of biosensors in food analysis. *Trends Biotechnol.* 1997;15(9):369–77.
222. Yu D, Blankert B, Viré JC, Kauffmann JM. Biosensors in drug discovery and drug analysis. *Anal Lett.* 2005;38(11):1687–701.
223. Cooper MA. Optical biosensors in drug discovery. *Nat Rev Drug Discov.* 2002;1(7):515–28.
224. Vigneshvar S, Sudhakumari CC, Senthilkumaran B, Prakash H. Recent advances in biosensor technology for potential applications – an overview. *Front Bioeng Biotechnol.* 2016;4:11.
225. Keusgen M. Biosensors: new approaches in drug discovery. *Naturwissenschaften.* 2002;89(10):433–44.
226. Zhang J, Zhang L. Nanostructures for surface plasmons. *Adv Opt Photon.* 2012;4(2):157–321.
227. Pitarka JM, Silkin VM, Chulkov EV, Echenique PM. Theory of surface plasmons and surface-plasmon polaritons. *Rep Prog Phys.* 2007;70(1):1.
228. Tang Y, Zeng X, Liang J. Surface plasmon resonance: an introduction to a surface spectroscopy technique. *J Chem Educ.* 2010;87(7):742–6.
229. Kretschmann E. Decay of non radiative surface plasmons into light on rough silver films. Comparison of experimental and theoretical results. *Opt Commun.* 1972;6(2):185–7.
230. Rich RL, Myszka DG. Spying on HIV with SPR. *Trends Microbiol.* 2003;11(3):124–33.
231. Cormier EG, Persuh M, Thompson DA, Lin SW, Sakmar TP, Olson WC, et al. Specific interaction of CCR5 amino-terminal domain peptides containing sulfotyrosines with HIV-1 envelope glycoprotein gp120. *Proc Natl Acad Sci U S A.* 2000;97(11):5762–7.
232. Hoffman TL, Canziani G, Jia L, Rucker J, Doms RW. A biosensor assay for studying ligand-membrane receptor interactions: binding of antibodies and HIV-1 Env to chemokine receptors. *Proc Natl Acad Sci U S A.* 2000;97(21):11215–20.
233. Misumi S, Nakajima R, Takamune N, Shoji S. A cyclic dodecapeptide-multiple-antigen peptide conjugate from the undecapeptidyl arch (from Arg(168) to Cys(178)) of extracellular loop 2 in CCR5 as a novel human immunodeficiency virus type 1 vaccine. *J Virol.* 2001;75(23):11614–20.
234. Stenlund P, Babcock GJ, Sodroski J, Myszka DG. Capture and reconstitution of G protein-coupled receptors on a biosensor surface. *Anal Biochem.* 2003;316(2):243–50.
235. Brigham-Burke M, Edwards JR, O'Shannessy DJ. Detection of receptor-ligand interactions using surface plasmon resonance: model studies employing the HIV-1 gp120/CD4 interaction. *Anal Biochem.* 1992;205(1):125–31.
236. Cormier EG, Tran DN, Yukhayeva L, Olson WC, Dragic T. Mapping the determinants of the CCR5 amino-terminal sulfopeptide interaction with soluble human immunodeficiency virus type 1 gp120-CD4 complexes. *J Virol.* 2001;75(12):5541–9.
237. Zhao H, Gorshkova II, Fu GL, Schuck P. A comparison of binding surfaces for SPR biosensing using an antibody-antigen system and affinity distribution analysis. *Methods (San Diego, Calif).* 2013;59(3):328–35.
238. Gu L, Sims B, Krendelchtkiv A, Tabengwa E, Matthews QL. Differential binding of the HIV-1 envelope to phosphatidylserine receptors. *Biochim Biophys Acta Biomembr.* 2017;1859(10):1962–6.
239. Hijazi K, Wang Y, Scala C, Jeffs S, Longstaff C, Stieh D, et al. DC-SIGN increases the affinity of HIV-1 envelope glycoprotein interaction with CD4. *PLoS One.* 2011;6(12):e28307.
240. Martin-Garcia J, Cocklin S, Chaiken IM, Gonzalez-Scarano F. Interaction with CD4 and antibodies to CD4-induced epitopes of the envelope gp120 from a microglial cell-adapted human immunodeficiency virus type 1 isolate. *J Virol.* 2005;79(11):6703–13.
241. Prigent J, Jarossay A, Planchais C, Eden C, Dufloo J, Kök A, et al. Conformational plasticity in broadly neutralizing HIV-1 antibodies triggers polyreactivity. *Cell Rep.* 2018;23(9):2568–81.

242. Derking R, Ozorowski G, Sliepen K, Yasmeen A, Cupo A, Torres JL, et al. Comprehensive antigenic map of a cleaved soluble HIV-1 envelope trimer. *PLoS Pathog.* 2015;11(3):e1004767.
243. Xu L, Pegu A, Rao E, Doria-Rose N, Beninga J, McKee K, et al. Trispecific broadly neutralizing HIV antibodies mediate potent SHIV protection in macaques. *Science.* 2017;358(6359):85–90.
244. Badamchi-Zadeh A, Tartaglia LJ, Abbink P, Bricault CA, Liu P-T, Boyd M, et al. Therapeutic efficacy of vectored PGT121 gene delivery in HIV-1-infected humanized mice. *J Virol.* 2018;92(7):e01925–17.
245. Lorin V, Malbec M, Eden C, Bruehl T, Porrot F, Seaman MS, et al. Broadly neutralizing antibodies suppress post-transcytosis HIV-1 infectivity. *Mucosal Immunol.* 2017;10(3):814–26.
246. Ahmed FE, Wiley JE, Weidner DA, Bonnerup C, Mota H. Surface Plasmon Resonance (SPR) spectrometry as a tool to analyze nucleic acid–protein interactions in crude cellular extracts. *Cancer Genomics Proteomics.* 2010;7(6):303–9.
247. Geuijen KPM, Oppers-Tiemissen C, Egging DF, Simons PJ, Boon L, Schasfoort RBM, et al. Rapid screening of IgG quality attributes – effects on Fc receptor binding. *FEBS Open Bio.* 2017;7(10):1557–74.
248. Shah NB, Duncan TM. Bio-layer interferometry for measuring kinetics of protein–protein interactions and allosteric ligand effects. *J Vis Exp.* 2014;18(84):e51383.
249. McCoy LE, van Gils MJ, Ozorowski G, Messmer T, Briney B, Voss JE, et al. Holes in the glycan shield of the native HIV envelope are a target of trimer-elicited neutralizing antibodies. *Cell Rep.* 2016;16(9):2327–38.
250. Dennison SM, Anasti KM, Jaeger FH, Stewart SM, Pollara J, Liu P, et al. Vaccine-induced HIV-1 envelope gp120 constant region 1-specific antibodies expose a CD4-inducible epitope and block the interaction of HIV-1 gp140 with galactosylceramide. *J Virol.* 2014;88(16):9406–17.
251. Dubrovskaya V, Guenaga J, de Val N, Wilson R, Feng Y, Movsesyan A, et al. Targeted N-glycan deletion at the receptor-binding site retains HIV Env NFL trimer integrity and accelerates the elicited antibody response. *PLoS Pathog.* 2017;13(9):e1006614.
252. Chabot DJ, Chen H, Dimitrov DS, Broder CC. N-linked glycosylation of CXCR4 masks coreceptor function for CCR5-dependent human immunodeficiency virus type 1 isolates. *J Virol.* 2000;74(9):4404–13.
253. Yen P-J, Herschhorn A, Haim H, Salas I, Gu C, Sodroski J, et al. Loss of a conserved N-linked glycosylation site in the simian immunodeficiency virus envelope glycoprotein V2 region enhances macrophage tropism by increasing CD4-independent cell-to-cell transmission. *J Virol.* 2014;88(9):5014–28.
254. Ozorowski G, Pallesen J, de Val N, Lyumkis D, Cottrell CA, Torres JL, et al. Open and closed structures reveal allostery and pliability in the HIV-1 envelope spike. *Nature.* 2017;547(7663):360–3.
255. Ingale J, Wyatt RT. Kinetic analysis of monoclonal antibody binding to HIV-1 gp120-derived hyperglycosylated cores. *Bio Protoc.* 2015;5(20):e1615.
256. Fera D, Schmidt AG, Haynes BF, Gao F, Liao HX, Kepler TB, et al. Affinity maturation in an HIV broadly neutralizing B-cell lineage through reorientation of variable domains. *Proc Natl Acad Sci U S A.* 2014;111(28):10275–80.
257. Kumar R, Ozorowski G, Kumar V, Holden LG, Shrivastava T, Patil S, et al. Characterization of a stable HIV-1 B/C recombinant, soluble and trimeric envelope glycoprotein (Env) highly resistant to CD4-induced conformational changes. *J Biol Chem.* 2017;292(38):15849–58.
258. Petersen RL. Strategies using bio-layer interferometry biosensor technology for vaccine research and development. *Biosensors (Basel).* 2017;7(4):49.
259. Broussard JA, Green KJ. Research techniques made simple: methodology and applications of Förster Resonance Energy Transfer (FRET) microscopy. *J Investig Dermatol.* 2017;137(11):e185–e91.
260. Piston DW, Kremers G-J. Fluorescent protein FRET: the good, the bad and the ugly. *Trends Biochem Sci.* 2007;32(9):407–14.

