Immunology
The Journal of cells, molecules, systems and technologies

British Society for immunology

IMMUNOLOGY REVIEW ARTICLE

Immunoinformatics: an integrated scenario

Namrata Tomar and Rajat K. De
Machine Intelligence Unit, Indian Statistical Institute, Kolkata, India

doi:10.1111/j.1365-2567.2010.03330.x
Received 8 December 2009; revised 12 June 2010; accepted 21 June 2010.
Correspondence: R. K. De, Machine Intelligence Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata 700108, India.
Email: rajat@isical.ac.in
Senior author: Rajat K. De

Summary

Genome sequencing of humans and other organisms has led to the accumulation of huge amounts of data, which include immunologically relevant data. A large volume of clinical data has been deposited in several immunological databases and as a result immunoinformatics has emerged as an important field which acts as an intersection between experimental immunology and computational approaches. It not only helps in dealing with the huge amount of data but also plays a role in defining new hypotheses related to immune responses. This article reviews classical immunology, different databases and prediction tools. It also describes applications of immunoinformatics in designing *in silico* vaccination and immune system modelling. All these efforts save time and reduce cost.

Keywords: allergy; B cells; *in silico* models; major histocompatibility complex/human leucocyte antigen; T cells

Introduction

The term 'immunity' was developed to describe individuals who had recovered from certain infectious diseases and were protected from the same diseases when they were re-encountered. An immune system and associated biological processes exist within these individuals, which are responsible for developing 'immunity'. The role of an immune system is to protect against diseases by identifying and killing pathogens. An immune system includes innate and adaptive components. According to the traditional dogma of immunology, vertebrates have both innate and adaptive immune systems whereas invertebrates possess only an innate immune system.[1] The innate immune system acts more rapidly, and is older and more evolutionarily conserved than the adaptive immune system. It provides the backbone on which the adaptive immune system was able to evolve. The innate immune system is less specific and works as a first line of defence.[2] It comprises four types of defensive barriers, namely, anatomic (e.g. skin and mucous membranes), physiological (e.g. temperature, low pH), phagocytic (e.g. blood monocytes, neutrophils, tissue macrophages) and inflammatory (e.g. serum proteins). An adaptive immune response occurs against a pathogen within 5 or 6 days after the initial exposure to the pathogen.[2] It has evolved in vertebrates as a defence system. Functionally, it accounts for two inter-related activities: recognition and response. It can discriminate between the body's own cells and proteins from foreign molecules, and can recognize chemical differences between two pathogens. It can also recognize altered self cells, such as virus-infected self cells, and distinguish between healthy and cancerous cells. However, it may not always recognize cancer cells as foreign or abnormal cells. As soon as the adaptive immune system recognizes a pathogen, an effector response is elicited to kill or neutralize it. The response is unique to defend against a particular type of pathogen. Later exposure to the same pathogen induces a heightened and more specific response because the adaptive immune system retains memory.

The adaptive immune system has two parts: the cellular immune response of T cells and the humoral response of B cells.[2,3] An antigen has a specific small part, known as the epitope, which is recognized by the corresponding receptor present on B or T cells. B-cell epitopes can be linear and discontinuous amino acids. T-cell epitopes are short linear peptides. Most of the T cells can be in either of the two subsets, distinguished by the presence of one or other of two glycoproteins on their surface, designated as CD8 or CD4. CD4 T cells function as T helper (Th) cells that recognize peptides displayed by major histocompatibility complex (MHC) class II molecules. On the other hand, CD8 T cells function as cytotoxic T (Tc) cells, which recognize peptides displayed by MHC class I molecules. A brief description of various components of the human immune system is provided as supplementary material. The idea that the immune response exists in an

organism is quite old. The earliest literary reference to immunology goes back to 430 BC by Thucydides.[2] In 1798, Edward Jenner found some milkmaids who were immune to smallpox because they had earlier contracted cowpox (a mild disease). The next major advancement in immunology came with the induction of immunity to cholera by Louis Pasteur. After applying weakened pathogen to animals, he administered (in 1885) a dose of vaccine to a boy bitten by a rabid dog and the boy survived. However, Pasteur could not explain its mechanism. In 1890, experiments of Emil Von Behring and Shibasabura Kitasato led to the understanding of the mechanism of immunity. Their experiments described how antibodies present in the serum provided protection against pathogens.

An immune system may be considered as a network of thousands of molecules, which leads to many intertwined responses. It is structurally and functionally diverse and this diversity varies both between individuals and temporally within individuals. Huge amounts of data related to immune systems are being generated. Immunologists have been using high throughput experimental techniques for a long time, which have generated a vast amount of functional, clinical and epidemiological data. The development of new computational approaches to store and analyse these data are needed. Recently, immunology-focused resources and software are appearing, which help in understanding the properties of the whole immune system.[4] This has given rise to a new field, called immunoinformatics. Immunogenomics, immunoproteomics, epitope prediction and *in silico* vaccination are different areas of computational immunological research. Recently, Systems Biology approaches have been applied to investigate the properties of the dynamic behaviour of an immune system network.

Immunoinformatics includes the study and design of algorithms for mapping potential B- and T-cell epitopes, which lessens the time and cost required for laboratory analysis of pathogen gene products. Using this information, an immunologist can explore the potential binding sites, which, in turn, leads to the development of new vaccines. This methodology is termed 'reverse vaccinology' and it analyses the pathogen genome to identify potential antigenic proteins.[5] This is advantageous because conventional methods need to cultivate pathogen and then extract its antigenic proteins. Although pathogens grow fast, extraction of their proteins and then testing of those proteins on a large scale is expensive and

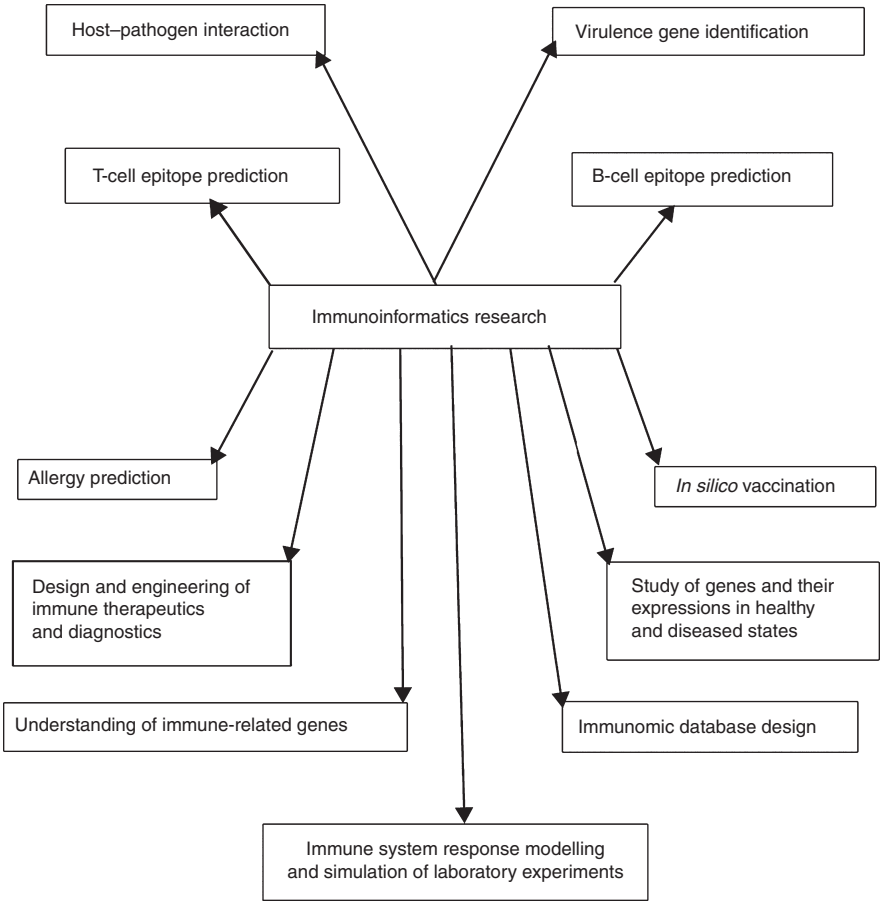


Figure 1. Immunoinformatics: research areas.

time consuming. Immunoinformatics is capable of identifying virulence genes and surface-associated proteins.

Figure 1 shows the different research areas of immunoinformatics. All of these areas are described in separate sections of this article. We describe various available information regarding classical immunology, different immunomic databases, and B-cell and T-cell epitope prediction tools and softwares. Several methods are now available that enable one to map epitopes and design therapeutic vaccines more quickly. Some of them are described in this article, which concludes with some applications of immunoinformatics.

Immunomics

The term 'immunome' corresponds to all the genes and proteins taking part in immune responses. It excludes genes and proteins that are expressed in cell types other than in immune cells.[6] According to Sette *et al.*[7] all immune reactions that are the result of interactions between the host and antigenic peptides are referred to as 'immunome reactions', and their study is entitled 'immunomics'. Like genomics and proteomics, immunomics is a new discipline that uses high throughput techniques to understand the immune system mechanism.[8,9]

Various datatypes and databases

In this section, we focus on various immune-system-related datatypes and databases. A brief description of these databases is provided. The section starts with some experimental techniques and results.

Experimental data

There has been an explosion in available experimental data in immunology as the result of the advent of high throughput molecular biology techniques. These techniques help in finding the structure and function of immune genes and their products.[10]

There are many immunological techniques that are used to understand the underlying mechanism of an immune system and its responses to various infections, diseases and drug administration, namely, affinity chromatography,[11] flow cytometry,[12] radioimmunoassay,[13] enzyme-linked immunosorbent assay,[13,14] competitive inhibition assay[15] and Coombs test.[16] Here, we present some experimental findings, which help to identify B-cell and T-cell epitopes and to study immune responses.

Experimental techniques for exploring immune system components The ability to identify epitopes in the immune response has important implications in the diagnosis of diseases. For this reason, epitopes for B and T cells need to be identified and mapped. In this context, Wanga *et al.*[17] mapped the B-cell epitopes present on non-structural protein 1 (NS1), i.e. NS1-18 and NS1-19 in Japanese encephalitis virus. For epitope mapping, a series of 51 partially overlapping fragments covering the entire NS1 protein were expressed with a glutathione S-transferase tag and then screened with a monoclonal antibody. They found that the motif of (146) EHARW (150) was the minimal unit of the linear epitope recognized by that monoclonal antibody.

Purification techniques like affinity chromatography are used to purify MHC–peptide from membrane MHC molecules, which can be analysed by capillary high-pressure liquid chromatography electrospray ionization-tandem mass spectrometry.[18] They can be further used to find new tumour-associated antigens. These are proteins that are not unique to cancer cells but are expressed in tumour cells. One approach to find tumour-associated antigens is based on transfection of the expression library made from complementary DNA into cells expressing the desired MHC haplotypes.[19] The clones are selected on the basis of their ability to provoke an immune response in T cells of the individuals with the same MHC type. MHC–peptide complexes are required for tumour therapeutics.

Dengue, a human viral disease transmitted by arthropod vectors, has an annual mortality rate of 25 000.[20] Dengue fever and dengue haemorrhagic fever are caused by the four dengue viruses, DEN-1, -2, -3 and -4, which are closely related antigenically. Random Peptide Libraries of peptides displayed on the phage help in selecting sequences that mimic epitopes from microorganisms. Amin *et al.*[21] used Random Peptide Libraries and identified two peptides, NS3 and NS4B. These two non-structural proteins resemble the antigenic structure of B-cell epitopes of dengue virus obtained from a phage-peptide library using human polyclonal antisera from patients who had recovered from dengue virus infection. These two peptides could be used for the development of a diagnostic kit and a potential vaccine.

Immunomic microarray technology and analysis

Using DNA microarray technology, one can measure the RNA expression of thousands of genes simultaneously in a single assay. The principle of all kinds of microarray technologies is binding and measurement of target biological specimens to complementary probes. Similar technology is used in functional immunomics and is referred to as 'immunomic microarray'. It includes dissociable antibody microarray,[22] serum microarray[23] and serological analysis of a complementary DNA expression library (SEREX).[24]

An antibody microarray consists of antibody probes and antigen targets, so that it can be used to measure concentration of antigen for a specific antibody probe, but peptide microarray has an opposite approach. It uses antigen peptides as fixed probes and serum antibodies as

targets. The recent technology is peptide–MHC microarray or the artificial antigen-presenting chip. In this technique, recombinant peptide–MHC complexes and co-stimulatory molecules are immobilized on a surface, and a population of T cells is incubated with the microarray. The T-cell spots act as artificial antigen-presenting cells[25] containing defined MHC-restricted peptides. The advantage of using peptide–MHC is that it can map the MHC-restricted T-cell epitope.

The proteins responsible for the normal functioning of the cellular machinery may have sequence similarity with various pathogenic microbes. They can induce autoimmunity and thereby are less useful for vaccine development. Microarray technology helps in selecting these proteins from genomic sequences.[26] It is being applied in autoimmune disease diagnosis and treatment,[27] allergy prediction,[28] T-cell and B-cell epitope mapping[29] and vaccination.[30] The immunomic and genomic microarray data differ in several ways, e.g. both of them have different designs. One can measure two or more signals simultaneously determined by a single feature, i.e. epitope in immunomic microarray.[31,32] DNA microarrays measure one response value for each gene per sample, i.e. messenger RNA concentration produced by the gene, but a single epitope can generate different response values corresponding to different epitopes in peptide–MHC chips. In the case of the B-cell epitope, it can be recognized by different isotypes of immunoglobulins, so here, one can measure both intensity and quality of the antibody response.

Immunomic databases

Knowledge of B-cell and T-cell epitope-mediated responses has been increased dramatically. Epitope information-related databases, bioinformatics tools and prediction algorithms help in understanding the structure and sequences of amino acids of epitopes. This knowledge is crucial for basic immunological studies, diagnosis and treatment of various diseases, and in vaccine research.[33] INNATEDB[34] (http://www.innatedb.ca) has been created to understand the complete network of pathways and interactions of innate immune system responses. It is an integrated biological database of the human and mouse molecules with 100 000 experimentally verified interactions and 2500 pathways involved in innate immunity. It has a newer version, called CEREBRAL,[35] which is a JAVA plugin for the CYTOSCAPE biomolecular interaction viewer[36] for automatically generating layouts of biological pathways. Table 1 lists some of the databases that deal with information related to B-cell epitopes, T-cell epitopes, allergy prediction and evolution of immune system genes and proteins.

B-cell epitope databases

Conformational epitopes have implicit structural information related to antigen and binding mode. It has been found that 90% of B-cell epitopes are conformational or discontinuous. BCIPEP[37] (http://www.imtech.res.in/raghava/bcipep) provides comprehensive information about experimentally verified B-cell epitopes and tools for mapping these epitopes on an antigen sequence. Immunogenicity of a peptide in Bcipep is divided into three dimensions: immunodominant, immunogenic and null immunogenic. Searches can be restricted to the basis of immunogenicity. Bcipep has some limitations such as, (i) it contains no discontinuous epitopes, (ii) it includes a limited number of unique peptides, and (iii) it provides information on

Table 1. Databases on B-cell epitopes, T-cell epitopes, allergen and molecular evolution of immune system components

Databases	Names	URLs	References
B-cell epitopes	CED	http://immunet.cn/ced/	[38]
	BCIPEP	http://www.imtech.res.in/raghava/bcipep	[37]
	EPITOME	http://cubic.bioc.columbia.edu/services/epitome/	[39]
	IEDB	http://www.immuneepitope.org/	[33]
	IMGT®	http://imgt.cines.fr	[43]
T-cell epitopes	JENPEP	http://www.darrenflower.info/jenpep/	[40]
	SYFPEITHI	http://www.syfpeithi.de	[41]
	IEDB	http://www.immuneepitope.org/	[33]
	FRED	http://www-bs.informatik.uni-tuebingen.de/Software/FRED	[42]
	IMGT®	http://imgt.cines.fr	[43]
Allergen	Database of IUIS	http://www.allergen.org	[47]
	ALLERGENPRO	http://www.niab.go.kr/nabic/	[48]
	SDAP	http://fermi.utmb.edu/SDAP/	[49]
Information related to molecular evolution of immune system components	IMMTREE	http://bioinf.uta.fi/ImmTree	[50]
	IMMUNOME database	http://bioinf.uta.fi/Immunome/	[51]
	IMMUNOMEBASE	http://bioinf.uta.fi/ImmunomeBase	[52]
	IMMUNOME Knowledge Base	http://bioinf.uta.fi/IKB/	[6]

peptides containing only natural amino acids. CED[38] (Conformational Epitope Database) can be used for the evaluation and improvement of existing epitope prediction methods. CED 0.03 release (http://immunet.cn/ced/) has 293 entries. It has a collection of B-cell epitopes from the literature, conformational epitopes defined by methods like X-ray diffraction, nuclear magnetic resonance, scanning mutagenesis, overlapping peptides and phage display. CED maintains well-defined conformational epitope information. It rejects conformational epitopes that are not defined clearly so the database is small. EPITOME[39] (http://cubic.bioc.columbia.edu/services/epitome/) contains all known antigen–antibody complex structures. A semi-automated tool has also been developed that identifies the antigenic interactions within the known antigen–antibody complex structures and compiled these interactions into EPITOME. None of the other databases can locate the complementary determining regions or identify the antigenic residues semi-automatically. EPITOME updating follows the updating of SCOP. Epitome is updated twice a year, as soon as SCOP is updated.

If we compare EPITOME and CED, we find that they are similar in size, the difference lies in the source of collection of B-cell epitopes. EPITOME collects B-cell epitopes only from Protein Data Bank (PDB) structures and includes information on complementary determining regions. In contrast, CED takes data from the literature and from the above-mentioned methods. As their sources are different, one can use the complementary information.

T-cell epitope databases

T-cell epitopes do not always have high affinity for MHC binders. A functional T-cell response requires MHC–peptide binding and a proper interaction of the MHC–peptide ligand with a specific T-cell receptor (TR). We need well-characterized data to model the process of binding of peptides to transfer associated protein (TAP) and MHCs, which function as T-cell epitopes. Some recent investigations include finding and mapping of potential epitopes. Epitope mapping leads to the design of effective vaccines. JENPEP[40] (latest updated version 2.0) (http://www.darrenflower.info/jenpep/) is a relational database with five types of data: a compilation of quantitative measures of binding for 12 336 entries of peptides to MHC I and II, an annotated list of 3218 entries of dominant and subdominant T-cell epitopes, and a set of over 441 records of quantitative data for peptide binding to TAP peptide transporter. In the latest update (i.e. in version 2.0), two new categories have been introduced: B-cell epitopes (816 entries) and peptide–MHC–TR complex formation (49 entries).

The SYFPEITHI database[41] (http://www.syfpeithi.de) has information on MHC class I and II anchor motifs, and binding specificity. It calculates a score based on the following rules: calculated score values differentiate among anchor, auxiliary anchor or preferred residues.

FRED[42] (http://www-bs.informatik.uni-tuebingen.de/Software/FRED) deals with methods of data processing. It also compares the performance of prediction methods by considering experimental values. It can handle polymorphic sequences. IMGT®[43] (the international IMMUNOGENETICS information system®; http://imgt.cines.fr) has a good collection of immunoglobin, T-cell receptor, MHC, and related proteins of the immune system of humans and other vertebrates. It has five databases and 15 interactive online tools for sequence, genome and three-dimensional structural analysis.

IEDB 2.0,[33] (Immune Epitope Database and Analysis Resource Database) (http://www.immuneepitope.org/), sponsored by the National Institute for Allergy and Infectious Diseases (http://www.niaid.nih.gov), has different tools to find B-cell and T-cell epitopes. It contained details of 75 056 peptide epitopes till July 2010.

It also facilitates the conversion of experimental data from text and figures in a journal publication into a computer-friendly format in the form of ONTIEs (Ontology of Immune Epitopes) (http://ontology.iedb.org). This module has been imported by the OBI (Ontology for Biomedical Investigations) Consortium (http://purl.obolibrary.org/obo/obi).[44]

Allergy prediction databases

Allergens are proteins or glycoproteins recognized by immunoglobulin E (IgE), which is produced by the immune system in allergic individuals. So far, 1500 allergenic structures have been identified.[45] Online allergen databases and allergy prediction tools are being used to find cross-reactivity between known allergens. Localization of B and T cells in the allergen may not coincide.[46] The differences between both kinds of epitopes present in an antigen are: T-cell epitopes are only linear (as mentioned earlier) and are distributed throughout the primary structure of the allergen, whereas B-cell epitopes can be either linear or conformational, recognized by IgE antibodies, and are located on the surface of the molecule accessible to antibodies. Moreover, in the case of B-cell epitopes, predicting allergenicity in a molecule based on known conformational epitopes is a difficult task.

The ALLERGEN NOMENCLATURE database of the International Union of Immunological Societies (IUIS) has an allergen database[47] (http://www/allergen.org). The ALLERGEN PRO database[48] (http://www.niab.go.kr/nabic/) contains information related to 2434 allergens, e.g. allergens in rice microbes (712 records), animals (617 records) and plants (1105 records). The web server ALLERGOME 4.0[45] (http://www.allergome.org) provides an exhaustive repository of IgE-binding compound data. It has a total 1736 allergen

sources (updated in March 2010). The Real-Time Monitoring of IgE sensitization module (ReTiME), in ALLERGOME 4.0, enables one to upload raw data from both *in vivo* and *in vitro* experiments. This is the first attempt where information technology has been applied to allergy data mining. SDAP[49] (Structural database of Allergenic Proteins) (http://fermi.utmb.edu/SDAP/) is a web server that provides cross-referenced access to the sequence and structure of the IgE epitope of allergenic proteins. Its algorithm is based on conserved properties of amino acid side chains. In its latest update, it has 887 allergenic proteins.

Databases related to molecular evolution of immune genes and proteins

To explore the molecular evolution of the human immune system, a reference set of genes and proteins must be defined. For this reason, Ortutay *et al.*[50] constructed a database IMMTREE (http://bioinf.uta.fi/ImmTree) for the evolutionary trees of proteins of the human immune system. It contains information for orthologues of the human genes in 80 species. The IMMUNOME database[51] (http://bioinf.uta.fi/Immunome/) is another database in which 847 genes and proteins are annotated and characterized according to their functions, protein domains and gene ontology terms from the human immunome.

A vast amount of molecular data for genes and proteins for the immune system has accumulated. The Immunome Knowledge Base (IKB)[6] is a single service access to many immune system databases and resources. It combines the other databases, namely IMMUNOME[51] and IMMUNOMEBASE,[52] and several additional data items in an integrated fashion. It has orthologue groups of 1811 metazoan immunity genes for studying the evolution of the immune system, and includes the evolutionary history of genes and proteins, orthologous genes, information on disease-causing mutations, alternatively spliced variants and copy number variations.

Various tools and algorithms

Here, we throw some light on available immunology-related tools and algorithms. Traditionally, determination of the binding affinity of MHC molecules and antigenic peptides is the main objective when predicting epitopes. The experimental techniques are found to be difficult and time consuming. As a result, several *in silico* methodologies are being developed and used to identify epitopes. These techniques include matrix-driven methods, finding structural binding motifs, a quantitative structure–activity relationship (QSAR) analysis, homology modelling, protein threading, docking techniques and design of several machine-learning algorithms and tools. In the past, computational techniques could only identify sequence characteristics but new improved algorithms and tools are being designed to increase the predictive performance. Table 2 lists some of the tools that deal with B-cell and T-cell epitope prediction, allergy prediction and *in silico* vaccination. Here, we describe different methodologies for epitope and allergy prediction, and the process of *in silico* vaccination briefly.

B-cell epitope prediction

B cells produce antibodies that are protein in nature. B-cell epitopes are antigenic determinants on the surface of pathogens that interact with B-cell receptors. The B-cell receptor binding site is hydrophobic with six hypervariable loops of variable length and amino acid composition. As described in ref.[53], B-cell epitopes are classified as continuous/linear and discontinuous/conformational. Most of the B-cell epitopes are discontinuous where distant residues are brought into spatial proximity by protein folding. Experiments are mostly based on linear epitopes. There are both sequence-based and structure-based prediction tools but prediction tools are limited for discontinuous B-cell epitopes.[37,54]

Prediction using amino acid propensity scale

Classically, amino acid propensity scales such as hydrophilicity and characteristic flexibility have been used to identify epitopes present in antigens. Pellequer *et al.*[55] compared several propensity scale methods using a dataset of 14 epitope annotated proteins and found that the scales of Parker *et al.*[56] Chou and Fasman,[57] Levitt,[58] and Emini *et al.*[59] provide better results than the other scales tested.[53] El-Manzalawy *et al.*[60] compared propensity-scale-based methods with a Naive Bayes classifier. They used three different representations of the classifier input: amino acid identities, position-specific scoring matrix profiles and dipeptide composition. They used two datasets, one is the propensity dataset and the other is from BCIPEP.[37] They considered 125 non-redundant antigens at 30% sequence similarity cut off from BCIPEP. The BEPITOPE tool[61] predicts continuous epitopes based on the prediction of protein turns. It is a newer version of PREDITOP[62] and uses more than 30 propensity scale values. The BCEPRED server[63] (http://www.imtech.res.in/raghava/bcepred/) predicts linear B-cell epitopes with 58·7% accuracy based on combined amino acid properties like accessibility, hydrophilicity, flexibility, polarity, exposed surface and turns.

Analyses of antigen–antibody interactions are performed on antibody-binding sites on proteins, which help in predicting the linear and conformational B-cell epitopes. Taking this into consideration, a database,

Table 2. Webservers for prediction of B-cell epitopes, T-cell epitopes, allergy and for *in silico* vaccination

Webservers and Tools	Names	URLs	References
B-cell epitope prediction	ABCpred	http://www.imtech.res.in/raghava/abcpred	[65]
	BEPITOPE	jlpellequer@cea.fr	[61]
	COBEpro	http://scratch.proteomics.ics.uci.edu/	[66]
	Bepipred	http://www.cbs.dtu.dk/services/BepiPred	[67]
	IMGT®	http://imgt.cines.fr	[43]
	Bcepred	http://www.imtech.res.in/raghava/bcepred/	[63]
	Discotope	http://www.cbs.dtu.dk/services/DiscoTope/	[70]
	CEP	http://bioinfo.ernet.in/cep.htm	[73]
	AgAbDb	http://202.41.70.51:8080/agabdb2/	[64]
	MIMOP	request from franck.molina@cpbs.univ-montp1.fr	[75]
	MIMOX	http://web.kuicr.kyoto-u.ac.jp/hjian/mimox	[76]
	Pepitope	http://pepitope.tau.ac.il/	[74]
	3DEX	http://www.schreiber-abc.com/3dex/	[78]
	IEDB	http://www.immuneepitope.org/	[33]
T-cell epitope prediction	MMBPred	http://www.imtech.res.in/raghava/mmbpred/	[80]
	NetCTL	http://www.cbs.dtu.dk/services/NetCTL/	[84]
	NetMHC 3.0	http://www.cbs.dtu.dk/services/NetMHC	[85]
	TAPPred	http://www.imtech.res.in/raghava/tappred/	[89]
	Pcleavage	http://www.imtech.res.in/raghava/pcleavage/	[90]
	ElliPro	http://tools.immuneepitope.org/tools/ElliPro	[99]
	MHCPred	http://www.darrenflower.info/mhcpred/	[100]
	Propred	http://www.imtech.res.in/raghava/proped	[106]
	EpiToolKit	http://www.epitoolkit.org	[108]
	Syfpeithi	http://www.syfpeithi.de	[41]
	IMGT®	http://imgt.cines.fr	[43]
	IEDB	http://www.immuneepitope.org/	[33]
Allergy prediction	AlgPred	http://www.imtech.res.in/raghava/algpred	[113]
	Allermatch	http://www.allermatch.org	[114]
	APPEL	http://jing.cz3.nus.edu.sg/cgi-bin/APPEL	[117]
	EVALLER	http://bioinformatics.bmc.uu.se/evaller.html	[118]
In silico vaccination	VaxiJen	http://www.darrenflower.info/VaxiJen/	[126]
	DyNAVacs	http://miracle.igib.res.in/dynavac/	[127]
	NERVE	http://www.bio.unipd.it/molbinfo	[128]
	VIOLIN	http://www.violinet.org	[129]
	Vaxign	http://www.violinet.org/vaxign/	[130]

AgAbDb[64] (http://202.41.70.51:8080/agabdb2/), has been developed that is based on molecular interactions of antigen–antibody co-crystal structures.

Prediction using machine learning methodologies

Several researchers used machine learning algorithms and tools to retrieve characteristics of an epitope through learning a dataset. For example, Saha and Raghava[65] used artificial neural networks (ANNs) in ABCpred (http://www.imtech.res.in/raghava/abcpred); Sweredoski and Baldi[66] presented COBEpro using a support vector machine (SVM).

Saha and Raghava[65] used feed forward and recurrent neural networks to predict continuous B-cell epitopes. They took 700 nr B-cell epitopes and equal number of non-epitopes from SWISSPROT database for training and testing. Sweredoski and Baldi[66] presented COBEpro, which is a two-step system for the prediction of continuous B-cell epitopes. In the first step, COBEpro assigns a fragment epitopic propensity score to protein sequence fragments using an SVM. In the second step, it calculates an epitopic propensity score for each residue based on the SVM scores of the peptide fragment in the antigenic sequence. It is incorporated into the SCARTCH prediction suite (http://scratch.proteomics.ics.uci.edu/). However, COBEpro cannot be used to distinguish antigen from non-antigen. It should be used with high-throughput technologies to increase efficacy. Larsen *et al.*[67] introduced Bepipred (http://www.cbs.dtu.dk/services/BepiPred). They constructed three datasets of linear B-cell epitopes, annotated proteins from literature, AntiJen database[68]

and Los Alamos human immunodeficiency virus (HIV) database (http://www.hiv.lanl.gov). They tested a number of propensity scale methods on the Pellequer *et al.* dataset,[55] and found the best scale to be by Levitt.[58] Then, they used a Hidden Markov model (HMM) to predict the location of linear B-cell epitopes and tested HMMs on the Pellequer *et al.* dataset to find optimal parameters. HMM was combined with one set of the two best propensity scale methods, i.e. Parker *et al.*[56] and Levitt[58] to get more accurate predictions.

Prediction methodology for discontinuous B-cell epitopes

As mentioned earlier, more than 90% of B-cell epitopes are discontinuous but they may comprise a linear amino acid chain of peptides, which is brought closure in three-dimensional space.[69] There is a specialized form of protein–protein interaction in these epitopes. Changes in protein folding may lead to changes in the number of epitopes.[46] The characterization and prediction of B-cell epitopes are mainly conformation dependent so the task of prediction is more difficult compared with that of T-cell epitopes. The most accurate way to identify the B-cell epitope is through X-ray crystallography. Anderson *et al.*[70] presented a method called DISCOTOPE, (http://www.cbs.dtu.dk/services/DiscoTope/), which is a combination of amino acid statistics, spatial information and surface exposure. It was trained on a dataset of discontinuous epitopes of 76 X-ray structures of antibody–antigen complexes. It detects 15·5% of residues located in discontinuous epitopes with a specificity of 95%. The conventional Parker hydrophilicity scale (for predicting linear B-cell epitopes) identifies only 11·0% of residues with 95% specificity. It is said to be the first method developed for prediction of discontinuous B-cell epitopes with better performance than methods based only on sequence data.

Bublil *et al.*[71] developed MAPITOPE for conformational B-cell epitope mapping. The hypothesis behind MAPITOPE is that the simplest meaningful fragment of an epitope is an amino acid pair of residues that lie within the epitope, which are the result of folding. A set of affinity isolated peptides was obtained by screening the phage display peptide libraries with the antibody of interest. This set was given as algorithm input, and one to three epitope candidates on the surface of the atomic structure of the antigens were obtained as output.

A computational method has been presented by Sollner *et al.*[72] to automatically select and rank peptides for the stimulation of otherwise functionally altered antibodies. They investigated the integration of B-cell epitope prediction with the variability of antigen, and the conservation of patterns for posttranslational modification prediction. By their observation, they found high antigenicity, low variability and low likelihood of posttranslational modification for the identification of biorelevant sites. Greenbaum *et al.*[53] assembled non-redundant datasets of repetitive three-dimensional structure of antigen and antigen–antibody complexes from the PDB. The CEP web interface[73] (http://bioinfo.ernet.in/cep.htm) predicts conformational and sequential epitopes, and also antigenic determinants. It uses structure-based approaches, solvent accessibility of amino acids and spatial distance cut-off to predict antigenic determinants. Less availability of the three-dimensional structure data of protein antigens limits the utility of this server.

Mimotope-based epitope prediction methodology

Phage display library has a large number (more than 109) of random peptides.[74] It is widely used for finding protein–protein interactions (especially in antibody–antigen interactions), protein function identification and in development of new drugs and vaccines. These libraries are screened to find the pool of peptides that can bind to desired antibody. These pools of peptides are called mimotopes.[69,74,75] Mimotopes and antigens are both recognized by the same antibody paratope. Mimotopes are said to be the imitated part of the epitope. So, it is possible that a mimotope may have some valuable information about the epitope. However, homology may not exist between the mimotope and the epitope of the native antigen. This mimicry exists because of similarities in physiochemical properties and spatial organization.[75] Considering these properties, mimotope pools are used to mine information to predict an epitope. Using this concept, the MIMOP tool[75] has been developed. MIMOP predicts linear and conformational epitopes based on two algorithms: MIMALIGN uses degenerated alignment analyses, and MIMCONS is based on consensus identification. MIMOX[76] (http://web.kuicr.kyoto-u.ac.jp/~hjian/mimox) comes in the same category, which maps a single mimotope or a consensus sequence of a set of mimotopes onto the corresponding antigen structure. Then, it searches for all of the clusters of residues that could be the native epitope. PEPITOPE[74] (http://pepitope.tau.ac.il/) (an advanced server for mimotope-based epitope prediction approaches) uses two algorithms: PEPSURF[77] and MAPITOPE.[71] It maps each mimotope so as to map them onto the solved structure of the antigen surface. Alignment of the mimotope is done first in MIMOX; this step is different in PEPITOPE. If we compare it with MIMOP, MIMOP aligns the peptides to the antigen at the sequence level rather than directly to the three-dimensional structure. The three-dimensional structure is considered only after the alignment stage.

Sometimes linear peptides mimic conformational epitopes. The same phage display peptide libraries for screening with the respective antibodies are used to select these mimotopes. Schreiber *et al.*[78] presented a software, 3DEX (3D-EPITOPE-EXPLORER) (http://www.schreiber-abc.

com/3dex/) that allows localizing of linear peptide sequences within three-dimensional structures of proteins. Its algorithm takes into account the physiochemical neighbourhood of C-α or C-β atoms of individual amino acids and surface exposure of the amino acids. Authors were able to localize mimotopes from the plasma of patients who were HIV-positive within the three-dimensional structure of gp120. The epitopes defined by 3DEX are not proven by mathematical calculations and energy minimizations.

T-cell epitope prediction

It is necessary to bind antigenic peptides with MHC so that cytotoxic T cells can recognize them. Hence, identification of MHC binding peptides is a central part of any algorithm that predicts T-cell epitopes. There exist several methodologies for the prediction of MHC binding peptides, which are based on the idea of quantitative matrices, HMM, ANN, SVM and structure of the peptides.

Prediction through matrix-driven methods

Huang and Dai[79] first investigated a new encoding scheme of peptides. This scheme used the BLOSUM matrix with the amino acid indicator vectors for direct prediction of T-cell epitopes. It replaced each non-zero entry in the amino acid indicator vector by the corresponding value appearing in the diagonal entries in the BLOSUM matrix. The MMBPred[80] (http://www.imtech.res.in/raghava/mmbpred/) server predicts the mutated promiscuous and high-affinity MHC binding peptide. It uses the matrix data in a linear prediction model and ignores peptide conformation. The prediction is based on the quantitative matrices of 47 MHC alleles.

Prediction through HMM

Transfer Associated Protein is an important component of the MHC I antigen-processing and presentation pathway. A TAP transporter can translocate peptides of 8–40 amino acids into endoplasmic reticulum. Zhang *et al.*[81] developed PRED^{TAP} (http://antigen.i2r.a-star.edu.sg/predTAP) for the prediction of peptide binding to hTAP. They used a three-layer back propagation network with the sigmoid activation function. The inputs were the binary strings, representing nonamer peptide. Second, they used second-order HMM. The results are both sensitive and specific.

Prediction through ANN

Neilsen *et al.*[82] described an improved neural network model to predict T-cell class I epitopes. They have a combination of sparse encoding, BLOSUM encoding and input derived from HMM. The dataset consists of 528 nonamer amino acid peptides for which the binding affinity to the HLA I molecule A*0204 has been measured in a method described by Buus *et al.*[83] NetCTL server[84] (http://www.cbs.dtu.dk/services/NetCTL/) uses a method to integrate the prediction of peptide MHC class I binding, proteasomal C-terminal cleavage and TAP transport efficiency. It has updated the version from 1.0 to 1.2 to improve the accuracy of MHC class I peptide-binding affinity and proteasomal cleavage prediction. NetMHC server 3.0[85] (http://www.cbs.dtu.dk/ services/NetMHC) is based on ANN and weight matrices. It has been trained on data from 55 MHC peptides (43 human and 12 non-human) and position-specific scoring matrices for a further 67 HLA alleles.

MHC class I molecule motifs are well defined but the prediction of MHC class II binding peptides is found to be difficult for a number of reasons, including variable length of reported binding peptides, undetermined core region for each peptide and number of amino acids as primary anchor. Brusic *et al.*[86] developed PERUN, a hybrid method for the prediction of MHC class II binding peptide. It uses available experimental data and expert knowledge of binding motifs, evolutionary algorithms and ANN. They used PlaNet package version 5.6[87] to design and train a three-layered fully connected feed-forward ANN.

Prediction using other machine learning methodologies

Nanni[88] demonstrated the use of SVM and SV (Support Vector) data description to predict T-cell epitopes. In the case of TAPPred[89] (http://www.imtech.res.in/raghav/tappred/), Bhasin and Raghava analysed nine features of amino acids to find the correlation between binding affinity and physiochemical properties. They developed an SVM-based method to predict the TAP binding affinity of peptides, and found cascade SVM to be more reliable. Cascade SVM has two layers of SVMs and its performance is better than the other available algorithms.

Computational techniques are found to be easier than experimental analysis for determining cleavage specificities of proteasomes. It is experimentally established that the immunoproteasome is involved in the generation of the MHC class I ligand. For this purpose, Pcleavage[90] (http://www.imtech.res.in/raghava/pcleavage/) has been developed to predict both kinds of cleavage sites in antigenic proteins. It uses SVM,[91] Parallel Exemplar based Learning[92] and Waikato Environment for Knowledge Analysis.[93]

Ant colony search systems have proved useful for solving combinatorial optimization problems and can be applied to the identification of a multiple alignment of a set of peptides. Basically, they[94] attempt to find an optimal alignment for a given set of peptides based on the search strategy.

Structure-based prediction

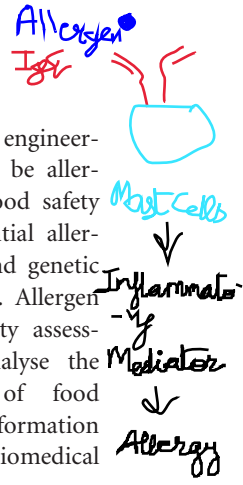
Peptide–MHC binding data are necessary to find T-cell epitopes. Current methods are mostly based on peptide binding affinity to MHC for predicting T-cell epitope. The three-dimensional QSAR technology CoMSIA has been applied to the problem of peptide–MHC binding.[95] It uses the interaction potential around aligned sets of three-dimensional peptide structures to describe binding. TEPITOPE[96] by Bian and Hammer is used to predict promiscuous and allele-specific HLA II restricted T-cell epitopes *in silico*. TEPITOPE's user interface has display and comparison of pocket profiles, and finds similar HLA II differing in their binding capacity for a given peptide sequence. Kangueane and Sakharkar[97] implemented a web server T-cell epitope designer for MHC peptide which uses a definition of virtual binding pockets to position specific peptide residue anchors and estimation of peptide residue virtual binding pocket compatibility.

Zhao *et al.*[98] described a novel predictive model using information from 29 human MHCp crystal structures. The overall binding between peptide and MHC provides a cumulative measure of the physical and chemical compatibility between each residue in the peptide and the residue forming the virtual pockets. ElliPro[99] (http://tools.immuneepitope.org/tools/ElliPro) is a web tool that implements a modified version of the Thorton method, residue clustering algorithm, the Modeller program and the Jmol viewer. It predicts and visualizes the antibody epitope in protein sequence and structure. It implements three algorithms for the approximation of the protein shape as an ellipsoid, calculation of the residue protrusion index and clustering of neighbouring residue based on their protrusion index values.

It is generally accepted that only peptides that bind to MHC with an affinity above a threshold value (typically 500 nM), function as T-cell epitopes. Guan *et al.*[100] in the Edward Jenner Institute for Vaccine Research, UK, introduced MHCPred (http://www.darrenflower.info/mhcpred/). It is a Perl implementation of two-dimensional QSAR application to peptide–MHC prediction and covers both class I and class II MHC allele peptide specificity models. Peptides that can bind to MHC on the tumour cell surface have potential to initiate a host immune response against the tumour. Schiewe and Haworth[101] developed an algorithm PeSSI (peptide–MHC prediction of structure through solvated interfaces) for flexible structure prediction of peptide binding to the MHC molecule. They used CT antigens (Cancer Testis), KU-CT-1, that have the potential to bind HLA-A2.

Jojic *et al.*[102] developed an improved structure-based model which used known three-dimensional structures of a small number of MHC–peptide complexes, the MHC class I sequence, known binding energies for MHC–peptide complexes, and a larger binary dataset with information about strong binders and non-binders. They used adaptive double threading, where the parameters of the threading model are learnable, and both MHC and peptide sequences can be threaded onto the structure of other alleles. Furman *et al.*[103] used an approach that can be applied to a wide range of MHC class I alleles. In this algorithm, peptide candidates are threaded, and their binding compatibility is evaluated by statistical pairwise potentials. They used the pairwise potential table of Miyazawa and Jernigan.[104]

Immunodominant peptides are being used for rational design of peptide vaccines focusing on T-cell immunity. Altuvia and Margalit[105] focused on antigenic peptides recognized by cytotoxic T cells. They applied the threading approach to screen a library of peptide sequences and identified those that optimally fitted within the MHC groove. ProPred[106] (http://www.imtech.res.in/ raghava/proped) is a graphical web tool for predicting MHC class II binding regions in antigenic protein sequences. They extracted the matrices for 51 HLA-DR alleles from a pocket profile database developed by Sturniolo *et al.*[107] The EpiToolkit[108] (http://www.epitoolkit.org) web server includes several prediction methods for MHC class I and class II ligands, and minor histocompatibility antigens. It can also investigate the effect of mutation on T-cell epitopes.



Allergy prediction

Food derived from biotechnology and genetic engineering contains some foreign proteins, which can be allergic to many human beings. Because of this, food safety is an important issue. Evaluation of the potential allergenicity of food derived from biotechnology and genetic engineering is a current food safety assessment. Allergen sequence databases are essential tools for safety assessments of bioengineered foods. They can analyse the structural and physiochemical properties of food allergen proteins. They focus on molecular information such as protein sequences, structures and biomedical information.

Allergy occurs by both extrinsic and intrinsic factors. A type I hypersensitive reaction is induced by certain allergens that elicit IgE antibodies.[2] Use of genetically modified food and therapeutics makes allergenic protein prediction necessary. According to the proposed guidelines of World Health Organization (WHO) and Food and Agriculture Organization (FAO) in 2001, a protein is considered an allergen when it has at least six contiguous amino acids the same or a window of 80 amino acids when compared with known allergens. It has already been established that allergens do not share common structural characteristics. Hence, allergen databases are being used as reference for finding the sequence similarity in allergenicity evaluation.[109] It is said that a protein is

considered an allergen if it has a region or peptides identical to a known IgE epitope.

The allergen prediction method proposed by Kong *et al.*[110] is based on the determination of a combination of two allergen motifs in a given protein sequence. They took 575 proteins for allergen dataset and 700 sequences for a non-allergen test set from the given reference.[111] They developed a database that has all possible combinations of two motifs from the set of allergenic motifs by using a motif length of 35 amino acids and motif number of 500. Zorzet *et al.*[112] introduced a computational approach for classifying the amino acid sequences in allergens and non-allergens. They identified 91 pre-processed food allergens from various specialized public repositories of food allergy and the SWALL database (SWISSPROT and TrEMBL).

Saha and Raghava[113] created AlgPred (http://www.imtech.res.in/raghava/algpred) using SVM and a similarity-based approach for analysis, and scanned all 183 IgE epitopes against all proteins of the dataset. The server allows use of a hybrid option to predict allergens using a combined approach (SVMc, IgE epitope, ARPs BLAST and MAST).

Stadler and Stadler[109] used the MEME motif discovery tool to identify the most relevant motif present in an allergen sequence. If the query finds an allergen motif or scores better than an E-value of 10^{-8} in the pairwise sequence alignment step, it is considered as the allergenic sequence. Then, these are compared with the FAO/WHO guidelines by performing allergenicity prediction for the sequence in SWISSPROT and a synthetic test database. Allermatch[114] (http://www.allermatch.org) is a webtool that uses a sliding window approach to predict potential allergenicity of proteins. It is done according to the current recommendations of the FAO/WHO Expert Consultation,[115] as outlined in Codex alimentarius.[116] But this method generates false-positive and false-negative hits so it is advised by the FAO/WHO that the outcomes should be combined with other allergenicity assessment methods.

The APPEL[117] (Allergen Protein Prediction E-Lab) tool uses SVM to identify novel allergen proteins. This tool correctly classified 93% of 229 allergens and 99·9% of 6717 non-allergens. It is based on a statistical method and has the potential to discover novel allergen proteins. The EVALLER[118] web server (http://bioinformatics.bmc.uu.se/evaller.html) uses a filtered length-adjusted allergen peptides (DFLAP) method[119] (via ulfh@slv.se) to identify the potential allergen proteins. DFLAP extracts variable length allergen sequence fragments and employs SVM. An uncertainty score has shown that the EVALLER is much more confident in identifying the 'presumably an allergen' category than that of non-allergens.

The EVALLER and APPEL servers assigned all calmodulins or calmodulin-like proteins as presumably non-allergens.[118] But a conventional alignment approach (e.g. 35% similarity over 80 amino acid segments) gives preference to finding sequence similarity between input proteins and known allergens and put the above-mentioned proteins the in allergen category. These proteins are presumably non-allergenic homologues to the polcalcin family (members being potential allergens involved in pollen–pollen cross-sensitization). Tools based on structural and physical characteristics are useful to identify potential cross-reacting proteins that may escape detection through the sequence similarity method alone.

Applications of immunoinformatics

In this section, we focus on applications of immunoinformatics. It includes *in silico* vaccine design and immune system modelling.

In silico vaccination

It is easy to apply new approaches for vaccine design, as genome sequencing, comparative proteomics and immunoinformatics tools are well developed. 'Reverse vaccinology', a new concept, analyses the entire genome to identify potentially antigenic extracellular proteins and so helps to save time and money. It was pioneered for *Neisseria meningitides*, which is responsible for sepsis and meningococcal meningitides. The vaccine type is conjugate and is based on capsular polysaccharide. These vaccines are available for pathogenic *N. meningitides* A, C, Y and W135.[120]

Microarray technique for vaccine design

Through microarray technology, it is easy to screen genes of various pathogens in different growth states and conditions for vaccine design.[121] It reduces the number of genes useful for vaccine in a given genome. Signal peptides derived from genomic sequences, structural motifs and immunogenicity are important for vaccine development.

Epitope-driven approaches for vaccine design

These are comparatively more useful as they have no lethal effect like the whole protein vaccines. It may induce an immune response against immunodominant epitopes.[122] This kind of vaccine has a single start codon with an epitope which can be inserted consecutively in the construct.[123] The prediction of promiscuous binding ligands is considered to be a prerequisite for most subunit vaccine design strategies.[124]

Peptide-based vaccine design

Small peptides derived from epitopes are used as peptide-based vaccines. These peptides are recognized by MHC class I and therefore boost the immune response. Florea

et al.[125] described three novel classes of methods to predict MHC binding peptides, and a voting scheme to integrate them for improved results. The first method is based on quadratic programming applied to quantitative and qualitative data. The second method uses linear programming and the third one considers sequence profiles obtained by clustering known epitopes to score candidate peptides. This method is found to be better than other sequence-based methods for finding the MHC binders.

Alignment-free approach for vaccine design

Earlier approaches for the identification of antigens were dependent on sequence alignment, which had several drawbacks. Some proteins have similar structure and biological properties, but they may lack sequence similarity. To get rid of these limitations, a new alignment-free approach for antigen prediction has been proposed for which Doytchinova and Flower[126] used three datasets, one each for bacteria, viruses and tumours. The models were validated using leave-one-out cross-validation (LOO-CV) on the whole sets and by external validation using test sets. These models were implemented in a server called VaxiJen (http://www.darrenflower.info/VaxiJen/).

DNA vaccines

It has already been found that DNA vaccines can produce both cell-mediated and humoral immune responses, and are very useful in defending intracellular pathogens. DyNAVacs[127] (http://miracle.igib.res.in/dynavac/) incorporates different modules like codon optimization for heterologous expression of genes in bacteria, yeast and plants, mapping restriction enzyme sites, primer design, Kozak sequence insertion, custom sequence insertion and design of genes for gene therapy.

The software NERVE[128] (http://www.bio.unipd.it/molbinfo) helps in designing subunit vaccines against bacterial pathogens. It combines automation with an exhaustive treatment of vaccine candidate selection tasks by implementing and integrating six different kinds of analyses. Xiang *et al.*[129] developed a web-based database system, VIOLIN (Vaccine Investigation and Online Information Network) (http://www.violinet.org), which curates, stores and analyses published vaccine data. It contains four integrated literature mining and search programs: Litsearch, Vaxpresso, Vaxmesh and Vaxlert. They have developed a web-based vaccine design system called Vaxign,[130] which predicts possible vaccine targets. Major predicted features include subcellular location of a protein, transmembrane domain, adhesion probability, sequence conservation among genomes, sequence similarity to host (human or mouse) proteome, and epitope binding to MHC class I and class II.

Immune system modelling

Immune system modelling provides an integrated view of the immune system in both qualitative and quantitative terms. These models can test and find out the antigen–antibody interactions and immune responses for a particular antigen, in case of drug administration or testing of a vaccine candidate. This helps in reducing time and cost. Peters *et al.*[33] developed a hepatitis C virus infection model that could predict the results of tumour necrosis factor-α acting by blocking *de novo* infection, blocking viral replication or effecting virion clearance. A model can calculate the likelihood of HIV developing a drug-resistant mutation, if provided with certain replication and mutation rates. Using the visual modelling application described by Gong and Cai,[131] one can understand the adaptive immune system effectively. The hierarchical immune system consists of an inherent immune tier, an adaptive immune tier and an immune cell tier. It is designed and visualized with the Java Applet technique for simulation. For further simulation purpose, the learning of the antibody is implemented through the evolutionary mechanism of the immune algorithm. ImmunoGrid (http://www.immunogrid.org) and Virolab (http://www.virolab.org:080/virolab) projects are working to simulate immune systems. ImmunoGrid tries to simulate immune processes by combining experiments and computational studies while Virolab is attempting to develop a virtual laboratory for infectious diseases by examining the genetic causes of human illnesses.[121] SIMISYS 0.3[132] is another example of a software that models and simulates the innate and adaptive components of the immune system, based on computational framework of cellular automata. It simulates healthy and disease conditions by interpreting interactions among the cells including, macrophages, dendritic cells, B cells, T helper cells and pathogenic bacteria.

Exclusive computational approaches like mathematical modelling generate enormous amounts of data, but there should be a balance between virtual and real experimental data. Computationally generated data need to be formally tested and translated into real knowledge. The post-genomic era needs to exchange data from wet laboratory to simulation and vice versa.[133] The model should be accurate, easy to use and understandable to both model designers and biologists, who can verify their hypothesis through *in silico* experiments.

Conclusions and discussions

This review considers useful online immunological databases, tools and webservers. It is described how immunoinformatics is useful in reducing the time and cost involved in the traditional study of immunology. Immunoinformatics may be placed at the junction point

between experimental and computational approaches. It complements wet laboratory immunology.

Most of the existing methods tend to predict epitopes with high affinity to MHC molecules. These methods are indirect as they predict MHC binders instead of T-cell epitopes, as opposed to the earlier methods. It is hypothesized that the T cell recognizes a peptide of amphipathic nature. The hydrophobic terminal of the antigenic peptide reacts with MHC while the hydrophilic end interacts with the TR. Earlier approaches used this phenomenon. Methods based on predicting structural binding motifs need structural data generated by molecular biology. This approach scans epitopic sequences to find MHC binders. However, these approaches become useless if motifs are not present. They need the three-dimensional structure of the MHC–peptide complex, which is again a limitation.

A matrix-driven method needs information about each residue of interacting peptide, and thereby gives better results. Machine-learning techniques are quite good as they can deal with non-linear data. Earlier approaches have some limitations in handling real data (non-linear data). SVM (a statistical learning methodology) is a learning technique that supports continuous and categorical variables. SVM is better than ANN because it attains a global minimum and is capable of working with fewer training patterns.[134] Hence both sequence characteristics and computational techniques should be integrated to acquire higher prediction accuracy. Recently, the prediction of promiscuous peptides (capable of binding to a wide array of MHC molecules) is being given much emphasis. Screening of large-scale pathogens and mapping of T-cell epitopes allow identification of the prime target of epitope-based T-cell vaccine designs.

'Reverse vaccinology' is a revolution in immunology because it uses the whole spectrum of antigens. This helps in using pools of vaccine candidates that otherwise would be missed (because of poor or no *in vitro* experimental information or problems in culturing the specific pathogen).[134] It makes the available pools of vaccine candidates easier to use when designing therapeutic vaccines. As of now, different groups are applying reverse vaccinology approaches that show promising pre-clinical results.

Immunoinformatics models are being used that are analogous to and that simulate the real behaviour of immune system processes. These models help in understanding the kinetics of cells during immune responses. They make understanding the biological pathways and underlying mechanisms easier. The models are engineered in such a way that they can be studied and interpreted easily, and can be rebuilt if new experimental data are introduced. These mathematical models remove the uncertainty of systems; as they are found to be close to wet laboratory experiments this leads to designing the path for refinement and modelling new experiments. Computational modelling of the immune system provides scientific solutions to several problems but it should not be forgotten that they rely on assumptions only, so they cannot be directly compared with real biological data. They can be improved by the availability of more data, significant parameters, or by modifying the underlying equations. These changes can better mimic the biological interactions in an organism. Currently, models are designed to simulate the biological data only over a fixed time period.[135] There are no data for extended time spans available to validate the models. This limits the accuracy of the results. An ability of these models to show the system's changes over an extended time period for immune response in case of antigen attack or drug administration would reduce the necessity for experimental research.

Exploration of the immune response to a specific drug can be a future research area in the modelling field. Drug response to a host's immune system can be better studied through computational models. The effect of drug administration can be added to model the immune system to find the drug efficacy.[135]

Moreover, the field of immune system modelling provides ideas about the dose composition, drug dosage duration, age of the patient and other parameters. It can give new suggestions for the study of immune system function and drug function to treat certain diseases. These modelling capabilities may lead to the invention of drugs that can treat a disease in a more effective way and without any side-effects. Diseases that are characterized by complex interactions between the host cellular immune system and evolving pathogens such as HIV infection can be investigated by such models.

Disclosures

The authors have no financial disclosures.

References

1 Kimbrell DA, Beutler B. The evolution and genetics of innate immunity. *Nat Rev Genet* 2001; **2**:256–67.

2 Thomas K, Goldsby J, Osborne RA, Barbara A, Kuby J. *Kuby Immunology*, 6th edn. New York: WH Freeman and Co, 2006.

3 Korber B, LaBute M, Yusim K. Immunoinformatics comes of age. *PLoS Comput Biol* 2006; **2**:0484–92.

4 Gardy JL, Lynn DJ, Brinkman FSL, Hancock REW. Enabling a systems biology approach to immunology: focus on innate immunity. *Trends Immunol* 2009; **30**:249–62.

5 Davies MN, Flower DR. Harnessing bioinformatics to discover new vaccine. *Drug Discov Today* 2007; **12**:389–95.

6 Ortutay C, Vihinen M. Immunome Knowledge base (IKB): an integrated service for immunome research. *BMC Immunol* 2009; **10**. doi:10.1186/1471-2172-10-3.

7 Sette A, Fleri W, Peters B, Sathiamurthy M, Bui HH. A roadmap for the immunomics of category A-C pathogens. *Immunity* 2005; **22**:155–61.

8 Groot ASDe. Immunomics: discovering new targets for vaccine and therapeutics. *Drug Discov Today* 2006; **11**:203–9.

9 Grainger DJ. Immunomics: principles and practice. *IRTL* 2004; **2**:1–6.

10 Yates A, Chan CCW, Callard RE, George AJT, Stark J. An approach to modelling in immunology. *Brief Bioinform* 2001; **2**:245–57.

11 Kaplan No, Everse J, Dixon Je, Stolzenbach Fe, Lee Cy, Lee Clt, Taylor Ss, Mosbach K. Purification and separation of pyridine nucleotide-linked dehydrogenases by affinity chromatography techniques. *Proc Natl Acad Sci USA* 1974; **71**:3450–4.

12 Davey HM. Flow cytometric techniques for the detection of microorganisms. *Methods Cell Sci* 2004; **24**:91–7.

13 Durkin MM, Patricia A, Connolly PA, Wheat LJ. Comparison of radioimmunoassay and enzyme-linked immunoassay methods for detection of *Histoplasma capsulatum* var. *capsulatum* antigen. *J Clin Microbiol* 1997; **35**:2252–5.

14 Ma H, Shieh KJ, Lee SL. Study of ELISA technique. *Nat Sci* 2006; **4**:36–7.

15 Nishimaki T, Sagawa K, Motogi S, Saito K, Morito T, Yoshida H, Kasukawa R. A competitive inhibition test of enzyme immunoassay for the anti-nRNP antibody. *J Immunol Methods* 1987; **100**:157–60.

16 Levine MA, Thornton P, Forman SJ *et al.* Positive Coombs test in Hodgkin's disease: significance and implications. *Blood* 1980; **55**:607–11.

17 Wanga B, Huaa RH, Tiana Z-J, Chena N-S, Zhaoa F-R, Liua T-Q, Wanga Y-F, Tong G-Z. Identification of a virus-specific and conserved B-cell epitope on NS1 protein of Japanese encephalitis virus. *Virus Res* 2009; **141**:90–5.

18 Admon A, Barnea E, Ziv T. Tumor antigens and proteomics from the point of view of the major histocompatibility complex peptides. *Mol Cell Proteomics* 2003; **2**:388–98.

19 Boon T, Coulie PG, den Eynde BV. Tumor antigens recognized by T cells. *Immunol Today* 1997; **18**:267–8.

20 Gubler DJ. Dengue and dengue hemorrhagic fever. *Clin Microbiol Rev* 1998; **11**:480–96.

21 Amin N, Aguilar A, Chamac ho F *et al.* Identification of dengue-specific B-cell epitopes by phage-display random peptide library. *Malaysian J Med Sci* 2009; **16**:4–14.

22 Wang Y. Immunostaining with dissociable antibody microarrays. *Proteomics* 2004; **4**:20–6.

23 Magdalena J, Odling J, Qiang PH, Martenn S, Joakin L, Uhlen M, Hammarstrom L, Nilsson P. Serum microarrays for large scale screening of protein levels. *Mol Cell Proteomics* 2005; **4**:1942–7.

24 Sahin U, Tureci O, Pfreundschuh M. Serological identification of human tumor antigens. *Curr Opin Immunol* 1997; **9**:709–16.

25 Oelke M, Maus MV, Didiano D, June CH, Mackensen A, Schneck JP. *Ex vivo* induction and expansion of antigen-specific cytotoxic T cells by HLA-Ig coated artificial antigen-presenting cells. *Nat Med* 2003; **9**:619–24.

26 Groot DeAS, Sbai H, Aubin CS, Mcmurry J, Martin W. Immuno-informatics: mining genomes for vaccine components. *Immunol Cell Biol* 2002; **80**:225–69.

27 Quintana FJ, Hagedorn PH, Gad E, Yifat M, Eutan D, Cohen IR. Functional immunomics: microarray analysis of IgG autoantibody repertoires predicts the future response of mice to induced diabetes. *Proc Natl Acad Sci USA* 2004; **101**:14615–21.

28 Sampson HA. Food allergy – accurately identifying clinical reactivity. *Allergy* 2005; **60**:19–24.

29 Vegvar de HEN, Robinson WH. Microarray profiling of antiviral antibodies for the development of diagnostics, vaccines, and therapeutics. *J Clin Immunol* 2004; **111**:196–201.

30 Henry ENdeV, RamaRao A, Lawrence S, Paul JU, Harriet LR, Robinson WH. Microarray profiling of antibody responses against simian-human immunodeficiency virus: postchallenge convergence of reactivities independent of host histocompatibility type and vaccine regimen. *J Virol* 2003; **77**:11125–38.

31 Nahtman T, Jernberg A, Mahdavifar S, Zerweck J, Schutkowski M, Maeurer M, Reilly M. Validation of peptide epitope microarray experiments and extraction of quality data. *J Immunol Methods* 2007; **328**:1–13.

32 Braga-Neto UM, Marques ETA. From functional genomics to functional immunomics: new challenges, old problems, big rewards. *PLoS Comput Biol* 2006; **2**:651–62.

33 Peters B, Sidney J, Bourne P *et al.* The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol* 2005; **3**:1361–70.

34 Lynn DJ, Winsor GL, Chan C *et al.* InnateDB: facilitating systems-level analyses of the mammalian innate immune response. *Mol Syst Biol* 2008; **4**:1–11.

35 Barsky S, Gardy JL, Hancock R, Munzer T. Cerebral: a Cytoscape plugin for layout of and interaction with biological networks using subcellular localization annotation. *Bioinformatics* 2007; **23**:1040–2.

36 Shanon P, Markiel A, Ozier O *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* 2003; **13**:2498–504.

37 Saha S, Bhasin M, Raghava GPS. Bcipep: a database of B-cell epitopes. *BMC Genomics* 2005; **6**. doi:10.1186/1471-2164-6-79.

38 Huang J, Honda W. CED: a conformational epitope. *BMC Immunol* 2006; **7**:7.

39 Schlessinger A, Ofran Y, Yachdav G, Rost B. Epitome: database of structure-inferred antigenic epitopes. *Nucleic Acids Res* 2006; **34**:D777–80.

40 Blythe MJ, Doytchinova IA, Darren R. JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics* 2002; **18**:434–9.

41 Rammensee HG, Bachmann J, Emmerich NPN, Bachor OA, Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 1999; **50**:213–9.

42 Feldhahn M, Donnes P, Thiel P, Kohlbacher O. FRED – a framework for T-cell epitope detection. *Bioinformatics* 2009; **25**:2758–9.

43 Lefranc M-P, Giudicelli V, Ginestoux C *et al.* IMGT®, the international ImMunoGeneTics information system®. *Nucleic Acids Res* 2009; **37**:D1006–12.

44 Lord P, Shah N, Sansone SA, Stephens S, Soldatova L (eds). The OBI Consortium. Modeling biomedical experimental processes with OBI, In: *Proceedings of the 12th Annual Bio-Ontologies 35 Meeting. International Society for Computational Biology*. Sweden: Stockholm 2009;41–4.

45 Mari A, Scalab E, Palazzob P, Ridolfib S, Zennarob D, Carabella G. Bioinformatics applied to allergy: allergen databases, from collecting sequence information to data integration. The Allergome platform as a model. *Cell Immunol* 2006; **244**:97–100.

46 Pomes A. Relevant B cell epitopes in allergic disease. *Int Arch Allergy Immunol* 2010; **152**:1–11.

47 Hoffman D, Lowenstein H, Marsh DG, Platts-Mills TAE, Thomas W. Allergen nomenclature. *Bull World Health Organ* 1994; **72**:796–806.

48 Kim C, Kwon S, Lee G, Lee H, Choi J, Kim Y, Hahn J. A database for allergenic proteins and tools for allergenicity prediction. *Bioinformation* 2009; **3**:344–5.

49 Ivanciuc O, Schein CH, Braun W. SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res* 2003; **31**:359–62.

50 Ortutay C, Siermala M, Vihinen M. ImmTree: database of evolutionary relationships of genes and proteins in the human immune system. *Immunome Res* 2007; **3**: doi:10.1186/1745-75803-4.

51 Ortutay C, Vihinen M. Immunome: a reference set of genes and proteins for systems biology of the human immune system. *Cell Immunol* 2006; **244**:87–9.

52 Rannikko K, Ortutay C, Vihinen M. Immunity genes and their orthologs: a multi-species database. *Int Immunol* 2007; **19**:1361–70.

53 Greenbaum JA, Andersen PH, Blythe M *et al.* Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools. *J Mol Recognit* 2007; **20**:75–82.

54 Tong JC, Ren EC. Immunoinformatics: current trends and future directions. *Drug Discov Today* 2009; **14**:684–9.

55 Pellequer J, Westhof E, Regenmortel MV. Predicting the location of structure of continuous epitopes in proteins from their primary structure. *Methods Enzymol* 1991; **203**:176–201.

56 Parker J, Guo D, Hodges R. New hydrophilicity scale derived from High-Performance Liquid Chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites. *Biochemistry* 1986; **25**:5425–32.

57 Chou PY, Fasman GD. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv Enzymol Relat Areas Mol Biol* 1978; **47**:45–148.

58 Levitt M. Conformational preferences of amino acids in globular proteins. *Biochemistry* 1978; **17**:4277–85.

59 Emini E, Hughes J, Perlow D, Boger J. Induction of hepatitis A virus-neutralizing antibody by a virus specific synthetic peptide. *J Virol* 1985; **55**:836–9.

60 El-Manzalawy Y, Dobbs D, Honavar V. Predicting protective linear B-cell epitopes using evolutionary information. In: *Proceedings of the 2008 IEEE International Conference on Bioinformatics and Biomedicine*, Washington: IEEE Computer Society 2008:289–92.

61 Odorico M, Pellequer JL. BEPITOPE: predicting the location of continuous epitopes and patterns in protein. *J Mol Recognit* 2003; **16**:20–2.

62 Pellequer JL, Westhof E. PREDITOP: a program for antigenicity predictions. *J Mol Graph* 1993; **11**:204–10.

63 Saha S, Raghava GPS. BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. In: Nicosia G, Cutello V, Bentley PJ, Timis J eds. *Artificial Immune Systems*. Berlin/Heidelberg:: ICARIS Springer, LNCS 2004; 3239:197–204.

64 Ghate AD, Bhagwat BU, Bhosle SG, Gadepalli SM, Kulkarni-Kale UD. Characterization of antibody-binding sites on proteins: development of a knowledge base and its applications in improving epitope prediction. *Protein Pept Lett* 2007; **14**:531–5.

65 Saha S, Raghava GPS. Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* 2006; **65**:40–8.

66 Sweredoski MJ, Baldi P. COBEpro: a novel system for predicting continuous B-cell epitopes. *Protein Eng Des Sel* 2009; **22**:113–20.

67 Larsen JEP, Lund O, Nielsen M. Improved method for predicting linear B cell epitopes. *Immunome Res* 2006; doi:10.1186/1745-7580-2-2.

68 Toseland CP, Clayton DJ, McSparron H *et al.* AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. *Immunome Res* 2005; **1**. doi:10.1186/1745-7580-1-4.

69 Evans MC. Recent advances in immunoinformatics: application of *in silico* tools to drug development. *Curr Opin Drug Discov Devel* 2008; **11**:233–41.

70 Anderson Ph, Nielsen M, Lund O. Prediction of residues in discontinuous B cell epitopes using protein 3D structures. *Protein Sci* 2006; **15**:2558–67.

71 Bublil EM, Mayrose NTFI, Penn O, Berman AR. Stepwise prediction of conformational discontinuous B-cell epitopes using the mapitope algorithm. *Proteins* 2007; **68**:294–304.

72 Sollner J, Grohmann R, Rapberger R, Perco P, Lukas A, Mayer B. Analysis and prediction of protective continuous B cell epitopes on pathogen proteins. *Immunome Res* 2008; **4**. doi:10.1186/1745-7580-4-1.

73 Kale KU, Bhosle S, Kolaskar AS. CEP: a conformational epitope prediction server. *Nucleic Acids Res* 2005; **33**:W168–71.

74 Mayrose I, Penn O, Erez E *et al.* Pepitope: epitope mapping from affinity-selected peptides. *Bioinformatics* 2007; **23**:3244–6.

75 Moreau V, Granier C, Villard S, Laune D, Molina F. Discontinuous epitope prediction based on mimotope analysis. *Bioinformatics* 2006; **22**:1088–95.

76 Huang J, Gutteridge A, Honda W, Kanehisa M. MIMOX: a web tool for phage display based epitope mapping. *BMC Bioinformatics* 2006; **7**. doi: 10.1186/1471-2105-7-451.

77 Mayrose I, Shlomi T, Rubinstein ND, Gershoni JM, Ruppin E, Sharan R, Pupko T. Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm. *Nucleic Acids Res* 2007; **35**:69–78.

78 Schreiber A, Humbert M, Benz A, Dietrich U. 3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins. *J Comput Chem* 2005; **26**:879–87.

79 Huang L, Dai Y. Direct prediction of T-cell epitopes using support vector machines with novel Sequence encoding schemes. *J Bioinform Comput Biol* 2006; **4**:93–107.

80 Bhasin M, Raghava GPS. Prediction of promiscuous and high-affinity mutated MHC binders. *Hybrid Hybridomics* 2003; **22**:229–34.

81 Zhang GL, Petrovsky N, Kwoh CK, August JT, Brusic V. Pred^{TAP} : a system for prediction of peptide binding to the human transporter associated with antigen processing. *Immunome Res* 2006; **2**. doi: 10.1186/1745-7580-2-3.

82 Neilsen M, Lundegaard C, Worning P, Lauemoller SL, Lamberth K, Buus S, Brunak S, Lund O. Reliable prediction of T-cell epitopes using networks with novel sequence representations. *Protein Sci* 2003; **12**:1007–17.

83 Buus S, Stryhn A, Winther K, Kirkby N, Pedersen LO. Receptor–ligand interactions measured by an improved spun column chromatography technique. A high efficiency and high throughput size separation method. *Biochim Biophys Acta* 1995; **1243**:453–60.

84 Larsen MV, Lundegaard C, Lamberth K, Buss S, Lund O, Nielsen M. Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction. *BMC Bioinformatics* 2007; **8**. doi: 10.1186/1471-2105-8-424.

85 Lundegaard C, Lamberth K, Harndahl M, Buus S, Lund O, Nielsen M. NetMHC-3.0: accurate web accessible predictions of human mouse and monkey MHC class I affinities for peptides of length 8–11. *Nucleic Acids Res* 2008; **36**:W509–12.

86 Brusic V, Rudy G, Honeyman M, Hammer J, Harrison L. Prediction of MHC class II-binding peptides using an evolutionary and artificial neural network. *Bioinformatics* 1998; **14**:121–30.

87 Miyata J. A User's Guide to PlaNet Version 5.6. Boulder: Computer Science Department, University of Colorado. 1991.

88 Nanni L. Machine learning algorithms for T-cell epitopes prediction. *Neurocomputing* 2006; **69**:866–8.

89 Bhasin M, Raghava GPS. Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci* 2004; **13**:596–607.

90 Bhasin M, Raghava GPS. Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences. *Nucleic Acids Res* 2005; **33**:W202–7.

91 Joachims T. Marking large-scale support vector machine learning practical. In: Scholkopf B, Burges CJC, Smola AJ, eds. *Advances in Kernel Methods: Support Vector Learning*. Cambridge, MA, USA: MIT Press, 1999:169–84.

92 Cost S, Salzberg S. A weighted nearest neighbor algorithm for learning with symbolic features. *Mach Learn* 1993; **10**:57–78.

93 Witten IH, Frank E. *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, 2nd edn. San Francisco: Morgan Kaufman, 1999.

94 Dorigo M, Maniezzo MV, Cambini AC. Ant system: optimization by a colony of cooperating agents. *IEEE Trans Syst Man Cybern Part B* 1996; **26**:29–41.

95 Flower DR. Towards *in silico* prediction of immunogenic epitopes. *Trends Immunol* 2003; **24**:667–74.

96 Bian H, Hammer H. Discovery of promiscuous HLA restricted T cell epitope with TEPITOPE. *Methods* 2004; **34**:468–75.

97 Kangueane P, Sakharkar MK. T epitope designer: HLA peptide binding prediction server. *Bioinformation* 2005; **1**:21–4.

98 Zhao B, Mathura VS, Ganapathy R, Moochhala S, Sakharkar MK, Kangneane P. A novel MHCp binding prediction model. *Hum Immunol* 2003; **64**:1123–43.

99 Ponomarenko JV, Bui HH, Li W, Fusseder N, Bourne PE, Sette A, Peters B. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics* 2008; **9**. doi: 10.1186/1471-2105-9-514.

100 Guan P, Doytchinova IA, Zygouri C, Flower DR. MHCPred: a server for quantitative prediction of peptide-MHC binding. *Nucleic Acids Res* 2003; **31**:3621–4.

101 Schiewe AJ, Haworth IS. Structure based prediction of MHC–peptide association: algorithm comparison and approach to cancer vaccine design. *J Mol Graph Model* 2007; **26**:667–75.

102 Jojic N, Gomez MR, Heckerman D, Kadle C, Furman OS. Learning MHC-I peptide binding. *Bioinformatics* 2006; **22**:e227–35.

103 Furman OS, Altuvia Y, Sette A, Margalit H. Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci* 2000; **9**:1838–46.

104 Miyazawa S, Jernigan RL. Residue–residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol* 1996; **256**:623–44.

105 Altuvia Y, Margalit H. A structure-based approach for prediction of MHC-binding peptides. *Methods* 2004; **34**:454–9.

106 Singh H, Raghava GPS. Propred: prediction of HLA-DR binding sites. *Trends Immunol* 2001; **17**:1236–7.

107 Sturniolo T, Bono E, Ding J *et al.* Generation of tissue-specific and promiscuous HLA ligand database using DNA microarrays and virtual HLA class II matrices. *Nat Biotechnol* 1999; **17**:555–61.

108 Feldhahn M, Thiel P, Schuler MM, Hillen N, Stevanovic S, Rammensee HG, Ohlbacher O. EpiToolKit – a web server for computational immunomics. *Nucleic Acids Res* 2008; **1**:W519–22.

109 Stadler MB, Stadler BM. Allergenicity prediction by protein sequence. *FASEB J* 2003; **17**:1141–3.

110 Kong W, Tan TS, Tham L, Choo KW. Improved prediction of allergenicity by combination of multiple sequence motifs. *In Silico Biol* 2006; **7**:77–86.

111 Bjorklund AK, Atmadja SD, Zorzet A, Hammerling U, Gustafsson MG. Supervised identification of allergen-representative peptides for *in silico* detection of potentially allergenic proteins. *Bioinformatics* 2005; **21**:39–50.

112 Zorzet A, Gustafsson M, Hammerling U. Prediction of food protein allergenicity: a bioinformatic learning systems approach. *In Silico Biol* 2002; **2**:525–34.

113 Saha S, Raghava GPS. AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res* 2006; **34**:W202–9.

114 Fiers MWEJ, Kleter GA, Nijland H, Peijnenburg AACM, Peter NJ, Ham RCHJV. Allermatch™, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinformatics* 2004; **5**. doi: 10.1186/1471-2105-5-133.

115 FAO/WHO. Allergenicity of Genetically Modified Foods. 2001; Available at http://www.who.int/foodsafety/publications/biotech/en/ec_jan2001.pdf.

116 FAO/WHO. Codex Principles and Guidelines on Foods Derived from Biotechnology. 2003; Available at ftp://ftp.fao.org/codex/standard/en/CodexTextsBiotechFoods.pdf.

117 Cui J, Han LY, Li H, Ung CY, Tang ZQ, Zheng CJ, Cao ZW, Chen YZ. Computer prediction of allergen proteins from sequence-derived protein structural and physicochemical properties. *Mol Immunol* 2007; **44**:514–20.

118 Barrio AM, Atmadja DS, Nistr A, Gustafsson MG, Hammerling U, Rudloff EB. EVALLER: a web server for *in silico* assessment of potential protein allergenicity. *Nucleic Acids Res* 2007; **35**:694–700.

119 Soeria-Atmadja D, Lundell T, Gustafsson MG, Hammerling U. Computational detection of allergenic proteins attains a new level of accuracy with *in silico* variable-length peptide extraction and machine learning. *Nucleic Acids Res* 2006; **34**:3779–93.

120 Pizza M, Scarlato V, Masignani V *et al.* Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 2000; **287**:1816–20.

121 Groot ASDe, Rappuoli R. Genome derived vaccines. *Expert Rev Vaccines* 2003; **3**:59–76.

122 Gallimore A, Hengartner H, Zinkernagel R. Hierarchies of antigen-specific cytotoxic T cell responses. *Immunol Rev* 1998; **164**:29–36.

123 Morris S, Kelly C, Howard A, X Li, Collins F. The immunogenicity of single and combination DNA vaccines against tuberculosis. *Vaccine* 2000; **18**:2155–63.

124 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**:127–40.

125 Florea L, Haldorsson B, Kohlbacher O, Schwarty R, Hoffman S, Istrail S. Epitope prediction algorithm for peptide-based vaccine design. In *Proceedings of the IEEE Computer Society Conference on Bioinformatics*, Washington: IEEE Computer Society. 2003:17–26.

126 Doytchinova IA, Flower DR. VaxiJen: a server for prediction of protective antigens, tumor antigens and subunit vaccines. *BMC Bioinformatics* 2007; **8**. doi: 10.1186/1471-2105-8-4.

127 Nagarajan H, Gupta R, Agarwal P, Scaria V, Pillai B. DyNAVacS: an integrative tool for optimized DNA vaccine design. *Nucleic Acids Res* 2006; **34**:W264–6.

128 Vivona S, Bernante F, Filippini F. NERVE: new enhanced reverse vaccinology environment. *BMC Biotechnol* 2006; **6**. doi: 10.1186/1472-6750-6-35.

129 Xiang Z, Todd T, Ku KP *et al.* VIOLIN: vaccine investigation and online information network. *Nucleic Acids Res* 2008; **36**:D923–8.

130 Xianga Z, He Y. Vaxign: a web-based vaccine target design program for reverse vaccinology. *Procedia in Vaccinology* 2009; **1**:23–9.

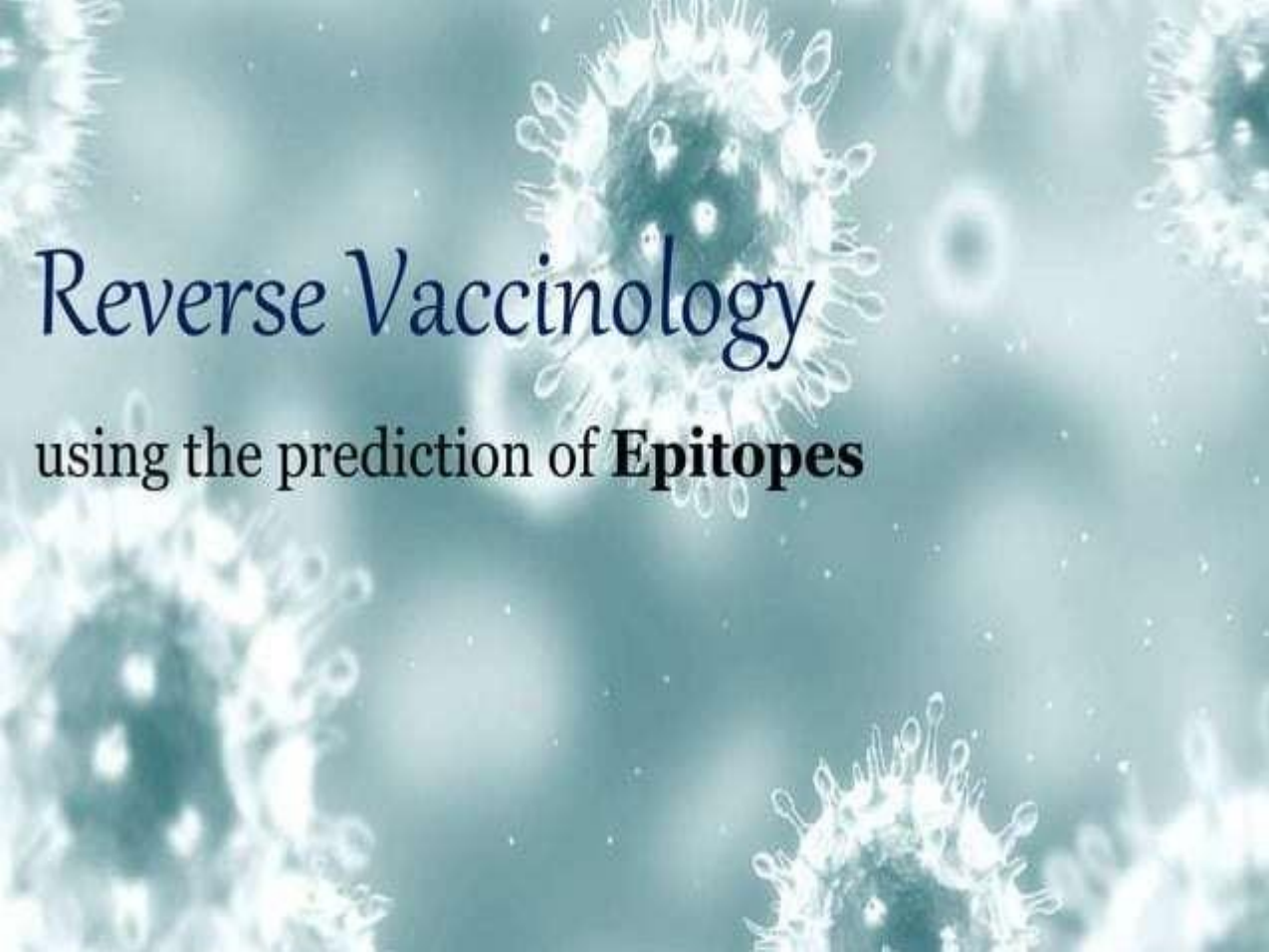
131 Gong T, Cai Z. Visual modeling and simulation of adaptive immune system. In *Proceedings of the 2005 IEEE Engineering in Medicine and Biology 27th Annual Conference*, Shanghai, China, 2005; 6:6116–9.

132 Kalita JK, Chandrashekar K, Hans R, Selvam P, Newell MK. Computational modelling and simulation of the immune system. *Int J Bioinform Res Appl* 2006; **2**:63–88.

133 Castiglione F, Liso A. The role of computational models of the immune system in designing vaccination strategies. *Immunopharmacol Immunotoxicol* 2005; **27**:417–32.

134 Vivona S, Gardy JL, Ramachandran S, Brinkman FSL, Raghava GPS, Flower DR, Filippini F. Computer-aided biotechnology: from immuno-informatics to reverse vaccinology. *Trends Biotechnol* 2008; **26**:190–200.

135 Daz P, Gillespie M, Krueger J, Prez J, Radebaugh A, Shearman T, Vo G, Wheatley C. A mathematical model of the immune system's response in obesity-related chronic inflammation. In: *McNair/MAOP Summer Research Symposium, Virginia Tech, Blacksburg VA*. 2008; 2:26–45.