261. Ha T, Ting AY, Liang J, Caldwell WB, Deniz AA, Chemla DS, et al. Single-molecule fluorescence spectroscopy of enzyme conformational dynamics and cleavage mechanism. *Proc Natl Acad Sci.* 1999;96(3):893–8.
262. Shrestha D, Jenei A, Nagy P, Vereb G, Szöllösi J. Understanding FRET as a research tool for cellular studies. *Int J Mol Sci.* 2015;16(4):6718.
263. Schuler B, Lipman EA, Eaton WA. Probing the free-energy surface for protein folding with single-molecule fluorescence spectroscopy. *Nature.* 2002;419(6908):743–7.
264. Hofmann H, Hillger F, Pfeil SH, Hoffmann A, Streich D, Haenni D, et al. Single-molecule spectroscopy of protein folding in a chaperonin cage. *Proc Natl Acad Sci.* 2010;107(26):11793–8.
265. Cole GB, Reichheld SE, Sharpe S. FRET analysis of the promiscuous yet specific interactions of the HIV-1 Vpu transmembrane domain. *Biophys J.* 2017;113(9):1992–2003.
266. Takagi S, Momose F, Morikawa Y. FRET analysis of HIV-1 Gag and GagPol interactions. *FEBS Open Bio.* 2017;7(11):1815–25.
267. Schroeder S, Kaufman JD, Grunwald M, Walla PJ, Lakomek NA, Wingfield PT. HIV-1 gp41 transmembrane oligomerization monitored by FRET and FCS. *FEBS Lett.* 2018;592(6):939–48.
268. Sharma KK, Przybilla F, Restle T, Godet J, Mely Y. FRET-based assay to screen inhibitors of HIV-1 reverse transcriptase and nucleocapsid protein. *Nucleic Acids Res.* 2016;44(8):e74.
269. Ma X, Lu M, Gorman J, Terry DS, Hong X, Zhou Z, et al. HIV-1 Env trimer opens through an asymmetric intermediate in which individual protomers adopt distinct conformations. *elife.* 2018;7:e34271.
270. Munro JB, Gorman J, Ma X, Zhou Z, Arthos J, Burton DR, et al. Conformational dynamics of single HIV-1 envelope trimers on the surface of native virions. *Science.* 2014;346(6210):759–63.
271. Stockmann H, Todorovic V, Richardson PL, Marin V, Scott V, Gerstein C, et al. Cell-surface receptor-ligand interaction analysis with homogeneous time-resolved FRET and metabolic glycan engineering: application to transmembrane and GPI-anchored receptors. *J Am Chem Soc.* 2017;139(46):16822–9.

# Index

## A

*Ab initio/de novo* protein modeling, 98  
Abiogenesis, 554  
Absent-in-melanoma (AIM)-like receptors (ALRs)  
    dsDNA, 171  
    IFI16, 171  
    inflammasomes, 171  
    POPs, 172  
    viral DNA, 170  
Absorption, distribution, metabolism, excretion (ADME), 65  
Acid-citrate-dextrose (ACD), 474  
Acquired immune deficiency syndrome (AIDS), 577, 626  
Activated protein-1 (AP-1), 183  
Activation function, 497  
Active immunotherapy, 240  
Adaptive immunity, 159, 160  
Adenosine deaminase (ADA), 534  
Adenoviruses, 196  
Adoptive cells transfer (ACT), 239  
Agent-based models (ABMs), 25, 254, 256  
Agent-oriented programming languages (AOP), 284  
AI legacy  
    codes/calculations, 610  
    DAI, 609  
    ESA, 609  
    molecular docking, 610  
    shared computer resources, 610  
    Swarm intelligence, 610  
    virology, 610  
AIDS dementia complex (ADC), 578  
Allergenicity, 595

Allostasis, 47–49  
Alternating Decision Tree (ADT), 120  
American Cancer Society (ACS), 157, 219  
Amino acids (AAs), 98  
AMP-activated protein kinase (AMPK), 536  
Anaxagoras, 561  
Androgen receptor (AR), 199  
Ant Colony Optimization algorithm (ACO), 25, 28, 32  
    applications in virology, 27  
    gene subsets, 26  
Ant Colony Optimization with continuous domain algorithm (ACOR), 75  
Ant colony system (ACS), 28  
Ant System (AS), 25  
Antibody engagement, 630  
Anticipated climate catastrophe  
    atmosphere, 581  
    climate solutions multiplicity, 583–587  
    components, 581  
    health outcomes/medical emergencies, 587–596  
    interventions, 583  
    macro-environment, 583  
    meteorological variables, 581  
    oceanic zones types, 582  
    variations, 581  
Antigen-presenting cells (APCs), 43, 158, 174, 212  
Antigen-specific vaccines, 237  
Anti-idiotype vaccines, 231  
AntiRetroScan (ANS), 318  
Anti-retroviral therapy (ART), 310, 425, 535  
Anti-retrovirals (ARVs), 426, 516–518, 520, 521, 524, 577

- Anti-viral drugs, 61
- Approximate Bayesian Computation (ABC) model, 493–495
- Area under curve (AUC) score, 640
- Area under the receiver operating characteristic (AUROC), 126
- Arrhenius, S., 561
- Artificial immune system (AIS), 252
- Artificial intelligence (AI), 51, 52, 346, 557, 600, 609–611, 618, 619
- algorithms, 412
  - ANN* (*see* Artificial Neural Networks (ANN))
  - biomarkers, 522
  - case-based, 413
  - development of HAND, 523
  - machine learning, 522, 523
  - medical applications, 412
  - nanotechnology, 522
  - neurovirulence, 523
  - NN models, 414
  - production rules, 413
  - rule-based, 413
- Artificial life (AL), 253, 254
- Artificial neural networks (ANN), 9, 64, 139, 140, 143, 315, 316, 353, 416, 418, 531, 532, 537, 642
- AI, 498–499
  - architecture, 11
  - benefits of, 531
  - chest x-rays, 531
  - cluster analysis, 75
  - computing systems, 82
  - feedforward neural networks, 496–497
  - FFBP, 77, 79, 89
  - imaging, 535
  - machine learning, 530, 534
  - MLP, 496
  - model-free approach, 495
  - Mycobacterium bacilli*, 532–533
  - numerical experiments, 81
  - pleural effusions, 534
  - smear microscopy, 531
  - symptoms, 531
  - tool, 530
  - train
    - loss function, 497, 498
    - optimization, 498
  - usage, 75
- Astrobiology, 564
- Astrobiology ethics, 563
- Astrovirology, 553, 564
- Asymptomatic neurocognitive impairment (ANI), 517
- Attribute selection methods, 369
- Autoinflammatory diseases, 170
- Automated machine learning methods, 503, 504
- Autophagy, 168
- Autoregressive integrated moving average (ARIMA) model, 9
- Azidothymidine 5-triphosphate (AZTTP), 519
- B**
- Basic immune simulator (BIS), 259
- Bayesian networks (BN), 143, 315, 346
- Bayesian space-time modeling
  - AI algorithms, 600
  - air pollution, 600
  - analytical biases, 600
  - global climate change, 599
  - latent models, 600
  - medical emergencies, 599
  - mortality, 599
  - proof-of-concept, 599
  - warming reflects, 599
- Bayesian tree models, 315
- B-cells receptors (BCR), 162
- Bedaquiline, 537
- Best Multiple Linear Regression (BMLR), 127
- Big data in virology, 498, 499, 503–505
- Bioinformatics, 52
- BioLayer interferometry (BLI), 646
  - HIV bNAbs, 646
  - SPR, 646
- Biological immune systems
  - adaptive immunity, 159
  - B-cells, 162
  - cytotoxic T-cells and B-cells, 160
  - innate immunity, 158
  - microorganisms, 160
  - monocytes, 160
  - Notch genes, 163
  - PRRs, 166
  - T-cell, 162
  - T-helper cells, 164
- Biological models, 494
- Biological systems, 153
- Biomarker pattern software [BPS], 6
- Biomarkers, 534
- Biosafety
  - complexity, 547
  - contamination, 547
  - dangerous/deleterious effects, 547
  - Ebola, 547
  - Mars, 547
  - National and international space programs, 548