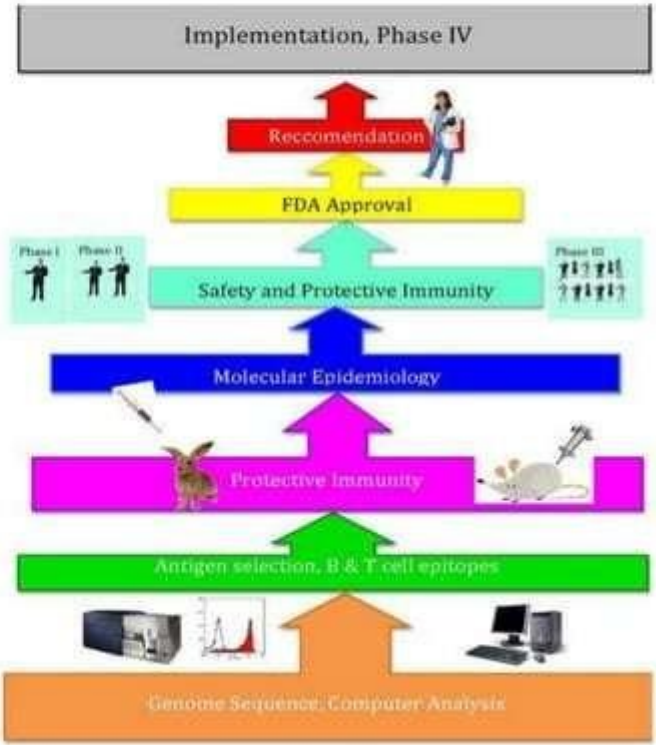

Reverse Vaccinology
using the prediction of **Epitopes**

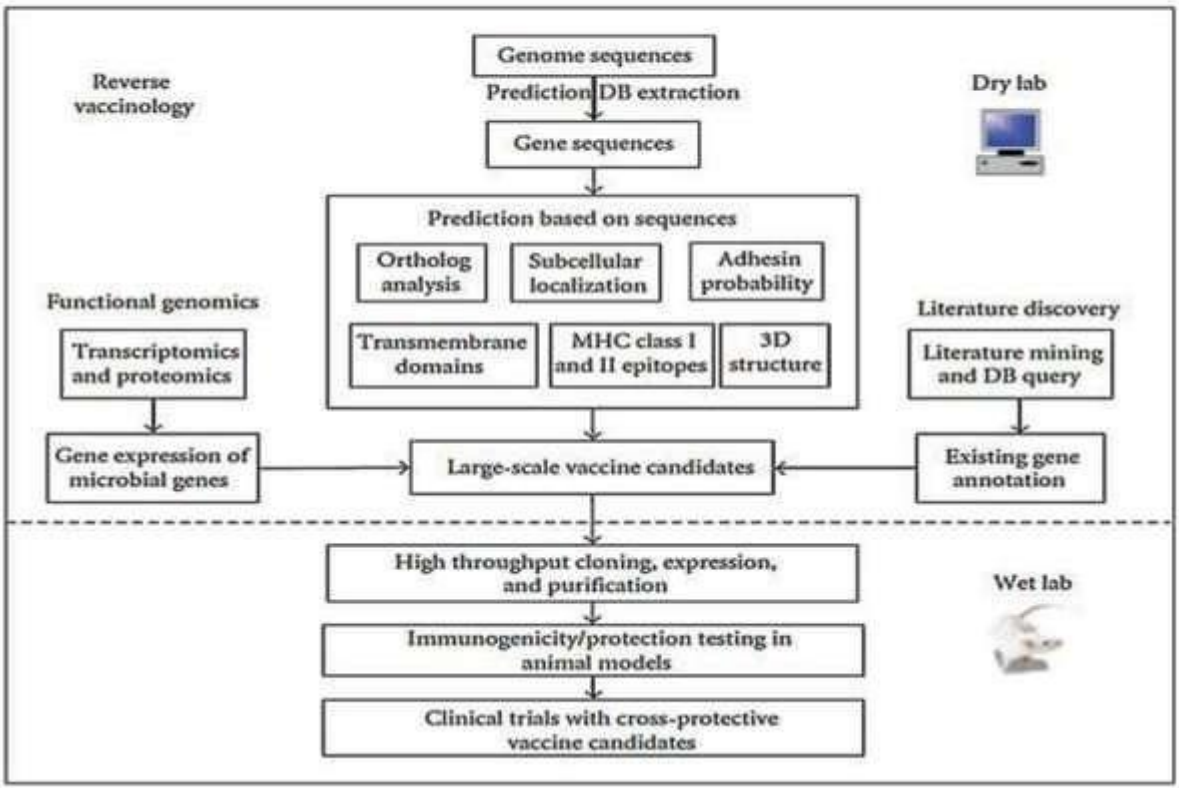
Objective

- To minimize the Laboratory based research.
- To develop a computational analysis of Antigens using Bioinformatics tools.
- To design a molecule that can replace an antigen in detection process.
- For the development of immunodiagnostic tests and vaccines.
- For detection of antibodies produced as a result of infections, allergies, autoimmune diseases, or cancers.

REVERSE VACCINOLOGY

- The technique of identifying the proteins that are exposed on the surface by using genome instead of the microorganism, this novel approach is known as "reverse vaccinology"
- Reverse vaccinology is an improvement on vaccinology that employs bioinformatics, pioneered by Rino Rappuoli and first used against Serogroup B meningococcus.




Reverse vaccinology
Genome sequences
Prediction DB extraction
Gene sequences
Dry lab
Prediction based on sequences
Ortholog analysis
Subcellular localization
Adhesin probability
Transmembrane domains
MHC class I and II epitopes
3D structure
Functional genomics
Transcriptomics and proteomics
Literature discovery
Literature mining and DB query
Gene expression of microbial genes
Large-scale vaccine candidates
Existing gene annotation
High throughput cloning, expression, and purification
Wet lab
Immunogenicity/protection testing in animal models
Clinical trials with cross-protective vaccine candidates

Conventional Vaccinology

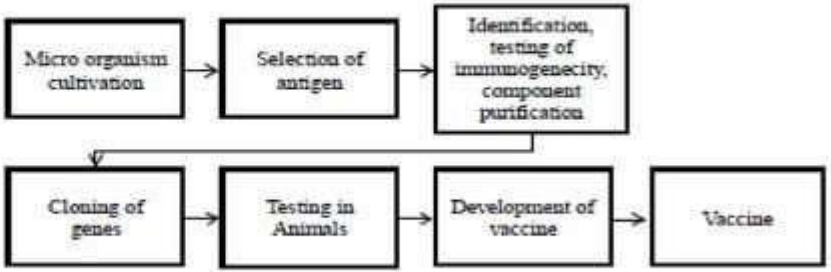


Figure 2: Pipeline for the development of the vaccines by conventional means.

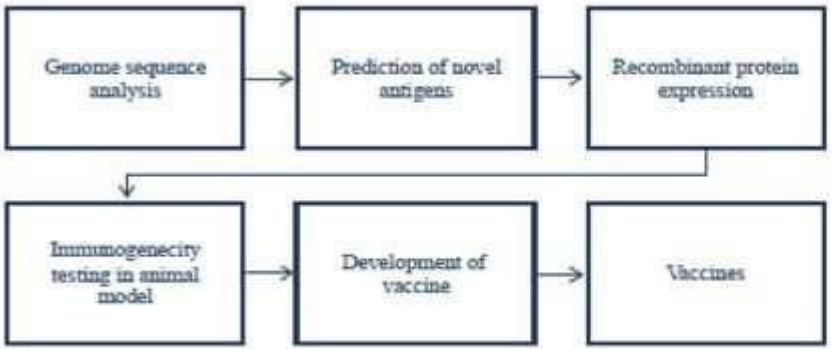
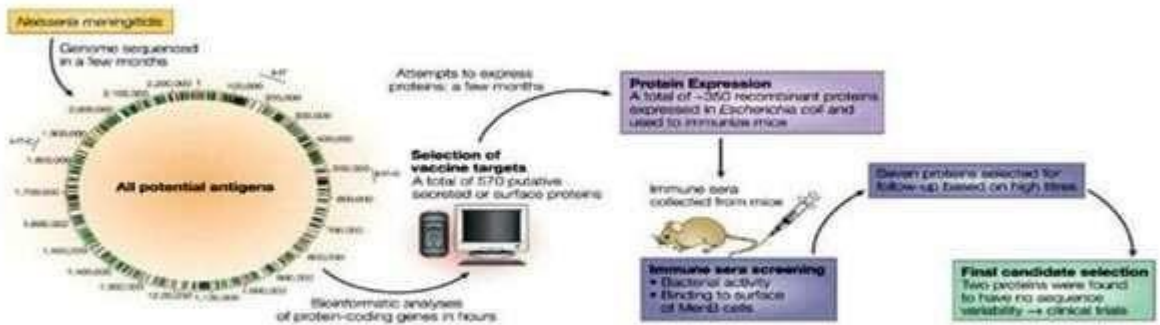


Figure 3: Pipeline for the development of the vaccines by reverse vaccinology.

Reverse Vaccinology

Reverse Vaccinology with Meningococcus B

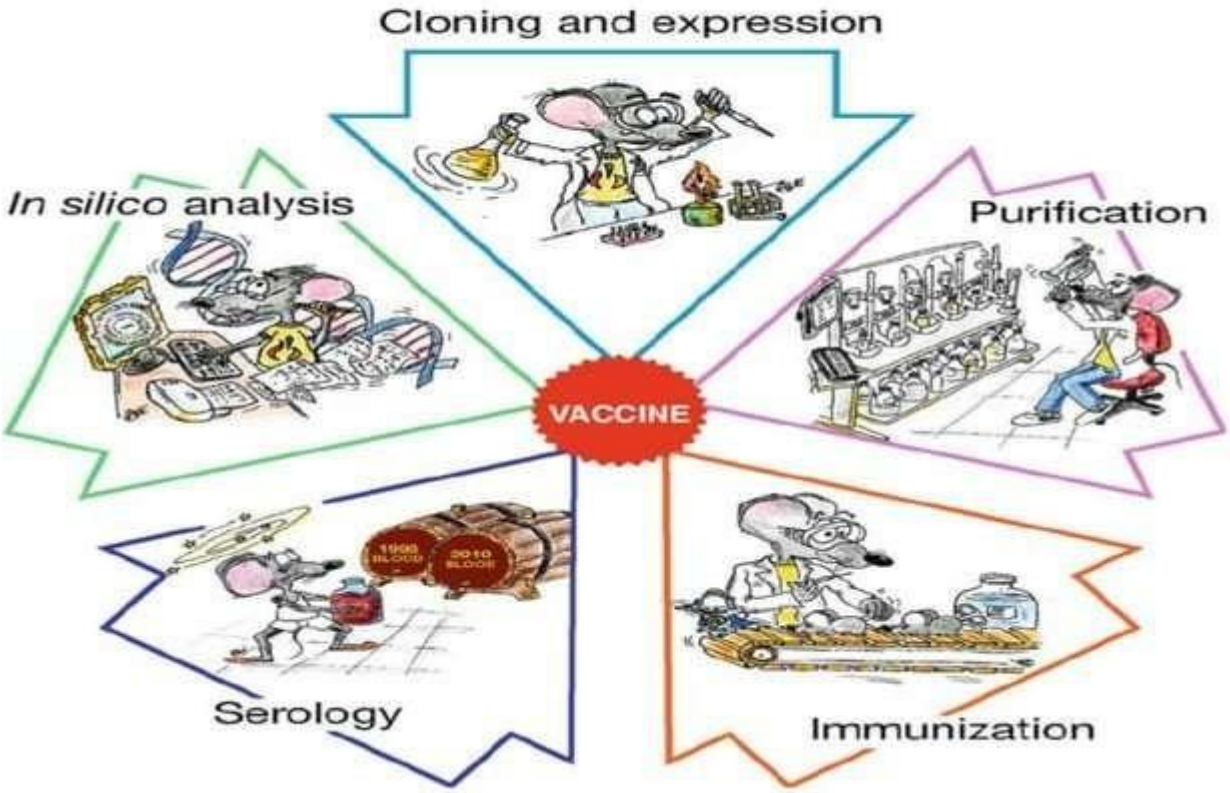
- Rappuoli and others at the J. Craig Venter Institute first sequenced the MenB genome.
- Then, they scanned the sequenced genome for potential antigens.
- They found over 600 possible antigens, which were tested by expression in Escherichia coli.
- Several proved to function successfully in mice, however, not alone for a good immune response.
- The addition of outer membrane vesicles that contain lipopolysaccharides (adjuvant) previously identified by using conventional vaccinology approaches enhanced immune response to the level that was required.
- Later, the vaccine was proven to be safe and effective in adult humans



Available Softwares

Computer Aided bioinformatics projects are extremely popular, as they help guide the laboratory experiments. Some of them popularly used for Reverse Vaccinology are-:

- NERVE - one relatively new dataprocessing program
- Vaxign- an even more comprehensive program, was created in 2008. Vaxign is web-based and completely public-access extremely accurate and efficient
- RANKPEP – an Online software, for the peptide bonding predictions.


Cloning and expression
In silico analysis
Purification
VACCINE
Serology
Immunization

EPITOPE PREDICTION

- Antigens are substances that can be recognised by Ab receptor of B-cells and T-cells when complexed with TCRs or MHC molecules.
- It can be synthesized or, in case of a protein, its gene can be cloned into an expression vector.
- Epitope prediction means to discover peptides that could mimic protein epitopes and possess the same immunogenicity as the whole protein.

Role of Epitope Prediction in Reverse Vaccinology

- An epitope is an antigenic determinant which are present on the surface of organisms that can be detected by the antibody.
- As the techniques are available for studying host- pathogen interactions, whole genome study and every unique gene, the work is now focused on the development of epitope driven vaccines that are target specific.
- Reverse vaccinology deals with computational analysis of genome that can be used for the prediction of the epitopes that are surface proteins. So the epitopes play an important role in development of a candidate vaccine.

- The major role played in immune system is by B and T lymphocyte.
- B cells are important in recognizing the epitopes of the antigens that can be identified by the paratopes of antibody.
- In some cases, T cells play a role in cell mediated immunity as the processed antigenic peptides interact with the T cell when they are presented in context of T cell.
- So the prediction of the epitopes of T and B cell plays an important role in determination of the candidate vaccine.
- The epitope prediction plays an important role in designing of epitope based vaccine.

Discontinuous Epitope Prediction

- Based on the knowledge of the protein three-dimensional structure.
- Discontinuous epitopes with a known 3D structure can be reconstituted from the antibody binding peptides selected from randomized peptide libraries.
- Several bioinformatics tools address the convenience for structural studies-
- 3D-Epitope-Explorer (3DEX), MIMOX, Epitope Mapping Tool (EMT)), EPIMAP, MIMOP, PepSurf and Mapitope.
- The MEPS server facilitates a structure-based design of peptides representing the whole surface or a particular region of a protein.

METHODS FOR B-CELL EPITOPE PREDICTION

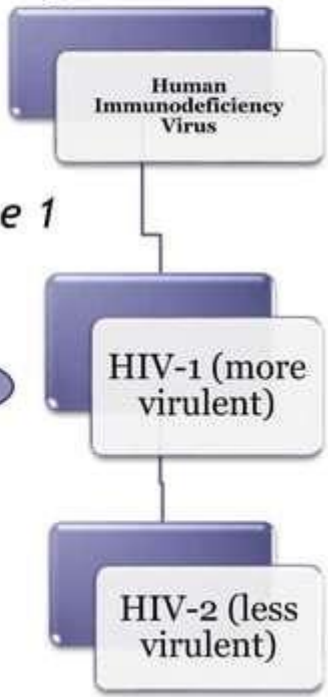
Structural (X-ray crystallography, nucleic magnetic resonance (NMR), and electron microscopy (EM) of antibody antigen complexes)

Functional (Mass Spectrometry as well as immunoassays, including ELISA, ELISPOT, Western blot)

Predicting T-cell epitopes in HIV

AIDS- a condition in humans that leads to a progressive failure of the Immune system.

- *Use of* Immunoinformatics *to identify and map the T-cell epitopes in proteins of HIVtype 1 is far efficient than the conventional culturing of HIV in labs.*
  - Time consuming
  - Difficult
- *Selected Protein sequence -*
  - *NEF*
  - *ENV*



- *Currently, diagnosis is done by Immunoassays like-*
  - ELISA
  - Western blotting

- *To avoid wet lab approach, we apply the use of Bioinformatics for prediction of epitopes and later, chemically synthesizing them for diagnosis and development of peptide based vaccine.*

- *In this study, used online Tools are-*
  - Propred
  - Propred- 1

PROS	CONS
• Theoretical prediction is a highly challenging task • In- silico approach • Fast and efficient • For detection of antibodies produced as a result of infections, allergies, autoimmune diseases, or cancers.	• Only proteins can be targeted using this process, no ther biomolecules like Polysaccharide

Recent Advancements

- Research efforts are also underway to develop synthetic peptide vaccines against-
  - HIV
  - Human T-cell leukemia virus type 1 (HTLV-1)
  - Streptococcus pyogenes infections
  - Malaria
  - Severe acute respiratory syndrome (SARS).

- Also, synthetic peptides are considered as therapeutic vaccines to cure cancer, Alzheimer, and autoimmune diseases.

Thank You

Reverse vaccinology

Rino Rappuoli

Biochemical, serological and microbiological methods have been used to dissect pathogens and identify the components useful for vaccine development. Although successful in many cases, this approach is time-consuming and fails when the pathogens cannot be cultivated *in vitro*, or when the most abundant antigens are variable in sequence. Now genomic approaches allow prediction of all antigens, independent of their abundance and immunogenicity during infection, without the need to grow the pathogen *in vitro*. This allows vaccine development using non-conventional antigens and exploiting non-conventional arms of the immune system. Many vaccines impossible to develop so far will become a reality. Since the process of vaccine discovery starts in silico using the genetic information rather than the pathogen itself, this novel process can be named reverse vaccinology.

Addresses
IRIS, Chiron S.p.A., Via Fiorentina 1, 53100 Siena, Italy;
e-mail: Rino_Rappuoli@biocine.it

Current Opinion in Microbiology 2000, **3**:445–450

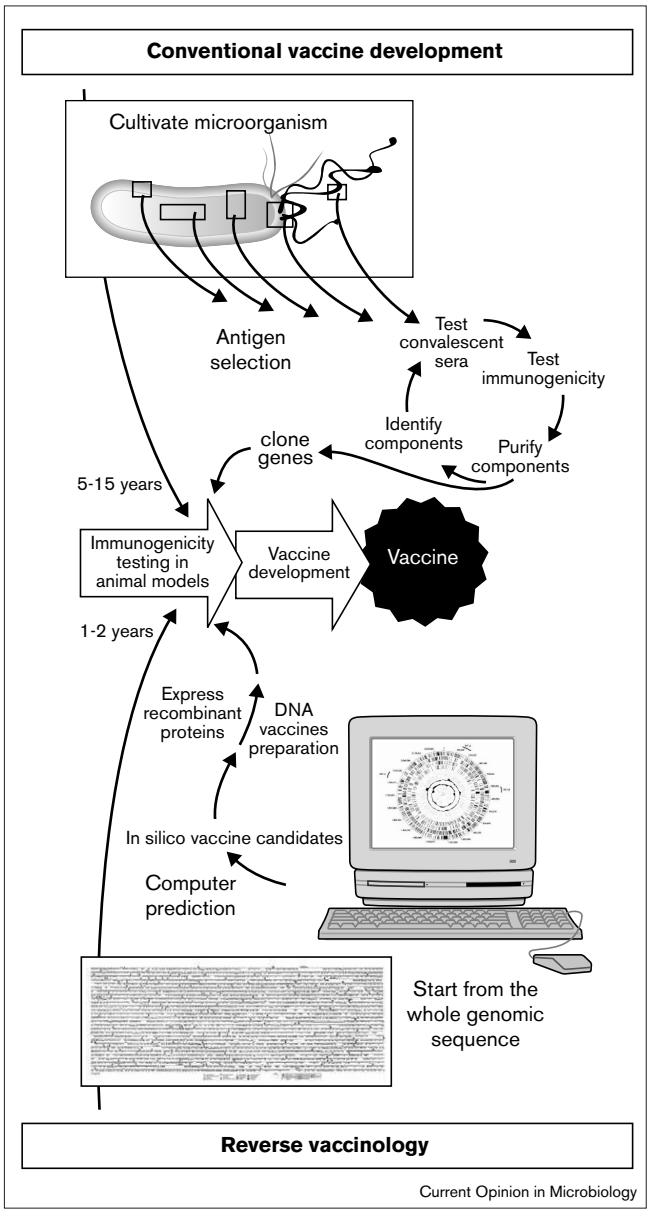
1369-5274/00/$ – see front matter

Abbreviations
HCV hepatitis C virus
Men B group B meningococcus

Introduction: conventional vaccinology

The conventional approach to vaccine development uses two methods: first, attenuation of pathogens by serial passages *in vitro* to obtain live-attenuated strains to be used as vaccines, and second, identification of protective antigens to be used in non-living, subunit vaccines [1]. In this review, we focus on subunit vaccines. The conventional way to develop these vaccines is summarized in Figure 1. In order to identify the components of the pathogen suitable for vaccine development, the pathogen is grown in laboratory conditions and the components building the pathogen are first identified one at a time, by biochemical, serological or genetic methods. The identification of protective antigens that could be potential vaccine candidates involves separating each component of the pathogen one by one. This approach is time-consuming and allows the identification only of those antigens that can be purified in quantities suitable for vaccine testing. Since the most abundant proteins are most often not suitable vaccine candidates, and the genetic tools required to identify the less abundant components may be inadequate or not available at all, this approach can take years or decades. For the bacterial and parasitic pathogens studied to date, the maximum number of potential vaccine antigens identified during a century of vaccine development is usually less than ten. This conventional method also means that vaccine development is not possible when the pathogen cannot be grown in laboratory conditions. An exception to this has been the hepatitis B vaccine where the pathogen, although unable to grow *in vitro*, could be recovered in large quantities from the plasma of infected people [2].

Figure 1



Schematic representation of the essential steps of vaccine development by the conventional approach and by reverse vaccinology.

Once a suitable antigen is identified, it needs to be produced in large scale, often by growing the pathogen itself. Cloning of the gene coding for the antigen is often

necessary in order to better characterize and produce the identified antigen(s). Finally, the new molecule can enter vaccine development. Although successful in many cases, this approach took a long time to provide vaccines against those pathogens for which the solution was easy and failed to provide a solution for those bacteria and parasites that did not have obvious immunodominant protective antigens [3•].

Reverse vaccinology

The reverse approach to vaccine development takes advantage of the genome sequence of the pathogen. The genome sequence provides at once a catalog of virtually all protein antigens that the pathogen can express at any time. As shown in Figure 1, this approach starts from the genomic sequence and, by computer analysis, predicts those antigens that are most likely to be vaccine candidates. The approach can, therefore, be very naïve, and poses the question of whether any of the potential antigen candidates can provide protective immunity without knowing whether the antigen is abundant, immunogenic during infection or expressed *in vitro*. This approach allows not only the identification of all the antigens seen by the conventional methods, but also the discovery of novel antigens that work on a totally different paradigm. Therefore, this method allows the discovery of novel mechanisms of immune intervention. The feasibility of the approach relies heavily on the availability of a high-throughput system to screen protective immunity. When this is available, in theory all genes of a pathogen can be tested, without any bias of any type. Unfortunately, owing to our limited knowledge of vaccine immunology, good correlates of protection are rare and, therefore, screening for protective immunity is the rate-limiting step of reverse vaccinology. The other limit of this approach is the inability to identify non-protein antigens such as polysaccharides, which are important components of many successful vaccines, and the identification of CD1-restricted antigens such as glycolipids, which represent new promising vaccine candidates.

Applications of reverse vaccinology

The publication of the complete genome sequence of many bacteria, parasites and viruses means that the reverse approach to vaccine development can be put into practice. Below we discuss the different approaches that are being used or potentially could be used to develop novel and effective vaccines against a variety of pathogens.

Group B meningococcus

Group B meningococcus (MenB) represents the first example of the successful application of reverse vaccinology. The conventional approach to vaccine development against this pathogen had been struggling for four decades without progress. On the one hand, the capsular polysaccharide used to develop conventional and conjugate vaccines against all other pathogenic meningococci could not be used because the MenB capsule, which is chemically identical to an α2–8 linked polysialic acid present in many of our tissues, is poorly immunogenic and a potential cause of autoimmunity. On the other hand, the protein-based approach had identified as protective antigens the most abundant proteins of the outer membrane [4]. However, these abundant surface-exposed proteins usually contain many amphipathic domains, which span the outer membrane several times and assume a β-barrel conformation (Figure 2a). The protective epitopes in these proteins are located in the loops that are exposed on the external surface and are usually formed by the precise conformation of a few amino acids. Therefore, in order to induce protective immunity these antigens need to be folded within the outer membrane (recombinant proteins do not induce protection) and any change in one of the few amino acids of the loop will result in a different epitope. Vaccines based on outer membrane vesicles (OMV) and containing the major outer membrane proteins have been developed and shown to be efficacious in clinical trials; however, owing to the high sequence variability of the external loops in different MenB strains, protection is induced only against the immunizing strain. As a consequence, the conventional approach to vaccine development has failed to deliver a universal vaccine.

Using reverse vaccinology, fragments of DNA were screened by computer analysis while the MenB nucleotide genome sequence was being determined [5•,6••]. Six hundred novel genes were predicted to code for surface-exposed or exported proteins. These were cloned and expressed in *Escherichia coli* as fusions to the glutatione transferase or to a histidine tag. Of these fusion proteins, 350 were successfully expressed, purified and used to immunize mice. The sera obtained were used to confirm the surface exposure of the proteins by ELISA and FACS analysis, and to test for the ability to induce complement-mediated *in vitro* killing of bacteria, a test that correlates with vaccine efficacy in humans. Within 18 months, while the nucleotide sequence was still being finalized, 85 novel surface-exposed proteins were discovered and 25 of these were shown to induce bactericidal antibodies [6••]. These numbers are impressive if one considers that during the past four decades no more than a dozen of such proteins had been identified. The surprising finding was not only the high number of the new proteins found but also the quality of the new proteins. In addition to the conventional outer membrane proteins with variable surface-exposed loops (as in Figure 2a), many of the new proteins were lipoproteins or other types of surface-associated proteins without membrane-spanning domains (Figure 2b). These were often conserved in sequence, and carried multiple protective epitopes conserved in most strains. These novel proteins provide an optimal basis for the development of a novel and effective vaccine against MenB [6••].

Malaria

Malaria, together with AIDS and tuberculosis, belongs to the triad of the most dangerous diseases that threaten

human health. The 500 million new infections each year and 2.5 million annual deaths indicate that all measures used so far to control the disease have failed [7]. Vaccination would be an effective way to control the spread of malaria, but vaccines are not available, despite many years of research [3•]. Approximately 20 antigens have been identified from the malaria parasite but none of them is good enough for a vaccine. The problem is further complicated by the different antigenic profiles expressed by sporozoites, merozoites and gametocytes, that the parasite assumes during its life cycle. The solution can only come from a genomic approach. The sequence of two of the 14 chromosomes of *Plasmodium falciparum* have been published [8•,9•] and provided the full set of genes contained in the two chromosomes. The complete sequence of the whole genome will soon provide information on the predicted 6000 genes. Analysis of the whole genome expression will show which genes are expressed by the sporozoite, liver and sexual life-stages of the parasite. Expression of genes predicted to be immunogenic as recombinant proteins delivered with adjuvants or as DNA vaccines will eventually provide the effective vaccine against malaria [10••,11].

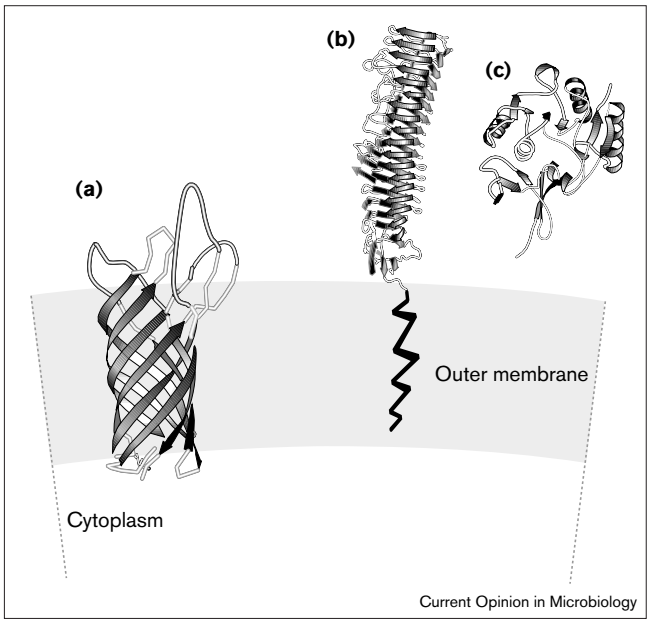
The task is a formidable challenge, however, it is doable. It is just a matter of resources and co-ordination. The most difficult task is the development of an *in vivo* or *in vitro* model that allows high-throughput screening of vaccine candidates.

Tuberculosis

Mycobacterium tuberculosis infects approximately two billion people worldwide and causes 1.5 million deaths annually [12]. The inability of AIDS patients to keep the infection under control and the appearance of multi-resistant strains make the disease an unrestrained danger. The available live-attenuated BCG vaccine is not a solution, because of the variable efficacy reported in the trials. Furthermore, subunit vaccines have not been developed because all the antigens identified by conventional vaccinology provide protection that in animal models is lower than that provided by BCG [12]. Also vaccine development and testing is complicated by the long time required for bacterial growth.

The sequence of the whole genome of *M. tuberculosis* [13•] has provided a list of all possible genes, which now can all be expressed as recombinant proteins or as DNA vaccines and tested for protective immunity [14]. The absence of a high-throughput screening for protective antigens makes the effort difficult but doable by means of a systematic approach. However, a number of genome- and proteome-based approaches are providing novel vaccine candidates, while at the same time the increased knowledge of this difficult bacterium makes it easier to approach [15,16,17••]. The combination of the genome and the use of the fast-growing *Mycobacterium marinum* is the winning combination to accelerate the discovery of an effective tuberculosis vaccine [18].

Figure 2



Examples of protective antigens of *Neisseria meningitidis*. **(a)** A schematic structure of a typical outer membrane protein that is mostly embedded within the membrane. These proteins contain one or two protective epitopes located at the tip of most external loops – change of one amino acid is enough to escape immunity. Outer membrane proteins represent the major part of the antigens of *N. meningitidis* identified by conventional vaccinology. Many novel outer membrane proteins have been identified by reverse vaccinology. **(b,c)** Schematic structures of membrane-anchored lipoproteins or secreted antigens from *N. meningitidis* that have been identified by reverse vaccinology.

Syphilis

During the past four centuries, syphilis has been a nightmare comparable to today's AIDS [19]. If untreated, this sexually-transmitted disease leads to neurological disorders, cardiovascular problems and death, but after the discovery of penicillin the disease became easy to control. However, today syphilis represents a new threat both in developed and developing countries because it causes genital ulcers, which facilitate the spread of HIV. There are approximately 9000 syphilis cases in the US and it is common in Africa and other developing countries, which are estimated to have 12.5 million cases [20].

Syphilis is caused by a bacterium, *Treponema pallidum*, that cannot be cultivated in the laboratory and, therefore, has been refractory to conventional approaches to vaccine development. Attempts to identify vaccine antigens using the bacterium grown in rabbits had identified approximately 20 different antigens. Once again, the sequence of the complete genome made available at once all the genes of the bacterium, which can all be expressed as recombinant proteins or as DNA vaccines [21•,22,23••]. Therefore, for the first time it is now possible to approach development of a syphilis vaccine in a systematic way. The absence of a high-throughput

Table 1

Comparison of conventional and genomic approaches to vaccine development.

Conventional vaccinology	Reverse vaccinology
Essential features	
Most abundant antigens during disease	All antigens immunogenic during disease
Antigens immunogenic during disease	Antigens even if not immunogenic during disease
Cultivable microorganism	Antigens even in non-cultivable microorganisms
Animal models essential	Animal models essential
Correlates of protection useful	Correlates of protection very important
	Correct folding in recombinant expression important
	High-throughput expression/analysis important
Advantages	
Polysaccharides may be used as antigens	Fast access to virtually every single antigen
Lipopolysaccharide-based vaccines are possible	Non-cultivable microorganisms can be approached
Glycolipids and other CD1-restricted antigens can be used	Non abundant antigens can be identified
	Antigens that are not immunogenic during infection can be identified
	Antigens that are transiently expressed during infection can be identified
	Antigens not expressed *in vitro* can be identified
	Non-structural proteins can be used
Disadvantages	
Long time required for antigen identification	Non proteic antigens cannot be used (polysaccharide, lipopolysaccharides,
Antigenic variability of many of the identified antigens	glycolipids and other CD1-restricted antigens)
Antigens not expressed *in vitro* cannot be identified	
Only structural proteins are considered	

animal model again makes the problem difficult but not impossible to solve.

Hepatitis C virus

Hepatitis C virus (HCV) [24] is perhaps the best example of a vaccine being developed entirely by reverse vaccinology. In this case the virus that causes the disease has never been cultivated *in vitro* (it grows only in humans and chimpanzees [25]) and has never been visualized by electron microscopy, making it impossible to use any conventional approach to vaccine development. The cloning and sequencing of the HCV genome allowed the identification of the etiological agent [26], the recombinant expression of its proteins, and the immediate development of diagnostic tools, which prevents hundreds of new infections each day ever since. The availability of the genome sequence also allowed the prediction of the envelope proteins that normally are used to develop vaccines against enveloped viruses [27]. These proteins (E1 and E2) have been expressed in many hosts, but so far only mammalian cells have been able to express them in a form that induces production of antibodies able to interfere with the binding of E2 to the host receptor [28]. These recombinant proteins have been able to protect chimpanzees from infection with the homologous HCV virus [29].

While vaccine development using the E1 and E2 conventional vaccine targets is making progress, perhaps the most interesting questions are whether we can take advantage of the knowledge of the genome to design totally non-conventional vaccine targets and whether proteins never used in conventional vaccines (i.e. non-structural proteins) can become effective vaccines. These proteins should be able to confer protection mostly through cell-mediated immunity and not rely on antibody neutralization of viral infection. The encouraging results obtained with some early proteins such as Tat and Rev [30••–32••] in the case of HIV suggest that this may be a novel way to protect against viruses.

Other pathogens

The pathogens described above are perhaps some of the most representative among those that can be approached by reverse vaccinology. However, the list of the pathogens where the conventional approaches to vaccine development have failed or provided only partial solutions is extensive. Among these we can list bacteria such as *Chlamydia* [33•,34,35], pneumococcus [36–39], *Streptococcus*, *Staphylococcus*, pseudomonas, *Borrelia* [40,41••], *Escherichia coli*, gonococcus, typhoid, *Brucella*, *Ricksettia* [42•] and *Bartonella* (the genome sequences of most of these pathogens are about to be completed and available on the website http://www.tigr.org), and parasites such as *Leishmania* and many others.

Conclusions

Conventional approaches to vaccine development are time consuming, identify only abundant antigens that may or

may not provide immunity, and fail when the pathogen cannot be cultivated under laboratory conditions. Reverse vaccinology (i.e. genomic-based approaches to vaccine development) can overcome these problems (see Table 1) and allow researchers to identify novel antigen vaccine candidates. The sequencing of the complete genome of many pathogens, such as group B meningococcus, has allowed the successful application of reverse vaccinology where conventional approaches have failed. With the genome sequences of many other bacteria, parasites and viruses to be completed in the near future, reverse vaccinology means that many vaccines that were impossible to develop will become reality, and novel vaccines, using non-conventional antigens (i.e. non-structural proteins) can be developed.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

1. Rappuoli R, Del Giudice G: **Identification of vaccine targets.** In *Vaccines: From Concept to Clinic*. Edited by Paoletti LC, McInnes PM. Boca Raton: CRC Press; 1999:1-17.

2. Buynak EB, Roehm RR, Tytell AA, Bertland AU II, Lampson GP, Hilleman MR: **Vaccine against human hepatitis B.** *JAMA* 1976, **235**:2832-2834.

3. • National Institutes of Health: **Jordan Report 2000. Accelerated Development of Vaccines.** 2000:i-173 [http://www.nih.gov].

A comprehensive report on the state-of-the-art of most vaccines in development.

4. Zollinger WD: **New and improved vaccines against meningococcal disease.** In *New Generation Vaccines*. Edited by Levine MM, Woodrow GC, Kaper JB, Cobon GS. New York: Marcel Dekker Inc; 1997:469-488.

5. • Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ *et al.*: **Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58.** *Science* 2000, **287**:1809-1815.

The genome sequence of serogroup B meningococcus is reported.

6. •• Pizza M, Scarlato V, Masignani V, Giuliani MM, Aricò B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B *et al.*: **Whole genome sequencing to identify vaccine candidates against serogroup B meningococcus.** *Science* 2000, **287**:1816-1820.

This paper provides the first example of a rational approach to vaccine development using reverse vaccinology. The whole genome sequence of *N. meningitidis* is screened by computer to identify vaccine candidates. The antigens identified in silico are expressed in *E. coli* that are used to immunize mice and the sera tested for bactericidal activity *in vitro*. Many novel vaccine candidates were discovered that had been missed by all other technologies used so far.

7. World Health Organization: **World malaria situation in 1994.** *Wkly Epidemiol Rec* 1997, **72**:269-274.

8. • Gardner MJ, Tettelin H, Carucci DJ, Cummings LM, Aravind L, Koonin EV, Shallom S, Mason T, Yu K, Fujii C *et al.*: **Chromosome 2 sequence of the human malaria parasite *Plasmodium falciparum*.** *Science* 1998, **282**:1126-1132.

The complete sequence of chromosome 2 of *Plasmodium falciparum*.

9. • Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher CM, Craig A, Davies RM, Devlin K, Feltwell T *et al.*: **The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*.** *Nature* 1999, **4000**:532-538.

The complete sequence of chromosome 3 of *Plasmodium falciparum*.

10. •• Hoffman SL, Rogers WO, Carucci DJ, Venter JC: **From genomics to vaccines: malaria as a model system.** *Nat Med* 1998, **4**:1351-1353.

The possibility of using reverse vaccinology against malaria is discussed. Malaria, for which a vaccine has been refractory to any other approach, may be an ideal target for reverse vaccinology, which may succeed where everything else failed.

11. Wang R, Doolan DL, Le TP, Hedstrom RC, Coonan KM, Charoenvit Y, Jones TR, Hobart P, Margalith M, Ng J *et al.*: **Induction of antigen-specific cytotoxic T lymphocytes in humans by a malaria DNA vaccine.** *Science* 1998, **282**:476-480.

12. Ridzon R, Hannan M: **Tuberculosis vaccines.** *Science* 1999, **286**:1298-1300.

13. • Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE III *et al.*: **Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence.** *Nature* 1998, **393**:537-544.

The complete sequence of *Mycobacterium tuberculosis*.

14. Doherty TM, Andersen P: **Tuberculosis vaccines: development work and future.** *Curr Opin Pulm Med* 2000, **6**:203-208.

15. Brosch R, Philipp WJ, Stravropoulos E, Colston MJ, Cole ST, Gordon SV: **Genomic analysis reveals variation between *Mycobacterium tuberculosis* H3 and the attenuated *M. tuberculosis* H37Ra strain.** *Infect Immun* 1999, **67**:5768-5774.

16. Jungblut PR, Schaible UE, Mollenkopf HJ, Zimny-Arndt U, Raupach B, Mattow J, Halada P, Lamer S, Hagens K, Kaufmann SH: **Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens.** *Mol Microbiol* 1999, **33**:1103-1117.

17. •• Behr MA, Wilson MA, Gill WP, Salamon H, Schoolnik GK, Rane S, Small PM: **Comparative genomics of BCG vaccines by whole-genome DNA microarray.** *Science* 1999, **284**:1520-1523.

Whole genome comparison using microchips is used to identify the differences between the live-attenuated strains of BCG, which are used as vaccines against tuberculosis, and the wild type, disease-causing strains. Several surprises (unexpected differences) emerge from this analysis.

18. Ramakrishnan L, Federspiel NA, Falkow S: **Granuloma-specific expression of mycobacterium virulence proteins from the glycine-rich PE-PGRS family.** *Science* 2000, **288**:1436-1439.

19. Singh AE, Romanowski B: **Syphilis: review with emphasis on clinical, epidemiologic, and some biologic features.** *Clin Microbiol Rev* **12**:187-209.

20. St Louis ME, Wasserheit JN: **Elimination of syphilis in the United States.** *Science* 1998, **281**:353-354.

21. • Fraser CM, Norris SJ, Weinstock GM, White O, Sutton GG, Dodson R, Gwinn M, Hickey EK, Clayton R, Ketchum KA *et al.*: **Complete genome sequence of *Treponema pallidum*, the syphilis spirochete.** *Science* 1998, **281**:375-388.

The complete sequence of *Treponema pallidum*, the causative agent of syphilis.

22. Pennisi E: **Genome reveals wiles and weak points of syphilis.** *Science* 1998, **281**:324-325.

23. •• Norris SJ, Weinstock GM: **The genome sequence of *Treponema pallidum*, the syphilis spirochete: will clinicians benefit?** *Curr Opin Infect Dis* 2000, **13**:29-36.

The authors discuss how the knowledge of the genome may help improve the approach to syphilis research, including vaccine development.

24. Sarbah SA, Younossi ZM: **Hepatitis C: an update on the silent epidemic.** *J Clin Gastroenterol* 2000, **30**:125-143.

25. Bradley DW, McCaustland KA, Cook EH, Ebert JW, McCaustland KA, Schable CA, Fields HA: **Posttrasfusion non-A non-B hepatitis in chimpanzees: physicochemical evidence that the tubule-forming agent is a small enveloped virus.** *Gastroenterol* 1985, **88**:773-779.

26. Choo Q-L, Kuo G, Weiner AJ, Overby LR, Bradley DW, Houghton M: **Isolation of a cDNA clone derived from a blood-borne non-A and non-B viral hepatitis genome.** *Science* 1989, **244**:359-362.

27. Reed KE, Rice CM: **Overview of hepatitis C virus genome structure, polyprotein processing, and protein properties.** *Curr Top Microbiol Immunol* 2000, **242**:55-84.

28. Rosa D, Campagnoli S, Moretto C, Guenzi E, Cousens L, Chin M, Dong C, Weiner AJ, Lau JY, Choo QL *et al.*: **A quantitative test to estimate neutralizing antibodies to the hepatitis C virus: cytofluorimetric assessment of envelope glycoprotein 2 binding to target cells.** *Proc Natl Acad Sci USA* 1996, **93**:1759-1763.

29. Choo QL, Kuo G, Ralston R, Weiner A, Chien D, Van Nest G, Han J, Berger K, Thudium K, Kuo C *et al.*: **Vaccination of chimpanzees against infection by the hepatitis C virus.** *Proc Natl Acad Sci USA* 1994, **91**:1294-1298.

30. Pauza CD, Trivedi P, Wallace M, Ruckwardt TJ, Le Baunce H, Lu W,
•• Bizzini B, Burny A, Zagury D, Gallo RC: **Vaccination with tat toxoid attenuates disease in simian/HIV-challenged macaques.** *Proc Natl Acad Sci USA* 2000, **97**:3515-3519.
The possibility of producing large quantities of recombinant non-structural proteins (which would be impossible by growing the virus using conventional vaccinology) allows researchers to test novel paradigms for vaccination. Here, along with [31••,32••], early viral proteins are used to vaccinate against HIV, trying to exploit cellular-mediated immunity against these antigens to protect from infection. This paradigm is novel in vaccinology.

31. Cafaro A, Caputo A, Fracasso C, Maggiorella MT, Goletti D,
•• Baroncelli S, Pace M, Sernicola L, Koanga-Mogtomo ML *et al.*: **Control of SHIV-89.6P-infection of cynomolgus monkeys by HIV-1 Tat pro vaccine.** *Nat Med* 1999, **5**:643-650.
See annotation for [30••].