- NIH and CDC, 548  
panoply/plethora, 548
- Biosafety biocontainment level (BSL-4), 543,  
547, 548, 565
- Biosensor technology, 644
- Biosensors, 475–479
- Biostatistical algorithm  
multiple regression, 49, 50  
predictive model, 50
- Blood brain barrier (BBB), 47, 424, 516,  
518–522, 524
- Boltzmann, L., 555, 612
- Bone marrow transplants (BMT), 133
- Boolean expressions, 318
- Bovine Viral Diarrhea Virus (BVVDV), 333
- Brain virus, 607
- Broadly neutralising antibody (bNAb), 646
- BSL-4 laboratories, 457
- Burkitt's lymphoma, 206
- C**
- cAMP response element binding protein  
(CREB), 183
- Cancer immunotherapy, 242  
inhibitory receptors, 222  
monoclonal antibodies, 221
- Cancer stem cells (CSC), 225
- Cancerous immunity, 214
- Cancers  
DNA tumor viruses, 207  
HPV, 210  
human cancer, 208  
immune system, 210  
immunosuppression, 207  
virus, 206
- Carbon-, 555
- Case-based reasoning (CBR), 30
- Catecholamine, 474
- CD4<sup>+</sup> T cell depletion, 630
- CD4<sup>+</sup> T cell subsets, 631
- CD4<sup>+</sup> T cells, 630, 632
- CD4-interaction inhibitors, 635–636
- Cell culture, 439
- Cell marker analysis  
high-dimensional imaging, 448, 449  
multiplexed, 447, 448
- Cell surface exclusion, 327
- Cell-to-cell transmission machinery, 329
- Cell-to-cell transmission modes, 326
- Cell zones, 275
- Cellular Immune Surveillance  
BBB, 47  
FAS, 46
- immune system  
APC, 44  
components, 42  
KIR, 45  
lymphatic vessels, 43  
MHC, 43  
neutrophils, 43
- PRR, 45  
retroviral, 45, 46
- TLRs, 45  
TNF, 45  
TRAF, 46
- Cellular immunology, 265
- Cellular microenvironment, 272, 282
- Cellular parameters, 472
- Cellular tropism, 632, 634, 635, 649
- Central nervous system (CNS), 424, 578  
AI (*see* Artificial Intelligence (AI))  
BBB, 516  
future directions, 521–522
- HIV (*see* Human immunodeficiency virus  
(HIV))
- machine learning, 524
- nanotechnology, 517, 523, 524
- neurovirulence, 516
- pathogen recognition, 516
- Cervical cancer, 212, 214  
diagnosis and AI, 416  
history, 407  
low-resource settings, 406, 408, 412,  
417, 418
- LSIL, 407
- resource-limited settings, 417
- screening and AI, 414  
CDSS, 415  
CEF and NegEx, 416  
DSSS, 416  
treatment and AI, 416, 417
- Cervical carcinogenesis, 213
- Cervical cytology screening, 408, 409
- Cervical intraepithelial neoplasia (CIN), 211
- Cervical screening methods, new, 411
- Checkpoints activation, 281
- Chemiluminescence, 477
- Chemoattractants, 336
- Chemokines, 180
- Chest x-rays, 531, 535
- Chi-square test, 371
- Cholera, 593
- Chromatin assembly, repair and remodeling  
(CARR), 580
- Chronic obstructive pulmonary disease  
(COPD), 454
- Class switch recombination (CSR), 157

- Classification and Regression Tree (CART)  
models, 350
- Classifying drug targets  
druggability, 119  
*in vitro* evaluation, 119  
liver fibrosis prediction, 120  
ML algorithms, 119, 120  
PPI, 119
- Claudius, R., 611
- Climate crisis  
anticipated climate catastrophe, 581–596  
global temperature, 576  
global warming, 577  
HIV-Disease/AIDS/IRIS/Neuro-AIDS,  
577–580  
IPCC, 576  
macro-environmental balance, 576  
medical emergencies, 597–600
- Climate solutions multiplicity  
energy requirements, 584  
evidence-based interventions, 586  
individual actions, 583  
natural gas, 585  
nuclear power, 584  
plastics, 585  
politics/treaties, 584  
principal classes of policies, 583  
stakeholder feedback analysis, 586  
systematic review, 586  
TER, 586
- Clinical decision support system (CDSS), 415
- Clinical Entity Finder (CEF), 416
- CNS pathologies  
CMV, 579  
CNS lymphomas, 579  
cryptococcal meningitis, 580  
herpes zoster virus, 579  
homeostasis, 580  
immune suppression, 580  
leukoencephalopathy, 580  
macro-environmental variables, 580
- CNS penetration effectiveness (CPE), 518, 524
- CO-detection by indexing (CODEX), 445
- Coding sequences, 237
- Combination anti-retroviral therapy (cART), 626
- Combined immunotherapy, 246
- Complex adaptive systems (CAS), 152
- Complex biological systems, 152  
characteristics, 155  
definition, 153  
functions, 156  
integration processes, 156  
interactions, 156  
processes, 154
- quantification, 154  
replication components, 156  
technological, 154  
transfer processes, 156
- Complex system models, 262
- Computational model, 352
- Computational structure-based methods, 318
- Computed tomography (CT), 440
- Computer-aided geometric optimization, 60
- Computer-aided programs (CAD), 535
- Conceptual model, 271
- Conditional Random Field (CRF), 122
- Conjugative plasmids, 333, 336
- Consolidated data, 286
- Construction phase, 283
- Continuous Wavelet Transform (CWT), 76,  
77, 79–81, 84, 85, 88–90, 93
- Conventional evolutionary models, 493
- Convolution neural network (CNN), 2, 15
- Coreceptor usage, 626, 629, 631, 635, 637–639
- Correlation feature selection (CFS), 370
- Corticosteroids, 536
- Creeper virus, 607
- CRISPR/Cas adaptive immune system, 492
- Cross Validation (CV), 19
- Cross validation measures, 374
- Crossover process, 21
- Cryotherapy, 225
- Cryptococcal meningitis, 580
- Cucumber green mottle mosaic virus  
(CGMMV), 386
- Cyclin-dependent kinase (CDK), 199
- Cytokine-anti-cytokine antibody complexes, 247
- Cytokine-based immunotherapy, 234
- Cytokine induction, 181
- Cytokines, 179, 180  
Cytokines IL-4 and IL-10, 215
- Cytomegalovirus (CMV), 579, 606
- Cytoskeleton, 472
- Cytotoxic T cells (CTLs), 630
- Cytotoxic T lymphocytes (CTLs), 48
- D**
- Damage-associated molecular patterns  
(DAMPs), 165
- ‘Dark Web’, 607
- Data-driven approach, 503
- Decision support scoring system (DSSS), 416
- Decision tree (DT), 124, 125, 127, 349
- Decision tree algorithm, 6  
application in virology, 5  
leaf node, 4  
root node, 3

- Deep convolutional neural network (DCNN), 16  
Deep learning, 64, 65, *see* Machine learning  
Deep learning approaches, 642  
Deep-learning networks, 13–15  
Deep meta-architectures, 17  
Deep neural networks  
    applications in virology, 15  
Defective interfering particles, 327  
Delaunay triangulation, 351  
Dendritic cells (DCs), 228, 229, 251, 263, 272, 273, 628  
Dengue fever/dengue haemorrhagic fever (DF/DHF), 5  
Depth of the network, 496  
Design of experiments (DOE), 456  
Detection  
    AI, 557  
    Fermi-paradox, 544  
    Goldilocks paradigm, 545  
    machine replication/error, 556  
    methanol, 555  
    microfossils, 546  
Differential equations of delay (DDE), 259  
Diffusion Limited Aggregation (DLA), 76  
Digital cervicography, 410, 417  
Digital radiography (DR), 535  
Direct-acting antivirals (DAAs), 426  
Disease modeling, 454  
    COPD, 454  
    *in silico* studies, 456, 457  
    *in vitro* systems, 453  
    *in vivo* systems, 454  
    organoids, 455, 456  
    Organs-on-chips systems, 454, 455  
    virus, 455  
Distributed AI (DAI), 609  
DNA damage response (DDR), 193  
DNA methylation machinery, 193  
DNA sequencing, 563, 608, 616, 617  
DNA-PK signaling pathway, 195  
Docking-based virtual screening  
    (DB-VS), 121  
Domain attributes, 362, 376  
Double-stranded DNA (dsDNA), 169  
Drug discovery, 60–62, 64, 65  
Drug discovery approaches, 643  
Drug resistance, 311, 319, 353, 536, 537  
Druggability, 119
- E**  
Ebola outbreak, 589  
Ebola virus (EBOV), 438  
Egoistic genome, 326, 328
- Einstein, A., 611  
El Niño, 594  
Embedded methods, 372  
Emerging technologies, 60  
Enceladus, 553, 555  
Energy matrices, 101  
Ensemble methods, 312  
Enthalpy, 611  
Entropy  
    concepts, 611  
    HA protein, 614  
    influenza virus, 613  
    molecular evolution, 613  
    noise, 614, 615  
    non-equilibrium reactions, 611  
    Shannon concepts, 615–617  
    thermodynamics, 611  
Env-receptor interactions, 626, 647  
Env sequences, 643  
Epitope prediction, 388  
    B-cell, 389, 390  
    servers, 390  
    SVM, 391–392  
    T-cell, 388, 389  
Europa Lander Mission report of 2016, 559  
European Space Agency (ESA), 609  
Evolution  
    carbon-containing molecules, 550  
    Darwinian, 563  
    Europa Lander Mission report, 559  
    thermodynamics, 551  
    toxic/pathogenic, 562  
Evolutionary programming  
    definition, 99  
    NP-complete problem, 102  
    pseudocode, 102  
Evolutionary search algorithm  
    hill climbing, 102  
    HIV-1 protein sequences, 111  
    2D square lattice, 99  
Exhaustive mutational analysis, 237  
Exobiology, 547  
Experimental evolution, 487, 490, 493, 504, 507  
    biochemical processes, 492  
    CRISPR/Cas adaptive immune system, 492  
    global public health, 492  
    human microbiota, 492  
    microbial communities, 492  
    states of reproduction, 492  
Experimentation phase, 285  
Extensible Mark-up Language (XML), 608  
Extracellular mechanisms, 327

Extra-solar system investigations  
 astrobiology ethics, 563  
 astrobiology stem cells, 562, 563  
 exploration, 559  
 extra-terrestrial life  
   (see Extra-terrestrial life)  
 Mars, 562  
 neutrinos and astrovirology, 564  
 NIH/NASA budgets, 560  
 pathogen paradigms, 562  
 point-counter-point paradigms, 561  
 space exploration/potential contamination, 560  
 technology development, 556–558

Extra-terrestrial life  
 abiogenesis, 554  
 extinction events, 554  
 hidden civilization survival hypothesis, 544, 545  
 life cycles  
   inheritance/propagatory systems, 546  
   temporal spectrum, 545  
   time scale, 545  
 microfossils/isotope-radioisotope quantification, 546–549  
 organic compound survival, 550, 551  
 organic molecules, 550  
 origin, 549, 550  
 ribosomes, 551, 552  
 thermodynamics, 552, 553

Extreme environments and caves, 561

Extreme Learning Machine (ELM), 75

*Ex-vivo* manipulation, 240

**F**

Feed Forward Back Propagation (FFBP)  
 application, 77, 89  
 attributes, 93  
 classification results, 92  
 network properties, 85  
 performance graph, 91

Fermi, E.S., 544

F-factor plasmids interfere, 333

Filter ranking methods, 369

First apoptosis signal (FAS), 46

First-generation sequencing, 501

Firstpapillomavirinae, 199

Five variable loops (V1–V5), 627

Fold recognition/threading, 98

Follicular zone, 276

Förster (Fluorescence) resonance energy transfer (FRET), 647

Futuristic methods  
 agricultural industry, 503  
 big data, 503  
 data-driven research, 503  
 global public health, 503  
 machine learning methods, 503  
 model-free approach, 503  
 novel computational approaches, 503  
 supervised learning, 503, 505–507  
 unsupervised learning, 503, 507–508  
 virology, 504–505

**G**

Gammapapillomavirus, 199

Gene delivery, 330

Generalized ACO Algorithm  
 artificial ants, 27  
 initialisation, 26  
 iterations, 26

Generalized GA  
 applications in virology, 22  
 crossover process  
   multi point, 21  
   single point, 21  
   uniform, 21  
 mutation, 21  
 PLS, 23  
 replication cycle, 23  
 selection  
   tournament, 20

Generalized PSO algorithm, 29

Generative adversarial networks (GANs), 441

Genetic algorithms (GA), 125, 127  
 attributes, 18  
 crossover, 18  
 mutation, 18  
 selection, 18

Genetic algorithm–multiple linear regression (GA–MLR), 23, 127

Genetic drift, 491

Genetic vaccines, 231  
 advantage, 232

Genetics  
 biological engineering progress, 563  
 diversity, 561  
 RNA, 552

Genome-wide association studies (GWAS), 500

Genotype-phenotype pairs, 313

Genotypic-based coreceptor prediction algorithms, 637, 638

Genotypic Interpretation Systems (GIS), 317

Germinal centre zone, 276

- Gibbs equation, 553  
Gibbs, A., 555, 611, 612  
Glaciers and permafrost, 505  
Global biogeochemical processes, 506  
Global electronic ‘webs’, 607  
Glyco-FRET, 649  
Goldilocks  
  abiogenesis, 555  
  conditions, 555  
  elements, 555  
  exobiology, 555  
  groups, 555  
  paradigm, 545  
  through the looking glass paradigm, 544  
Graft versus host disease (GvHD), 133  
Gram-positive bacteria, 335  
Graphic interchange format (GIF), 608  
Graphical user interface (GUI), 284  
Gravitational lens, 557
- H**
- HA Polynomial Dataset (HAPD), 16  
Haemagglutinin (HA), 613  
Hamiltonian Monte Carlo (HMC), 494  
Health outcomes/medical emergencies  
  air pollution, 590, 595  
  allergic diseases, 595  
  *Anopheles* mosquitoes, 594  
  biological system, 589  
  cholera, 593  
  climate-related health issues, 596  
  Ebola outbreak, 589  
  ecosystem dynamics, 588  
  El Niño, 594  
  epigenetic alteration, 589  
  flooding and rainfall, 594  
  gargantuan macro-environmental  
    alterations, 588  
  heatwaves and mortality, 594  
  human microbiome, 590  
  imminent threats, 590–593  
  macro-environmental stressors, 588  
  microenvironment, 589  
  psycho-emotional/physiological stress, 596  
  public health threats, 589  
  rising temperatures, 596  
  social exclusion, 596  
  thermal environment, 596  
  vector-borne diseases, 594  
  water quality/availability, 595  
  weather events, 595  
Zika virus outbreak, 590
- Health Resources and Services Administration (HRSA), 5  
Hemagglutinin protein (HA) receptor, 8  
Hepatitis B virus (HBV)  
  chronic, 198  
  DNA vaccines, 234  
  DNMT1 and DNMT3A, 199  
  double-stranded DNA virus, 197  
  HBx protein, 197, 198  
  HepG2 cells, 197  
  infections, 198  
  proteins, 197  
  therapeutic vaccines, 234  
Hepatitis C virus (HCV), 424  
  drug designing (*see* Machine learning (ML) approaches)  
  drug discovery (*see* Machine learning (ML) algorithms)  
  NS5B polymerase, 120  
Hepatitis virus  
  ANN, 75, 82–84  
  attributes, 76  
  confusion matrices, 92  
  deterioration, 77  
  diagnosis, 75  
  experimental results, 82, 84  
  FFBP application, 92  
  methods, 78, 79  
  Modulus Maxima, 86  
  numerical experiments, 81, 82  
  patient details, 77  
  stages, 76  
  WTMM analyses, 85  
  WTMM approaches, 76  
Hepatocellular carcinoma (HCC), 208  
  cell proliferation, 208  
  development, 208  
  HBx, 209  
  infection, 208  
  MLL4 and TERT genes, 209  
Herpes simplex virus type 1 (HSV-1), 473  
Heuristic Method (HM), 127  
Hidden Markov Model (HMM), 143  
High-grade Squamous Intraepithelial Lesion (HSIL), 415  
High performance computing (HPC), 64  
Highly active antiretroviral therapy (HAART), 309, 580  
HIS-HPV16 interactions, 266  
Histological reactions, 533  
HIV-associated dementia (HAD), 517, 523, 578