32. Osterhaus AD, van Baalen CA, Gruters RA, Schutten M,
•• Siebelink CH, Hulskotte EG, Tijhaar EJ, Randall RE, van Amerongen G, Fleuchaus A *et al.*: **Vaccination with Rev and Tat against AIDS.** *Vaccine* 1999, **17**:2713-2714.
See annotation for [30••].

33. Stephens RS, Kalman S, Lammel C, Fan J, Marathe R, Aravind L,
• Mitchell W, Olinger L, Tatusov RL, Zhao Q *et al.*: **Genome sequence of an obligate intracellular pathogen of humans: *Chlamydia trachomatis*.** 1998, **282**:754-759.
The complete sequence of *Chlamydia trachomatis*.

34. Read TD, Brunham RC, Shen C, Gill SR, Heidelberg JF, White O, Hickey EK, Peterson J, Utterback T, Berry K *et al.*: **Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39.** *Nucleic Acids Res* 2000, **28**:1397-1406.

35. Kalman S, Mitchell W, Marathe R, Lammel C, Fan J, Hyman RW, Olinger L, Grimwood J, Davis RW, Stephens RS: **Comparative genomes of *Chlamydia pneumoniae* and *C. trachomatis*.** *Nat Genet* 1999, **21**:385-399.

36. Klein DL, Ellis RW: **Conjugate vaccines against *Streptococcus pneumoniae*.** In *New Generation Vaccines*. Edited by Levine MM, Woodrow GC, Kaper JB, Cobon GS. New York: Marcel Dekker Inc; 1997:504-525.

37. Toumanen E: **Molecular and cellular biology of pneumococcal infection.** *Curr Opin Microbiol* 1999, **2**:35-39.

38. Baltz RH, Norris FH, Matsushima P, DeHoff BS, Rockey P, Porter G, Burgett S, Peery R, Hoskins J, Braverman L *et al.*: **DNA sequence sampling of the *Streptococcus pneumoniae* genome to identify novel targets for antibiotic development.** *Microb Drug Resist* 1998, **4**:1-9.

39. Polissi A, Pontiggia A, Feger G, Altieri M, Mottl H, Ferrari L, Simon D: **Large-scale identification of virulence genes from *Streptococcus pneumoniae*.** *Infect Immun* 1998, **66**:5620-5629.

40. Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK *et al.*: **Genomic sequence of a Lyme disease spirochete, *Borrelia burgdorferi*.** *Nature* 1997, **390**:580-586.

41. Nordstrand A, Barbour AG, Bergström S: ***Borrelia* pathogenesis
•• research in the post-genomic and post-vaccine era.** *Curr Opin Microbiol* 2000, **3**:86-92.
Although a vaccine based on OspA had been developed against Lyme disease before the genome sequence of the bacterium had been determined, the efficacy of this vaccine is limited to the East Coast of the USA. The authors discuss how sequencing the genome changed the approach to the problem and how it may help improve the vaccine.

42. Andersson SG, Zomorodipour A, Andersson JO,
• Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG: **The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria.** *Nature* 1998, **396**:133-140.
The complete sequence of the *Rickettsia prowazekii*.

Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

The rational design of vaccines

Vincent W. Bramwell and Yvonne Perrie

This review provides an insight into the various opportunities for vaccine intervention, analysis of strategies for vaccine development, vaccine ability to modulate immune responses and resultant rational vaccine design. In addition, wider aspects are considered, such as biotechnological advances, advances in immunological understanding and host–pathogen interactions. The key question addressed here is, with all our research and understanding, have we reached a new echelon in vaccine development, that of rational design?

The goal of any vaccination is the induction of an appropriate and effective immune response, but precisely what constitutes an effective immune response for many diseases is still unclear. Specific levels of antibody are considered protective in vaccine efficacy for vaccines against hepatitis, diphtheria and tetanus; for example, anti-hepatitis surface antigen antibody levels >10 mIU/ml, anti-diphtheria antibody levels >100 mIU/ml and anti-tetanus antibody levels >100 mIU/ml are quoted as protective levels of antibody in humans [1,2]. However, for many other diseases, correlations of efficacy are less obvious or yet to be agreed. By definition of perceived need, we are most acutely aware of the requirement of effective vaccines against infectious agents, pathogens ancient, re-emergent and new, yet the opportunities for manipulation of immune responses offer potential in the prevention and treatment of a far larger diversity of diseases. From these, immunomodulation in cancer, allergy and autoimmune disease are also considered here.

Recent developments have brought vaccines against infectious diseases to the forefront of the scientific community. The 'wake up call' that came from SARS (severe acute respiratory syndrome) has undoubtedly fuelled present fears of a flu pandemic. The 1918 Spanish flu pandemic provides an extreme historical precedent, causing an estimated 50 million deaths after infecting one billion people worldwide [3]. As to the impact of such an outbreak, today there is serious concern over the capacity of the existing infrastructure to limit fatalities by providing suitable healthcare [4]. Although development of a suitable vaccine can take a minimum of six months, in an organized and prepared scenario this could provide the major weapon that will ultimately contain such a catastrophe.

The latest polio outbreaks look set to delay the long hope for eradication of this disease [5]. In the UK, the substitution of the live oral polio vaccine with an inactive vaccine should provide a step towards safe cessation of polio vaccination when the time is right. The live vaccine has the capacity to generate infectious polio virus through mutation, and altered immunological properties of mutants present concern about virus spread in immunized as well as non-immunized populations [6,7].

Vincent W. Bramwell
Yvonne Perrie*
Medicines Research Unit, School of Life and Health Sciences, Aston University, Aston Triangle, Birmingham B4 7ET, UK
*e-mail: y.perrie@aston.ac.uk

1359-6446/05/$ – see front matter PII: S1359-6446(05)03600-7

Taken as a whole, vaccine research and development encompasses a wide diversity of potential applications, developmental strategies and opportunities. The ongoing elucidation of immunological mechanisms of pathogen recognition and adjuvant action, as well as the knowledge of pathogen mediated mechanisms of immune evasion and gene expression, have transformed our understanding of immunology and disease in recent years. In this complex but increasingly focussed environment of vaccine development, are there opportunities for new solutions for vaccine design? And can new and refined approaches succeed where more conventional approaches appear to have failed?

Vaccines against key diseases

Infectious diseases

There are several reasons as to why the bacillus Calmette–Guérin (BCG) vaccine against tuberculosis (TB) is perceived as ineffective, including manipulation of immune responses (as for virulent TB [8,9]) and, interestingly, variation between propagated strains of BCG [10]. In the case of TB, recent literature has correlated the production of interleukin 4 (IL-4) and IL-4δ2 (a splice variant of IL-4 and an IL-4 antagonist) with the propensity of a latently infected host to succumb to active tuberculosis. The production of IL-4δ2 is correlated with the control of the disease [11]. Focus on the type of immune response needed to engender protection against TB seems to be orientated (convincingly) towards suppression of IL-4 and the long standing dogma advocating the importance of Th1 type cytokines (such as interferon gamma, IFN-γ) in TB resistance [12,13].

In the case of HIV, only one HIV candidate vaccine has completed clinical trials Phases I, II and III in more than 20 years of the epidemic. The Phase III trial was based on the use of recombinant envelope proteins, with the aim of evoking virus neutralizing antibodies. However, results from this were disappointing.

Currently available vaccines are mostly effective against viral diseases that are normally cleared during natural infection [influenza, smallpox, polio and all three viruses in the measles, mumps and rubella (MMR) immunisation]. HIV, in common with Epstein–Barr virus, cytomegalovirus, hepatitis (B and C) and herpes simplex virus, normally establishes chronic infection. Opinion is divided as to what type of immune response is required in the control of HIV. Neutralizing antibodies might be efficient in blocking virus particles but poorly effective against cell-associated virus, whereas some cytotoxic T lymphocytes (CTLs) are effective against virally infected cells but not against free virus particles. Latent infection facilitates immune evasion and the rate of mutation inherent in HIV also facilitates evasion (by escape from responding T cells [14]). Promising vaccine strategies will likely depend upon highly conserved epitopes and upon vaccine strategies that induce potent immune responses capable of driving cytotoxicity and producing broadly effective neutralizing antibodies [15].

Allergy and autoimmune diseases

Allergic and autoimmune diseases represent undesired hypersensitive immune responses to normally harmless environmental antigens, in the case of allergy, or normally tolerated self-proteins, in the case of autoimmunity. The understanding of the pathogenesis of allergic and autoimmune diseases has led to several proposed associations and causes for these diseases. It has been shown that high-affinity T cells can out-compete lower-affinity T cells during responses to antigen *in vivo* [16]. Competition between T cells provides the basis for the argument that maintaining a balanced peripheral immune system might also be a side effect of normal competition for shared resources within an intact immune system [17], a mechanism proposed to be responsible for the increased occurrence of these diseases, where hypersensitivity is normally passively controlled by expansion of immune cells specific for prevalent infectious agents. Autoimmunity can have strong associations with exposure to crossreactive proteins, and genes encoding major histocompatibility complex (MHC) are strongly associated with some autoimmune diseases, but this is markedly less so for allergy. Allergy and autoimmunity are thought to differ by virtue of their regulation through central and peripheral tolerance. Genetic predisposition and environmental exposure are believed to play key roles in the development of asthma and atopy [18], with the significant increase in the incidence of these disorders thought to provide evidence for the involvement of various environmental factors, especially in developed countries [19].

The rationale for immunological therapeutic intervention is strongly supported by the present lack of curative treatments for autoimmune and allergic disorders, a growing unmet need, where the largely palliative treatment is not without problems [20].

Evidence that pathology associated with allergy and autoimmunity is reversible is provided by transient and long-term resolution of disease as well as by the success of specific allergen immunotherapy, where subcutaneous or sublingual administration of the sensitizing protein(s) modifies immunity and reduces allergen sensitivity [20].

Cancer

Serological identification of antigens by recombinant expression cloning has led to the identification of a multitude of new tumour antigens [21]. It is thought that most or all paraneoplastic neurological disorders (neurological disorders remote from the site of a malignant neoplasm or its metastases) are immune mediated [22] and this has been cited as evidence for the involvement of specific immune responses in cancer suppression [23]. Indeed, failure to find the relevant antigen in the cancer of a patient should prompt a search for a second cancer

TABLE 1

The main players: oncogenic infectious agents and their associated malignancies

Oncogenic infectious agents	Associated malignancies
Human papillomaviruses	Cervical carcinoma
Human polyomaviruses	Mesotheliomas, brain tumours
Epstein–Barr virus	B-cell lymphoproliferative diseases and nasopharyngeal carcinoma
Herpesvirus	Kaposi's sarcoma and primary effusion lymphomas
Hepatitis B and hepatitis C viruses	Hepatocellular carcinoma
Human T-cell leukemia virus-1	T-cell leukemias
Helicobacter pylori	Gastric carcinoma

[24]. Immunotherapy has recently achieved much interest as a possible addition to chemotherapeutic treatment [25] and this could open up a new and innovative avenue for clinical trials. Effective chemotherapy against *Schistosoma mansoni* with praziquantel is dependent on the presence of antibodies recognizing schistosome glycoprotein epitopes [26]. Another interesting approach is the therapeutic vaccination against cancer by targeting anti-apoptotic molecules [27].

Infectious agents are implicated as causes of cancer and contribute to a variety of malignancies worldwide. Some of the major players in this role are shown in Table 1 and they account for several of the most common malignancies and up to 20% of malignancies around the globe [28]. As such, these agents offer increased potential for prevention of cancer by prophylactic vaccination.

Passive immunisation

Administration of antisera as a treatment against snake venom dates back to the 1890s and the work of Albert Calmette. Although this article largely focuses on active immunisation and the generation of host antibodies by the host immune system itself, the strategy of passive immunisation is worth consideration, as epitopes identified using modern molecular biology techniques might be targeted by the use of passively administered monoclonal antibodies. Examples of immunotherapy using monoclonal antibodies are Campath® (alemtuzumab), manufactured by Genzyme for the treatment of B-cell chronic lymphocytic leukaemia, and Rituxan® (Rituximab), a chimeric murine–human monoclonal antibody that targets the CD20 antigen found on the surface of normal and malignant B lymphocytes and used in the treatment of non-Hodgkin's lymphoma. This strategy is not restricted to cancers; for example Remicade®, a chimeric monoclonal antibody specific for tumour necrosis factor α (TNF-α, is used in the treatment of autoimmune disease.

Vaccine development strategies

Historical outline

The implementation of rational design was first evident when Pasteur reproducibly manufactured attenuated cultures of chicken cholera vaccines, thereby routinely preventing cholera in vaccinated chickens. Extrapolation of this strategy for anthrax vaccines in livestock in the 1880s was a success with significant economic benefits. Despite the elucidation of an attenuated vaccination strategy against rabies and other inactivated whole-organism vaccines towards the end of the nineteenth century, it wasn't until the cell-culture revolution in 1950–1980 that attenuated viral vaccines really took off, together with the isolation and use of extracts and subunits of infectious agents. Pasteur's approaches of attenuation and inactivation still today provide the two poles of vaccine technology [29]. A chronology of important developments and achievements in vaccine research and implementation is shown in Figure 1. Recent developments in vaccine technology and application include combination vaccines, new adjuvants and delivery systems, reverse vaccinology and vaccines against noninfectious diseases. The history of vaccine development is very much related to technological advances [30].

Exploiting vaccine delivery systems and adjuvants

Subunit vaccines can be produced in bulk, safely and reproducibly, using recombinant DNA technology. Much research has already been accomplished in the development of suitable carrier systems able to engender enhanced immune responses to entrapped peptide and protein epitopes. The identification of potentially useful antigens is greatly facilitated by improved molecular biology techniques, such as microarray technology (discussed further below), but the elucidation of their full potential might rely on the development of effective delivery systems and adjuvants. Many adjuvants are based on microbial components but others, traditionally aluminium salts and more recently emulsions and surfactant based formulations, exploit different mechanisms of action. Particulate delivery can be mediated by polymer [often poli(lactide-co-glycolide)] microparticles, immunostimulatory complexes, liposomes and virosomes [31]. Possible advantages of particulate delivery are that antigen and coadjuvant can be delivered to the same cell and particulates have excellent potential for targeting cells of the immune system [32]. The extensive diversity of adjuvants and delivery systems is more comprehensively reviewed elsewhere [31–34] and this review will, where appropriate, focus on the application and other aspects of this technology.

Date	Event or vaccine introduction	Major development
1796	Edward Jenner uses cowpox vaccine against smallpox	Related animal virus used to prevent disease in humans; can be seen as the birth of 'vaccinology'
1870s	Chicken cholera vaccine created by Pasteur	First live attenuated bacterial vaccine
1880s	Rabies and anthrax vaccines	Chemical attenuation; first live attenuated viral vaccine (rabies) (Pasteur)
1890s	Typhoid, cholera, plague vaccines	Inactivation of whole organisms
1920s	Whole cell pertussis vaccine (inactivated); attenuated BCG; Tetanus and diphtheria toxoid vaccines	*In vitro* passage; Use of inactivated toxins
1930s	Inactivated influenza vaccine; attenuated yellow fever	Chick embryo and tissue culture (yellow fever)
1940s	Japanese encephalitis; also war support for vaccine development	Use of pathogen derived extracts and subunits; DPT (tri-valent diphtheria/pertussis/tetanus) recommended by the AAP for routine use
1950s	Inactivated polio virus vaccine (IPV)	The cell culture 'revolution' between the 1950s and 1980s provided the foundation for the attenuated viral vaccines in use today
1960s	Live oral polio (OPV) and measles virus vaccines licensed	
1970s	Pneumococcal and meningococcal vaccines; licensure of rubella vaccine (earlier in the USA); adenovirus vaccines; Influenza	Use of capsular polysaccharides; last indigenous case of smallpox (Somalia), cessation of smallpox vaccination; use of ressortants (influenza)
1980s	Hepatitis B vaccine; *H. influenzae* type b protein conjugated polysaccharide; typhoid	Hepatitis B vaccines produced by recombinant DNA technology replaced the (plasma derived) vaccine licensed earlier; development of auxotrophic vaccines (typhoid)
1990s	Varicella, lyme disease vaccines; dtap (diphtheria, tetanus, acellular pertussis) licensed; supersedes whole-cell DPT; various combination vaccines licensed	Development of safer vaccines (Dtap); elucidation of the possibility of DNA vaccines
	Rotavirus vaccine recommended, licensed and withdrawn in less than two years (1998-99 USA)	
2000 onwards	Live influenza vaccine; bovine-human rotavirus vaccine; polio: centers for Disease Control recommends use of IPV instead of OPV; withdrawal of lyme disease (Lymerix) vaccine; penta-valent DtaP/HepB/IPV (Pediarix) licensed; serious consideration of biowarfare vaccines (e.g. smallpox); papillomavirus vaccines against cancer	

Drug Discovery Today

FIGURE 1

A historical outline of important developments and achievements in vaccine research. The figure outlines many of the important developments and achievements that have underpinned the development of successful vaccines from 1796. The variolation against Smallpox is believed to originate in China or India, later spreading to the Middle East, Africa, Turkey and Great Britain.

The 300 million doses of influenza vaccine needed annually worldwide require more than 350 million chicken eggs and six or more months. The development of cell-culture technology that could replace the egg-based manufacturing process might facilitate faster production of much-higher antigen yields [4]. Live vaccines generally possess a natural adjuvant capability that is built on the evolved ability of our immune system to recognize many facets of potentially dangerous microbes and they have contributed immeasurably to the control of disease. Although some live vaccines might have undesirable characteristics, such as the mutation of polio virus and persistence of viruses such as varicella zoster virus, their use has been, and indeed remains, able to significantly reduce the impact of associated diseases [35,36]. Viruses and bacteria as carriers of heterologous antigens and genes have received considerable attention over previous decades, and generation of auxotrophic mutants, as well as the use of other replication incompetent vectors and of reassortant viruses generated by the addition of RNA segments to viruses with segmented genomes, such as influenza and rotavirus, indicates the potential inherent in these efficient types of vaccine. In an interesting utilization of viral replication, a self-replicating RNA vector encoding a model antigen (β-gal), which was shown to protect mice against subsequent tumour challenge with colon carcinoma cells engineered to express β-gal, was also shown to effectively induce apoptosis in transduced cells [37]. The observed apoptosis also facilitated enhanced uptake by dendritic cells and was likely mediated by the presence of double-stranded RNA in replication of the virally derived genome. Double-stranded RNA is recognized by Toll-like receptor (TLR) binding. Different TLRs recognize different surface and intracellular components of microorganisms. The interaction between a TLR and a microbial component triggers the activation of the innate immune system and

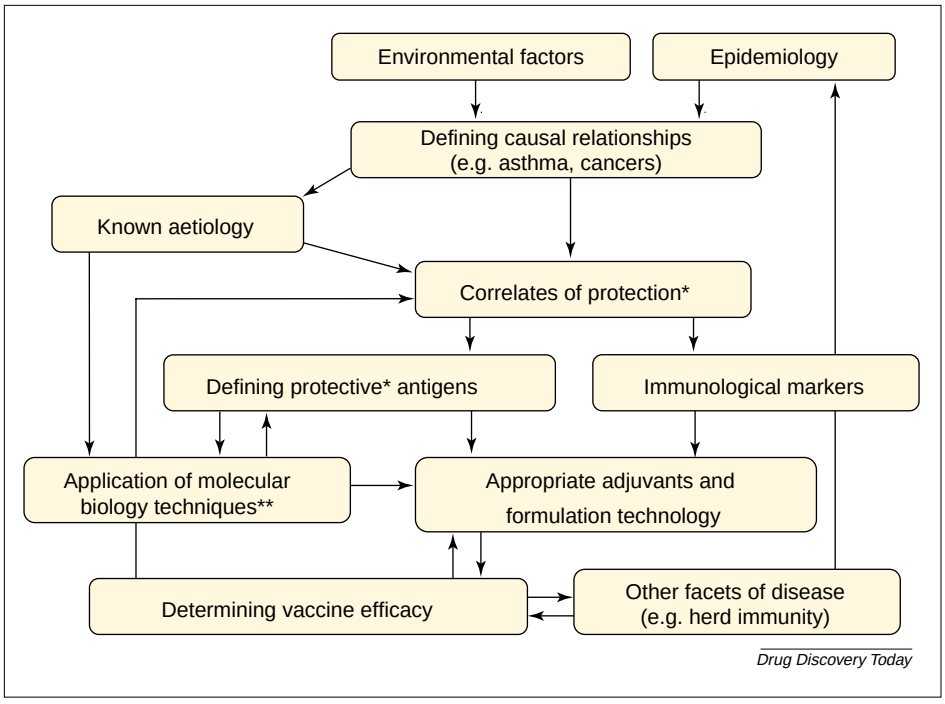


FIGURE 2

Important factors in the rational design of vaccines and the cyclical generation of knowledge. The level of involvement of any of these factors depends on the aetiological agent of disease. For example, the level of protection required in a population (herd immunity) will be different and this could allow theoretical flexibility in vaccine efficacy. The application of molecular biology techniques can be crucial in the identification of new candidate antigens and subsequent determination of vaccine efficacy using adjuvants can feed knowledge back to correlates of protection in terms of immunological markers. This knowledge can then be used in choice of appropriate adjuvants and formulation. The key implication projected by this schematic is that for the greatest challenges in vaccine development the cyclical generation of knowledge provides a strong role for rational design.
* Can be protective or therapeutic. ** Including reverse vaccinology and associated technologies.

the initiation of acquired immunity [38]. The elucidation of these pathways sheds light on mechanisms of adjuvant action and provides a basis for the inclusion of such potent agents and their analogues in vaccine design.

The role of rational design

Looking for associations

It is important to consider how we classify or view rational design in the context of vaccines. A picture springs to mind of the rapid analysis of pathogens, the identification and allocation of immunological interactions for specific genes, as well as the generation of a superimposed choreography of the infection or disease process, resulting in the identification of credible candidate target antigens. Following this, appropriate formulation with adjuvants and delivery systems might elicit immune responses of the type and magnitude predictive of providing protection or therapeutic action. The truth is that we are still very far from this idyll.

Recent articles have intimated the minor contribution of our knowledge of immunology to the development of vaccines [39,29]. However, the development of improved and large-scale cellular immunity techniques for the analysis of immune responses, such as ELISPOT for cytokine induction, tetramer staining for peptide specific CTLs, along with analysis of the immunological basis for the efficacy of successful vaccines, facilitates the role of immunology in predicting the appropriate context for antigen delivery. Extensive characterisation of adjuvant action and the potential to target desired immune responses through the exploitation of this knowledge is key for rational vaccine design [29,31,32,39,40]. The interaction between different elements involved in rational vaccine design is outlined in Figure 2.

The use of microarray technology can allow identification of virulence genes and vaccine targets [41] and is one of the most powerful tools for the study of the transcriptome, the complete set of transcripts of an organism. Indeed, in conjunction with proteomics and comparative genome analysis, interpretation of whole-genome sequences through bioinformatics (or genomic mining) can be used to assign putative gene functions to each open reading frame on the basis of homology to known proteins [42]. Systematic identification of potential antigens of a pathogen using this information, without the need for cultivation of the pathogen, is termed 'reverse vaccinology' [43] and represents a significant departure towards the idyll of rational vaccine design described above, at least in terms of dissection of potential pathogen-related antigens.

A new era of vaccine design?

Knowledge of pathogen interactions during infection can provide an invaluable insight into potentially successful targets for vaccines. For example, knowledge of gene expression in *Mycobacterium tuberculosis* enables us to see why vaccines based on early secreted proteins, such as Ag85 and ESAT-6, can generate effective pre-exposure vaccines and helps us to predict that an effective post-exposure vaccine will utilize dormancy induced proteins, such as α-crystallin, heat shock protein 'HspX' [44].

In the case of HIV, the implication of lipid rafts in viral entry and budding processes might have provided a rationale for the evaluation of peptides, based on highly conserved caveolin-1 binding domains of HIV-1 glycoprotein gp41 in candidate vaccine formulations [45]. Encouragingly, this has resulted in peptides capable of the elicitation of neutralizing antibody responses able to inhibit different clades of HIV-1. It is thought that the poor ability of some specific antibodies to neutralize primary isolates is due, at least in part, to steric factors that limit antibody access to the gp120 epitopes [46].

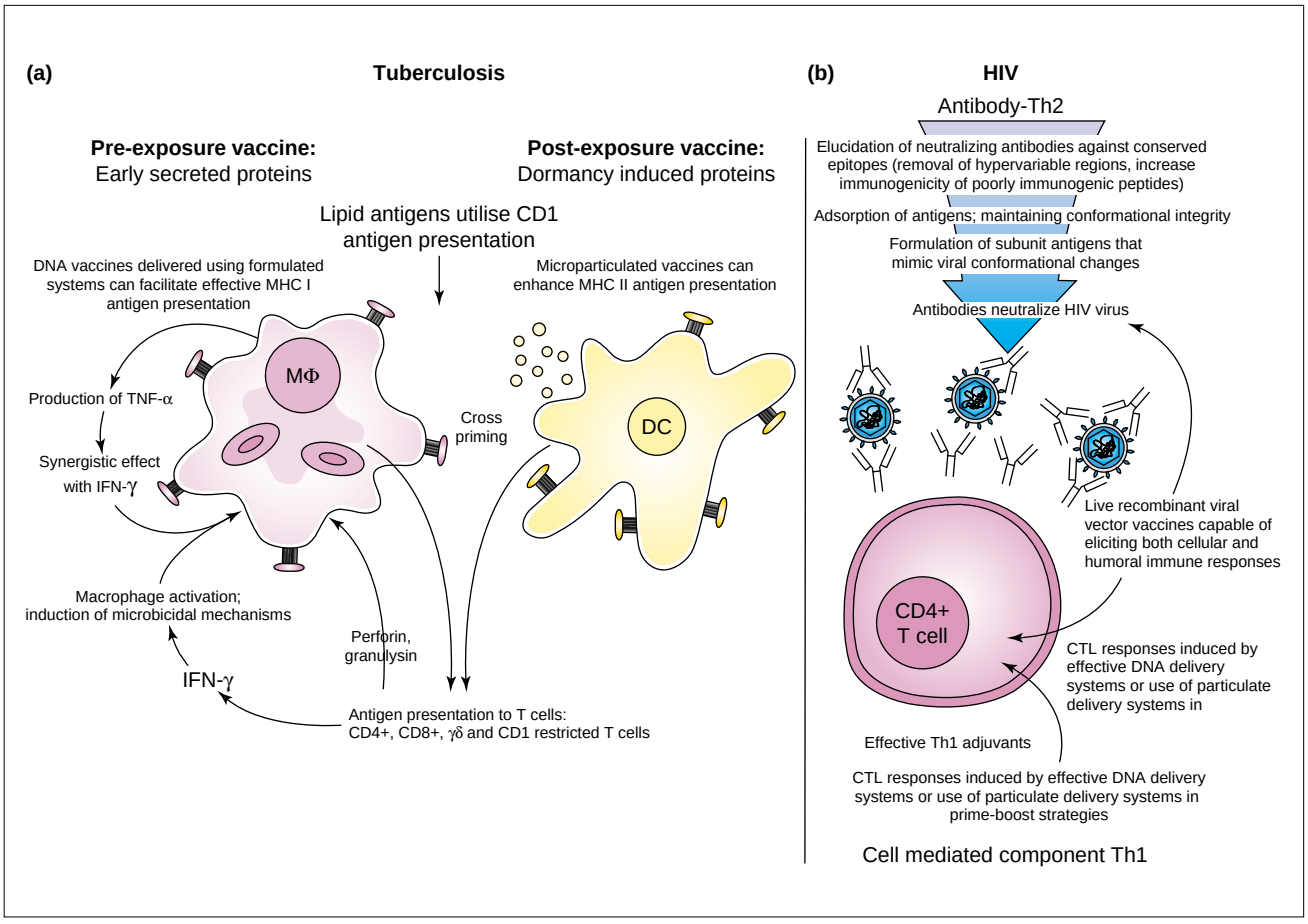


FIGURE 3

Proposed opportunities for vaccination against TB and HIV. In both cases, the utilization of adjuvants capable of driving the required type of immune response can be used. This can be based on our current understanding of TLR interaction, and studies that manipulate this understanding might also serve to elucidate complex interactions during infection and disease. **(a)** According to recent models, antigen specific CD4+ T cells, CD8+ T cells, γδ T cells and CD1 restricted T cells, all participate in protection against *M. tuberculosis* infection in response to antigen presentation by macrophages (MΦ) and dendritic cells (DC), probably also involving crosspriming, where mycobacterial antigens are transferred from infected MΦ to DC. It is unlikely that the induction of IFN-γ or the production of inducible nitric oxide synthase alone is enough to control TB infection, and evidence indicates that there might be another role for CD4+ T cells. The development of pre- and post-exposure vaccines will likely require antigens expressed by *M. tuberculosis* under pre- and post-exposure conditions, respectively. Interestingly, nonpolymorphic CD1 restricted T cells are thought to have evolved as a result of prolonged interaction with mycobacteria, indicating the extent of the impact of mycobacteria on its mammalian host. **(b)** For vaccines against HIV, it seems increasingly likely that strategies will depend upon highly conserved epitopes and on vaccine strategies that induce potent immune responses capable of driving cytotoxicity, as well as broadly reactive, highly effective neutralizing antibodies. Target antigens might very probably be different for each of these strategies. Figure adapted, with permission, from Ref. [32].

Antibodies against the caveolin-1 binding domains are rare in many HIV infected individuals, possibly because of the presence of hypervariable immunodominant epitopes or the location and lack of exposure of conserved regions. The proposed opportunities for vaccines against HIV and TB are summarized in Figure 3; the rationale for elicitation of successfully neutralizing antibodies against HIV is centred upon increasing the immunogenicity of poorly immunogenic peptides, modification and/or removal of hypervariable regions and mimicking viral conformational changes.

One of the hottest areas of interest regarding new vaccine development surrounds the Phase III efficacy trials and prospective licensure of two vaccines against human papillomavirus (HPV) [47]. The vaccines, one from Merck and one from GSK, are effectively vaccines against cervical cancer; the (now) unequivocal link between this common virus, which rarely causes serious disease in itself, and cervical cancer found credence as long ago as the 1970s. The production of both vaccines (the Merck vaccine produced in yeast administered with an aluminium adjuvant and the GSK vaccine in baculovirus with the adjuvant AS04; aluminium and bacterial lipid already approved for use in Europe) has its basis in work that elucidated the production and self-assembly of virus-like particles of the HPV outer L1 and L2 coat proteins [48,49] in various cell types.

Vaccines against most *Neisseria meningitidis* serogroups have been developed using traditional approaches. However, group B meningococcus has represented a particular challenge because of the sequence variation of surface proteins and crossreactivity of the capsular polysaccharide

with host polysialic acid [50]. In this case, reverse vaccinology was convincingly applied to overcome these problems by alternative means [51]. Screening of DNA fragments, discarding likely cytoplasmic protein sequences and known *Neisseria* antigens, extensive cloning and expression of identified candidate genes allowed selection of antigens that were found not only to offer significant potential for protection against group B meningococcus but also to facilitate crossprotection against heterologous strains.

Also worth mentioning is the technique of molecular breeding or gene shuffling, where reassembled chimeric genes are created from a selection of homologous gene sequences. The related sequences are denatured, annealed and subsequently extended in what is in effect a self-priming polymerase chain reaction (PCR) that takes advantage of crossovers, deletions, insertions, inversions and point mutations, as occurs in natural evolution (hence the term 'molecular breeding'). Proteins and genes with enhanced activity are then selected for inclusion in further rounds of molecular breeding. This has a proven application for the generation of enzymes with markedly enhanced activity [52] and the enhancement of the functions of a diverse range of proteins, such as green fluorescent protein [53] and IFN-α [54]. This technique is thought to be superior to other methods that employ random mutagenesis, such as error-prone PCR (hence the phrase 'the rational basis for irrational design' [55]), but the screening of candidate antigen genes is necessarily expensive in terms of resources and time if it is compared for example to the screening of an enzyme. Their potential has been noted with relevance to allergy vaccines [56], as pre-existing immunity is the therapeutic target and screening of allergen variants might be somewhat more practicable.

Conclusions

Bringing together developmental strategies with the knowledge gleaned from host pathogen interactions represents a highly evolved design strategy. Although each individual molecular biology technique appears to have something to contribute to vaccine development, it is the holistic application of a combined approach that is closest to a premeditated and planned rational vaccine design. Pathogens such as *Streptococcus pneumoniae*, *Porphyromonas gingivalis*, *Staphylococcus aureus*, *Chlamydia pneumoniae* and *Bacillus anthracis* have all been investigated using bioinformatics techniques in the context of reverse vaccinology [43], with continued interesting results. In terms of antigen identification, the National Institute of Allergy and Infectious Diseases (NIAID) recently announced initiation of the Large-Scale Antibody and T Cell Epitope Discovery Program; utilizing complementary methods for epitope discovery, it is designed to identify immune epitopes from selected infectious agents and will make this information freely available to scientists worldwide through the Immune Epitope Database and Analysis Resource (IEDB), currently under development [57].

Recent discoveries have resulted in a wealth of options, when considering what we can do in terms of vaccine design and production. Mechanisms of adjuvant action have been elucidated, pattern recognition receptors including Toll-like receptors, NOD1, NOD2, scavenger receptors, mannose and other receptors are becoming increasingly well defined [38] and intracellular events related to these, such as activation of transcription factors and expression of inflammatory cytokines, shed further light on how these ligands and adjuvants that bind to pattern recognition receptors mediate their immunological actions. The ability of infectious agents to induce rapid, innate immune responses to molecular patterns provides the basis for adjuvant immunotherapy of cancers. Studies, such as the recent work examining the adjuvant activity of BCG cell-wall skeleton, peptidoglycan and lipopolysaccharide by the induction of genes in dendritic cells [58], will very possibly help to achieve definition of efficient effector output with minimal toxicity.

The easier and safer production of vaccines and vaccine components facilitated by modern biological techniques underpins a rational approach that is an integral part of the rational design of vaccines. The extensive knowledge of immunological mechanisms and host pathogen interactions is able to contribute to the design of effective vaccines in addition to the elucidation of the mechanisms of action of many candidate vaccine agents. In fact, rational design of vaccines represents a driving force born in the first revelations that components from or relating to a microbial pathogen can be used to protect against that disease and present throughout development of vaccines in the modern era, honing our understanding and preparing the way for the next steps in the battle to combat today's significant pathogens and diseases, such as cancer, HIV, tuberculosis and malaria. In this era, we begin to explore the very limits of immunotherapeutic and prophylactic intervention and the tools that have been developed to elucidate gene expression and function might have at last begun to find their application in our most significant of challenges.

Acknowledgements

The authors thank Graham Smith at Aston University, Birmingham B4 7ET, for his rendering of Figure 3.

References

1 Dentico, P. *et al.* (2002) Anamnestic response to administration of purified non-adsorbed hepatitis B surface antigen in healthy responders to hepatitis B vaccine with long-term non-protective antibody titres. *Vaccine* 20, 3725–3730

2 Schmitt, H.J. *et al.* (2003) The safety, reactogenicity and immunogenicity of a 7-valent pneumococcal conjugate vaccine (7VPnC) concurrently administered with a combination DTaP-IPV-Hib vaccine. *Vaccine* 21, 3653–3662

3 Ritvo, P. *et al.* (2005) Vaccines in the public eye. *Nat. Med.* 11 (Suppl. 4), S20–S24

4 Osterholm, M.T. (2005) Preparing for the Next Pandemic. *N. Engl. J. Med.* 352, 1839–1842

5 Dyer, O. (2005) WHO's attempts to eradicate polio are thwarted in Africa and Asia. *BMJ* 330, 1106

6 Cherkasova, E.A. *et al.* (2005) Spread of vaccine-derived poliovirus from a paralytic case in an immunodeficient child: an insight into the natural evolution of oral polio vaccine. *J. Virol.* 79, 1062–1070
7 Martin, J. *et al.* (2004) Long-term excretion of vaccine-derived poliovirus by a healthy child. *J. Virol.* 78, 13839–13847
8 Gagliardi, M.C. *et al.* (2004) Bacillus Calmette-Guerin shares with virulent Mycobacterium tuberculosis the capacity to subvert monocyte differentiation into dendritic cell: implication for its efficacy as a vaccine preventing tuberculosis. *Vaccine* 22, 3848–3857
9 Gutierrez, M.G. *et al.* (2004) Autophagy is a defense mechanism inhibiting BCG and mycobacterium tuberculosis survival in infected macrophages. *Cell* 119, 753–766
10 Behr, M.A. (2002) BCG–different strains, different vaccines? *Lancet Infect. Dis.* 2, 86–92
11 Demissie, A. *et al.* (2004) Healthy individuals that control a latent infection with mycobacterium tuberculosis express high levels of Th1 cytokines and the IL-4 antagonist IL-4δ2. *J. Immunol.* 172, 6938–6943
12 Cooper, A.M. *et al.* (1993) Disseminated tuberculosis in interferon gamma gene-disrupted mice. *J. Exp. Med.* 178, 2243–2247
13 Flynn, J.L. *et al.* (1993) An essential role for interferon gamma in resistance to Mycobacterium tuberculosis infection. *J. Exp. Med.* 178, 2249–2254
14 Price, D.A. *et al.* (1997) Positive selection of HIV-1 cytotoxic T lymphocyte escape variants during primary infection. *Proc. Natl. Acad. Sci. U. S. A.* 94, 1890–1895
15 Heeney, J.L. (2004) Requirement of diverse T-helper responses elicited by HIV vaccines: induction of highly targeted humoral and CTL responses. *Expert Rev. Vaccines* 3, S53–S64
16 Kedl, R.M. *et al.* (2002) T cells down-modulate peptide-MHC complexes on APCs *in vivo*. *Nat. Immunol.* 3, 27–32
17 Barthlott, T. *et al.* (2003) T Cell Regulation as a side effect of homeostasis and competition. *J. Exp. Med.* 197, 451–460
18 Weiss, S.T. and Raby, B.A. (2004) Asthma genetics. *Hum. Mol. Genet.* 13, R83–R89
19 Yamada, R. and Ymamoto, K. (2005) Recent findings on genes associated with inflammatory disease. *Mutat. Res.* 573, 136–151
20 Larche, M. and Wraith, D.C. (2005) Peptide-based therapeutic vaccines for allergic and autoimmune diseases. *Nat. Med.* 11, S69–S76
21 Preuss, K.D. *et al.* (2002) Analysis of the B-cell repertoire against antigens expressed by human neoplasms. *Immunol. Rev.* 188, 43–50
22 Darnell, R.B. and Posner, J.B. (2003) Paraneoplastic Syndromes Involving the Nervous System. *N. Engl. J. Med.* 349, 1543–1554
23 Stevenson, F.K. *et al.* (2004) DNA vaccines to attack cancer. *Proc. Natl. Acad. Sci. U. S. A.* 101, 14646–14652
24 Graus, F. *et al.* (2001) Anti-Hu-associated paraneoplastic encephalomyelitis: analysis of 200 patients. *Brain* 124, 1138–1148
25 Lake, R.A. and Robinson, B.W. (2005) Immunotherapy and chemotherapy–a practical partnership. *Nat. Rev. Cancer* 5, 397–405
26 Brindley, P.J. *et al.* (1989) Role of host antibody in the chemotherapeutic action of praziquantel against Schistosoma mansoni: identification of target antigens. *Mol. Biochem. Parasitol.* 34, 99–108
27 Andersen, M.H. *et al.* (2005) Regulators of apoptosis: suitable targets for immune therapy of cancer. *Nat. Rev. Drug Discov.* 4, 399–409
28 Pagano, J.S. *et al.* (2004) Infectious agents and cancer: criteria for a causal relation. *Semin. Cancer Biol.* 14, 453–471
29 Plotkin, S.A. (2005) Vaccines: past, present and future. *Nat. Med.* 11, S5–S11
30 Plotkin, S.A. (2005) Six revolutions in vaccinology. *Pediatr. Infect. Dis. J.* 24, 1–9
31 Pashine, A. *et al.* (2005) Targeting the innate immune response with improved vaccine adjuvants. *Nat. Med.* 11, S63–S68
32 Bramwell, V.W. and Perrie, Y. (2005) Particulate delivery systems for vaccines. *Crit. Rev. Ther. Drug Carrier Syst.* 22, 151–214
33 Bramwell, V.W. *et al.* (2005) Particulate delivery systems for biodefense subunit vaccines. *Adv. Drug Deliv. Rev.* 57, 1247–1265
34 Pink, J.R. and Kieny, M.P. (2004) 4th Meeting on novel adjuvants currently in/close to human clinical testing: World Health Organization–Organisation Mondiale de la Sante Fondation Merieux, Annecy, France, 23–25 June 2003. *Vaccine* 22, 2097–2102
35 Minor, P.D. (2002) Eradication and cessation of programmes: vaccination and public health care. *Br. Med. Bull.* 62, 213–224
36 Oxman, M.N. *et al.* (2005) A Vaccine to prevent herpes zoster and postherpetic neuralgia in older adults. *N. Engl. J. Med.* 352, 2271–2284
37 Ying, H. *et al.* (1999) Cancer therapy using a self-replicating RNA vaccine. *Nat. Med.* 5, 823–827
38 Akira, S. and Takeda, K. (2004) Toll-like receptor signalling. *Nat. Rev. Immunol.* 4, 499–511
39 Lambert, P.H. *et al.* (2005) Can successful vaccines teach us how to induce efficient protective immune responses? *Nat. Med.* 11, S54–S62
40 Alpar, H.O. and Bramwell, V.W. (2002) Current status of DNA vaccines and their route of administration. *Crit. Rev. Ther. Drug Carrier Syst.* 19, 307–383
41 Singh, U. *et al.* (2004) DNA content analysis on microarrays. *Methods Mol. Biol.* 270, 237–248
42 Serruto, D. *et al.* (2004) Biotechnology and vaccines: application of functional genomics to Neisseria meningitidis and other bacterial pathogens. *J. Biotechnol.* 113, 15–32
43 Mora, M. *et al.* (2003) Reverse vaccinology. *Drug Discov. Today* 8, 459–464
44 Kaufmann, S.H.E. and McMichael, A.J. (2005) Annulling a dangerous liaison: vaccination strategies against AIDS and tuberculosis. *Nat. Med.* 11, S33–S44
45 Hovanessian, A.G. *et al.* (2004) The caveolin-1 binding domain of HIV-1 glycoprotein gp41 is an efficient B cell epitope vaccine candidate against virus infection. *Immunity* 21, 617–627
46 Labrijn, A.F. *et al.* (2003) Access of antibody molecules to the conserved coreceptor binding site on glycoprotein gp120 is sterically restricted on primary human immunodeficiency virus type 1. *J. Virol.* 77, 10557–10565
47 Cohen, J. (2005) Public health. High hopes and dilemmas for a cervical cancer vaccine. *Science* 308, 618–621
48 Rose, R.C. *et al.* (1993) Expression of human papillomavirus type 11 L1 protein in insect cells: *in vivo* and *in vitro* assembly of viruslike particles. *J. Virol.* 67, 1936–1944
49 Zhou, J. *et al.* (1991) Expression of vaccinia recombinant HPV 16 L1 and L2 ORF proteins in epithelial cells is sufficient for assembly of HPV virion-like particles. *Virology* 185, 251–257
50 Jodar, L. *et al.* (2002) Development of vaccines against meningococcal disease. *Lancet* 359, 1499–1508
51 Pizza, M. *et al.* (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287, 1816–1820
52 Stemmer, W.P. (1994) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature* 370, 389–391
53 Crameri, A. *et al.* (1996) Improved green fluorescent protein by molecular evolution using DNA shuffling. *Nat. Biotechnol.* 14, 315–319
54 Chang, C.C. *et al.* (1999) Evolution of a cytokine using DNA family shuffling. *Nat. Biotechnol.* 17, 793–797
55 Tobin, M.B. *et al.* (2000) Directed evolution: the 'rational' basis for 'irrational' design. *Curr. Opin. Struct. Biol.* 10, 421–427
56 Punnonen, J. (2000) Molecular breeding of allergy vaccines and antiallergic cytokines. *Int. Arch. Allergy Immunol.* 121, 173–182
57 Sette, A. *et al.* (2005) A roadmap for the immunomics of category A-C pathogens. *Immunity* 22, 155–161
58 Ishii, K. *et al.* (2005) Gene-inducing program of human dendritic cells in response to BCG cell-wall skeleton (CWS), which reflects adjuvancy required for tumor immunotherapy. *Immunol. Lett.* 98, 280–290

Cell Host & Microbe

CellPress

Forum

Rational Vaccine Design in the Time of COVID-19

Dennis R. Burton[1,2,*] and Laura M. Walker[3,*]
[1]The Scripps Research Institute, La Jolla, CA 92037, USA
[2]Ragon Institute of MGH, MIT and Harvard, Cambridge, MA 02139, USA
[3]Adimab, Lebanon, NH 03766, USA
*Correspondence: burton@scripps.edu (D.R.B.), laura.walker@adimab.com (L.M.W.)
https://doi.org/10.1016/j.chom.2020.04.022

As scientists consider SARS-CoV-2 vaccine design, we discuss problems that may be encountered and how to tackle them by what we term "rational vaccine design." We further discuss approaches to pan-coronavirus vaccines. We draw on experiences from recent research on several viruses including HIV and influenza, as well as coronaviruses.