- HIV-associated neurocognitive disorders (HAND), 517, 518, 522–524, 578
- HIV-associated TB, 530
- HIV CD4<sup>+</sup> T cell pathogenesis, 630
- HIV Co-receptor
- AIDS, 626
  - cART, 626
  - CCR5 or CXCR4, 627
  - CD4 receptor, 626
  - CD4<sup>+</sup> T cell, 626
  - CD4<sup>+</sup> T cells and macrophages, 626
  - mechanisms, 626
  - phenotypes, 628
- HIV-Disease/AIDS/IRIS/neuro-AIDS
- CNS pathologies, 579–580
  - metabolic status/immunity, 577
  - related pathologies, 578, 579
- HIV Drug Resistance Database (HIVDB), 309
- HIV lifecycle, 644
- HIV Neurobehavioral Research Center (HNRC), 523
- HIV prevention, 432
- HLA-peptide binding prediction
- immunological assays, 143
  - ML techniques, 140, 142
  - short peptide epitopedenign, 140
  - tools and techniques, 141
- Hodgkin's lymphoma, 206
- Homology-based modeling, 98, 318
- Hosmer-Lemeshow (H-L), 126
- Host tropism
- definition, 8
  - determinant, 8
  - vectors, 8
- Host-directed therapies, 536, 537
- HPV vaccine, prevention, 411, 412
- HPV16 domain, 264, 269, 281
- Human beings immune system, 157
- Human immune virus 1 (HIV-1)
- correlations, 111
  - evolutionary programming, 99
  - genome, 98
  - HP model, 99–103
  - modeling coarse protein structure (*see* HP model)
  - PSP methods, 98, 104–109
  - statistical analysis, 110
- Human immunodeficiency virus (HIV), 424, 486, 491, 535, 577, 606, 626
- age range, 308
  - AI (*see* Artificial Intelligence (AI))
  - AIDS pandemic, 309
  - ARVs, 309
  - categories of HAND, 517
- CPE, 518
- drug resistance, 311
- drug resistance databases, 309
- genotyping, 311
- HIVDB, 311
- infection, 309
- liposomes, 519, 520
- machine learning, 518
- ML algorithms, 311
- mortalities and infections, 309
- nanotechnology, 518, 519, 521
- NNRTIs, 314
- proteins, 309
- resistance mutations, 309
- semi-supervised learning, 313
- SLN, 518–521
- supervised learning methods, 312
- SVM, 314
- treatment of, 518
- unsupervised learning, 312
- use of ARVs, 517
- viral escape, 517
- Human leukocyte antigens (HLA)
- ML techniques, 139, 140
  - peptide binding data, 137, 138
  - peptide binding groove, 137, 138
  - peptide binding patterns, 134
  - peptide binding prediction, 140–143
  - peptide structure complexes, 134
  - short antigen peptides, 133, 134
  - super-types, 139
  - typing and alleles
  - IPD-IMGT®/HLA, 133
  - polymorphic, 133
  - serological cytotoxicity, 133
  - transplants, 133
- Human papillomavirus (HPV)
- cervical carcinogenesis, 407
  - DNA testing, 410
  - high-risk genotypes, 406
  - nucleic acid testing, 410
  - testing, 408, 409, 415, 417
  - vaccination, 406, 412, 417, 418
- Human T-cell leukemia virus (HTLV), 606
- Hydrophobic-polar (HP) model
- mathematical formulation, 102, 103
  - modeling of proteins
  - folding pattern, 99
  - straight chain, 100
  - 2D-square lattice, 99
  - scoring function
  - energy matrices, 101
  - HhPN, 101

- interactions, 101  
lattice model, 100  
search algorithm, 101  
1-[2-Hydroxyethoxy)methyl]-6-(phenylthio)-thymine (HEPT), 28  
Hypothesis-based modeling, 158
- I**
- Imaging Mass Cytometry (IMC), 446  
Immune evasion purposes, 627  
Immune reconstitution inflammatory syndrome (IRIS), 536, 577  
Immune system (IS), 156, 257  
Immune system simulator (IMMSIM), 260  
Immune-tweening, 52–54  
Immuno Polymorphism Database - ImMunoGeneTics Database/Human Leukocyte Antigen (IPD-IMGT®/HLA), 133  
Immunoanalytical approach, 477  
ImmunoGrid project, 260  
Immunohistochemistry (IHC), 446  
Immunomodulatory antibodies, 222  
Immunophenotyping, 477  
Immunotherapy, 223  
*in situ* hybridization (ISH), 446  
In-Bag data, 7  
Infectious disease imaging, AI,  
    440, 441  
    BSL-4 facility, 441  
    EBOV infection, 442, 443  
    machine-learned features, 443  
    RG-4 pathogens, 442  
Infectious diseases, 562  
Inflammasomes, 169  
Inflight analyses  
    biosensors, 479  
    cellular parameters, 472  
    deep-space exploration missions, 473  
    extreme conditions, 472  
    hypergravity, 472  
    microgravity, 472  
    monitoring and diagnostic  
        biosensors, 475–478  
        chemiluminescence, 477  
        classic sampling methods, 476  
        fiber-optic cytometer technology, 477  
        immunoanalytical approach, 477  
        immunophenotyping, 477  
        microbead-based multiplexed  
            immunoassays, 477  
        “omics” techniques, 476  
        proteins/DNA, 479
- multiplex approach, 479  
    physiological dysregulations, 471  
    spaceflight, 473–475  
    stress hormones, 473  
Influenza A virus (IAV), 332, 486, 499  
Influenza virus, 613  
Information gain, 369  
Innate immunity, 158  
Innovative approaches, 530, 537  
Integrated development environment (IDEs), 283  
Intelligent self-scattering  
    cell population, 337  
    cell surface markers, 340  
    extracellular factors, 340  
    gene delivery, 337  
    gene transmission systems, 338  
    gene vectors, 336, 338  
    MOI, 337, 338  
    non-viral vectors, 337  
    self-focusing vector, 338  
    self-scattering approaches, 340  
    strict non-admittance, 339  
    therapeutic particles, 339  
    therapeutic transgene copy number, 340  
    transgene expression dosage, 339  
Inter-agent signaling, 332  
    chemorepellents, 336  
    conjugative gene transmission, 336  
    plasmid-encoded rejection factor, 336  
Interferons type-I (IFN-I), 215  
Intergovernmental Panel on Climate Change (IPCC), 576  
International Agency for Research on Cancer (IARC), 206  
International Aids Society (IAS), 317  
International Committee on Taxonomy of Virus (ICTV), 189, 190  
International Papillomavirus Society (IPVS), 412  
International Space Station (ISS), 471, 472,  
    475, 477–479  
Intracellular homologous interference  
    mechanisms, 328  
Intra-swarm signaling, 330  
Irradiated tumor cells, 239  
Isotope effects, 550
- J**
- John Cunningham (JC) virus, 580
- K**
- Kaposi sarcoma, 206  
Keratinocytes (KCs), 159, 243, 274