The COVID-19 pandemic is impinging upon the lives of billions of people worldwide. Meanwhile, laboratories across the globe are working intensively on developing small-molecule drugs, antibodies, and vaccines to counter the virus SARS-CoV-2. It is likely that while antibodies to the virus will be available relatively quickly, vaccines will take much longer, and the availability of small molecule drugs is more uncertain. However, most agree that the most affordable long-term solution to the problem posed by the virus is the development of a safe and effective vaccine. The development of such a vaccine could be straightforward, perhaps being solely antibody-based and requiring only the presentation of the surface S protein as a recombinant molecule, a genetic construct, or expressed from a suitable viral vector to induce a long-lived protective antibody response. It is also possible that development will encounter roadblocks that dictate greater sophistication in the design of immunogens and immunization strategies. As a single example of the kind of roadblock that can be encountered, the development of a vaccine for respiratory syncytial virus (RSV) has been held back more than 50 years, fundamentally because of a lack of understanding of the appropriate conformation of the surface F glycoprotein to be presented to the immune system, which has only recently resolved from detailed molecular data. Even if a straightforward approach is effective for a SARS-CoV-2 vaccine, ideally, we would like to develop a vaccine capable of containing multiple betacoronaviruses or at least sarbecoviruses (i.e., "pan-coronavirus" vaccines). Such vaccines would hopefully be effective in reducing disease not only due to current known coronaviruses but also to those that may emerge or re-emerge in the future. This approach would undoubtedly require a great deal of immunogen design work, but there are some hopeful indications from antibody responses to SARS-CoV-1 and SARS-CoV-2.

The COVID-19 Vaccine Landscape

Currently, more than 70 vaccine candidates to SARS-CoV-2 are at some stage of development. Many seek to induce neutralizing antibodies (nAbs) to the spike (S) protein on the surface of the virus, given the association of nAbs with protection for many successful viral vaccines (Figure 1). For a respiratory pathogen such as SARS-CoV-2, a vaccine might seek to induce systemic nAbs and prevent lower respiratory tract infection, as for respiratory syncytial virus (RSV) antibodies and vaccines. The prevention of upper respiratory tract infection, likely mediated by mucosal Abs, may be more difficult to sustain through vaccination. A number of factors may contribute to the development of a successful nAb-based vaccine, including 1) the ability of the vaccine to induce nAbs in most vaccinees, 2) the level of nAbs required to provide protection from disease, 3) the durability of the vaccine-induced nAb response, 4) the durability of memory B cells that might differentiate into Ab-producing cells upon virus exposure, 5) the dependence of nAb protection on the ability of vaccine-induced Abs to activate Fc-mediated effector functions, 6) complicating adverse events that may be associated with induction of weakly or non-neutralizing antibodies (antibody-dependent enhancement [ADE] or enhanced respiratory disease [ERD]), and 7) the ability of the vaccine to induce cellular immunity that may be required, together with nAbs, to provide optimal protection.

Data on these factors is expected to accumulate rapidly as human vaccine trials progress. Meanwhile, preliminary animal protection studies provide some evidence of protection against re-infection with SARS-CoV-2 (Bao et al., 2020). For SARS-CoV-1 and MERS, animal models provide evidence of vaccine protection, including in nonhuman primates (NHPs) (Wang et al., 2015). There is data for SARS-CoV-1 showing antibody enhancement of infection in NHPs, mice, ferrets, but in almost all cases, vaccination is associated with greater survival and reduced virus titers (Roper and Rehm, 2009).

Overall, there are currently grounds for optimism that one of the strategies being investigated to generate a SARS-CoV-2 vaccine will be effective, at least in the shorter term. However, there are some potential red flags, including the apparent failure of some individuals to generate nAbs following natural infection (frequently, but not always, natural infection is more effective at the induction of nAbs than vaccination), the durability of nAb responses observed in other coronavirus infections, and some observations of immune enhancement effects in coronavirus infections.

A Rational Approach to a SARS-CoV-2 Vaccine

If current approaches to a SARS-CoV-2 vaccine are less than optimal, rational vaccine design strategies could be

employed to direct the immune response toward protective epitopes on the S protein. In principle, focusing the B cell response toward epitopes associated with potent neutralizing activity would lead to longer term vaccine protection due to the lower concentrations of antibody required for neutralization. This strategy would also minimize the induction of non- or weakly- neutralizing antibodies, which would mitigate the potential for immune enhancement. Given that the vast majority of anti-CoV nAbs have been shown to target the receptor binding domain (RBD), one way to accomplish this may be to immunize only with the RBD rather than the whole S protein. Indeed, this approach seems to show initial success in induction of nAbs in an animal model (Quinlan et al., 2020).

Another approach would begin by identifying potent nAbs from naturally infected donors and structurally defining the epitopes recognized by these antibodies. This process of designing vaccine antigens based on an exploration of the interactions between potent nAbs and their target epitopes has been termed "reverse vaccinology 2.0" (Burton, 2017). However, for many viruses, one historical hurdle to this approach has been the difficulties associated with generating potent nAbs from naturally infected donors. Studies have shown that such antibodies, particularly the most potent, can be rare in anti-viral memory B cell repertoires, and their identification has often required (1) screening of many donors to identify those that have mounted the most potent neutralizing serum responses and (2) high-throughput single B cell technologies that enable deep mining of antigen-specific memory B cell repertoires (Walker and Burton, 2018). Over the past decade, these two advances have led to the discovery of highly potent, and in some cases broadly cross-reactive, nAbs to a plethora of viral pathogens (Corti and Lanzavecchia, 2013). It is too early (mid-April 2020) to definitively state how difficult it will be to isolate potent nAbs to SARS-CoV-2, but early data suggests this may not be as difficult as it has been for some viruses.

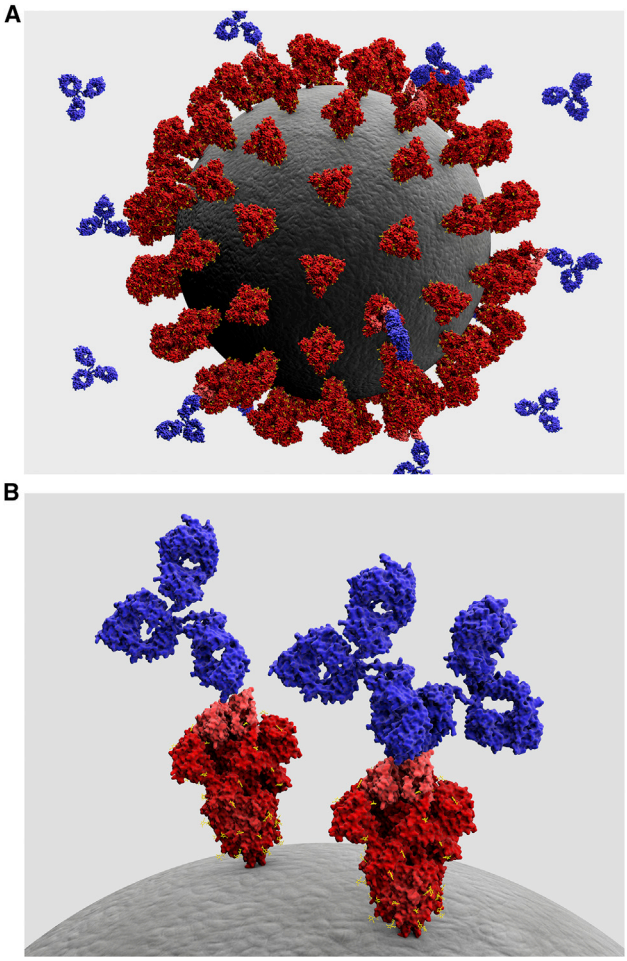


Figure 1. Graphical Visualization of Antibodies Binding to Coronavirus Spike Proteins on the Virion Surface

(A) Coronavirus particle studded with S glycoprotein molecules (red) and antibody IgG molecules (purple), bound and free. The E and M proteins are not shown in this representation.

(B) Two S glycoprotein molecules on the virus surface, one with one IgG molecule (purple) bound, one with two IgG molecules bound. Only the first two glycan residues of each glycan chain are shown.

Once panels of potent nAbs are identified, structural studies of these antibodies in complex with full-length or sub-domains of S will be required to inform the generation of immunogens that optimally present these neutralizing epitopes to the immune system. Various antigen engineering strategies have been employed to focus the antibody response on protective epitopes, including germline targeting, epitope-based protein scaffolds and structure-guided stabilization of full-length envelope proteins. The latter antigen engineering strategy, first applied to RSV F protein, was employed to generate a stabilized MERS S protein, which induced higher titers of neutralizing antibodies than wild-type S in a mouse model (Pallesen et al., 2017). Recently, these same mutations proved to be effective in stabilizing the SARS-CoV-2 S trimer, which allowed for the rapid determination of the structure by cryo-EM (Wrapp et al., 2020). Immunogenicity studies in animal models and humans will reveal how effective this molecule is in inducing nAbs to SARS-CoV-2. If nAbs are not high in the overwhelming majority of vaccinees, then study of the binding of nAbs and non-neutralizing Abs (nnAbs) to the immunogen can help identify potential flaws in design. Various immunogen design strategies could then be employed to reduce the immunogenicity of these non-desired epitopes. For example, in the case of Zika virus, a stabilized E-dimer-based subunit vaccine was recently engineered to abrogate the induction of antibodies to the immunodominant fusion loop and precursor membrane protein, which are antigenic sites associated with weakly neutralizing but enhancing activity (Slon-Campos et al., 2019). Immunization of mice with this stabilized Zika E antigen resulted in the induction of protective antibodies that did not cross-react with DENV or induce ADE of DENV infection. Indeed, ADE has been raised as a potential complication of a SARS-CoV-2 vaccine, although opinion remains sharply divided as to whether this is a real or hypothetical complication. Immune enhancement could be associated with Fc-receptor-enhanced uptake of viruses into FcR-bearing cells, as for flaviviruses (ADE), or

could possibly result from Abs, perhaps particularly nnAbs, forming immune complexes with viral proteins and depositing in capillaries of the lung, leading to complement activation and tissue damage (ERD). In either case, more precise immunogen design than typically used in vaccine design would be indicated. Finally, glycan masking is another strategy that has shown some success in limiting antibody responses to non-desired epitopes.

Another scenario that may arise is that traditional SARS-CoV-2 vaccines induce nAb responses that wane rapidly over time. This is of particular concern given previous studies showing that a large proportion of individuals exposed to SARS and MERS failed to generate long-lived nAb responses (Choe et al., 2017; Wu et al., 2007). Furthermore, in the case of MERS, many patients with mild or asymptomatic disease did not even mount short-lived nAb responses, and emerging data suggests the same may be true for SARS-CoV-2 (Choe et al., 2017; Wu et al., 2020). It may then be that a vaccine has to improve upon natural infection. Strategies to improve on the immunogenicity of the viral glycoprotein are being explored for HIV. One example is "slow delivery" immunization, whereby the dose of antigen is escalated over days to weeks. Recent studies have shown that this method of immunogen delivery leads to prolonged retention of antigen in lymph nodes and increased germinal center B cell numbers compared to traditional "single bolus" immunization (Tam et al., 2016). Furthermore, in the context of HIV immunization, slow delivery immunization was shown to enhance nAb development in NHPs by modulating the immunodominance of non-neutralizing epitopes (Cirelli et al., 2019). In another example, it was recently shown that site-specific immobilization of HIV trimer antigens on alum prolonged immunogen bioavailability, resulting in enhanced germinal center and nAb responses compared to traditional alum-absorbed antigens (Moyer et al., 2020). Finally, of course, the development of potent adjuvants such as GSK's AS01B can greatly enhance Ab responses to glycoproteins.

A Rational Approach to Pan-coronavirus Vaccines

A new coronavirus will likely emerge in the future, just as the current virus has followed SARS and MERS. The same pattern is true of other pathogens. A Zika virus epidemic followed the endemic dengue and yellow fever viruses, West Nile virus spread across the United States, and other flaviviruses may emerge. Ebola Bundibungyo emerged after Ebola Zaire and Ebola Sudan had been described. Hantaviruses such as Puumala and Seoul were well known in Eurasia, but then more recently Sin Nombre and Andes viruses emerged in the Americas. The list goes on to reflect the fact that pathogens mutate. Thus, in many ways, many currently envisaged countermeasures to COVID-19 can be viewed as temporary fixes to a long-term problem that has existed for decades.

Ideally, vaccines should provide protection not only against current versions of pathogens but also against those likely to emerge in the future. Are such "pan-pathogen" vaccines possible? The discovery of broadly neutralizing antibodies (bnAbs) first against highly antigenically variable viruses, such as HIV and influenza virus, and then against related paramyxoviruses, flaviviruses, lyssaviruses, orthopoxviruses, and filoviruses provides the proof of principle for pan-pathogen vaccines (Walker and Burton, 2018). It seems that, if one searches hard enough, one can find antibodies that recognize relatively conserved epitopes even against a background of considerable variation. We attribute this observation to the ability of antibodies to recognize every nook and cranny of a pathogen surface, albeit in any given case in a minority of individuals and as a minor part of the response. Early data describing a potent mAb that cross-neutralizes SARS-CoV-2 and SARS-CoV-1 provide support for attempts to design a pan-sarbecovirus vaccine (Pinto et al., 2020). The cross-reactive nAb binds to the receptor binding domain (RBD) of the two viruses, suggesting this domain as a target for vaccine design. The generation of pan-betacoronavirus vaccines is likely to be more difficult. For example, the S proteins of MERS-CoV and SARS-CoV-2 share only 35% identity. Nevertheless, structural studies suggest that the fusion peptide is accessible and highly conserved and may be an appropriate vaccine target (Walls et al., 2016).

Another less appreciated advantage of designing a pan-coronavirus vaccine that elicits cross-reactive antibodies is that such antibodies may be more effective against variants of a single virus compared to strain-specific antibodies. For instance, a highly cross-reactive antibody against a functional site on both SARS-CoV-1 and SARS-CoV-2 is more likely to target a limited set of critical conserved residues that will show less propensity for mutation than a SARS-CoV-2-specific antibody to that site. This feature might be accomplished by antibodies targeting a smaller conserved footprint within the functional binding site. In any case, there is inevitably concern that a new virus infecting millions, or even billions, of people worldwide will generate variants that will require a vaccine able to induce antibodies with some degree of flexibility in their precise recognition properties.

Finally, it should be noted that the antibodies themselves, initiating the types of vaccine design efforts described above and dubbed "super-antibodies" for their outstanding attributes of cross-reactivity and/or potency, represent promising agents for prophylaxis and therapy of viral infections (Walker and Burton, 2018). High potency means that less antibody is required for efficacy, reducing the frequently quoted high cost of antibodies as drugs. Other factors that can reduce costs include antibody half-life extension and advances in antibody delivery and production. The path to FDA approval is also typically much faster for antibodies than for vaccines and small-molecule drugs. Overall, the combination of these features may enable antibodies to be one of the front-line treatments in responding to future outbreaks of infectious disease.

Conclusions

An unprecedented effort is ongoing to develop a SARS-CoV-2 vaccine. It is strongly hoped that this effort will be immediately and fully successful. However, some caution is prudent, and the gains that have been made in the last decade in understanding the interplay between the humoral immune system and viruses and in rational vaccine design should be fully exploited without waiting to see the results of the first vaccine efforts. Such endeavors will provide a "Plan B" in the event of problems or complications and, in any case, will likely contribute to an optimal vaccine. Further, the promise of developing pan-coronavirus vaccines to cope with future emerging pathogens should be explored.

REFERENCES

Bao, L., Deng, W., Gao, H., Xiao, C., Liu, J., Xue, J., Lv, Q., Liu, J., Yu, P., Xu, Y., et al. (2020). Reinfection could not occur in SARS-CoV-2 infected rhesus macaques. bioRxiv. https://doi.org/10.1101/2020.1103.1113.990226.

Burton, D.R. (2017). What Are the Most Powerful Immunogen Design Vaccine Strategies? Reverse Vaccinology 2.0 Shows Great Promise. Cold Spring Harb. Perspect. Biol. *9*, 030262.

Choe, P.G., Perera, R.A.P.M., Park, W.B., Song, K.H., Bang, J.H., Kim, E.S., Kim, H.B., Ko, L.W.R., Park, S.W., Kim, N.J., et al. (2017). MERS-CoV Antibody Responses 1 Year after Symptom Onset, South Korea, 2015. Emerg. Infect. Dis. *23*, 1079–1084.

Cirelli, K.M., Carnathan, D.G., Nogal, B., Martin, J.T., Rodriguez, O.L., Upadhyay, A.A., Enemuo, C.A., Gebru, E.H., Choe, Y., Viviano, F., et al. (2019). Slow delivery immunization enhances HIV neutralizing antibody and germinal center responses via modulation of immunodominance. Cell *177*, 1153–1171.e1128..

Corti, D., and Lanzavecchia, A. (2013). Broadly neutralizing antiviral antibodies. Annu. Rev. Immunol. *31*, 705–742.

Moyer, T.J., Kato, Y., Abraham, W., Chang, J.Y.H., Kulp, D.W., Watson, N., Turner, H.L., Menis, S., Abbott, R.K., Bhiman, J.N., et al. (2020). Engineered immunogen binding to alum adjuvant enhances humoral immunity. Nat. Med. *26*, 430–440.

Pallesen, J., Wang, N., Corbett, K.S., Wrapp, D., Kirchdoerfer, R.N., Turner, H.L., Cottrell, C.A., Becker, M.M., Wang, L., Shi, W., et al. (2017). Immunogenicity and structures of a rationally designed prefusion MERS-CoV spike antigen. Proc. Natl. Acad. Sci. USA *114*, E7348–E7357.

Pinto, D., Park, Y.J., Beltramello, M., Walls, A.C., Tortorici, M.A., Bianchi, S., Jaconi, S., Culap, K., Zatta, F., De Marco, A., et al. (2020). Structural and functional analysis of a potent sarbecovirus neutralizing antibody. bioRxiv. https://doi.org/10.1101/2020.1104.1107.023903.

Quinlan, B.D., Mou, H., Zhang, L., Guo, Y., He, W., Ojha, A., Parcells, M.S., Luo, G., Li, W., Zhong, G., et al. (2020). The SARS-CoV-2 receptor-binding domain elicits a potent neutralizing response without antibody-dependent enhancement. bioRxiv. https://doi.org/10.1101/2020.1104.1110.036418.

Roper, R.L., and Rehm, K.E. (2009). SARS vaccines: where are we? Expert Rev. Vaccines *8*, 887–898.

Slon-Campos, J.L., Dejnirattisai, W., Jagger, B.W., López-Camacho, C., Wongwiwat, W., Durnell, L.A., Winkler, E.S., Chen, R.E., Reyes-Sandoval, A., Rey, F.A., et al. (2019). A protective Zika virus E-dimer-based subunit vaccine engineered to abrogate antibody-dependent enhancement of dengue infection. Nat. Immunol. *20*, 1291–1298.

Tam, H.H., Melo, M.B., Kang, M., Pelet, J.M., Ruda, V.M., Foley, M.H., Hu, J.K., Kumari, S., Crampton, J., Baldeon, A.D., et al. (2016). Sustained antigen availability during germinal center initiation enhances antibody responses to vaccination. Proc. Natl. Acad. Sci. U S A *113*, E6639–E6648.

Walker, L.M., and Burton, D.R. (2018). Passive immunotherapy of viral infections: 'super-antibodies' enter the fray. Nat. Rev. Immunol. *18*, 297–308.

Walls, A.C., Tortorici, M.A., Bosch, B.J., Frenz, B., Rottier, P.J.M., DiMaio, F., Rey, F.A., and Veesler, D. (2016). Cryo-electron microscopy structure of a coronavirus spike glycoprotein trimer. Nature *531*, 114–117.

Wang, L., Shi, W., Joyce, M.G., Modjarrad, K., Zhang, Y., Leung, K., Lees, C.R., Zhou, T., Yassine, H.M., Kanekiyo, M., et al. (2015). Evaluation of candidate vaccine approaches for MERS-CoV. Nat. Commun. *6*, 7712.

Wrapp, D., Wang, N., Corbett, K.S., Goldsmith, J.A., Hsieh, C.L., Abiona, O., Graham, B.S., and McLellan, J.S. (2020). Cryo-EM structure of the 2019-nCoV spike in the prefusion conformation. Science *367*, 1260–1263.

Wu, L.P., Wang, N.C., Chang, Y.H., Tian, X.Y., Na, D.Y., Zhang, L.Y., Zheng, L., Lan, T., Wang, L.F., and Liang, G.D. (2007). Duration of antibody responses after severe acute respiratory syndrome. Emerg. Infect. Dis. *13*, 1562–1564.

Wu, F., Wang, A., Liu, M., Wang, Q., Chen, J., Xia, S., Ling, Y., Zhang, Y., Xun, J., Lu, L., et al. (2020). Neutralizing antibody responses to SARS-CoV-2 in a COVID-19 recovered patient cohort and their implications. medRxiv. https://doi.org/10.1101/2020.1103.1130.20047365.

OPEN ACCESS Freely available online

PLOS | PATHOGENS

Review

Vaccines: From Empirical Development to Rational Design

Christine Rueckert, Carlos A. Guzmán*

Department of Vaccinology and Applied Microbiology, Helmholtz Centre for Infection Research, Braunschweig, Germany

Abstract: Infectious diseases are responsible for an overwhelming number of deaths worldwide and their clinical management is often hampered by the emergence of multi-drug-resistant strains. Therefore, prevention through vaccination currently represents the best course of action to combat them. However, immune escape and evasion by pathogens often render vaccine development difficult. Furthermore, most currently available vaccines were empirically designed. In this review, we discuss why rational design of vaccines is not only desirable but also necessary. We introduce recent developments towards specifically tailored antigens, adjuvants, and delivery systems, and discuss the methodological gaps and lack of knowledge still hampering true rational vaccine design. Finally, we address the potential and limitations of different strategies and technologies for advancing vaccine development.

Introduction

Scourges of humanity, such as smallpox, polio, and measles, have been controlled by vaccination. Other epidemics, for instance tuberculosis, have yet to be sufficiently restrained by immunization. Accordingly, policy makers have given a high priority to the development of novel vaccines to induce protective immunity against selected pathogens. Most human vaccines contain attenuated or killed pathogens and were developed empirically, such as the yellow fever vaccine [1,2]. Safety concerns were associated with undefined vaccine preparations based on whole pathogens (e.g., inactivated or attenuated bacteria or viruses). Thus, novel subunit vaccines are based on a restricted number of individual components (i.e., antigens) of the specific pathogen, which are able to confer protective immunity. Obviously, the chances of finding effective components of subunit vaccines empirically are low. Immunogenic parts of pathogens that provide antigens for B cell receptors (BCRs) and antigenic peptides that are presentable by MHC molecules to T cell receptors (TCRs) have to be identified. It is critical to compensate for excluded pathogen-associated molecular patterns (PAMPs), which activate the innate immune system to induce an appropriate adaptive immune response. Finally, vaccine delivery systems may be needed. Hence, the rational design of vaccines is mandatory.

Rationally designed vaccines are composed of antigens, delivery systems, and often adjuvants that elicit predictable immune responses against specific epitopes to protect against a particular pathogen. In many cases a vaccine cannot be successfully designed due to insufficient knowledge about the mechanisms of protection. Although the repertoire of immune clearance mechanisms to fight pathogens is known, the specific contributions of different effector mechanisms are well-characterized for only a few pathogens. It is also largely unclear what determines the immunogenicity and selection of particular epitopes among all possible antigenic options offered by a pathogen. Which factors determine dominant or balanced immune responses? What are the mechanisms leading to long-term protection? Investigation of immune responses to known effective and ineffective vaccines and of pathogens' strategies of immune escape and evasion generates the basis to tackle these open questions. The approach relies on data from studies with empirically developed vaccines—for now and in the near future.

There are no universally accepted strategies and tools to rationally design vaccines. Vaccine development is still generally a tedious and costly empiric process. This review focuses on approaches to overcome empirical vaccine development and addresses their potential and limitations. It will become clear that even the latest developments are mostly first steps. Reports may sometimes sound too optimistic with regard to a prompt implementation of the introduced methods. Nevertheless, multiscale interdisciplinary efforts are strongly needed to reach this goal.

Antigen Selection and Optimization

Selecting the optimal antigen represents the cornerstone in vaccine design. With the advent of genomics, the traditional process of selecting candidate antigens one by one has been replaced by reverse vaccinology approaches. Namely, the coding potential of a pathogen's genome is exploited by *in silico* selection, high throughput screenings, and profiling technologies (e.g., genomics, proteomics) to define promising antigens in relation to in vivo expressed genes and clonal variation [3–6]. Importantly, this approach is not suitable for nonproteinaceous antigens. Depending on the desired response, the antigenic protein should contain appropriate BCR epitopes and peptides that can be recognized by the TCR in a complex with MHC molecules. Synthetic peptides produced at comparably low cost can also be incorporated in subunit vaccines. This is relevant especially in epidemic situations when large amounts of vaccine doses need to be produced in a very limited period of time. A peptide-based

Citation: Rueckert C, Guzmán CA (2012) Vaccines: From Empirical Development to Rational Design. PLoS Pathog 8(11): e1003001. doi:10.1371/journal.ppat.1003001

Editor: Tom C. Hobman, University of Alberta, Canada

Published November 8, 2012

Funding: This work was supported in part by grants from the EU (PANFLUVAC, TRANSVAC); BMBF in the context of the programs Gerontosys 2 (Gerontoshield), EuroNanoMed (HCVAX) and ERANetRUS (HCRUS), and the Helmholtz Association (IG-SCID). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: carlos.guzman@helmholtz-hzi.de

vaccine meets high safety standards due to the possibility of excluding allergens, toxins, or other functional molecular domains of the pathogen. Restricting the immune response to defined antigenic regions can, furthermore, help avoid effects such as autoimmune responses, dominant responses against epitopes prone to antigenic drift, or responses against epitopes with specificity for a particular strain rather than multiple strains of the pathogen. However, the identification of immunogenic peptide sequences requires a considerable amount of experimental effort. Computational prediction methods can strongly reduce time and costs for vaccine development. Nevertheless, clonal variability and in vivo selection resulting in immune escape could render ineffective a vaccine based on short peptides encompassing a limited number of epitopes. Furthermore, there are technological constraints associated with this approach (e.g., synthesis of long polypeptides).

To elicit antibody responses, vaccines should include BCR epitopes. Their prediction is particularly challenging, though, and most B cell epitopes are discontinuous; that is, they are comprised of distant parts of the protein's primary structure. In addition, they are of variable length (3–30 amino acids) and conformation-dependent [7]. BCR epitopes do not possess physico-chemical patterns in their amino acid sequences that can be used for *in silico* prediction [8]. Some epitopes change conformation when interacting with the cognate antibody's paratope, making even 3-D structure-based prediction difficult [7]. The use of learning machines that depend on quantitative data on known antibody epitopes led to the development of prediction tools for linear epitopes such as BCPREDS [9,10] and IMMUNOPRED [11,12]. In contrast, PEPOP [13,14] is based on 3-D structural data on antigen–antibody complexes, and it predicts discontinuous epitopes, their antigenicity, and immunogenicity, and suggests peptide constructs for synthesis. However, these methods have not yet reached sufficient predictive accuracy to be routinely applied in vaccine design.

The proteins or peptides of a subunit vaccine should also display sequences that allow T cell epitope formation in a complex with MHC molecules. MHC class I and II come in hundreds of alleles that are differentially combined between individuals. Choosing immunogenic peptides presented by MHC faces the challenge of not only predicting sequences appropriate for complexing with a particular MHC allele but also finding peptides that can reliably build epitopes in the diverse genetic background within a human population. Drawbacks of in vitro assay-based TCR epitope identification are (i) time consuming procedures for combinatorial coverage of relevant MHC alleles and candidate pathogenic antigens, (ii) high costs for peptide synthesis and reagents, and (iii) limited sensitivity when using naïve T cell populations. These efforts can be reduced extremely when combined with computational TCR epitope prediction [15,16].

In silico prediction of T cell epitopes cannot be based on physico-chemical properties of presented peptides but depends on the application of learning machines on data sets of known MHC allele–peptide pairs. The development and maintenance of databases is absolutely essential to constantly improve predictions [17,18]. Examples for such databases are IEDB [19,20] or SYFPEITHI, which only lists experimentally validated natural MHC–peptide complexes [21,22]. The tools OptiTope [23,24] and NetMHCcons [25,26] select for epitope peptides from specific MHC alleles or sets of MHC alleles as they occur naturally in individuals of a certain population. This is achieved by choosing promiscuous peptides that can be presented by several different MHC alleles of a supertype (i.e., universal peptides presented by most known alleles or a mixture of peptides binding to the most prevalent alleles within a population). The final goal is to provide suitable tools to generate immunogenic peptide sequences from any input antigen sequences. However, the broad applicability of these approaches towards rational vaccine design still remains to be proven.

Diversity also occurs at the level of the antigen. Immune escape of pathogen variants through mutation of immunogenic sequences has to be considered when selecting or designing antigens [27]. *In silico* generation of mosaic polyvalent antigens tackles this problem [28]. Immunization experiments with primates demonstrated the advantage of mosaic constructs over consensus or natural sequences to elicit T cell responses covering a broad selection of viral clades as well as antigenic immune escape variants that may evolve [29,30]. The repertoire of possible immunogens can also be widened by exploring glycan antigens [31]. Whenever constructs are designed, one has to ensure their stability and thus bioavailability. For example, HIV-derived peptides display quite variable half-lifes in the cytosol of human cells and this has an impact on their recognition by $CD8^+$ cells [32]. Structural vaccinology is a powerful emerging approach to optimize immunogens based on atomic-level structural information on requirements for conferring protective immunity [33–35]. Upon identification of immunogenic domains, it is possible to design constructs that lack decoy or masking portions of the antigen, such as epitope scaffolds that are able to elicit antibody responses against otherwise immune-recessive, cryptic, or transient epitopes [36]. It is also possible to engineer an optimized structure to enable broadly cross-protective responses. As example, chimeric proteins to effectively vaccinate against group B streptococci or *Neisseria meningitidis* were generated [37,38].

Adjuvants

Subunit vaccines are likely to lack the molecular cues needed for efficient activation of the innate immune system, thereby failing to induce vigorous adaptive immunity. PAMPs can act as adjuvants, however many pathogen-derived products might exhibit toxic activity [39]. The only globally approved adjuvant for humans is alum. It facilitates T_H2-dependent immune responses but promotes less effective cytotoxic responses and can cause side effects. A number of other adjuvants have been recently approved for use in defined human vaccines, such as MF59 and monophosphoryl lipid A-containing formulations [40,41], and there are other candidates in the pipeline. Adjuvants are not licensed per se, but as part of vaccine formulations. This together with stringent requirements for reagents used on healthy individuals raise the costs of clinical development [41]. Considerable effort was invested in the development of adjuvants for mucosal immunization [42]. Vaccination via mucosal routes is known to elicit both mucosal and systemic immunity [43], fighting pathogens at the site of entry. However, safety issues were observed following intranasal vaccination with the heat labile toxin of *Escherichia coli* and its attenuated derivative [42,44]. This will need to be considered for current candidate mucosal adjuvants, among them compounds with well-defined molecular targets, such as PAMPs, cytokines, and cyclic di-nucleotides [45–47]. For example, the TLR9-agonist CpG enhanced immune responses after vaccination against hepatitis B, anthrax, influenza, and malaria [48–51] and proved promising in vaccination of otherwise nonresponsive immune-compromised organisms [52]. However, many molecular mechanisms of adjuvanticity are still elusive. First insights were gained in receptors and signaling pathways involved in the recognition and processing of pathogenic factors and adjuvants in cells of the innate immune system [53–55]. Nevertheless, the discovered

mechanisms of adjuvanticity do not translate to generally applicable strategies for rationally designed vaccines (see also [4]). Hence, to date, adjuvantation requires an additional solid theoretical background for systematic implementation in rational vaccine design.

Antigen Delivery Systems

Delivery systems become necessary when antigens are not efficiently transported to the inductive sites or presented to the immune system. For example, rapid degradation can result in weak or virtually absent responses to otherwise immunogenic antigens. The coding sequence of an antigen can be integrated into a live virus-vector, which infects antigen-presenting cells (APCs), preferentially dendritic cells (DCs) [56,57]. The antigen is then directly presented by MHC molecules and can be recognized by TCRs. The continuous antigen expression leads to its persistent exposure to immune cells. Recombinant viral vectors can be modified with regard to effector cell targeting, expression promoters, and the type of antigenic transgene. Lentiviral vectors with improved safety and efficiency parameters have a comparatively high capacity for encoding transgenes, high transduction efficiency, low anti-vector host immunity, low genotoxicity, and persistent gene expression [58]. They proved promising in vaccination of mice with HIV-derived antigens and in nonhuman primates with SIV-derived antigens [59,60]. In spite of the adenoviral vaccine vector's known limited efficacy due to preexisting immunity in large populations [61,62], it still induces protective immune responses with characteristic induction of $CD8^+$ T cells in humans [63]. Recombinant adenoviral vectors derived from uncommon human serotypes, chimpanzee or human/chimpanzee chimeras can circumvent the problem of host immunity [64–66]. Human cytomegalovirus (hCMV) vaccine vectors are based on the ability of hCMV strains to superinfect individuals with persistent hCMV infection and immunity. Rhesus macaques developed specific $CD4^+$ and $CD8^+$ responses against SIV antigens delivered by a recombinant CMV vector [67,68]. Elucidation of the molecular mechanisms leading to memory inflation during chronic hCMV infections might even lead to hCMV-based strategies to trigger life-long responses. Attenuated recombinant poxviruses are also intrinsically immunogenic, and insights in the promoted innate immune responses have accumulated [69]. The above-described vectors have considerable potential in human vaccination, especially in prime-boost regimens aimed at fine-tuning responses [70]. Different attenuated or commensal bacteria have also been successfully exploited for delivering vaccine antigens and biologicals [71–75].

The delivery to DCs can be achieved by coupling antigens to antibodies specific for surface molecules, such as Clec9A. This method leads to antigen uptake and activation of T and B cells [76]. Similarly, fusion proteins of HIV antigens and antibody fragments targeting the DC surface molecule DEC205 elicited potent cellular immunity in nonhuman primates [77]. The risks related to live vectors in immune-compromised individuals can be eliminated by the application of virus-like particles (VLPs) that are reduced to the structures and antigenic components necessary for delivery and immunogenicity. VLPs are able to elicit efficient humoral immune responses [78–80], contributing to the control of infection [78,81]. Plasmid DNA vectors can be delivered to cells and elicit humoral responses [82], as proven by DNA vaccines against seasonal influenza in phase I trials [83]. Synthetic delivery

Table 1. Needs and challenges for the rational design of vaccines.

Subunit Vaccine Component	Focus of Future Developments	Benefit Toward Rational Design
Antigens	Knowledge on the most effective immune response against a particular pathogen	Selection of antigens and formulations evoking those responses
	Antibody epitope database	Basis for development of computational prediction tools
	Prediction of sequences that should be excluded due to (i) risk of autoimmune responses, (ii) immune escape by antigenic drift, and (iii) responses to only selected strains or clades of the pathogen	Design of antigens capable of eliciting potent cross-reactive immune responses with minimal risk for side effects
	Continuous survey and registration of evolving pathogenic strains and clades	Improved coverage for selected antigens
	Investigation of protein/peptide degradation rules for different vaccination routes	Improved stability of designed antigens
	Extension of MHC allele–peptide complex databases, especially for MHC class II	Increased reliability of epitope prediction with already available tools
Delivery systems	Advancement of nanotechnologies	Improved synthetic delivery systems
	Investigation of mechanisms to overcome preexisting immunity or persistent virus superinfection	Maximizes potential of live vectors derived from pathogens causing common human chronic infections
	Understanding the basis for eliciting memory responses	Design of vaccines triggering long-lasting protection
	Investigation of the interface between innate and adaptive immunity	Exploitation of optimal APC targets and intrinsic adjuvant properties of the delivery system
Adjuvants	Knowledge on the most effective immune response against a particular pathogen	Selection of adjuvants facilitating those responses
	Investigation of vaccination route-dependent adjuvant effects	Optimized use of adjuvants and vaccine design
	Elucidation of molecular mechanisms of adjuvanticity	Optimizes adjuvant use and forecasts potential side effects
	Investigation of the basis of immune stimulation in different population groups	Development of personalized vaccines

doi:10.1371/journal.ppat.1003001.t001

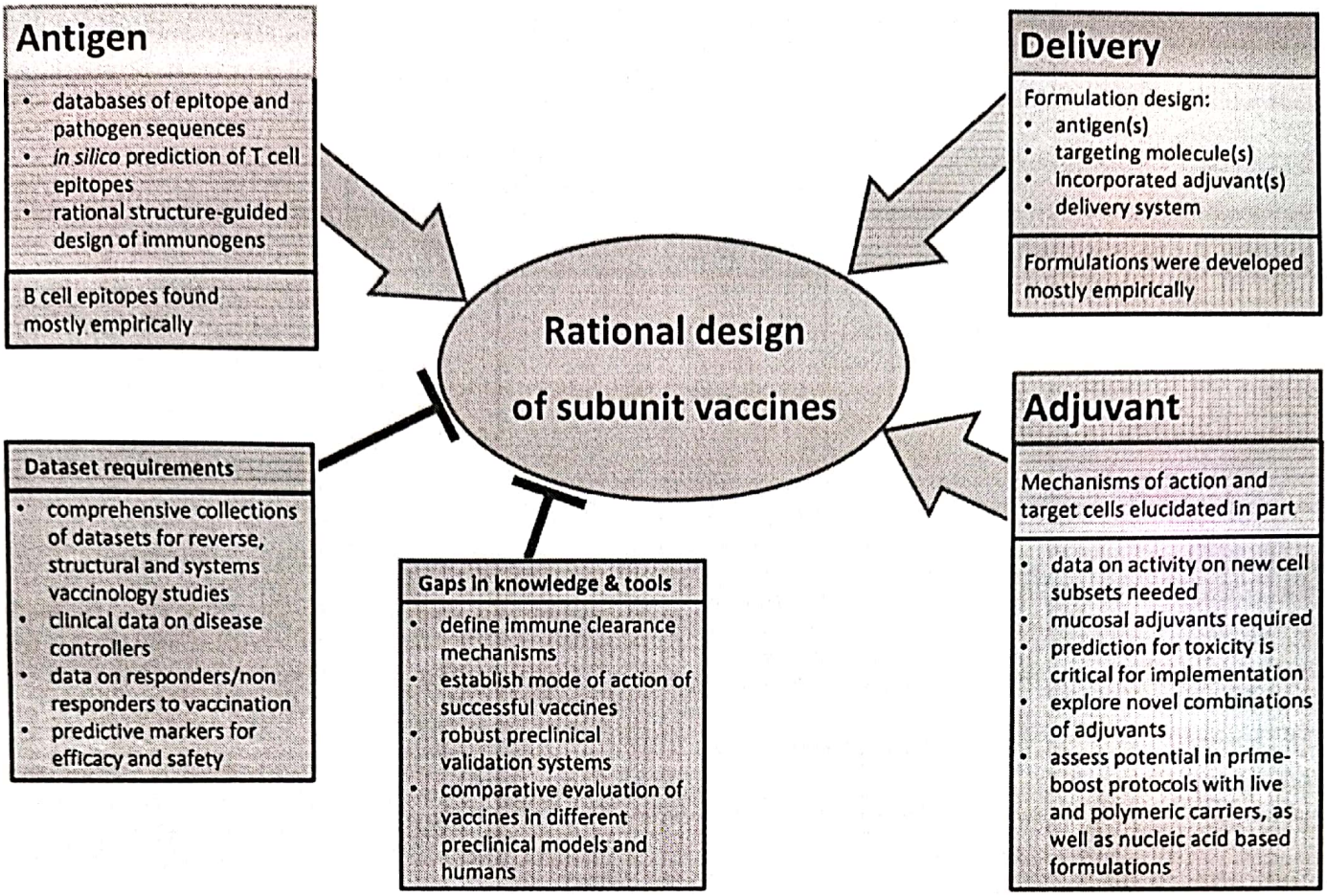


Figure 1. Optimizing the design for more efficient vaccines. Modern vaccinology focuses on the development of subunit vaccines to maximize efficacy and minimize risks in healthy and immune-compromised individuals. Different enabling technologies and knowledge contribute towards the rational design of formulations that would not only exhibit improved performance but also reduce the time and costs associated with preclinical and clinical development. Promising approaches/enabling factors and roadblocks are highlighted in green and pink, respectively.
doi:10.1371/journal.ppat.1003001.g001

systems, such as nanoparticles, block-copolymers, DNA nanostructures, and nanogels [84–86], can be loaded or coated with specific antigens and adjuvants. In addition, they can be tailored and functionalized according to specific needs (e.g., transcutaneous or mucosal delivery) [87,88]. Trials with nanoparticle vaccines for hepatitis B, leishmaniasis, and malaria demonstrated that they enhance immune responses [87,89,90]. Although often developed on an empirical base, the given examples are a proof-of-principle essential to rationally design such delivery vehicles in the future.

Immune Response Prediction

Understanding what is needed to confer protection without side effects is a prerequisite to develop a tailored intervention. To date, characterization of human responses to vaccination relies mainly on measuring antibody titers or cellular responses from peripheral blood samples. This does not allow a comprehensive analysis of responses with regard to the effector cells or mechanisms stimulated and the status in all relevant compartments for acquired immunity. Efforts to tackle this problem link the regulation of transcription or protein activity to the prediction of vaccination outcomes [91]. Recent reports suggest the potential of systems vaccinology for the analysis of gene expression profiling experiments to identify patterns or signatures linked to a desired outcome of vaccination [92–94]. Human studies showed correlations of gene expression profiles or protein expression patterns with immune system activation upon vaccination against yellow fever and influenza in responders and nonresponders [95–97]. Others characterized transcription profiles after treatment of mice or murine DCs with adjuvant molecules [98,99]. Correlations between successful immunization or toxic events and cellular expression profiles can be predictive for a particular vaccine. However, no general unambiguous markers were identified that would allow accurate prediction of efficacy or safety for vaccines in trials (introduced, for example, in [5]).

A quite different approach to predict immune responses upon exposure to potential immunogens is realized by the *in silico* immune system simulator C-ImmSim [100–102]. This model features simulation of different classes of B and T lymphocytes, innate immune cells (e.g., DCs and macrophages), and different immune compartments (e.g., bone marrow, thymus and tertiary lymphoid organs). *In silico* experiments simulate primary immune responses as well as challenge with a particular antigen in different definable MHC allele backgrounds. The proof-of-principle was performed with antigens of HIV or influenza virus that simulate immunization. The simulations could indeed predict observations in humans, for example that affinity maturation and antigenic dominance evolve, and that MHC diversity can have an impact on immune defense [100]. C-ImmSim can be updated whenever improved versions of the incorporated BCR and TCR epitope prediction methods become available. Though currently not successfully applied, the simulator has potential in vaccine

development by testing the immunogenicity of antigens and the potency to induce a robust immune response upon challenge with the antigen. It can be also used as a research tool to elucidate mechanisms of immune responses to fill gaps in knowledge that slow down rational design efforts.

A prevalent problem in vaccine translation is the delayed and costly transition from preclinical to clinical development due to difficulties in predicting human immune responses. Although closer to humans, primate models are associated with ethical, logistic, and financial constraints. An emerging alternative is the use of mice humanized for the immune system. Although they still need to be improved, they can be foreseen as powerful tools to predict human-specific immune responses to vaccines, as well as to investigate vaccine efficacy against pathogens with human tropism [103–105].

Concluding Remarks

In this review we elaborate on recent achievements that facilitate rational vaccine design. There are many visions on the expected impact of reverse vaccinology, epitope prediction, structural vaccinology, systems vaccinology, and personalized medicine on the rational design of effective vaccines [3,5,6,106]. However, the implementation of these concepts towards the development of new and more potent vaccines requires time and considerable financial investment. Rational vaccine design will rely strongly on the availability of clinical data on individuals with different clinical forms of disease or response to vaccination to learn what is needed for protection [107]. The gaps in knowledge on the immune system's specific clearance mechanisms against many pathogens slow down the identification of the immune response that should be evoked by tailored vaccines in different population groups (Table 1). Many aspects of the host pathogen interaction and host immune status during persistent infection are also poorly understood, thereby hindering the development of therapeutic vaccines [108]. Further data from trials with empiric formulations are required to identify patterns or biomarkers that can reliably guide prediction of vaccine efficacy and safety at reasonable success rates (Figure 1). A widely accepted goal in vaccine development is the applicability to huge populations, if not all humankind. Nevertheless, there are reasons for more personalized approaches that consider specific preconditions in recipients, such as genetic background, pre-exposure to pathogens or vaccines, unique physiological background related to local culture/habits, age, and immunodeficiency.

Implementation of rational development concepts in vaccinology demands patience, and advances will be incremental. Realization will depend on the application of flanking logistic and regulatory measures and the awareness of the strong impact of vaccine development to solve global health problems. Funding is also required for the basic research needed to provide the basis for rationally developed vaccines. However, we expect to see the advent of new and more efficient vaccines in the coming years as a result of the implementation of this emerging knowledge and enabling technologies.

Acknowledgments

We are grateful to Dr. Pablo D. Becker and Dr. Blair Prochnow for critical reading of the manuscript and helpful discussions.

References

1. Theiler M, Smith HH (1937) The use of yellow fever virus modified by in vitro cultivation for human immunization. J Exp Med 65: 787–800.
2. Poland JD, Calisher CH, Monath TP, Downs WG, Murphy K (1981) Persistence of neutralizing antibody 30–35 years after immunization with 17D yellow fever vaccine. Bull World Health Organ 59: 895–900.
3. Sette A, Rappuoli R (2010) Reverse vaccinology: developing vaccines in the era of genomics. Immunity 33: 530–541.
4. Bagnoli F, Baudner B, Mishra RP, Bartolini E, Fiaschi L, et al. (2011) Designing the next generation of vaccines for global public health. OMICS 15: 545–566.
5. Poland GA, Kennedy RB, Ovsyannikova IG (2011) Vaccinomics and personalized vaccinology: is science leading us toward a new path of directed vaccine development and discovery? PLoS Pathog 7: e1002344. doi:10.1371/journal.ppat.1002344
6. Kennedy RB, Poland GA (2011) The top five "game changers" in vaccinology: toward rational and directed vaccine development. OMICS 15: 533–537.
7. Ponomarenko JV, Regenmortel MHVv (2009) B-cell epitope prediction. In: Gu J, Bourne PE, editors. Structural bioinformatics, second ed. John Wiley & Sons, Inc.
8. Blythe MJ, Flower DR (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. Protein Sci 14: 246–248.
9. El-Manzalawy Y, Dobbs D, Honavar V (2008) Predicting linear B-cell epitopes using string kernels. J Mol Recognit 21: 243–255.
10. El-Manzalawy Y, Dobbs D, Honovar V (2008) BCPREDS: B-cell epitope prediction server. Artificial Intelligence Research Laboratory, Department of Computer Science, Iowa State University of Science and Technology. Available: http://ailab.cs.iastate.edu/bcpreds/. Accessed August 2012.
11. Wee LJ, Simarmata D, Kam YW, Ng LF, Tong JC (2010) SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction. BMC Genomics 11 Suppl 4: S21.
12. Wee LJ, Simarmata D, Kam YW, Ng HL, Tong JC (2010) BayesB: server for SVM prediction of linear B-cell epitopes using Bayes Feature Extraction. Available: http://www.immunopred.org/bayesb/index.html. Accessed August 2012.
13. Moreau V, Fleury C, Piquer D, Nguyen C, Novali N, et al. (2008) PEPOP: computational design of immunogenic peptides. BMC Bioinformatics 9: 71.
14. Moreau V, Fleury C, Piquer D, Nguyen C, Novali N, et al. (2008) PEPOP. Available: http://pepop.sysdiag.cnrs.fr/PEPOP/. Accessed August 2012.
15. Lundegaard C, Lund O, Buus S, Nielsen M (2010) Major histocompatibility complex class I binding predictions as a tool in epitope discovery. Immunology 130: 309–318.
16. Li Pira G, Ivaldi F, Moretti P, Manca F (2010) High throughput T epitope mapping and vaccine development. J Biomed Biotechnol 2010: 325720.
17. Wang P, Sidney J, Kim Y, Sette A, Lund O, et al. (2010) Peptide binding predictions for HLA DR, DP and DQ molecules. BMC Bioinformatics 11: 568.
18. Tung CW, Ziehm M, Kamper A, Kohlbacher O, Ho SY (2011) POPISK: T-cell reactivity prediction using support vector machines and string kernels. BMC Bioinformatics 12: 446.
19. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, et al. (2010) The immune epitope database 2.0. Nucleic Acids Res 38: D854–D862.
20. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, et al. (2009) Immune Epitope Database and Analysis Resource. Available: http://www.immuneepitope.org/. Accessed August 2012.
21. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50: 213–219.
22. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S (1999) Available: http://www.syfpeithi.de/. SYFPEITHI: a database of MHC ligands and peptide motifs. Accessed August 2012.
23. Toussaint NC, Kohlbacher O (2009) OptiTope–a web server for the selection of an optimal set of peptides for epitope-based vaccines. Nucleic Acids Res 37: W617–W622.
24. Toussaint NC, Kohlbacher O (2009) Available: http://www.epitoolkit.org/optitope. Accessed August 2012.
25. Karosiene E, Lundegaard C, Lund O, Nielsen M (2012) NetMHCcons: a consensus method for the major histocompatibility complex class I predictions. Immunogenetics 64: 177–186.
26. Karosiene E, Lundegaard C, Lund O, Nielsen HM (2012) NetMHCcons 1.0 server. Available: http://www.cbs.dtu.dk/services/NetMHCcons/. Accessed August 2012.
27. Kaur K, Sullivan M, Wilson PC (2011) Targeting B cell responses in universal influenza vaccine design. Trends Immunol 32: 524–531.
28. Fischer W, Perkins S, Theiler J, Bhattacharya T, Yusim K, et al. (2007) Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants. Nat Med 13: 100–106.
29. Barouch DH, O'Brien KL, Simmons NL, King SL, Abbink P, et al. (2010) Mosaic HIV-1 vaccines expand the breadth and depth of cellular immune responses in rhesus monkeys. Nat Med 16: 319–323.
30. Santra S, Liao HX, Zhang R, Muldoon M, Watson S, et al. (2010) Mosaic vaccines elicit CD8+ T lymphocyte responses that confer enhanced immune coverage of diverse HIV strains in monkeys. Nat Med 16: 324–328.

Mol Biotechnol (2008) 40:101–111
DOI 10.1007/s12033-008-9062-7

REVIEW

IMGT®, the International ImMunoGeneTics Information System® for Immunoinformatics

Methods for Querying IMGT® Databases, Tools, and Web Resources in the Context of Immunoinformatics

Marie-Paule Lefranc

Published online: 8 May 2008

Abstract IMGT®, the International ImMunoGeneTics information system® (http://imgt.cines.fr), was created in 1989 by the Laboratoire d'ImmunoGénétique Moléculaire (LIGM) (Université Montpellier 2 and CNRS) at Montpellier, France, in order to standardize and manage the complexity of immunogenetics data. IMGT® is recognized as the international reference in immunogenetics and immunoinformatics. IMGT® is a high quality integrated knowledge resource, specialized in (i) the immunoglobulins (IG), T cell receptors (TR), major histocompatibility complex (MHC) of human and other vertebrates; (ii) proteins that belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF); and (iii) related proteins of the immune systems (RPI) of any species. IMGT® provides a common access to standardized data from genome, proteome, genetics, and three-dimensional (3D) structures for the IG, TR, MHC, IgSF, MhcSF, and RPI. IMGT® interactive on-line tools are provided for genome, sequence, and 3D structure analysis. IMGT® Web resources comprise 10,000 HTML pages of synthesis and knowledge (IMGT Scientific chart, IMGT Repertoire, IMGT Education, etc.) and external links (IMGT Biocnotes and IMGT other accesses).

Keywords IMGT · Ontology · Immunoglobulin · T cell receptor · MHC · IgSF · MhcSF

M.-P. Lefranc (✉)
IMGT, The International ImMunoGeneTics Information System, Laboratoire d'ImmunoGénétique Moléculaire, Université Montpellier 2, Institut de Génétique Humaine, IGH, UPR CNRS 1142, 141 rue de la Cardonille, 34396 Montpellier Cedex 5, France
e-mail: Marie-Paule.Lefranc@igh.cnrs.fr

Introduction

The number of genomics, genetics, three-dimensional (3D), and functional data published in the immunogenetics field is growing exponentially and involves fundamental, clinical, veterinary, and pharmaceutical research. The number of potential protein forms of the antigen receptors, immunoglobulins (IG), and T cell receptors (TR) is almost unlimited. The potential repertoire of each individual is estimated to comprise about 10^{12} different IG (or antibodies) and TR, and the limiting factor is only the number of B and T cells that an organism is genetically programmed to produce. This huge diversity is inherent to the particularly complex and unique molecular synthesis and genetics of the antigen receptor chains. This includes biological mechanisms such as DNA molecular rearrangements in multiple loci (three for IG and four for TR in humans) located on different chromosomes (four in humans), nucleotide deletions and insertions at the rearrangement junctions (or N-diversity), and somatic hypermutations in the IG loci (for review, see [1, 2]).

IMGT®, the International ImMunoGeneTics Information System® (http://imgt.cines.fr) [3, 4], was created in 1989 by the Laboratoire d'ImmunoGénétique Moléculaire (LIGM) (Université Montpellier 2 and CNRS) at Montpellier, France, in order to standardize and manage the complexity of the immunogenetics data. IMGT® is recognized as the international reference in immunogenetics and immunoinformatics. IMGT® is a high quality integrated knowledge resource, specialized in (i) the IG, TR, major histocompatibility complex (MHC) of human and other vertebrates, (ii) proteins that belong to the immunoglobulin superfamily (IgSF) and to the MHC superfamily (MhcSF), and (iii) related proteins of the immune systems (RPI) of any species. IMGT® provides a common access to standardized

data from genome, proteome, genetics, and 3D structures for the IG, TR, MHC, IgSF, MhcSF, and RPI [3, 4].

The IMGT® information system consists of databases, tools, and Web resources [3]. IMGT® databases include one genome database, three sequence databases, and one 3D structure database. IMGT® interactive on-line tools are provided for genome, sequence, and 3D structure analysis. IMGT® Web resources comprise 10,000 HTML pages of synthesis and knowledge (IMGT Scientific chart, IMGT Repertoire, IMGT Education, IMGT Index, etc.) and external links (IMGT Bloc-notes and IMGT other accesses) [4]. Despite the heterogeneity of these different components, all data in the IMGT® information system are expertly annotated. The accuracy, the consistency, and the integration of the IMGT® data, as well as the coherence between the different IMGT® components (databases, tools, and Web resources), are based on IMGT-ONTOLOGY [5], the first ontology in immunogenetics and immunoinformatics. IMGT-ONTOLOGY provides a semantic specification of the terms to be used in the domain and, thus, allows the management of immunogenetics knowledge for all vertebrate species.

Standardization: IMGT-ONTOLOGY and IMGT Scientific Chart

IMGT-ONTOLOGY axioms and the concepts generated from them are available, for the biologists and IMGT® users, in the IMGT Scientific chart [4] and have been formalized, for the computing scientists, in IMGT-ML [6, 7]. The IMGT Scientific chart [4] comprises the controlled vocabulary and the annotation rules necessary for the immunogenetics data identification, description, classification, and numerotation and for knowledge management in the IMGT® information system. All IMGT® data are expertly annotated according to the IMGT Scientific chart rules. Standardized keywords, labels and annotation rules, standardized IG and TR gene nomenclature, the IMGT unique numbering, and standardized origin/methodology are defined, respectively, based on the main axioms of IMGT-ONTOLOGY [5] (Table 1). The IMGT Scientific chart is available as a section of the IMGT® Web resources. Examples of IMGT® expertised data concepts derived from the IMGT Scientific chart rules are summarized in Table 1.

IDENTIFICATION Axiom: IMGT® Standardized Keywords

IMGT® standardized keywords for IG and TR include the following: (i) general keywords—indispensable for the sequence assignments, they are described in an exhaustive and non-redundant list, and are organized in a tree structure and (ii) specific keywords—they are more specifically associated to particularities of the sequences (orphon, transgene, etc.). The list is not definitive and new specific keywords can easily be added if needed. IMGT/LIGM-DB standardized keywords have been assigned to all entries.

DESCRIPTION Axiom: IMGT® Standardized Labels

A total of 270 feature labels are necessary to describe all structural and functional subregions that compose IG and TR sequences, whereas only seven of them are available in EMBL, GenBank, or DDBJ [14–16]. Levels of annotation have been defined, which allow the users to query sequences in IMGT/LIGM-DB even though they are not fully annotated. Prototypes represent the organizational relationship between labels and give information on the order and expected length (in number of nucleotides) of the labels. This provides rules to verify the manual annotation and to design automatic annotation tool. A total of 285 additional feature labels have been defined for the 3D structures. Annotation of sequences and 3D structures with these labels (in capital letters) constitutes the main part of the expertise. Interestingly, 65 IMGT®-specific labels have been entered in the newly created Sequence Ontology [17].

CLASSIFICATION Axiom: IMGT® Standardized IG and TR Gene Nomenclature

The objective is to provide immunologists and geneticists with a standardized nomenclature per locus and per species which allows extraction and comparison of data for the complex B cell and T cell antigen receptor molecules. The concepts of classification have been used to set up a unique nomenclature of human IG and TR genes, which was approved by the Human Genome Organization (HUGO) Nomenclature Committee HGNC in 1999 [9]. All the human IG and TR genes [1, 2, 18, 19] have been entered by the IMGT Nomenclature Committee in Genome Database GDB [8], LocusLink and Entrez Gene at NCBI, USA, and in IMGT/GENE-DB [20]. IMGT reference sequences have been defined for each allele of each gene based on one or, whenever possible, several of the following criteria: germline sequence, first sequence published, longest sequence, and mapped sequence. They are listed in the germline gene tables of the IMGT Repertoire. The IMGT Protein displays show the translated sequences of the alleles *01 of the functional or ORF genes [1, 2].

NUMEROTATION Axiom: The IMGT Unique Numbering

A uniform numbering system for IG and TR sequences of all species has been established to facilitate sequence

Table 1 IMGT-ONTOLOGY main axioms, IMGT Scientific chart rules, and examples of IMGT® expertised data concepts

IMGT-ONTOLOGY main axioms [5]	IMGT Scientific chart rules [4]	Examples of IMGT® expertised data concepts
IDENTIFICATION	Standardized keywords	Species, molecule type, receptor type, chain type, gene type, structure, functionality, and specificity
DESCRIPTION	Standardized labels and annotations	Core (V-, D-, J-, C-REGION)
		Prototypes
		Labels for sequences
		Labels for 2D and 3D structures
CLASSIFICATION	Reference sequences	Nomenclature of the human IG and TR genes [1, 2] [entry in 1999 in GDB [8], HGNC [9], and LocusLink and Entrez Gene at NCBI]
	Standardized IG and TR gene nomenclature (group, subgroup, gene, and allele)	Alignment of alleles
		Nomenclature of the IG and TR genes of all vertebrate species
NUMEROTATION	IMGT unique numbering for:	Protein displays
	V- and V-LIKE-DOMAINs [10]	Colliers de Perles [13]
	C- and C-LIKE-DOMAINs [11]	FR-IMGT and CDR-IMGT delimitations
	G- and G-LIKE-DOMAINs [12]	Structural loops and beta strands delimitations
ORIENTATION	Orientation of genomic instances relative to each other	Chromosome orientation
		Locus orientation
		Gene orientation
		DNA strand orientation
OBTENTION	Standardized origin and methodology	

comparison and cross-referencing between experiments from different laboratories whatever the antigen receptor (IG or TR), the chain type, or the species [21, 22].

This numbering results from the analysis of more than 5,000 IG and TR variable region sequences of vertebrate species from fish to human. It takes into account and combines the definition of the framework (FR) and complementarity determining region (CDR) [23], structural data from X-ray diffraction studies [24], and the characterization of the hypervariable loops [25]. In the IMGT unique numbering, conserved amino acids from FR always have the same number whatever the IG or TR variable sequence and whatever the species they come from, for example cysteine 23 (in FR1-IMGT), tryptophan 41 (in FR2-IMGT), leucine (or other hydrophobic amino acid) 89, and cysteine 104 (in FR3-IMGT). Tables and two-dimensional (2D) graphical representations designated as IMGT Colliers de Perles are available on the IMGT® Web site at http://imgt.cines.fr and in the works of M.-P. Lefranc and G. Lefranc [1, 2]. The IMGT Collier de Perles of a variable domain or V-DOMAIN of an IG light chain is shown, as an example, in Fig. 1.

This IMGT unique numbering has several advantages:

1. It has allowed the redefinition of the limits of the FR and CDR of the IG and TR variable domains. The FR-IMGT and CDR-IMGT lengths become in themselves crucial information, which characterize variable regions belonging to a group, a subgroup, and/or a gene.

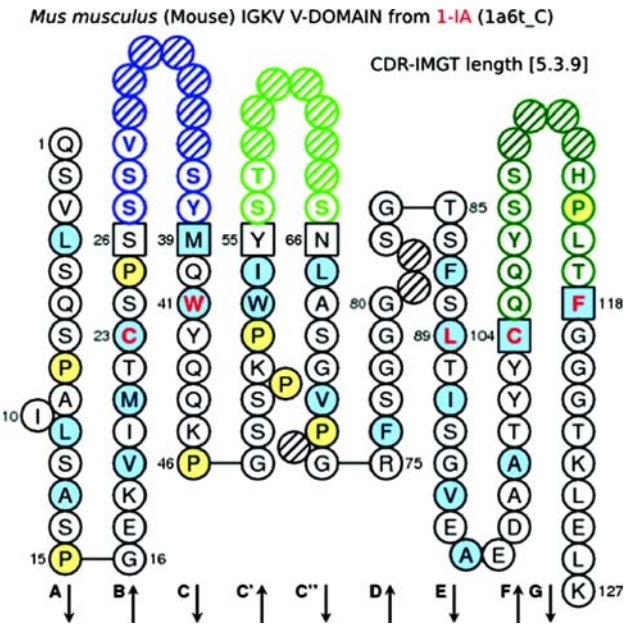


Fig. 1 IMGT Collier de Perles of a V-DOMAIN. The IMGT Collier de Perles of V-DOMAIN is based on the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN [10]. Amino acids are shown in the one-letter abbreviation. The CDR-IMGT are limited by amino acids shown in squares, which belong to the neighboring FR-IMGT. The CDR3-IMGT extends from position 105 to position 117. Hatched circles correspond to missing positions according to the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN [10]. *Arrows* indicate the direction of the nine beta strands that form the two beta sheets of the immunoglobulin (IG) fold [1, 2]

2. FR amino acids (and codons) located at the same position in different sequences can be compared without requiring sequence alignments. This also holds for amino acids belonging to CDR-IMGT of the same length.
3. The unique numbering is used as the output of the IMGT/V-QUEST alignment tool. The aligned sequences are displayed according to the IMGT unique numbering and with the FR-IMGT and CDR-IMGT delimitations.
4. The unique numbering has allowed a standardization of the description of mutations and the description of IG and TR allele polymorphisms [1, 2]. The mutations and allelic polymorphisms of each gene are described by comparison to the IMGT reference sequences of the allele *01 [1, 2].
5. The unique numbering allows the description and comparison of somatic hypermutations of the IG variable domains.

By facilitating the comparison between sequences and by allowing the description of alleles and mutations, the IMGT unique numbering represents a big step forward in the analysis of the IG and TR sequences of all vertebrate species. Moreover, it gives insight into the structural configuration of the domains and opens interesting views on the evolution of these sequences, as this numbering can be used for all sequences belonging to the V-set and C-set of the IgSF. Structural and functional domains of the IG and TR chains comprise the V-DOMAIN (9-strand beta-sandwich) (Fig. 2), which corresponds to the V-J-REGION or V-D-J-REGION and is encoded by two or three genes [1, 2], and the constant domain or C-DOMAIN (7-strand beta-sandwich) (Fig. 2). The IMGT unique numbering has been initially defined for the V-DOMAINs of the IG and TR and for the V-LIKE-DOMAINs of IgSF proteins other than IG and TR, for example in vertebrates human CD4 and *Xenopus* CTXg1 and in invertebrates *Drosophila* amalgam and *Drosophila* fasciclin II [10, 26]. It has been extended to the C-DOMAINs of the IG and TR and to the C-LIKE-DOMAINs of IgSF proteins other than IG and TR [11, 26, 27]. More recently, the IMGT unique numbering has also been defined for the groove domain or G-DOMAIN (four beta-strand and one alpha-helix) (Fig. 2) of the MHC classes I and II chains and for the G-LIKE-DOMAINs of MhcSF proteins other than MHC, for example MICA [12, 28].

ORIENTATION Axiom: Orientation of Instances Relative to Each Other

The ORIENTATION axiom and concepts allow to set up genomic orientation (for chromosome, locus, and gene) and DNA strand orientation. It is particularly useful in large genomic projects to localize a gene in a locus and/or a sequence (or a clone) in a contig or on a chromosome.

OBTENTION Axiom: Controlled Vocabulary for Biological Origin and Experimental Methodology

The OBTENTION axiom, and the generated concepts that are still in development, will be particularly useful for clinical data integration. This will help us to compare the repertoires of the IG antibody recognition sites and of the TR recognition sites in normal and pathological situations (autoimmune diseases, infectious diseases, leukaemias, lymphomas, and myelomas).

IMGT® Genomic, Genetic, and Structural Approaches

To extract knowledge from IMGT® standardized immunogenetics data, three main IMGT® biological approaches have been developed: genomic, genetic, and structural approaches (Table 2). The IMGT® genomic approach is gene-centred and mainly orientated towards the study of the genes within their loci and on the chromosomes. The IMGT® genetic approach refers to the study of the genes in relation to their sequence polymorphisms and mutations, their expression, their specificity, and their evolution. The genetics approach heavily relies on the DESCRIPTION axiom (and particularly on the V-, D-, J-, and C-REGION core concepts for the IG and TR), on the CLASSIFICATION axiom (IMGT® gene and allele names), and on the NUMEROTATION axiom [IMGT unique numbering [10–12]). The IMGT® structural approach refers to the study of the 2D and 3D structures of the IG, TR, MHC, and RPI and to the antigen- or ligand-binding characteristics in relationship with the protein functions, polymorphisms, and evolution. The structural approach relies on the CLASSIFICATION axiom (IMGT® gene and allele names), DESCRIPTION axiom (receptor and chain description and domain delimitations), and NUMEROTATION axiom (amino acid positions according to the IMGT unique numbering [10–12]).

For each approach, IMGT® provides databases [one genome database (IMGT/GENE-DB), three sequence databases (IMGT/LIGM-DB, IMGT/MHC-DB, and IMGT/PRIMER-DB), one 3D structure database (IMGT/3Dstructure-DB)], interactive tools (ten on-line tools for genome, sequence, and 3D structure analysis), and IMGT Repertoire Web resources (providing an easy-to-use interface to carefully and expertly annotated data on the genome, proteome, and polymorphism and structural data of the IG and TR, MHC and RPI) (Table 2). These databases, tools, and Web resources are detailed in the following sections. Other IMGT® Web resources include:

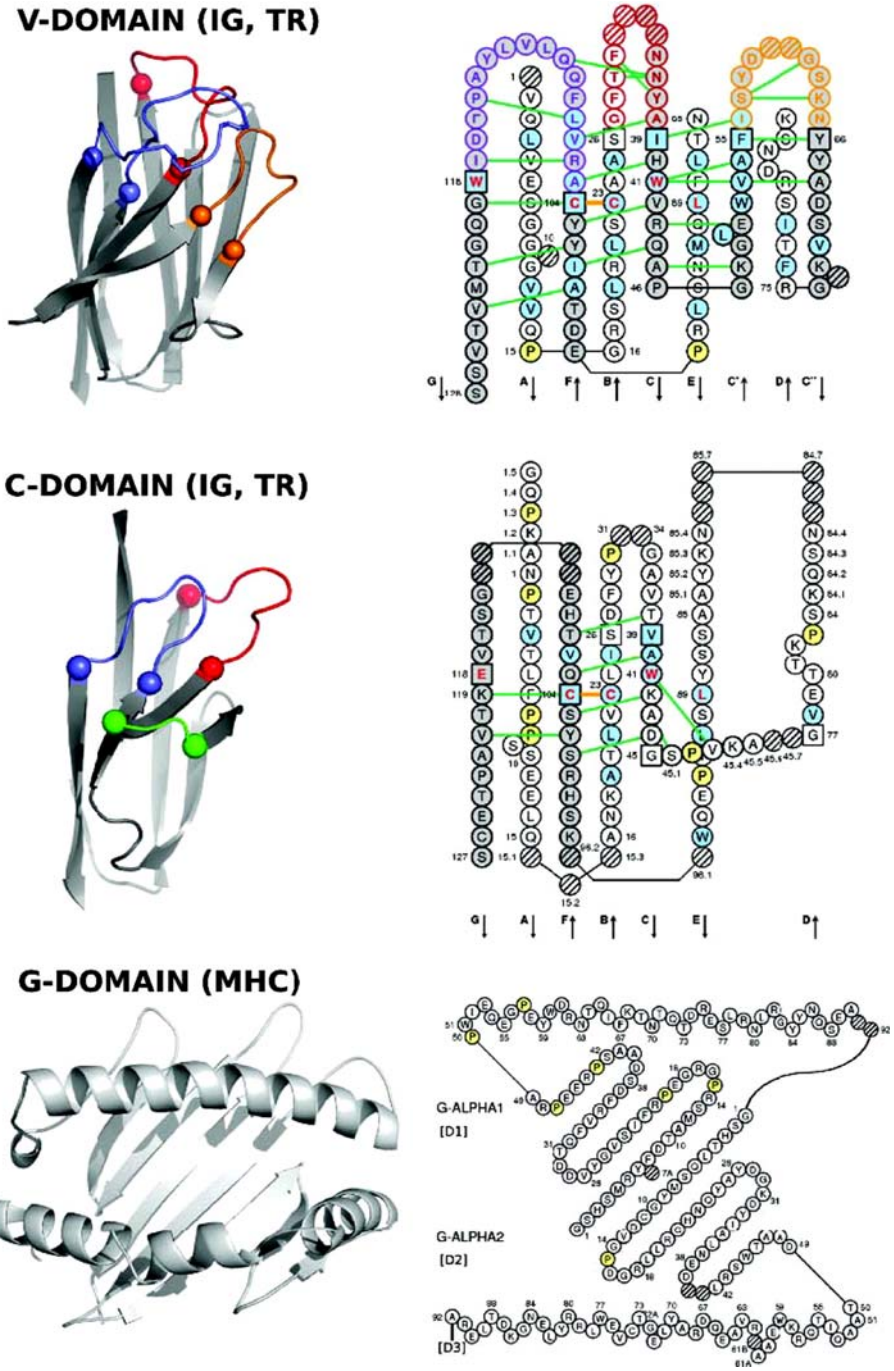


Fig. 2 Three-dimensional structures and IMGT Collier de Perles of a V-DOMAIN, a C-DOMAIN and G-DOMAINs. (**a**) V-DOMAIN. The IMGT Collier de Perles is based on the IMGT unique numbering for V-DOMAIN and V-LIKE-DOMAIN [10]. The V-DOMAIN chosen as an example is a human immunoglobulin (IG) variable heavy domain or VH (IMGT/3Dstructure-DB: 1aqk_H). *Arrows* indicate the direction of the nine beta strands of the V-DOMAIN that form the two beta sheets of the IG fold [1, 2]. (**b**) C-DOMAIN. The IMGT Collier de Perles is based on the IMGT unique numbering for C-DOMAIN and C-LIKE-DOMAIN [11]. The C-DOMAIN chosen as an example is a human IG constant light lambda domain or C-LAMBDA (IMGT/3Dstructure-DB: 1mcd_B). *Arrows* indicate the direction of the seven beta strands of the C-DOMAIN that form the two beta sheets of the IG fold [1, 2]. (**c**) G-DOMAINs. The IMGT Colliers de Perles are based on the IMGT unique numbering for G-DOMAIN and G-LIKE-DOMAIN [12]. The G-DOMAINs chosen as examples are human major histocompatibility complex (MHC) class I alpha groove domains or G-ALPHA1 and G-ALPHA2 (IMGT/3Dstructure-DB:1agb_A). Amino acids are shown in the one-letter abbreviation. *Hatched circles* correspond to missing positions according to the IMGT unique numbering [10–12]

1. IMGT Bloc-notes (Interesting links, etc.) provides numerous hyperlinks towards the Web servers specializing in immunology, genetics, molecular biology, and bioinformatics (associations, collections, companies, databases, immunology themes, journals, molecular biology servers, resources, societies, tools, etc.) [38].
2. IMGT Lexique.
3. The IMGT Immunoinformatics page.
4. The IMGT Medical page.
5. The IMGT Veterinary page.
6. The IMGT Biotechnology page.
7. IMGT Education (Aide-mémoire, Tutorials, Questions, answers, etc.) provides useful biological resources for students and includes figures and tutorials (in English and/or in French) in immunogenetics.
8. IMGT Aide-mémoire provides an easy access to information such as genetic code, splicing sites, amino acid structures, and restriction enzyme sites.
9. IMGT Index is a fast way to access data when information has to be retrieved from different parts of the IMGT site. For example, “allele” provides links to

Table 2 IMGT® databases, tools, and Web resources for genomic, genetic, and structural approaches

Approaches	Databases	Tools	Web resources[a]
Genomic	IMGT/GENE-DB [20]	IMGT/GeneView IMGT/LocusView IMGT/CloneSearch IMGT/GeneSearch IMGT/GeneInfo [29]	IMGT Repertoire "Locus and genes" section: • Chromosomal localizations [1, 2] • Locus representations [1, 2] • Locus description • Gene tables, etc. • Potential germline repertoires • Lists of genes • Correspondence between nomenclatures [1, 2]
Genetic	IMGT/LIGM-DB [30] IMGT/PRIMER-DB [31] IMGT/MHC-DB [32]	IMGT/V-QUEST [33] IMGT/JunctionAnalysis [34] IMGT/Allele-Align IMGT/PhyloGene [35]	IMGT Repertoire "Proteins and alleles" section: • Alignments of alleles • Protein displays • Tables of alleles, etc.
Structural	IMGT/3Dstructure-DB [36]	IMGT/StructuralQuery [36]	IMGT Repertoire "2D and 3D structures" section: • IMGT Colliers de Perles (2D representations on one layer or two layers) • IMGT® classes for amino acid characteristics [37] • IMGT Colliers de Perles reference profiles [37] • 3D representations

[a] Only Web resources examples from the IMGT Repertoire section are shown

the IMGT Scientific chart rules for the allele description and to the IMGT Repertoire "Alignments of alleles" and "Tables of alleles" (http://imgt.cines.fr).

IMGT® Databases, Tools, and Web Resources for Genomics

Genomic data are managed in IMGT/GENE-DB, which is the comprehensive IMGT® genome database [20]. In February 2007, IMGT/GENE-DB contained 1,512 IG and TR genes and 2,461 alleles from human and mouse IG and TR genes. Based on the IMGT® CLASSIFICATION axiom, all the human IMGT® gene names [1, 2], approved by the HUGO Nomenclature Committee HGNC in 1999, are available in IMGT/GENE-DB [20] and in Entrez Gene at NCBI (USA) [39]. All the mouse IMGT® gene and allele names and the corresponding IMGT reference sequences were provided to Mouse Genome Informatics MGI Mouse Genome Database MGD in July 2002 and were presented by IMGT® at the 19th International Mouse Genome Conference IMGC 2005, in Strasbourg, France and entered in IMGT/GENE-DB [20]. IMGT-GENE-DB allows a query per gene and allele name. IMGT/GENE-DB interacts dynamically with IMGT/LIGM-DB [30] to download and display human and mouse gene-related sequence data. This is the first example of an interaction between IMGT® databases using the CLASSIFICATION axiom.

The IMGT® genome analysis tools manage the locus organization and gene location and provide the display of physical maps for the human and mouse IG, TR, and MHC loci. They allow to view genes in a locus (IMGT/GeneView and IMGT/LocusView) to search for clones (IMGT/CloneSearch), to search for genes in a locus (IMGT/GeneSearch and IMGT/GeneInfo) based on IMGT® gene names, functionality or localization on the chromosome, to provide information on the clones that were used to build the locus contigs (accession numbers are from IMGT/LIGM-DB and gene names from IMGT/GENE-DB) or to display information on the human and mouse IG and TR potential rearrangements.

The IMGT Repertoire genome data include chromosomal localizations, locus representations, locus description, germline gene tables, potential germline repertoires, lists of IG and TR genes and links between IMGT®, HGNC, GDB, Entrez Gene, and OMIM, and correspondence between nomenclatures [1, 2].

IMGT® Databases, Tools, and Web Resources for Genetics

IMGT/LIGM-DB [30] is the comprehensive IMGT® database of IG and TR nucleotide sequences from human and other vertebrate species, with translation for fully annotated sequences, created in 1989 by LIGM, Montpellier, France,

on the Web since July 1995. IMGT/LIGM-DB is the first and the largest IMGT® database. In April 2008, IMGT/LIGM-DB contained 122,425 nucleotide sequences of IG and TR from 222 species. The unique source of data for IMGT/LIGM-DB is EMBL that shares data with the other two generalist databases GenBank and DDBJ. IMGT/LIGM-DB sequence data are identified by the EMBL/GenBank/DDBJ accession number. Based on expert analysis, specific detailed annotations are added to IMGT flat files.

Since August 1996, the IMGT/LIGM-DB content closely follows the EMBL one for the IG and TR, with the following advantages: IMGT/LIGM-DB does not contain sequences that have previously been wrongly assigned to IG and TR; conversely, IMGT/LIGM-DB contains IG and TR entries that have disappeared from the generalist databases [for example, the L36092 accession number that encompasses the complete human TRB locus is still present in IMGT/LIGM-DB, whereas it has been deleted from EMBL/GenBank/DDBJ due to its very large size (684,973 bp); in 1999, IMGT/LIGM-DB detected the disappearance of 20 IG and TR sequences that inadvertently had been lost by GenBank, and allowed the recuperation of these sequences in the generalist databases].

The IMGT/LIGM-DB annotations (gene and allele name assignment, labels) allow data retrieval not only from IMGT/LIGM-DB, but also from other IMGT® databases. For example, the IMGT/GENE-DB entries provide the IMGT/LIGM-DB accession numbers of the IG and TR cDNA sequences that contain a given V, D, J, or C gene. The automatic annotation of rearranged human and mouse cDNA sequences in IMGT/LIGM-DB is performed by IMGT/Automat [40], an internal Java tool that implements IMGT/V-QUEST and IMGT/JunctionAnalysis.

Standardized information on oligonucleotides (or primers) and combinations of primers (Sets and Couples) for IG and TR are managed in IMGT/PRIMER-DB [31], the IMGT® oligonucleotide database on the Web since February 2002. IMGT/MHC-DB [32] hosted at EBI comprises IMGT/HLA for human MHC (or HLA) and IMGT/MHC-NHP for MHC of non-human primates.

The IMGT® tools for the genetics approach comprise IMGT/V-QUEST [33, 41] for the identification of the V, D, and J genes and of their mutations, IMGT/JunctionAnalysis [34, 41] for the analysis of the V-J and V-D-J junctions that confer the antigen receptor specificity, IMGT/Allele-Align for the detection of polymorphisms, and IMGT/Phylogene [35] for gene evolution analyses. IMGT/V-QUEST (V-QUEry and STandardization) (http://imgt.cines.fr) is an integrated software for IG and TR [33, 41]. This tool, which is easy to use, analyses an input IG or TR germline or rearranged variable nucleotide sequence. IMGT/V-QUEST results comprise the identification of the V, D, and J genes and alleles and the nucleotide alignment by comparison with sequences from the IMGT reference directory, the delimitations of the FR-IMGT and CDR-IMGT based on the IMGT unique numbering, the protein translation of the input sequence, the identification of the JUNCTION, the description of the mutations and amino acid changes of the V-REGION, and the 2D IMGT Collier de Perles representation of the V-REGION or V-DOMAIN. The set of sequences from the IMGT reference directory, used for IMGT/V-QUEST, can be downloaded in FASTA format from the IMGT® site.

IMGT/JunctionAnalysis [34, 41] is a tool developed by LIGM, complementary to IMGT/V-QUEST, which provides a thorough analysis of the V-J and V-D-J junction of IG and TR rearranged genes. The JUNCTION extends from 2nd-CYS 104 to J-PHE or J-TRP 118 inclusive. J-PHE or J-TRP is easily identified for in-frame rearranged sequences when the conserved Phe/Trp-GlyX-Gly motif of the J-REGION is present. The length of the CDR3-IMGT of rearranged V-J-GENEs or V-D-J-GENEs is a crucial piece of information. It is the number of amino acids or codons from position 105–117 (J-PHE or J-TRP non-inclusive). CDR3-IMGT amino acid and codon numbers are according to the IMGT unique numbering for V-DOMAIN [10]. IMGT/JunctionAnalysis identifies the D-GENE and allele involved in the IGH, TRB, and TRD V-D-J rearrangements by comparison with the IMGT reference directory and delimits precisely the P, N, and D regions [1, 2]. Results from IMGT/JunctionAnalysis are more accurate than those given by IMGT/V-QUEST regarding the D-GENE identification. Indeed, IMGT/JunctionAnalysis works on shorter sequences (JUNCTION) and with a higher constraint because the identification of the V-GENE and J-GENE and alleles is a prerequisite to perform the analysis. Several hundreds of junction sequences can be analysed simultaneously.

Other IMGT® Tools for sequence analysis comprise IMGT/Allele-Align that allows the comparison of two alleles highlighting the nucleotide and amino acid differences and IMGT/PhyloGene [35], an easy-to-use tool for phylogenetic analysis of IMGT standardized reference sequences.

The IMGT Repertoire polymorphism data are represented by "Alignments of alleles," "Tables of alleles," "Allotypes," "Protein displays," particularities in protein designations, IMGT reference directory in FASTA format, correspondence between IG and TR chain, and receptor IMGT designations [1, 2].