Killer cell immunoglobulin receptors (KIR), 45  
 k-Nearest neighbors (KNN), 125–127, 317, 352  
 Kretschmann configuration, 644  
 Kyoto Encyclopedia of Genes and Genomes (KEGG), 608

**L**

Langerhans cells, 250  
 Large-scale pandemic models, 487  
 Laser communications, 557  
 Latency reversing agents (LRAs), 635  
 Least Square Support Vector Machine (LSSVM), 375, 415  
 Lipopolysaccharides (LPS), 175  
 Liposomes, 518–521, 524  
 Lipschitz exponent (LE), 76, 77, 80, 81, 89  
 Liver cancer, 208  
 Logistic regression model, 534  
 Long-read DNA/RNA sequences, 501, 502  
 Low-density lipoprotein (LDL), 521  
 Low-grade squamous epithelial lesions (LSIL), 407, 415  
 Lymphoid tissues zone, 275

**M**

Machine learning (ML), 346, 406, 413, 640  
 algorithms, 356  
 CART model, 350  
 CD4 cell, 348  
 DT, 349  
 optimization methods, 347  
 PR and RT inhibitors, 353  
 QOL, 349  
 RBFs, 347  
 SVM, 347  
 Machine learning (ML) algorithms  
 decision tree, 124, 125  
 genetic algorithms, 125  
 KNN classification, 125, 126  
 MLR models, 127  
 Naive Bayesian classifier, 126  
 PSO, 126  
 RF, 122, 123  
 SVM, 123  
 Machine learning (ML) approaches, 642  
 ADMET prediction, 121  
 binding site, 122  
 classifying drug targets, 119–120  
 docking/virtual screening, 122  
 novel inhibitors discovery, 120–121  
 QSAR, 118  
 secondary structure prediction, 122

Machine learning methods, 64, 499, 503  
 Machine learning systems, 311  
 classification, 312  
 data quality, 313  
 Machine-learned features, 443  
 Macrophages, 273, 633, 634  
 alveolar, 634  
 antigen, 633  
 CD4<sup>+</sup> T cells, 633, 635  
 evolution, 634  
 functions, 633  
 HIV infection, 633  
 inflammatory, 633  
 phenotype, 633  
 receptor expression, 634  
 T cells, 635  
 Magnetic resonance imaging (MRI), 440  
 Major histocompatibility complex (MHC), 16, 43, 159  
 Majorana, E., 564  
 Mammalian thymus, 163  
 Maraviroc (MVC) binds, 636  
 Margulis, L., 552  
 Markov Chain Monte Carlo (MCMC), 494  
 Mathematical and computational models, 256  
 Mathematical model, 533  
 Maximum Entropy Markov Model (MEMM), 122  
 Merkel cell carcinoma (MCC), 206, 216  
 Merkel cell polyomavirus (MCPyV)  
 ALTO protein, 205  
 clonal integration pattern, 217  
 features, 217  
 human immune system, 218  
 ICTV, 203  
 infection, 204, 216  
 MCC, 216  
 mTOR pathway, 219  
 NCRR, 204  
 PyV, 205  
 replication, 205  
 ST-antigen, 217  
 therapeutic vaccination, 235  
 tumor genome, 217  
 VP1 protein, 204  
 Message Passing Interface (MPI), 25  
 Metagenomic NGS (mNGS), 449, 450  
 Methyl-methacrylate-sulfo-propyl-methacrylate (MMA-SMP), 520, 521  
 Microarray, 396  
 Microbead-based multiplexed immunoassays, 477  
 Microenvironments, 279

- Microfossils/isotope-radioisotope quantification  
biosafety, 547–549  
robots, 546  
viruses, 549
- Mild neurocognitive disorder (MND), 517
- Miniaturization, 475
- Minimum-inhibitory-concentration (MICs), 536
- Minor histo-compatibility (miHA) antigen, 133
- Model-free approach, 495, 499, 503
- Modeling lymphoid precursors, 274
- Modulus Maxima, 86
- Molecular biology, 618, 619
- Molecular docking, 121, 354
- Molecular imaging  
CT and MRI, 443  
RG-4 pathogen, 445
- Monocytes, 160, 161
- MOTIVATE trials, 637
- Multi point crossover, 21
- Multidrug-resistant TB (MDR TB), 530
- Multifractal Analysis, 76, 79–81, 84, 92
- Multifractal Spectrum, 77, 79, 93
- Multilayer artificial neural networks (MLP), 431
- Multilayer Perceptron (MLP), 316, 496
- Multilinear regression (MLR), 121, 127
- Multiple Sclerosis (MS), 75
- Multiple sequence alignments (MSA), 24
- Multiplex approach, 479
- Multiplexed ion beam imaging (MIBI), 446
- Multiplexed tissue imaging tools  
fluorescence-based, 445  
metal tag-based, 446
- Multiplicity of infection (MOI), 329
- Murine Leukemia Virus (MLV), 332
- Mutual information, 370
- Myeloid population, 161
- N**
- NA Binary Image Dataset (NABID), 16
- Naïve Bayes model, 16
- Naive Bayesian classifier, 124, 126, 127
- Nanomaterials, 242
- Nanoparticles (NPs), 240
- Nanopore, 501, 502
- Nanotechnology, 517–524
- Nanovaccines, 242
- NARX neural networks, 12
- National Institute of Allergy and Infectious Diseases (NIAID-USA), 260
- Natural Killer (NK) cells, 44, 251
- Naturally occurring viruses, 330, 337
- Nearest Neighbor rule (NN), 125
- Needle and syringe exchange programs (NSEP), 424, 426
- Negative for Intraepithelial Lesion or Malignancy (NILM), 415
- Negative predictive value (NPV), 110
- Negentropy, 612
- Neoantigens, 236
- Neural Network algorithm, 2, 10  
applications in virology, 12  
deep, 13–15
- Neural network, graphical representation, 414
- Neural networks, 643
- Neurocognitive Impairment (NCI), 350
- Neurological manifestations of AIDS (Neuro-AIDS), 577
- New chemical entity (NCE), 62
- Next-generation sequencing (NGS), 133, 381–383, 449, 478
- NOD-like receptors (NLRs), 166  
autophagic pathway, 168  
diversity, 167  
inflammasomes, 169  
NLRP6 and NLRP12, 170  
NOD1 and NOD2, 168  
PAMPs and DAMPs, 167  
roles, 167  
signal transduction, 167
- Non-canonical signaling, 186
- Non-coding regulatory region (NCRR), 204
- Nonnucleoside reverse transcriptase inhibitor (NNRTI), 519
- Novel computational approaches, 503
- Novel inhibitors discovery  
e-pharmacophore models, 120, 121  
molecular docking, 121  
NS5B-polymerase, 120  
virtual screening, 121
- NP*-complete problem, 102
- N-terminal region, 187
- Nuclear factor kappa B (NF- $\kappa$ B), 167, 184  
activation, 185, 186  
activators, 187  
canonical pathway, 185  
constitutive activation, 187  
DNA binding sequences, 184  
enzymatic activity, 186  
extracellular matrix, 188  
gene expression, 184  
 $IKK\beta$  pathways, 185  
intrinsic/extrinsic factors, 188  
non-canonical pathway, 186  
p50 and p52 proteins, 185  
role, 187  
TCR signaling pathways, 184
- Nuclear magnetic resonance (NMR), 98