IMGT® Databases, Tools, and Web Resources for Structural Analysis

Structural data are compiled and annotated in IMGT/3Dstructure-DB [36], the IMGT® 3D structure database,

created by LIGM, on the Web since November 2001. IMGT/3Dstructure-DB comprises IG, TR, MHC, and RPI with known 3D structures. In April 2008, IMGT/3Dstructure-DB contained 1,423 atomic coordinate files. These coordinate files, extracted from the Protein Data Bank (PDB) [42], are renumbered according to the standardized IMGT unique numbering [10–12]. The IMGT/3Dstructure-DB cards provide IMGT® annotations (assignment of IMGT® genes and alleles, IMGT® chain and domain labels, and IMGT Colliers de Perles on one layer and two layers), downloadable renumbered IMGT/3Dstructure-DB flat files, visualization tools, and external links. IMGT/3Dstructure-DB residue cards provide detailed information on the inter-and intra-domain contacts of each residue position (Fig. 3).

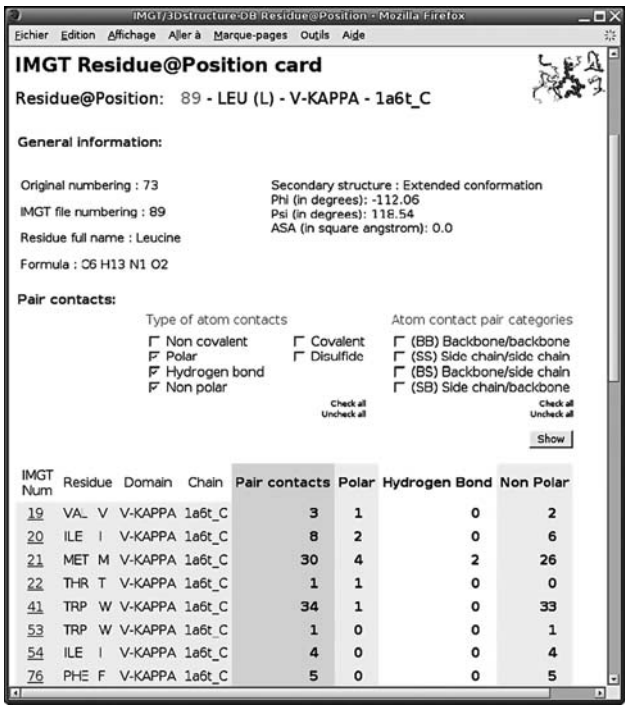


Fig. 3 IMGT Residue@Position card. The identification of a "IMGT Residue@ Position" comprises the position number according to the IMGT unique numbering [10–12], the residue name (with three letters and eventually one letter abbreviation), the domain description, and the IMGT/3Dstructure-DB chain ID. The example shows the contacts of position 89, occupied by a leucine LEU (L), in the V-KAPPA domain of the 1a6t_C chain. The original number in the PDB file is indicated. The secondary structure, the phi and psi angles (in *degrees*) and accessible surface area (ASA) (in *square angstroms*), are provided. The user can select, for the result display, the types of contacts (non-covalent, polar, hydrogen bond, non-polar, covalent bond, or disulfide bond) and the atom contact pair categories (backbone/backbone, side chain/side chain, backbone/side chain, and side chain/backbone atoms). The results are shown as a table with a list of the IMGT Residue@Position which are in contact with the IMGT Residue@Position at the top of the card, and for each of them, the total number of atom pair contacts and the detailed description of the contacts as selected by the user are also indicated

The IMGT/StructuralQuery tool [36] analyses the intra-molecular interactions for the V-DOMAINs. The contacts are described per domain (intra-and inter-domain contacts) and annotated in terms of IMGT® labels (chains and domain), positions (IMGT unique numbering), backbone or side-chain implication. IMGT/StructuralQuery allows to retrieve the IMGT/3Dstructure-DB entries, based on specific structural characteristics: phi and psi angles, accessible surface area (ASA), amino acid type, distance in angstrom between amino acids, and CDR-IMGT lengths [36].

To appropriately analyse the amino acid resemblances and differences between IG, TR, MHC, and RPI chains, 11 IMGT® classes were defined for the amino acid "chemical characteristics" properties and used to set up IMGT Colliers de Perles reference profiles [37]. The IMGT Colliers de Perles reference profiles allow to easily compare amino acid properties at each position whatever the domain, the chain, the receptor, or the species [37]. The IG and TR variable and constant domains and the MHC groove domains represent a privileged situation for the analysis of amino acid properties in relation with 3D structures, by the conservation of their 3D structure despite divergent amino acid sequences and by the considerable amount of genomic (IMGT Repertoire), structural (IMGT/3Dstructure-DB), and functional data available. These data are not only useful to study mutations and allele polymorphisms but are also needed to establish correlations between amino acids in the protein sequences and 3D structures, to analyse the IgSF and MhcSF domain interactions [43], and to determine amino acids potentially involved in the immunogenicity. One of the key elements in the adaptive immune response is the presentation of peptides by the MHC to the TR at the surface of T cells. The characterization of the TR/peptide/MHC trimolecular complexes (TR/pMHC) is crucial to the fields of immunology, vaccination, and immunotherapy. In IMGT/3Dstructure-DB, TR/pMHC molecular characterization and pMHC contact analysis have been standardized, based on the IMGT unique numbering for G-DOMAIN, and 11 IMGT pMHC contact sites (C1–C11) have been defined [44]. The IMGT pMHC contact sites represent the MHC amino acid positions that have contacts with the peptide side chains. They are particularly useful to compare pMHC interactions whatever the MHC classes or chains, whatever the species and whatever the peptide sequence or length [44]. There are no C2, C7, and C8 contact sites for MHC-I with 8-amino acid peptides and no C2 and C7 for MHC-I 3D structures with 9-amino acid peptides. In contrast, for MHC-II, C2 is present but there are no C7 and C8 [44]. The IMGT pMHC contact sites are provided dynamically for the pMHC and the TR/pMHC 3D structures available in IMGT/3Dstructure-DB. For example, the IMGT pMHC contact sites of a MHC-I (human HLA-A*0201) and a

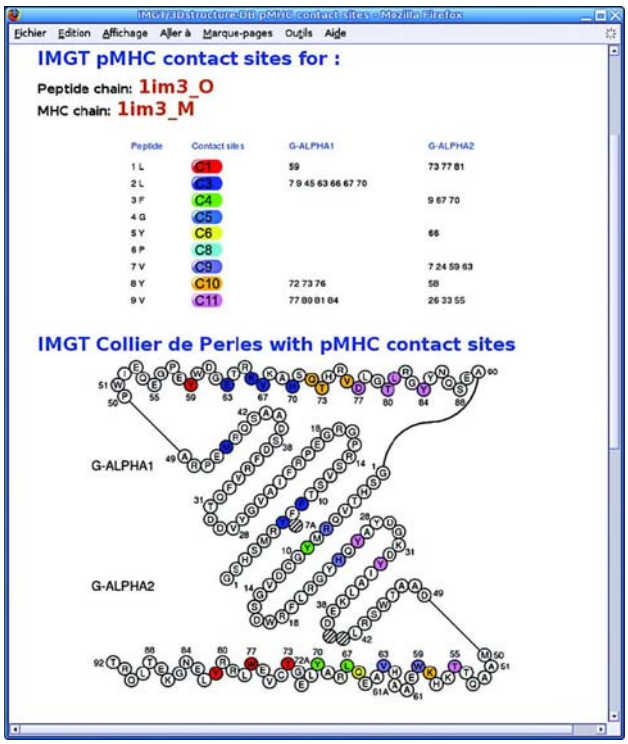


Fig. 4 IMGT peptide major histocompatibility complex (pMHC) contact sites of human HLA-A*0201 MHC-I and a 9-amino acid peptide side chains (IMGT/3Dstructure-*DB*: 1im3). The numbers 1–9 refer to the numbering of the peptide amino acids P1–P9. C1–C11 refer to the 11 pMHC contact sites defined by IMGT® [44]. There are no C2 and C7 in MHC-I 3D structures with 9-amino acid peptides. There are no C5 and C8 in this 3D structure as P4 and P6 do not contact MHC amino acids. The view of the IMGT Collier de Perles is from above the cleft, with G-ALPHA1 on top and G-ALPHA2 on bottom of the figure

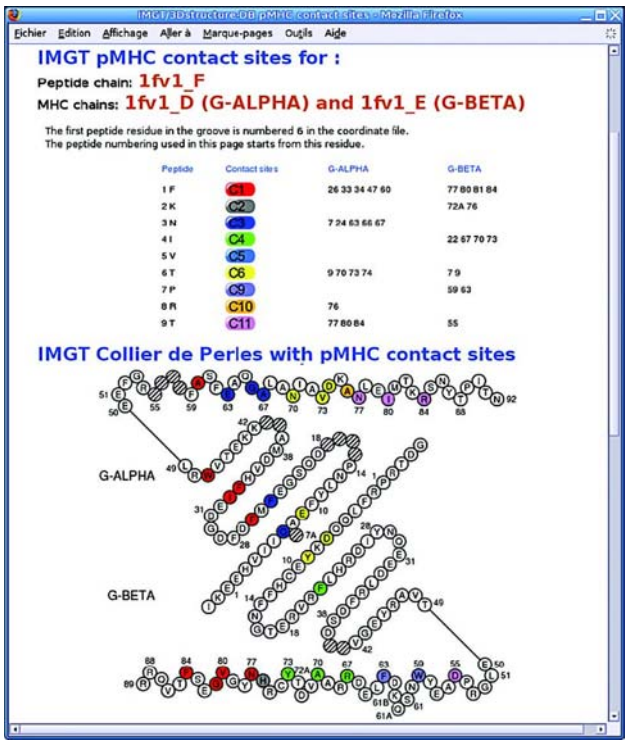


Fig. 5 IMGT peptide major histocompatibility complex (pMHC) contact sites of human HLA-DRA*0101 and HLA-DRB5*0101 MHC-II and the peptide side chains (9 amino acids located in the groove) (IMGT/3Dstructure-DB: 1fv1). The numbers 1–9 refer to the numbering of the peptide amino acids 1–9 located in the groove. C1–C11 refer to the 11 pMHC contact sites defined by IMGT® [44]. There are no C7 and C8 in MHC-II 3D structures with peptide of 9 amino acids located in the groove. There is no C5 in this 3D structure as 5 does not contact MHC amino acids. The view of the IMGT Collier de Perles is from above the cleft, with G-ALPHA on top and G-BETA on bottom of the figure

9-amino acid peptide side chain are shown in Fig. 4 (IMGT/3Dstructure-DB: 1im3), and the IMGT pMHC contact sites of a MHC-II (human HLA-DRA*0101 and HLA-DRB5*0101) binding nine amino acids of the peptide in the groove are shown in Fig. 5 (IMGT/3Dstructure-DB: 1fv1).

The IMGT Repertoire Structural data comprise IMGT Colliers de Perles [1, 2, 10–12], FR-IMGT and CDR-IMGT lengths, and 3D representations of IG and TR variable domains. This visualization permits rapid correlation between protein sequences and 3D data retrieved from the PDB.

Conclusion

Since July 1995, IMGT® has been available on the Web at http://imgt.cines.fr. IMGT® has an exceptional response with more than 150,000 requests a month. The information is of much value to clinicians and biological scientists in general. IMGT® databases, tools, and Web resources are extensively queried and used by scientists from both academic and industrial laboratories, who are equally distributed between the United States, Europe, and the remaining world. IMGT® is used in very diverse domains: (i) fundamental and medical research (repertoire analysis of the IG antibody recognition sites and of the TR recognition sites in normal and pathological situations such as autoimmune diseases, infectious diseases, AIDS, leukemias, lymphomas, and myelomas), (ii) veterinary research (IG and TR repertoires in farm and wildlife species), (iii) genome diversity and genome evolution studies of the adaptive immune responses, (iv) structural evolution of the IgSF and MhcSF proteins, (v) biotechnology related to antibody engineering [single chain Fragment variable (scFv), phage displays, combinatorial libraries, chimeric, humanized, and human antibodies], (vi) diagnostics (clonalities, detection, and follow-up of residual diseases), and (vii) therapeutical approaches (grafts, immunotherapy, and vaccinology). The creation of dynamic interactions between the IMGT® databases and tools, using Web services and IMGT-ML, and the design of IMGT-Choreography [4], represents novel and

major developments of IMGT®, the international reference in immunogenetics and immunoinformatics. The IMGT-ONTOLOGY axioms constitute the Formal IMGT-ONTOLOGY, also designated as IMGT-Kaleidoscope [45]. IMGT-ONTOLOGY represents a key component in the elaboration of Formal Ontologies in Life Sciences, and in the setting of standards of the European ImmunoGrid project (http://www.immunogrid.org) whose aim is to define the essential concepts for modelling of the immune system.

Citing IMGT®

Authors who make use of the information provided by IMGT® should cite [3] as a general reference for the access to and content of IMGT® and quote the IMGT® home page URL, http://imgt.cines.fr.

Acknowledgments I thank Véronique Giudicelli, Patrice Duroux, Quentin Kaas, Joumana Jabado-Michaloud, Géraldine Folch, Chantal Ginestoux, Denys Chaume, and Gérard Lefranc for helpful discussions. I am deeply grateful to the IMGT® team for its expertise and constant motivation. IMGT® is a registered mark of the Centre National de la Recherche Scientifique (CNRS). IMGT® has received the National Bioinformatics Platform RIO label since 2001 (CNRS, INSERM, CEA, and INRA). IMGT® was funded in part by the BIOMED1 (BIO-CT930038), Biotechnology BIOTECH2 (BIO4CT960037) and 5th PCRDT Quality of Life and Management of Living Resources (QLG2-200001287) programmes of the European Union (EU). IMGT® is currently supported by the CNRS, the Ministère de l'Education Nationale, de l'Enseignement Supérieur et de la Recherche (MENESR) (Université Montpellier 2 Plan Pluri-Formation, Institut Universitaire de France), Réseau National des Génopoles, the Région Languedoc-Roussillon, the Agence Nationale de la Recherche ANR (BIO-SYS06_135457, FLAVORES), and the EU ImmunoGrid (IST-028069).

References

1. Lefranc, M.-P., & Lefranc, G. (2001). *The Immunoglobulin FactsBook*. London, UK: Academic Press, 458 p. ISBN: 012441351X.
2. Lefranc, M.-P., & Lefranc, G. (2001). *The T cell Receptor FactsBook*. London, UK: Academic Press, 398 p. ISBN: 0124413528.
3. Lefranc, M.-P., Giudicelli, V., Kaas, Q., Duprat, E., Jabado-Michaloud, J., Scaviner, D., Ginestoux, C., Clément, O., Chaume, D., & Lefranc G. (2005). IMGT, the International ImMunoGeneTics information system. *Nucleic Acids Research, 33*, D593–D597.
4. Lefranc, M.-P., Clément, O., Kaas, Q., Duprat, E., Chastellan, P., Coelho, I., Combres, K., Ginestoux, C., Giudicelli, V., Chaume, D., & Lefranc, G. (2005). IMGT-Choreography for immunogenetics and immunoinformatics. Epub *In Silico Biology* 5 0006, http://www.bioinfo.de/isb/2004/05/0006/24 December 2004. *In Silico Biology, 5*, 45–60.
5. Giudicelli, V., & Lefranc, M.-P. (1999). Ontology for immunogenetics: The IMGTONTOLOGY. *Bioinformatics, 12*, 1047–1054.
6. Chaume, D., Giudicelli, V., & Lefranc, M.-P. (2001). IMGT-ML a language for IMGT-ONTOLOGY and IMGT/LIGM-DB data. In: CORBA and XML: *Towards a Bioinformatics Integrated Network Environment, Proceedings of NETTAB 2001, Network tools and Applications in Biology*, May 17–18, Gchoa, Italy, pp. 71–75.
7. Chaume, D., Giudicelli, V., Combres, K., & Lefranc, M.-P. (2003) IMGTONTOLOGY and IMGT-ML for Immunogenetics and immunoinformatics. In: *Abstract book of the Sequence Databases and Ontologies Satellite Event*, European Congress in Computational Biology ECCB'2003, September 27–30, Paris, France, pp. 22–23.
8. Letovsky, S. I., Cottingham, R. W., Porter, C. J., & Li, P. W. (1998). GDB: The human genome database. *Nucleic Acids Research, 26*, 94–99.
9. Wain, H. M., Bruford, E. A., Lovering, R. C., Lush, M. J., Wright, M. W., & Povey, S. (2002). Guidelines for human gene nomenclature. *Genomics, 79*, 464–470.
10. Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., & Lefranc, G. (2003). IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains. *Development and Comparative Immunology, 27*, 55–77.
11. Lefranc, M.-P., Pommié, C., Kaas, Q., Duprat, E., Bosc, N., Guiraudou, D., Jean C., Ruiz, M., Da Piedade, I., Rouard, M., Foulquier, E., Thouvenin, V., & Lefranc, G. (2005). IMGT unique numbering for immunoglobulin and T cell receptor constant domains and Ig superfamily C-like domains. *Development and Comparative Immunology, 29*, 185–203.
12. Lefranc, M.-P., Duprat, E., Kaas, Q., Tranne, M., Thiriot, A., & Lefranc, G. (2005). IMGT unique numbering for MHC groove G-DOMAIN and MHC superfamily (MhcSF) G-LIKE-DOMAIN. *Development and Comparative Immunology, 29*, 917–938.
13. Ruiz, M., & Lefranc, M.-P. (2002). IMGT gene identification and Colliers de Perles of human immunoglobulins with known 3D structures. *Immunogenetics, 53*, 857–883.
14. Cochrane, G., Aldebert, P., Althorpe, N., Andersson, M., Baker, W., Baldwin, A., Bates, K., Bhattacharyya, S., Browne, P., van den Broek, A., Castro, M., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Kanz, C., Kulikova, T., Lee, C., Leinonen, R., Lin, Q., Lombard, V., Lopez, R., McHale, M., McWilliam, H., Mukherjee, G., Nardone, F., Garcia Pastor, M. P., Sobhany, S., Stoehr, P., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W., & Apweiler, R. (2006). EMBL nucleotide sequence database: developments in 2005. *Nucleic Acids Research, 34*, D10–D15.
15. Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Wheeler, D. L. (2006). GenBank. *Nucleic Acids Research, 34*, D16–D20.
16. Okubo, K., Sugawara, H., Gojobori, T., & Tateno, Y. (2006). DDBJ in preparation for overview of research activities behind data submissions. *Nucleic Acids Research, 34*, D6–D9.
17. Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., & Ashburner, M. (2005). The sequence ontology: A tool for the unification of genome annotations. *Genome Biology, 6*(5), R44. Epub 29 Apr 2005.
18. Lefranc, M.-P. (2000). Nomenclature of the human immunoglobulin genes. In J. E. Coligan, B. E. Bierer, D. E. Margulies, E. M. Shevach, & W. Strober (Eds.), *Current protocols in immunology* (pp. A.1P.1–A.1P.37). Hoboken, NJ: Wiley and Sons.
19. Lefranc, M.-P. (2000). Nomenclature of the human T cell receptor genes. In J. E. Coligan, B. E. Bierer, D. E. Margulies, E. M. Shevach, & W. Strober (Eds.), *Current protocols in immunology* (pp. A.1O.1–A.1O.23). Hoboken, N.J.: Wiley and Sons.
20. Giudicelli, V., Chaume, D., & Lefranc, M.-P. (2005). IMGT/GENE-DB: A comprehensive database for human and mouse immunoglobulin and T cell receptor genes. *Nucleic Acids Research, 33*, D256–D261.
21. Lefranc, M.-P. (1997). Unique database numbering system for immunogenetic analysis. *Immunology Today, 18*, 509.

22. Lefranc, M.-P. (1999). The IMGT unique numbering for immunoglobulins, T cell receptors and Ig-like domains. *The Immunologist, 7*, 132–136.
23. Kabat, E. A., Wu, T. T., Perry, H. M., Gottesman, K. S., & Foeller, C. (1991). *Sequences of proteins of immunological interest*. Washington, DC, USA: National Institute of Health Publications Publication no. 91-3242.
24. Satow, Y., Cohen, G. H., Padlan, E. A., & Davies, D. R. (1986). Phosphocholine binding immunoglobulin Fab McPC603. *Journal of Molecular Biology, 190*, 593–604.
25. Chothia, C., & Lesk, A. M. (1987). Canonical structures for the hypervariable regions of immunoglobulins. *Journal of Molecular Biology, 196*, 901–917.
26. Duprat, E., Kaas, Q., Garelle, V., Lefranc, G., & Lefranc, M.-P. (2004). IMGT standardization for alleles and mutations of the V-LIKE-DOMAINs and C-LIKE-DOMAINs of the immunoglobulin superfamily. In: Pandalai, S. G. (Ed.), *Recent research developments in human genetics* (Vol. 2, pp. 111–136). Research Signpost: Trivandrum, Kerala, India, .
27. Bertrand, G., Duprat, E., Lefranc, M.-P., Marti, J., & Coste, J. (2004). Characterization of human FCGR3B*02 (HNA-1b, NA2) cDNAs and IMGT standardized description of FCGR3B alleles. *Tissue Antigens, 64*, 119–131.
28. Frigoul, A., & Lefranc, M.-P. (2005) MICA: Standardized IMGT allele nomenclature, polymorphisms and diseases. In Pandalai, S. G., (Ed.), *Recent research developments in human genetics* (Vol. 3, pp. 95–145). Research Signpost: Trivandrum, Kerala, India.
29. Baum, T. P., Pasqual, N., Thuderoz, F., Hierle, V., Chaume, D., Lefranc, M.-P., Jouvin-Marche, E., Marche, P. N., & Demongeot, J. (2004). IMGT/GeneInfo: Enhancing V(D)J recombination database accessibility. *Nucleic Acids Research, 32*, D51–D54.
30. Giudicelli, V., Duroux, P., Ginestoux, C., Folch, G., Jabado-Michaloud, J., Chaume, D., & Lefranc, M.-P. (2006). IMGT/LIGM-DB, the IMGT® comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Research, 34*, D781–D784.
31. Folch, G., Bertrand, J., Lemaitre, M., & Lefranc, M.-P. (2004). IMGT/PRIMER-DB. In M. Y. Galperin (Ed.), *Database listing*. The Molecular Biology Database Collection: 2004 update. *Nucleic Acids Research, 32*, D3–D22.
32. Robinson, J., Waller, M. J., Parham, P., de Groot, N., Bontrop, R., Kennedy, L. J., Stoehr, P., & Marsh, S. G. (2003). IMGT/HLA and IMGT/MHC sequence databases for the study of the major histocompatibility complex. *Nucleic Acids Research, 31*, 311–314.
33. Giudicelli, V., Chaume, D., & Lefranc, M.-P. (2004). IMGT/V-QUEST, an integrated software program for immunoglobulin and T cell receptor V-J and V-D-J rearrangement analysis. *Nucleic Acids Research, 32*, W435–W440.
34. Yousfi Monod, M., Giudicelli, V., Chaume, D., & Lefranc, M.-P. (2004). IMGT/JunctionAnalysis: The first tool for the analysis of the immunoglobulin and T cell receptor complex V-J and V-D-J JUNCTIONs. *Bioinformatics, 20*, i379–i385.
35. Elemento, O., & Lefranc, M.-P. (2003). IMGT/PhyloGene: An on-line tool for comparative analysis of immunoglobulin and T cell receptor genes. *Development and Comparative Immunology, 27*, 763–779.
36. Kaas, Q., Ruiz, M., & Lefranc, M.-P. (2004). IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Research, 32*, D208–D210.
37. Pommié, C., Sabatier, S., Lefranc, G., & Lefranc, M.-P. (2004). IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *Journal of Molecular Recognition, 17*, 17–32.
38. Lefranc, M.-P. (2006). Web sites of interest to immunologists. In J. E. Coligan, B. E. Bierer, D. E. Margulies, E. M. Shevach, & W. Strober (Eds.), *Current protocols in immunology* (pp. A.1J.1–A.1J.74). Hoboken, NJ: Wiley and Sons.
39. Maglott, D., Ostell, J., Pruitt, K. D., & Tatusova, T. (2007). Entrez Gene: Gene-centered information at NCBI. *Nucleic Acids Research, 35*, D26–D31.
40. Giudicelli, V., Chaume, D., Jabado-Michaloud, J., & Lefranc, M.-P. (2005). Immunogenetics sequence annotation: The strategy of IMGT based on IMGTONTOLOGY. *Studies in Health Technology and Informatics, 116*, 3–8.
41. Lefranc, M.-P. (2004). IMGT, The International ImMunoGeneTics Information System®, http://imgt.cines.fr. In B. K. C. Lo (Ed.), *Antibody engineering: Methods and protocols*. Totowa, NJ: Humana. *Methods in Molecular Biology, 248*, 27–49.
42. Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*, 235–242.
43. Duprat, E., Lefranc, M.-P., & Gascuel, O. (2006). A simple method to predict protein binding from aligned sequences—Application to MHC superfamily and beta2-microglobulin. *Bioinformatics, 22*, 453–459.
44. Kaas, Q., & Lefranc, M.-P. (2005). T cell receptor/peptide/MHC molecular characterization and standardized pMHC contact sites in IMGT/3Dstructure-DB. Epub *In Silico Biology* 5 0046, 20 October 2005. *In Silico Biol. 5*, 505–528.
45. Duroux, P., Kaas, Q., Brochet, X., Lane, J., Ginestoux, C., Lefranc, M.-P., & Giudicelli, V. (2008). IMGT-Kaleidoscope, the formal IMGT-ONTOLOGY paradigm. *Biochimie, 90*, 570–583.

Published online 09 October 2014 *Nucleic Acids Research, 2015, Vol. 43, Database issue* ***D405–D412***
doi: 10.1093/nar/gku938

The immune epitope database (IEDB) 3.0

Randi Vita[1,*], James A. Overton[1], Jason A. Greenbaum[2], Julia Ponomarenko[3], Jason D. Clark[4], Jason R. Cantrell[4], Daniel K. Wheeler[4], Joseph L. Gabbard[5], Deborah Hix[5], Alessandro Sette[1] and Bjoern Peters[1]

[1]Division of Vaccine Discovery, La Jolla Institute for Allergy and Immunology, La Jolla, 9420 Athena Circle, CA 92037, USA, [2]Bioinformatics Core, La Jolla Institute for Allergy and Immunology, La Jolla, CA 92037, USA, [3]San Diego Supercomputer Center, University of California, San Diego, CA 92093, USA, [4]Leidos Health, LLC, San Diego, CA 92121, USA and [5]Grado Department of Industrial and Systems Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA

Received August 12, 2014; Revised September 24, 2014; Accepted September 25, 2014

ABSTRACT

The IEDB, www.iedb.org, contains information on immune epitopes—the molecular targets of adaptive immune responses—curated from the published literature and submitted by National Institutes of Health funded epitope discovery efforts. From 2004 to 2012 the IEDB curation of journal articles published since 1960 has caught up to the present day, with >95% of relevant published literature manually curated amounting to more than 15 000 journal articles and more than 704 000 experiments to date. The revised curation target since 2012 has been to make recent research findings quickly available in the IEDB and thereby ensure that it continues to be an up-to-date resource. Having gathered a comprehensive dataset in the IEDB, a complete redesign of the query and reporting interface has been performed in the IEDB 3.0 release to improve how end users can access this information in an intuitive and biologically accurate manner. We here present this most recent release of the IEDB and describe the user testing procedures as well as the use of external ontologies that have enabled it.

INTRODUCTION

The IEDB was established in 2004, and over the past 10 years our team has manually curated almost 16 000 published manuscripts and processed 200 direct submissions. As a result, detailed experimental data regarding more than 120 000 epitopes are now freely and easily accessible to the scientific community via most web browsers as a web-based interface. In addition, if one wishes to view 3D structural data using the Epitope Viewer application, Java 6 or 7 is required. The IEDB's primary curation focus is on data from scientific publications available in PubMed (1) focused on infectious diseases, allergy, autoimmunity and transplantation. Excluded from the primary scope are HIV-derived epitopes captured in the LANL database (www.hiv.lanl.gov/content/immunology) and cancer epitopes for which there is no resource currently available due to lack of support for such a resource by the National Institutes of Health. As an exception, all publications describing the 3D structure of an epitope in complex with its adaptive immune receptor or major histocompatibility complex (MHC) molecule are included regardless of origin of the epitope in order to provide a complete dataset of this particularly valuable type of information. Details describing the curation process put in place and followed by the curation team, including quality controls for accuracy and consistency, have been discussed previously (2).

The IEDB houses epitope-specific experimental assays. That is, every assay reflects the binding of an epitope-specific T cell receptor (TCR), antibody or MHC molecule to an experimentally tested antigen or epitope. The structure entered as the epitope is limited to the exact entity that was actually tested in the assay or was clearly deduced to be the epitope by the authors. In many cases this is not the minimal epitope and may not be limited to the contact residues of the epitope, but is rather a region containing the epitope. The fields of the IEDB describe the details of these experiments in great detail. First, the epitope structure is designated as either peptidic or non-peptidic. Peptidic epitopes are described by their linear amino acid sequence or as discontinuous amino acids by position within their source protein. Peptidic epitopes having 3D structural data are described by the residues found to contact the antibody, TCR or MHC molecule. Non-peptidic epitopes are manually curated by staff from the ChEBI team (3) who annotate the complete molecular structures using SMILES annotation. If the epitope was derived from a protein or a larger non-peptidic structure, these are also pro-

*To whom correspondence should be addressed. Tel: +1 858 752 6912; Fax: +1 858 752 6987; Email: rvita@liai.org

vided along with the organism in which these structures are found. For example, the linear epitope FEIKCTKPEACS is derived from the *Phleum pratense* (Timothy grass) protein Phl p 1. All experimental assays that characterize the epitope or its recognition by immune receptors are entered into the IEDB, including all negative data. For example, the FEIKCTKPEACS epitope has 15 assays curated from three different references. Details on the gender and age of the host who made the immune response and on the process that led to it (e.g. immunization, infection or other exposure) are also captured. Important aspects of antibodies are presented such as isotype, antibody name, clonality, etc. The processes and/or purification steps used to generate epitope-specific T cells, including *in vitro* restimulation steps are stored. Additionally, the type of assay used and every antigen studied are curated. As often as possible, external authoritative resources are utilized to provide standardized nomenclature and additional richness to the data. Examples include: use of NCBI Taxonomy (4) to describe organisms, GenBank (5) and UniProt (6) for proteins, ChEBI (3) for non-peptidic structures, the Ontology for Biomedical Investigations (OBI) (7) for assay types, Gazetteer (http://purl.bioontology.org/ontology/GAZ) for geographic location and the Human Disease Ontology (8) for diseases.

Figure 1A details the breakdown of the IEDB content currently as a function of the various main categories of epitopes and references. The category of infectious disease predominates at the level of numbers of epitopes and references. The category of 'Other' includes many references describing the 3D structure of an epitope with its adaptive immune receptor and, accordingly, these tend to have fewer epitopes. The IEDB reached the milestone of being current with >95% of relevant published literature at the end of 2012, as shown in Figure 1B. Since then, the IEDB curation team has been dedicated to remaining current with newly published literature meeting the goal of making new data available to users within eight weeks of publication. As the demands of curating the backlog of previous publication have now eased, the focus of the IEDB team has shifted to improving the user experience and providing new and useful functionalities.

RESULTS

The driving force behind the IEDB 3.0 redesign has been user feedback accumulated since the 2.0 release of the IEDB in 2009 (2). Feedback was obtained from external immunological experts and database users through various channels, including *ad hoc* forums, help-desk requests/suggestions, a formal IEDB booth at major research conferences, and the annual IEDB user workshops. All feedback was compiled in 2013, in parallel with newly initiated mass appeals for feedback from web site users, contributors, developers and curators. Feedback from developers of related knowledge resources such as the PDB (9) was also solicited. In parallel, web site metrics were analyzed and analyses of queries made were performed to identify the most used search parameters. All gathered feedback was analyzed and redesign goals were set to incorporate as much feedback as possible and to make common queries easier to perform. Table 1 shows example feedback representative of the most commonly made requests. As shown, feedback was summarized into categories and actionable conclusions were drawn.

The two main repeated requests made for the search interface were the ability to refine search results and to ensure that the restriction of the original search would not get lost when drilling down into search results. To enable this, a completely new approach for search was formulated. This plan is analogous to the search on a travel web site, whereby a typical user first enters very simple key search parameters, such as where they wish to travel from and to and on what dates. Once the results meeting these criteria are displayed, the user has the ability to further limit the results based upon additional parameters, for example, further limiting to only nonstop flights, and then to decide upon a specific flight choice. Following this model, the home page was redesigned to present users with the most commonly used search parameters on the home page followed by a results summary page that adds additional filters allowing further refinement of the dataset. In parallel, we sought to improve the presentation and utility of the data displayed on the web site.

The form-based travel web site model was favored over the even simpler 'Google-style' interface because the IEDB houses well-structured data and the same search term can be found in different database fields, leading to unintuitive results if all are returned. For example, 'human' could refer to the host mounting the immune response or the source protein of the epitope, such as when human insulin is tested for immunogenicity in rats. At the same time, it is crucial to not overwhelm the users with too many options for their search as the IEDB contains more than 400 searchable fields, most of which are included in the 'specialized' search of the IEDB that only relatively few expert users are employing. To avoid overwhelming users with this large number of potential parameters, the home page search fields were limited to only include the most commonly used search parameters which are now placed prominently in the center of the home page. As shown in box A of Figure 2, the home page search fields are organized as discrete search sections, typical of the most common queries performed on the IEDB's site, with explanatory icons and help links embedded into the page. New icons were designed to highlight the main search components used for epitope related data. Icons were chosen based on a survey of scientists asked to identify the most relevant icon from a set to represent each major search parameter. These icons are similar to ones for hotel, airfare, or car rental on a travel web site, distinguishing the major types of searches possible. These search sections also serve to restrict search terms to specific database fields and help guide the user as to the types of data that the IEDB contains. For example, in the 'Host' section, a variety of hosts including humans, rodents, non-human primates, and an additional nine commonly studied species are presented.

Once a query such as the one populated in Figure 2 has been executed, the search results are presented on a new page with the current search filters displayed at the top of the results table (Figure 3, box A). Any filter can be removed by a single click on the 'X' next to each parameter. The amount of data present within each of the result set types of Epitopes, Antigens, Assays and References are conveyed

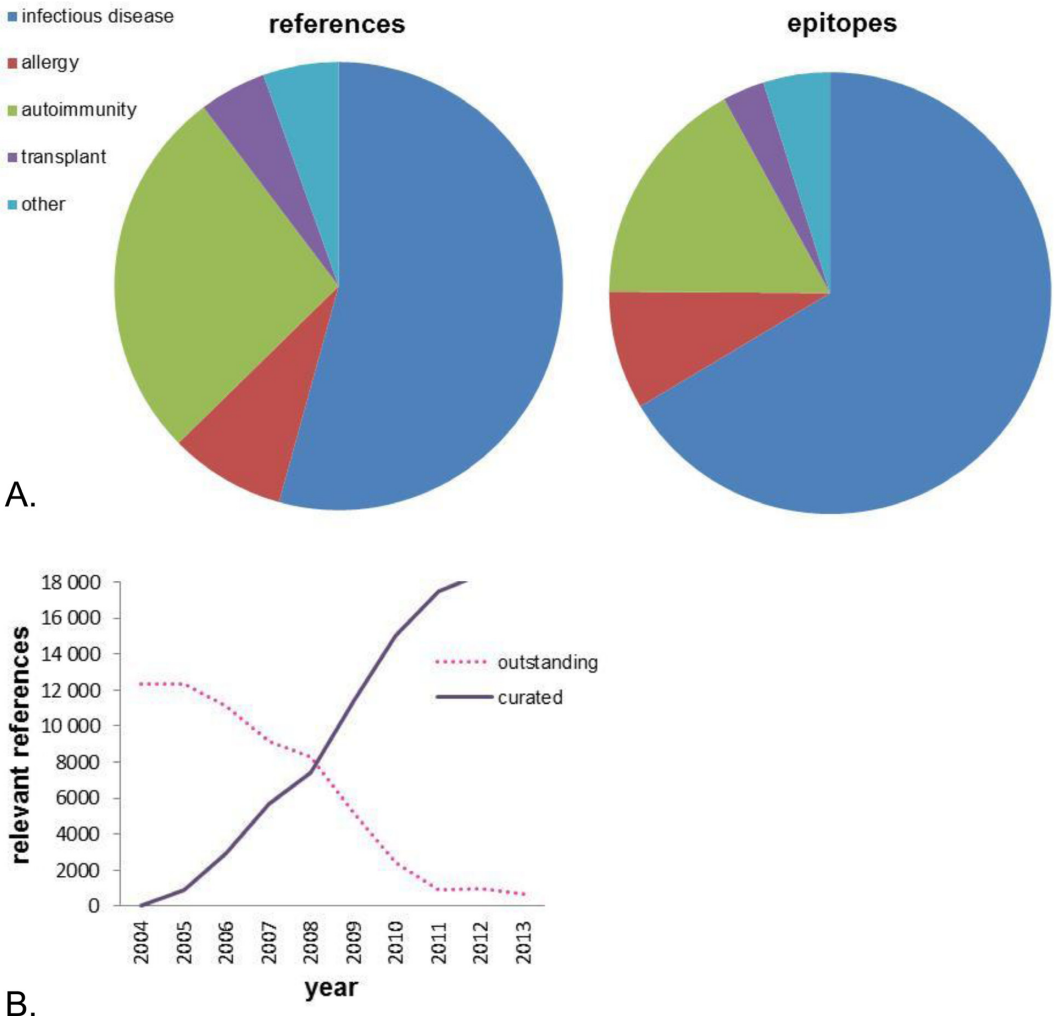


Figure 1. (**A**) The distribution of data in the IEDB by scientific field. (**B**) Curation of relevant references over time.

Table 1. Examples of the types of feedback gathered and the actions taken

Feedback	Sources	Category	Action
Allow one to further refine a query without having to use 'back' button	User observation, user help requests	New feature request	Added this ability
Provide downloadable graphics of immunome browser results	User help requests	New feature request	Added this ability
Provide pop-up hints where the user interface is not intuitive	User help requests	New feature request	Added this feature
Many links on home page rarely used	Web site metrics	Existing feature little used	Made lesser used links less prominent
Protein branch of the molecule tree better needs better nomenclature and synonyms	User observation, user help requests	Existing feature too complicated	Protein branch of the molecule tree was enhanced with these features
Make clearing selections easier	User observation	Existing feature too complicated	Simplified the interface
Analysis resource tools highly used, but hidden on home page	Web site metrics, user help requests	Existing feature difficult to find	These features were made more prominent
Add the ability to save queries	Workshop	New feature request	Not yet added, future
Confirm that the assay names are generally accepted in the immunological community	User observation	Existing feature too complicated	Not yet completed, future
Add cancer epitopes	Workshop, IEDB booth	New scope request	Will not do, out of scope

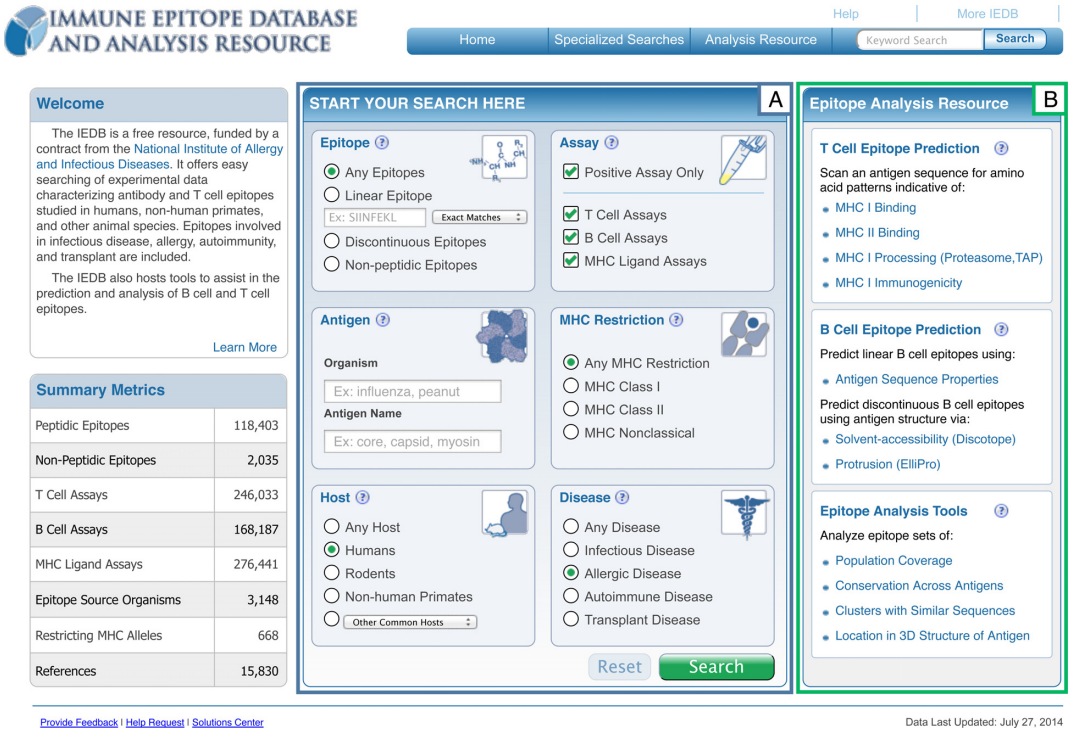


Figure 2. The IEDB 3.0 home page has the most commonly used search parameters centered on the page, shown in box (**A**), with the highly used analysis tools made more prominent, shown in box (**B**).