- Nucleic acid amplification (NAA), 534  
 Nucleoside reverse transcriptase inhibitor (NRTI), 519
- O**  
 Oceanic zones, 582  
 “Omics” techniques, 475, 476  
 Oncogenic viruses, 210  
 Oncolytic vaccine therapy, 226  
 Oncolytic virotherapy (OVT), 227  
 Oncoproteins, 280  
 One class SVM, 375  
 Open ocean zone types, 582  
 Open reading frame (ORF), 237  
 Opioid agonist therapy (OAT), 424  
 Optimal T-cell priming, 248  
 Organ Procurement and Transplantation Network (OPTN), 5  
 Orthohepadnavirus, 198  
 Out Of Bag (OOB) data, 120
- P**  
 Papanicolaou (Pap) test, 408  
 Papillomavirus (PVs)  
     cell culture systems, 199  
     cervical cancers, 200  
     classification, 200  
     DNA virus, 199  
     E2 protein, 202  
     E6 and E7 proteins, 202  
     E6/LCR sequences, 201  
     ICTV, 199  
     infection, 201  
     LCR region, 200  
     productive infection, 201  
 PARIMM model, 258  
 Particle Swarm Optimization (PSO), 3, 28, 32, 124, 126, 127  
     advantages, 29  
     applications in virology, 30  
     disadvantages, 29  
     generalized, 29  
 Pathogen-associated molecular patterns (PAMPs), 45  
 Pathogen detection, tissue sections, 446, 447  
 Pathological imaging, 445  
     animal models, 447  
 Patient sequencing  
     bioinformatic tools, 452  
     biomolecule quantification, 451, 452  
     mNGS, 449, 450  
     NGS, 449
- nucleic acid sequence, 450, 451  
 pathogen detection, 449, 450  
 third generation sequencing, 452, 453  
 Pattern recognition receptors (PRRs), 45, 158, 161, 165  
     DAMPs and PAMPs, 166  
 People who inject drugs (PWID)  
     CNS, 424  
     HCV, 424  
     HIV, 424  
         AI, 430  
         digital technology, 427, 428  
         epidemiology, 424, 425  
         phylogenetic/phylodynamic analyses, 429  
         prevention and treatment, 425–427  
 Peptide vaccine, 229  
 Performance measure, 373, 374  
 Peripheral nervous system (PNS), 578  
 Personal Genome Machine (PGM), 478  
 Personalized cancer vaccines, 238, 251  
 Personalized immunotherapy  
     components, 236  
     KCs, 243  
     nanoparticles, 241  
     neoantigens, 237  
     TAAs, 236  
 Personalized medicine, 476  
 Personalized vaccines, 247  
 Pharmacokinetics (PK), 121, 426  
 Phenoseq, 638  
 Phenotypic testing, 311  
 PI3K-Akt-mTOR signaling pathway, 218  
 PICOTS question, 587  
 Picture archiving and communication systems (PACS), 535  
 Plasma cells, 162  
 Plasmacytoid dendritic cells (pDCs), 175, 178  
 Poisson distribution, 337, 338  
 Polybutyl-cyanoacrylate (PBCA), 520, 521  
 Polymerase chain reaction (PCR), 133, 532  
 Polyomaviruses (PyVs), 203, 204  
 Position specific scoring matrix (PSSM), 368  
 Positive predictive value (PPV), 110  
 Positron emission tomography (PET), 440  
 Poxviridae family, 333  
 Pre-exposure prophylaxis (PrEP), 424  
 Principal component analysis (PCA), 356, 508  
 Probability distribution function (PDF), 143  
 Progressive multifocal leukoencephalopathy (PML), 580  
 Proinflammatory cytokines, 181  
 Proof-of-concept, 599  
 Protein data bank (PDB), 98

- Protein folding, 99, 100, 104, 110  
Protein–Ligand ANT System (PLANTS), 28  
Protein-ligand docking problem (PLDP), 28  
Protein structure prediction (PSP) methods  
    capsid protein (3H47), 109  
    envelope protein (PDB: 2EZ0), 109  
    integrase (5KRS), 108  
    matrix protein (1HIW), 107  
    Nef protein (1AVV), 107  
    protease (PDB 1HPV), 104  
    statistical parameters, 104, 107  
    virion infectivity factor (3DCG), 104  
    virus protein U (PDB 2N28), 104  
Protein–protein interaction (PPI), 119, 390, 393–395, 644  
Proteins/DNA, 479  
Proteochemometric (PCM)  
    model, 354  
    workflow, 355  
Psycho-neuroendocrine-osteo-immune algorithms, 53
- Q**  
Quality of life (QOL), 349  
Quantitative affinity matrices (QAM), 388  
Quantitative Matrices (QM), 143  
Quantitative real time PCR (qPCR), 396  
Quantitative structure activity relationships (QSAR), 23, 65, 118, 123, 125, 127, 143, 351, 398  
    descriptors, 367, 377  
    illustrative examples, 380–381  
    virology, 376, 377, 379, 380  
Quantum computers (QC), 557, 609, 610, 619
- R**  
Radial basis function (RBF), 347, 366  
Random forest (RF), 120, 122, 124, 127, 317, 350, 351, 642  
    advantage, 351  
    applications in virology, 8  
    features, 7  
    growth, 6  
    predictive power, 9  
    randomness, 2  
    variable selection, 7  
Rapid and cost-effective model, 534  
Rapid eye movement (REM), 48  
Raymond’s algorithm, 637–639  
Receiver operation characteristics (ROC), 5, 314  
Receptor-Env interaction, 649  
Regression algorithms, 640  
Regression function identification, 367  
Regression learning methods, 313  
Respiratory syncytial virus (RSV), 606  
Response Database Initiative (RDI) database, 352  
Retroviral Immune Surveillance, 45  
RF-based virtual screening (RB-VS), 121  
Risk Group (RG-) 4 pathogens, 438  
    cell culture methods, 439  
    EBOV, 438  
Risk Group (RG-) 4 research, 457  
RNA interference (RNAi), 521  
RNA interfering signaling complex (RISC), 580  
ROC analysis, 91  
Rule-based algorithms, 318  
Rule-based methods, 313
- S**  
Sagan, C., 550, 551  
Sanger sequencing, 501  
Schrodinger, E., 552  
Second-generation sequencing, 501  
Secondpapillomavirinae, 199  
Self-avoiding walk (SAW), 101, 103  
Self-Organizing Map (SOM), 75  
Self-scattering mechanisms, 330  
Semi-supervised learning, 640  
Semi-supervised method, 347  
Serological cytotoxicity, 133  
Shannon concepts  
    calculations, 617  
    finite fractal, 616  
    fractal dimension analysis, 616  
    Higuchi methods, 617  
    mammalian molecular sequences, 616  
    methods/discoveries, 615  
    normalization, 617  
    2D maps, 616  
Short interfering RNAs (siRNA), 522  
Signal transduction knowledge environment (STKE), 608  
Silicon, 555  
Simulation-based methods, 495  
Single molecule FRET (smFRET) analysis, 647  
Single Nucleotide Polymorphisms (SNPs), 125  
Single photon emission computed tomography (SPECT), 440  
Single point crossover, 21  
Single-Molecule-Real-Time (SMRT) sequencing, 501  
Single-nucleotide polymorphisms (SNPs), 426  
Singularity, 80, 85  
Sirtuin 1 (SIRT1), 536  
Smear microscopy, 531

- Solid lipid nanoparticles (SLN), 518–521  
 Somatic hypermutation (SHM), 157, 280  
 Stanford drug resistance database, 348  
 Statistical mechanics, 612, 613  
 Statistical models, 499  
 Stigmergy, 330  
 Strict non-admittance mechanisms, 327, 328  
 Submicron lipid emulsions (SLEs), 519  
 Sum-of-pair (SP) function, 24  
 Superinfection interference, 326, 327
  - circulating agents, 332
  - intra-swarm signaling, 330
  - RNA, 332
  - strict non-admittance, 330
  - viral and plasmid genomes, 330
  - viral/plasmid particles/genomes, 331
 Supervised learning, 362, 503, 505–507, 640  
 Supervised learning methods, 312, 640  
 Support vector machine (SVM), 2, 122–124, 127, 139, 140, 142, 143, 314, 347, 638, 640–641
  - epitope prediction, 386, 388, 391–392
  - least-squares, 396
  - linear, 363
  - NGS, 381–384
  - non-linear, 364
  - PPI, 390, 393–395
  - regression, 366
  - soft margin, 365
  - spectroscopic techniques, 383–387
  - virology, 376, 379, 380
  - web-servers, 397
 Support vectors, 348  
 Surface Plasmon resonance (SPR), 644
  - antibody characterisation, 645–646
  - benefits, 646
  - HIV research, 644
  - optical biosensor technology, 646
  - receptor-Env interactions, 644
  - role, 646
 Swarm intelligence, 328, 329, 332, 610
  - advantage, 329
  - artificial gene vectors, 329
  - functional types, 334
  - side-effect-free gene therapy, 329
  - therapeutic gene delivery, 329
  - therapeutic gene vectors, 334
 Swarm-level behavior, 335  
 Synthetic immune system (SIS), 258  
 Systematic review, 587
- T**  
 T cell immunoglobulins (TIM), 645  
 T-cell population, 265
- T-cell-based immunotherapy, 247  
 T-cells zone, 275  
 Technological innovation, 328  
 Technology development
  - 1I/2017 U1 (Oumuamua) transit, 558
  - AI and QC, 557
  - AI robotics, 556
  - constructor machines, 556
  - interplanetary lasers, 557
  - machine learning computation approaches, 557
 Therapeutic gene vectors, 337, 341  
 Therapeutic Target Database (TTD), 120  
 Therapeutic vaccines, 233  
 Therapeutic vaccines domain (TVC), 268  
 Thermal environment, 596  
 Thermodynamics
  - enthalpy, 553
  - entropy, 553
  - fractal, 553
  - Gibbs equation, 553
  - human genome, 553
  - life analysis, 552
  - prions, 553
 Third-generation DNA sequencing
  - experimental evolution, 500
  - first-generation sequencing, 501
  - GWAS, 500
  - host-pathogen interactions/intra-/inter-species, 500
  - long-read sequencing, 502–503
  - second-generation sequencing, 501
  - third-generation sequencing, 501
 Three-dimensional crystal structure data, 643  
 Time Of Flight mass CYtometry (CyTOF), 447  
 T-lymphocytes, 163  
 TNFR-associated factor (TRAF), 46  
 TNFRSF1A-associated via death domain [TRADD], 46  
 Tobacco Mosaic Virus (TMV), 606  
 Toll-IL-1R (TIR), 176  
 Toll-like receptors (TLRs), 45, 166, 172, 267, 277
  - active signal, 277
  - adaptive-immunity, 175
  - adaptor molecules, 175
  - amino acids and genomic structures, 174
  - cellular location, 174
  - deficiency, 179
  - definition, 172
  - deregulation processes, 178
  - HPV16, 277
  - and IL-1R, 173
  - innate and adaptive immune responses, 173

- MAPK-p38 pathway, 178  
NF- $\kappa$ B, 172  
PAMPs, 174  
signaling pathways, 176  
TRAM and TRIF adaptors, 178  
TRIF-dependent pathway, 177  
Tournament Selection, 20  
Traditional model-based approaches, 499  
Trans-activating transcriptor (TAT), 520, 522  
Transcription activator-like effector nuclease (TALENs), 456  
Transcription factors, 182  
Translational Environmental Restoration (TER) paradigm  
description, 586  
PICOTS question, 587  
public health issues, 587  
translational research, 587  
Translational health care  
allostasiomic profile, 597  
best evidence base, 597  
disease outbreaks, 597  
HIV management, 598  
HIV-seropositive patients, 597  
neuro-AIDS, 598  
obstacles, 598  
pharmacological interventions, 598  
TER outcomes, 597  
TER paradigm, 597  
Transmitted/founder (T/F) viruses, 629  
Traveling Salesman Problem (TSP), 25  
Tree-augmented Naïve (TAN), 639  
Trofile assay, 637  
Tuberculosis (TB)  
*AI* (*see* Artificial Intelligence (AI))  
contagious nature of, 530  
detection of tubercle bacilli, 537  
host-directed therapies, 537  
innovative approaches, 530, 537  
machine learning, 530, 537  
treatment  
Bedaquiline, 537  
challenging public health, 536  
corticosteroids, 536  
drug resistance, 536, 537  
host-directed therapies, 536  
IRIS, 536  
machine learning, 537  
medications, 537  
MICs, 536  
SIRT1-activating compounds, 536  
Tumor antigens, 219  
Tumor-associated antigens (TAAs), 219, 236, 239  
Tumor-associated macrophages (TAMs), 188  
Tumor-mediated immunosuppression, 243  
Tumor microenvironment, 228  
Tumor necrosis factor (TNF), 45  
Tumor necrosis factor receptor superfamily member 9 (TNFRSF9), 46  
Tumor neoantigens approach, 238  
Tumor viruses, 192  
Two-dimensional (2D) maps, 616
- U**
- Uniform crossover, 21  
United Nation Program on HIV and AIDS (UNAIDS), 309, 425  
Unsupervised learning, 312, 362, 503, 507–508  
Unsupervised methods, 640  
US Food and Drug Administration (FDA), 478
- V**
- Vaccination, 241  
Vaccination/adjuvant therapy, 220  
Vaccinia virus, 333  
van der Waals (vdW), 134, 137  
Varicella zoster virus (VZV), 472, 473, 475, 477  
Viral biology, attributes used, 367  
Viral databases and datasets, 61  
Viral diseases, 65  
Viral escape, 517  
Viral genomes, 22  
Viral immune surveillance, 45, 48, 49  
Viral life cycle, 191–192  
Viral Load (VL) test, 30  
viral outbreaks, 60  
Viral proteins, 207  
Viral replication, 201  
Viremic phase, 211  
Virions, 193  
Virology  
big data, 498, 499  
computational power, 498  
data-driven research, 499  
digital epidemiology, 499  
global human migration patterns, 499  
machine learning approaches, 499  
model-free approach, 499  
predicting viral epidemics, 499  
statistical models, 499  
traditional model-based approaches, 499  
Virtual screening, 121, 122  
Virus containing compartments (VCC), 635  
Virus particles, 193  
Virus related data, 60

- Viruses**
- AI legacy, 609–610
  - ANN (*see* Artificial Neural Networks (ANN))
  - ATM kinase, 194
  - biogeochemical processes, 486
  - biological, 606
  - CNS (*see* Central nervous system (CNS))
  - databases, 608, 609
  - DNA, 190
  - DNA-PK kinase, 194
  - entropy, 611, 613–614
  - evolution, 606
  - experimental evolution, 487
  - genome, 606
  - genome evolution
    - ABC model, 493, 495
    - conventional evolutionary models, 493
    - ecology systems, 493
    - experimental evolution (*see* Futuristic methods)
    - gene networks, 493
    - mutation rate, 490
    - phylogenetic trees, 493
    - population size, 488, 489
    - recombination rate, 491
    - selection, 489, 490
    - simulation-based methods, 495
  - Wright-Fisher model, 488
- global health, 486
- global virology (*see* Experimental evolution)
- history, 607
  - ICTV, 189
  - infectivity, 607
  - molecular virology, 617–619
  - oncoproteins, 191
  - program code probity, 617–619
- published literature, 608
- radical pluralism, 189
- statistical mechanics, 612, 613
- third-generation DNA sequencing (*see* Third-generation DNA sequencing)
- transmission and infection, 189
  - vaccines, 618
  - viral life cycle, 191
  - viral nucleic acids, 192
- Virus-induced interferons, 224
- Virus-infected cells, 211
- Virus-mediated carcinogenesis, 207
- Visual inspection, 408
- Volatile organic compounds (VOCs), 533
- von Neumann, J., 556, 563, 609, 611
- W**
- Wavelet Transform, 79, 80
- Wavelet Transform Modulus Maxima (WTMM) approach
- aspects, 76, 79, 84
  - hepatitis dataset, 77, 79–81, 84–87, 89–93
  - LE, 81
  - multifractal data analysis, 74
  - numerical experiments, 81
- Web-servers, 397
- West Nile virus (WNV), 5, 606
- White spot disease (WSD), 27
- Wrapper methods, 371, 372, 641
- Wright-Fisher model, 488, 495, 508
- Z**
- Zidovudine (AZT), 519
- Zika virus infection, 12
- Zika virus outbreak, 590
- ZINC™ database, 64