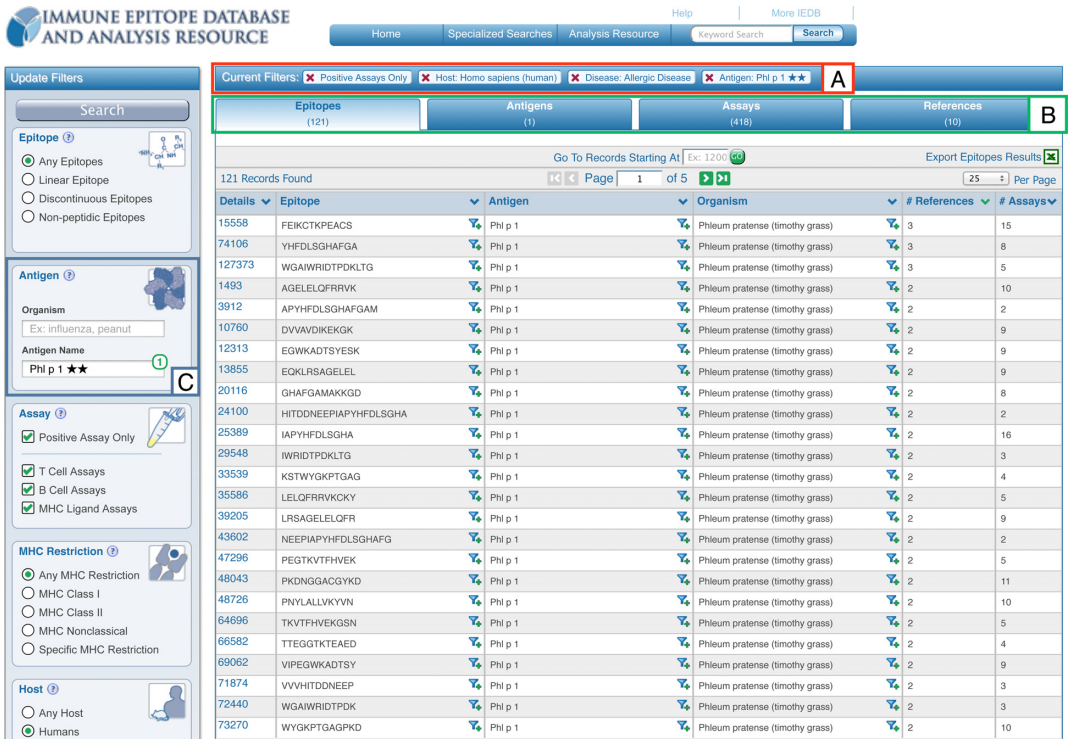


Figure 3. New results presentation format shows current search filters in box (**A**), counts returned per data type in box (**B**) and the new left search panel allowing for continued refinement or editing of one's query, such as by the epitope source, in box (**C**).

by counts and displayed as tabs that allow the user to easily navigate between them (Figure 3, box B). As shown in Figure 3 box C, a search panel added to the left side of the page allows the current result set to be further refined by adding search parameters or to run a new query entirely. These search panels contain the functionality present on the home page plus several additional search features, some of which were previously only present in the IEDB 2.0 'Advanced Search', such as the 'Assay Types.' We plan to continuously monitor the usage of each search parameter to identify additional fields that should be added to or removed from the search panel on the results page.

In addition to the query interface, the presentation of the results has been modified as well. Query results are grouped in four tabs: Epitopes, Antigens, Assays and References that match the current search criteria (Figure 3, box B). These different units of information reflect that some users want to utilize the IEDB as, for example, a way to explore the literature (on the reference tab), while others want to see which specific proteins in an organism have been studied for immune reactivity (on the antigen tab). The amount of data hosted in the IEDB has grown dramatically in the last few years, so that typical queries retrieve a very large number of epitopes. To make sure the most relevant epitopes are immediately visible, results are now sorted by how much information is available, such as the number of references with relevant data, as shown in Figure 3, rather than alphabetically, as was previously done. In addition to the left search panel, users can click on an epitope structure or its source to further narrow the result, using a new 'filter' icon present in the results table. Another noteworthy enhancement in the IEDB 3.0 is a new 'Antigen' tab which displays all epitopes that belong to the same antigen in one row. The Antigen table also provides information on how often epitopes from each antigen were studied with counts for number of epitopes, assays and references relevant to each antigen. Users may further narrow their results to a single antigen using the 'filter' icon present in the results table, or use another updated feature, the 'Immunome Browser,' which is discussed below.

Other search features that have been redesigned throughout are the 'Finder' elements, most notably the Molecule Finder in the antigen search panel. Accessed as shown in Figure 3, box C, this finder provides a hierarchical organization of proteins that allows narrowing the search to epitopes derived from a specific antigen, such as the common allergen Phl p 1. Navigation of proteins within the Molecule Finder was identified as being overly difficult based on user feedback, so it was redesigned to simplify this process. As each protein is derived from an organism, this redesign process began with a major simplification of the organism tree.

The organism tree is based on the NCBI Taxonomy (4), which contains hundreds of thousands of taxa in a hierarchy up to 39 nodes deep. The new organism tree uses a carefully selected subset of the full NCBI Taxonomy that covers all the taxa used in the IEDB, and reduces the depth of the hierarchy for easier navigation by immunologists. We tested the ability of users to correctly classify organisms using either the original NCBI Taxonomy or the revised IEDB organism tree and found that more correct classifications and more certainty in the choices made was obtained using the IEDB organism tree (Figure 4). This was accomplished by presenting users with 50 pairs of a species label with a higher taxon label. The pairs were randomly selected, half from the NCBI Taxonomy and half from the IEDB organism tree. In answer to the question 'Is [species] a [higher taxon]?' (e.g. 'Is "*Narcine timlei* (blackspotted numbfish)" a "Serpent (snake)"?') users could choose 'true,' 'false' or 'I would have to guess.' The users had a range of taxonomic knowledge, but all made more correct classifications and showed more certainty when classifying pairs from the IEDB organism tree as compared to the NCBI Taxonomy.

The Molecule Finder has two top-level branches for peptidic and non-peptidic epitopes. Non-peptidic epitopes are assigned to sources in ChEBI (3) and displayed using the ChEBI hierarchy. Peptidic epitopes derived from proteins occurring in nature have their specific source protein identified by GenPept (5) entries. The variety of distinct sequences represented in GenPept (e.g. the five versions of Phl p I shown in Figure 5) is necessary and reflective of the heterogeneity of proteins within individual species; however, the large number of entries and lack of standardized nomenclature previously overwhelmed users, and made it difficult to obtain all epitopes belonging to a single antigen.

To simplify the representation of proteins within the IEDB we now use UniProt (6) reference proteomes for each species (whenever possible) and use them as parent nodes under which GenPept entries are distributed based on sequence similarity. The use of reference proteomes ensures that each antigen is present just once in the tree, that a consistent nomenclature is utilized, and that additional information such as synonyms and protein classifications can be utilized. If no UniProt reference proteome was available, we constructed alternative reference proteomes in a semi-automated fashion. These are meant to serve as placeholders until the corresponding reference proteomes become available from UniProt. The quality and completeness of proteomes are indicated to the users using a system of stars. As shown in Figure 5, three stars indicate a UniProt reference proteome, two stars indicate a complete UniProt proteome that has not yet been reviewed, and a single star is used for proteins that are not part of a proteome.

The improvements made to the Molecule Finder benefit the Immunome Browser, which can utilize the reference proteomes as mapping targets. Previously available as an on-demand visualization tool in the IEDB (10), the Immunome Browser has now been tightly integrated within the antigen tab, redesigned and enhanced based on user feedback. Conceptually, the Immunome Browser is the first analytical tool integrated into the IEDB database, as it does not simply display information as stored in the database, but maps epitopes onto a reference antigen. Similar to the now commonly used genome browsers, this allows for the aggregation of information derived from different sources and their display in a common reference. On the antigen tab, the Immunome Browser can now be used to immediately visualize linear peptidic epitopes retrieved by a query along the length of the parent antigen based on sequence similarity. This displays how often each protein region has been studied in immune assays and in how many assays the immune response was positive or negative. Figure 6 shows the Immunome Browser output for the epitopes from the Timo-

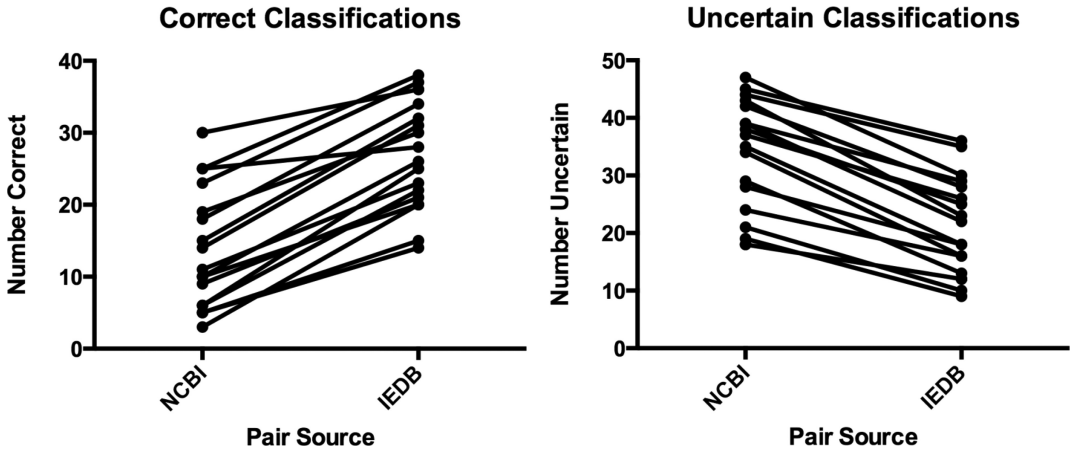


Figure 4. Comparison of classifications of pairs between NCBI Taxonomy and IEDB organism tree. Left: correct classifications. Right: uncertain classifications. The Wilcoxon signed rank test shows that both results are statistically significant with a *P*-value of <0.0001.

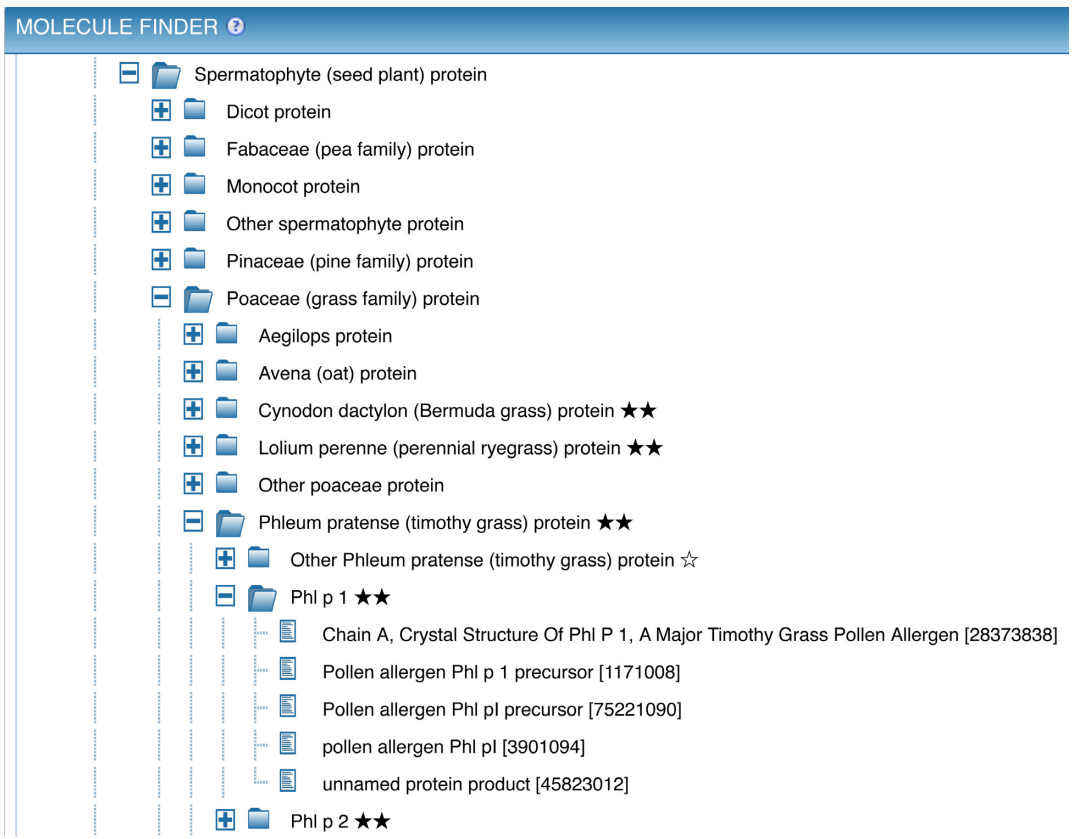


Figure 5. The Molecule Finder provides a hierarchical organization of proteins that allows narrowing the search to epitopes derived from a specific antigen, such as the common allergen Phl p 1. The reference proteome protein 'Phl p 1' is the parent of five individual GenPept entries for this protein from Timothy grass.

thy grass allergen Phl p 1 recognized in the human T cell response. The upper plot renders the lower and upper bounds of the 95% confidence interval of the response frequency for each target protein position, averaged over all epitopes mapped to that position and calculated as the number of positively responded subjects (or individuals in this case) relative to the total number of those tested. The bottom plot shows the number of positive and negative assays averaged over epitopes mapped to each position in the protein sequence. A table below the graphs (not shown) presents results for each epitope and each protein position in a tabular format that can be saved, along with the graph images, for further analysis and publication. The user can interactively zoom in and out the plots to a specific protein region and the table will update accordingly.

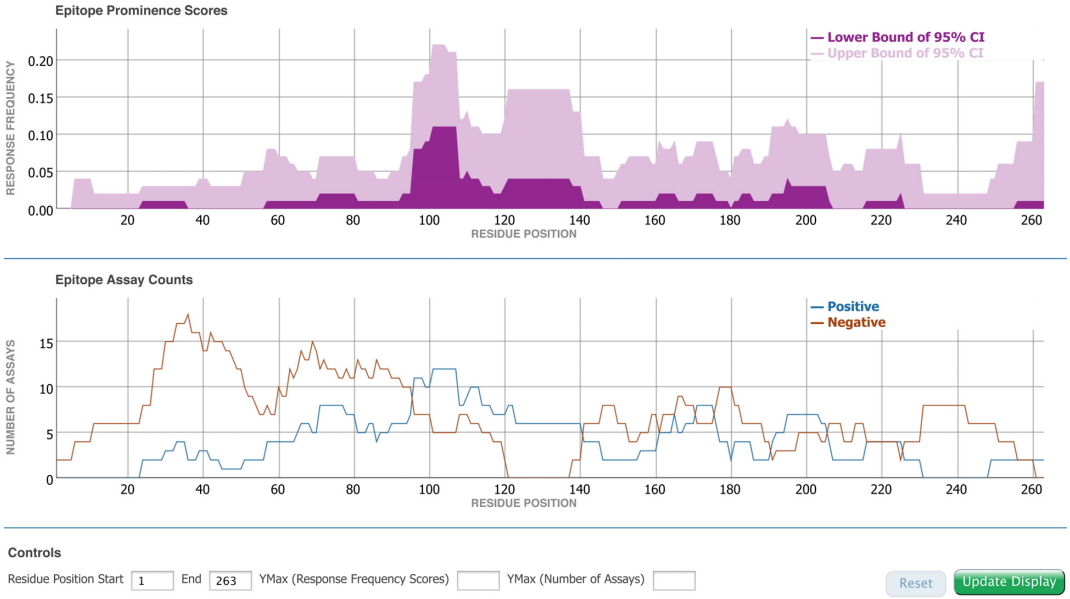


Figure 6. Immunome Browser plots for the epitopes from the Timothy grass allergen Phl p 1 recognized in the human T cell response.

In addition to the most common application of mapping epitopes from different protein variants to a common reference antigen, the Immunome Browser can also map any set of epitopes retrieved by a query to any user-specified protein or protein set. This enables analyses such as which viral epitopes have homologs in the human proteome. Controls to change the mapping criteria and the target protein/proteome to which the epitopes are to be mapped are also provided.

It is worth noting that the entire process of redesigning the IEDB was performed under the expert guidance of several consultants, including two usability experts and a graphic artist. The usability experts compared the user interface design to established design guidelines and successful extant interaction metaphors (11). This identifies critical usability problems early in the development cycle, so that these design issues can be addressed as part of the iterative design process (12). Iterative 'design, implement and evaluate' cycles were used as the IEDB user interface continually evolved. Feedback from the graphic artist and the usability experts was implemented regarding the color scheme, font and style with changes being made across all aspects of the graphics in order to update the general look of the web site, make it self-consistent and visually direct the users toward the most applicable features. The design, placement and functionality of links, search boxes, radio buttons and drop down lists were discussed with users and experts and each was redesigned. For example, the search panels on the results page filters were logically grouped and organized along the left-hand side of the page to be more consistent with commercial web retailer filtering metaphors, which have become de facto 'standards.'

CONCLUSION

After catching up on the curation of in-scope journal articles from the past, the focus of IEDB development for the 3.0 release has shifted toward improving query and reporting interfaces. The goal of this release was to provide intuitive ways to extract biologically accurate information from the large amounts of data now stored in the IEDB. We have here described the main new elements of the 3.0 release, all of which were motivated by user feedback gathered over the years. We believe that such development focusing on the usability of the web site is equally important to the introduction of new capabilities which—while often more exciting to implement from a web site developer's perspective—have little value if they are not actually utilized by the user community.

ACKNOWLEDGEMENT

We wish to acknowledge graphic artist Ben Hannam, Accomplish Studios, Chapel Hill, North Carolina, for his much appreciated contribution and expertise.

FUNDING

National Institutes of Health [HHSN272201200010C]. Funding for open access charge: National Institutes of Health [HHSN272201200010C].
Conflict of interest statement. None declared.

REFERENCES

1. Roberts,R.J. (2001) PubMed Central: the GenBank of the published literature. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 381–382.
2. Vita,R., Zarebski,L., Greenbaum,J.A., Emami,H., Hoof,I., Salimi,N., Damle,R., Sette,A. and Peters,B. (2010) The immune epitope database 2.0. *Nucleic Acids Res.*, **38**, D854–D862.
3. Hastings,J., de Matos,P., Dekker,A., Ennis,M., Harsha,B., Kale,N., Muthukrishnan,V., Owen,G., Turner,S., Williams,M. *et al.* (2013) The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Res.*, **41**, D456–D463.

4. Sayers,E.W., Barrett,T., Benson,D.A., Bryant,S.H., Canese,K., Chetvernin,V., Church,D.M., DiCuccio,M., Edgar,R., Federhen,S. *et al.* (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **37**, D5–D15.
5. Benson,D.A., Clark,K., Karsch-Mizrachi,I., Lipman,D.J., Ostell,J. and Sayers,E.W. (2014) GenBank. *Nucleic Acids Res.*, **42**, D32–D37.
6. The UniProt Consortium (2014) Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **42**, D191–D198.
7. Brinkman,R.R., Courtot,M., Derom,D., Fostel,J.M., He,Y., Lord,P., Malone,J., Parkinson,H., Peters,B., Rocca-Serra,P. *et al.* (2010) Modeling biomedical experimental processes with OBI. *J. Biomed. Semantics*, **22**(Suppl. 1), S1–S7.
8. Schriml,L.M., Arze,C., Nadendla,S., Chang,Y.W., Mazaitis,M., Felix,V., Feng,G. and Kibbe,W.A. (2012) Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Res.*, **40**, D940–D946.
9. Rose,P.W., Bi,C., Bluhm,W.F., Christie,C.H., Dimitropoulos,D., Dutta,S., Green,R.K., Goodsell,D.S., Prlic,A., Quesada,M. *et al.* (2013) The RCSB Protein Data Bank: new resources for research and education. *Nucleic Acids Res.*, **41**, D475–D482.
10. Kim,Y., Vaughan,K., Greenbaum,J., Peters,B., Law,M. and Sette,A. (2012) A meta-analysis of the existing knowledge of immunoreactivity against hepatitis C virus (HCV). *PLoS One*, **7**, e38028.
11. Hartson,R. and Pyla,P.S. (2012) *Modeling Biomedical Experimental Processes with OBI*. Morgan Kaufmann, Burlington, MA.
12. Nielsen,J. (1993) Iterative user-interface design. *Computer*, **26**, 32–41.

BMC Genomics

BioMed Central

Database

Open Access

Bcipep: A database of B-cell epitopes

Sudipto Saha, Manoj Bhasin and Gajendra PS Raghava*

Address: Institute of Microbial Technology Chandigarh, India

Email: Sudipto Saha - saha@imtech.res.in; Manoj Bhasin - bhasin@imtech.res.in; Gajendra PS Raghava* - raghava@imtech.res.in

* Corresponding author

Published: 29 May 2005

BMC Genomics 2005, **6**:79 doi:10.1186/1471-2164-6-79

Received: 30 September 2004

Accepted: 29 May 2005

This article is available from: http://www.biomedcentral.com/1471-2164/6/79

Abstract

Background: Bcipep is a database of experimentally determined linear B-cell epitopes of varying immunogenicity collected from literature and other publicly available databases.

Results: The current version of Bcipep database contains 3031 entries that include 763 immunodominant, 1797 immunogenic and 471 null-immunogenic epitopes. It covers a wide range of pathogenic organisms like viruses, bacteria, protozoa, and fungi. The database provides a set of tools for the analysis and extraction of data that includes keyword search, peptide mapping and BLAST search. It also provides hyperlinks to various databases such as GenBank, PDB, SWISS-PROT and MHCBN.

Conclusion: A comprehensive database of B-cell epitopes called Bcipep has been developed that covers information on epitopes from a wide range of pathogens. The Bcipep will be source of information for investigators involved in peptide-based vaccine design, disease diagnosis and research in allergy. It should also be a promising data source for the development and evaluation of methods for prediction of B-cell epitopes. The database is available at http://www.imtech.res.in/raghava/bcipep.

Background

The antigenic regions of protein recognized by the binding sites of immunoglobulin molecules are called B-cell epitopes [1]. These epitopes can be classified into two categories; i) conformational/discontinuous epitope, where residues are distantly separated in the sequence and brought into physical proximity by protein folding and, ii) linear/continuous epitope, comprised of a single continuous stretch of amino acids within a protein sequence that can react with anti-protein antibodies [2,3]. Most of the B-cell epitopes were thought to be discontinuous. However, in late 1980s it was shown that this conformational restriction is not a necessary condition for the production of protein-reactive anti-peptide antibodies [4]. The designing of the conformational epitopes is difficult and so experimental B-cell epitopes largely include linear epitopes. These linear epitopes can be exploited in the development of synthetic vaccines and disease diagnosis. A number of vaccines based on B-cell epitopes are currently under clinical phase trials against viruses [5], bacteria [6] and cancer [7]. These epitopes are also important for allergy research and in determining cross-reactivity of IgE-type epitopes of allergens [8].

A large number of B-cell epitopes have been reported in the literature in last two decades. There is a need to collect and compile these epitopes to evaluate the performance of existing B-cell epitope prediction methods and to further develop better methods [9,10]. We have observed that the performance of the existing physico-chemical

scales is not very high [11]. Recently, Blythe and Flower examined 484 amino acid propensity scales in predicting of B-cell epitopes and found that even the best set of scales and parameters performed only marginally better than random [12]. The evaluation of the existing scales indicates that there is a need to develop better methods by using artificial intelligence techniques. The collection of B-cell epitopes should help in deriving new scales for accurate in silico prediction of linear epitopes. Also it will help the immunologist to understand the complex nature of immunogenic peptides and for the development of vaccines. There are many databases available on T-cell epitopes [13-15]. In contrast, there are limited number of databases on B-cell epitope for example JenPep [16,17] and HIVDB [18]. Recently, JenPep has been superseded by AntiJen 2.0, which has included peptides bound to MHC ligand, TCR-MHC Complexes, T cell epitope, TAP, B cell epitope molecules and immunological protein-protein interactions. AntiJen also contains peptide library, copy numbers and diffusion coefficient data. Though AntiJen provides information about different types of peptides (>24000 entries) from a single source but it provides limited information about B-cell epitopes and tools to analyze and retrieve the data. Recently, we have created a comprehensive database, Bcipep, of B-cell epitopes. Latest version of Bcipep contains 3031 entries, where each entry provides detailed description of a B-cell epitope. Currently, we have covered only the continuous B-cell epitopes. The aim of this database is to assist the scientific community working in the areas of synthetic peptide vaccines (based on B-cell epitopes) and allergy research. The database will complement the existing databases such as AntiJen [16,17].

Availability

The database is available at http://www.imtech.res.in/raghava/bcipep, http://bioinformatics.uams.edu/mirror/bcipep/ (Mirror Site) and http://srs.ebi.ac.uk/srs6bin/cgbin/wgetz?-page+LibInfo+-id+1X2XW1JU5_L+-lib+BCIPEP (SRS version).

Database construction

The PostgreSQL relational database management system (RDBMS) has been used for storing, retrieval and managing the data. The scripts, which provide interface between user and database, were written in PERL, CGIPerl and Pgperl. B-cell epitopes were collected from the literature (PubMed, http://www.ncbi.nlm.nih.gov/pubmed/; ScienceDirect, http://www.sciencedirect.com/). A large number of HIV epitopes were extracted from a book [18].

Database description

The aim of Bcipep database is to provide; i) comprehensive information about B-cell epitopes, ii) tools for extraction and analysis of this information and, iii) hyperlinks to related databases. The overall architecture of Bcipep database is shown in Figure 1.

Database information

General

This database provides compehensive information about the linear B-cell epitopes which includes; i) amino acid sequence of epitope, ii) source protein from which epitopes were obtained, iii) experimental methods used in accessing the immunogenic potential of epitopes, iv) pathogen group (e.g., bacteria, virus, fungi, protozoa) of source protein and, v) miscellaneous information in the '*Comment*' field.

Immunogenicity and model organisms

The immunogenicity of a peptide is a semi-quantitative measure of its immunogenic potency. In Bcipep it is divided into three categories; i) immunodominant, if it increases 2–3 folds anti-peptide antibodies in comparison to reference or control (carrier protein, eg., BSA or KLH), ii) immunogenic, if it enhances anti-peptide antibodies by one-fold in comparison to reference, and iii) null-immunonogenic, where no difference was observed when compared to reference. This information is very important for developing B-cell epitope prediction method. The database also provides information about '*Model organism*' used for immunization.

Antibodies and neutralization

The database provides full information about monoclonal or polyclonal antibodies produced against an epitope. The information includes, isotypes of immunoglobin and name/number of monoclonal antibodies. The database also contains information about neutralization potential of anti-peptide antibody, which is crucial for considering a peptide for synthetic vaccine design.

Links to databases

The Bcipep provides hyperlinks to various sequence databases in order to provide detailed information about peptides in database. The '*database reference*' field consists name/code of protein available in SWISS-PROT [19]. The '*Antigenic structure*' field consists of PDB codes [20] of protein structures having matching peptides. These PDB codes are linked to OCA browser http://pdb.tau.ac.il/ in order to provide detailed structure information of these proteins. It also provides structure information about 242 antibodies, where each antibody is hyperlinked to PDB database through the OCA browser. The '*Publication reference*' field provides full information about related publications with link to PUBMED [21]. Bcipep is also linked to MHCBN database [15] in order to identify the peptides that are B-cell as well as T-cell epitopes.

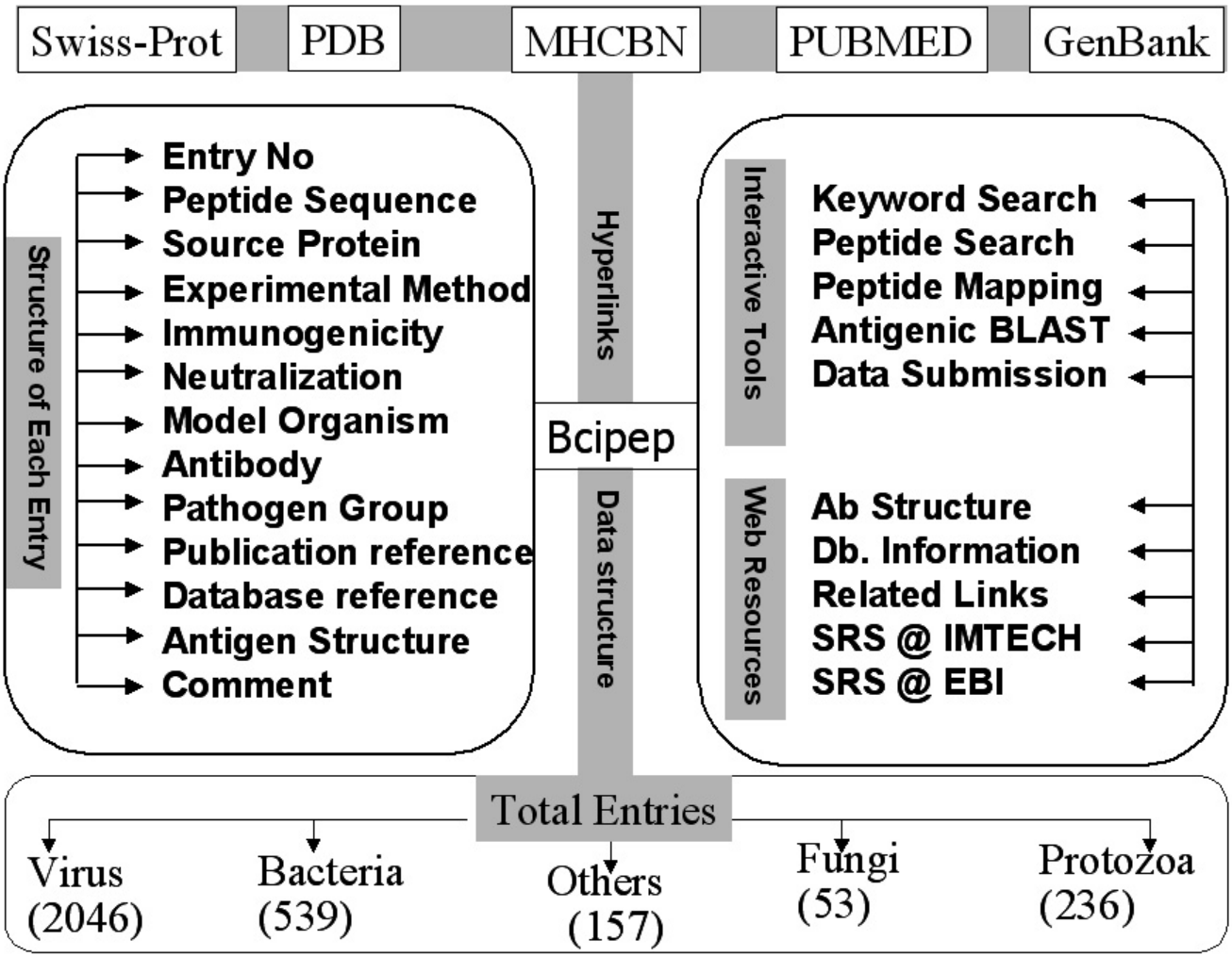


Figure 1
A schematic representation of Bcipep database.

Web tools

Bcipep has following major web-based tools for retrieval and analysis of information in Bcipep database (Figure 1). These web tools have been designed to facilitate the user in retrieving information from database.

Keyword search

This option allows users to perform search on all fields of the database ('*Peptide Sequence*', '*Source Protein*', '*Publication Reference*', '*Database Reference*'). One can restrict the keyword search on any specific field. It also allows users to select the fields to be displayed. An example of keyword search is shown in Figure 2a, where key word 'P26694' is searched in any filed of database. The output/result of this keyword search is shown in Figure 2b.

Peptide search

The database provides option to search a peptide in Bcipep. The tool will display full information about the peptides included in Bcipep. The server also permits users to search their query sequence in any pathogen group. Search can be restricted on the basis of immunogencity that is immunodominant, immunogenic or null-immunogenic. An example of input and output of peptide search is shown in Figures 3a and 3b respectively.

Mapping of T-cell epitopes

This server allows searching of peptide in Bcipep against MHCBN database. The MHCBN database provides information about components of cell-mediated immunity like MHC binders/non-binders, T-cell epitopes and TAP

Figure 2
The typical display of Bcipep database for keyword search; a) input page of keyword search; and b) output of keyword search.

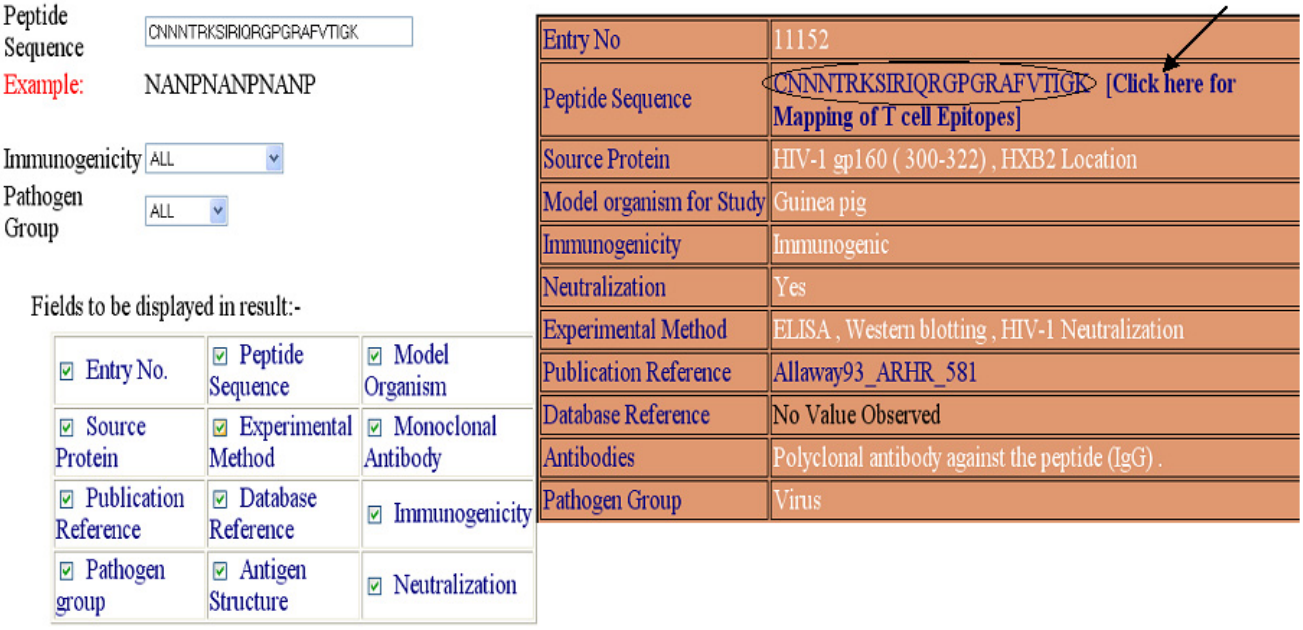


Figure 3
This example illustrate peptide search on Bcipep; a) Peptide search page; and b) result of peptide search.

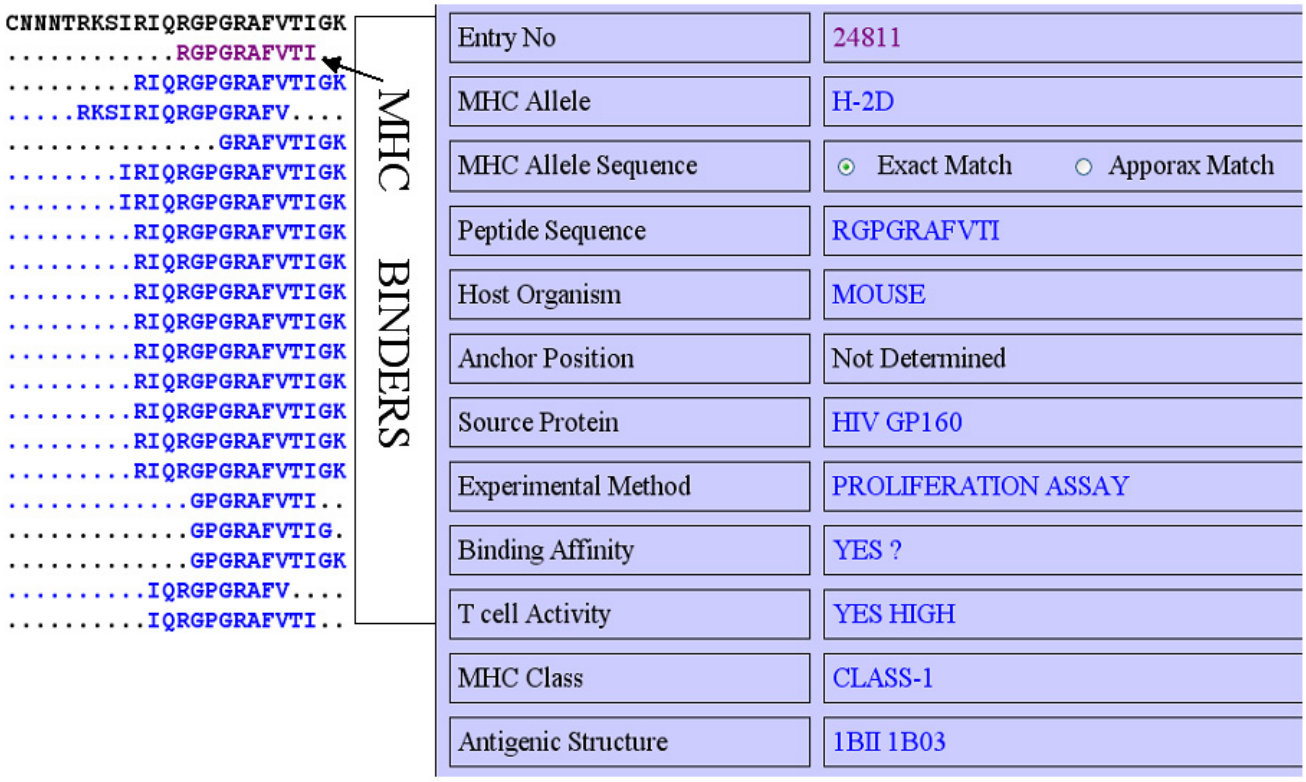


Figure 4
Mapping of peptide in MHCBN database; a) mapping of MHCBN peptides on B-cell epitope, and b) full information about a MHCBN peptide.

binders [14]. The peptides related to cell-mediated immunity can be mapped on resultant B-cell epitopes obtained from keyword/peptide search by clicking on '*Peptide Sequence*' field (Figures 2b and 3b). Thus, the server is useful in identifying the potential B-cell epitopes having T-cell epitopes (or MHC binders). The example of mapping of MHCBN peptides on B-cell epitope (by clicking on peptides sequence field of Figure 3b) is shown in Figure 4a. The full information of each map peptide can be obtained by clicking on the mapped sequence. One such example is shown in Figure 4a and 4b. The mapping allows user to detect the regions in B-cell epitope having promiscuous MHC binders (peptides that can bind to large number of MHC alleles) or T-cell epitopes.

Peptide mapping

The peptides of Bcipep can be mapped on query sequence using this option. The full information about mapped peptide can be obtained by clicking on it. The tool will assist the researchers in gaining knowledge about the known immunogenic or non-immunogenic regions in target protein of interest. The example input and output of peptide mapping is shown in Figures 5a and 5b respectively. The users can specify the pathogen group and/or immunogenicity level of peptides to be mapped on query sequence. As shown in Figure 5b, the graphical mapping of peptides/epitopes on query allows one to easily detect the regions that bind to large number of B-cell peptides.

BLAST search

This tool allows users to search their query protein against antigenic proteins maintained at Bcipep. The sequence of 1070 antigenic proteins has been obtained from SWISS-PROT. The similarity search is performed using the GWBLAST server http://www.imtech.res.in/raghava/gwblast/. The GWBLAST also allows users to analyze the BLAST output like multiple alignments, phylogenetic analysis.

Online data submission

The database has the facility to submit data to Bcipep online via Internet. Users can submit information about

Figure 5
Mapping of B-cell epitopes on antigen sequence, a) submission page of B-cell epitope mapping, and b) mapping results

their experimentally determined B-cell epitopes. We hope that immunologists will submit their information on B-cell epitope in Bcipep, similar to the sequence data at GenBank and SWISS-PROT. The Bcipep team will add more data from literature to maintain B-cell data up-to-date and cross check the data submitted by users to maintain quality of the data.

Potential utility and limitations

One of the major challenges in the field of subunit vaccine design is to identify the antigenic regions (B and T cell epitopes) that can generate antigen specific memory cells. Thus, the identification of regions/stretches on an antigen from the data pool of known epitopes is an important step in vaccine design. The Bcipep database would be very useful as it consists of comprehensive information about experimentally verified linear B-cell epitopes and tools for mapping these epitopes on an antigen sequence. In case query antigen contains known epitopes, this database might aid in the wet experimentation and lower the cost by reducing the overlapping repeats. This strategy is frequently used for the screening of transgenic proteins by searching linear IgE-binding epitopes [22]. Unlike discontinuous epitopes, the linear epitopes are easy to design, as they do not require tertiary structure information. Bcipep also provides information on neutralizing B-cell epitopes where an antibody generated against a B-cell epitope neutralizes the parent antigen. The current version of Bcipep provides neutralizing information on about 1309 such B-cell epitopes. This information is very important for selecting functional B-cell epitopes. This database also provides a link with MHCBN to search for overlapping regions of MHC binders and T-cell epitopes in the B-cell epitopes. Thus, the user can identify both antigenic regions that can activate B-cell and T-cell, which can lead to the development of better vaccine. The epitopes in Bcipep can be used to derive the rules for predicting B-cell epitopes.

The aim of designing synthetic linear peptides as epitope-vaccine is to induce neutralizing antibodies against the pathogen [23]. There are many reports that the linear B-cell epitopes were characterized as neutralizing antibodies as in Clostridium botulinum neurotoxin type A (Btx A)[24]. In Bcipep, there are 748 neutralizing anti-peptide antibodies entries. However, in some cases these linear epitope(s) fail to produce neutralizing antibodies and do not give protective immunity. For instance, it has been shown in the past that the antibodies against the synthetic peptides and short recombinant proteins of approximately 100 amino acids of hepatitis E virus (HEV) do not neutralize, suggesting that the HEV neutralization epitope(s) is conformation dependent [25]. The elicitation of a bactericidal and protective immune response to *Borrelia burgdoferi* decorin binding protein requires a properly folded conformation for the production of functional antibodies [26]. Recently, Corcoran *et al*, 2004, observed that B-cell memory is established and maintained against conformational epitopes of Parvovirus VP2 and against linear epitopes of VP1 but not against linear epitope VP2 [27]. Thus, it is not necessary that the linear B-cell epitope will always give rise to memory cells. One should also check the neutralizing information of B-epitopes, as only 748 B-cell epitopes out of 1309 in Bcipep were able to neutralize the parent protein.

For an effective use of Bcipep, it is important to understand the limitation of linear B-cell epitopes and data in

Bcipep. The few limitations of current version of Bcipep are; i) it does not cover discontinuous epitopes, ii) it has limited number of unique peptides (1590) in 3031 entries and, iii) it contains peptides having only natural amino acids. One should be careful in using linear B-cell epitopes in developing epitope based subunit vaccine. The organism used for immunization (information included in the database) should also be taken into consideration, since immune response is T-helper cell (MHC-II-peptide complex) dependant and B-cell epitope alone may not generate protective antibodies [28]. In some cases, the nature of the adjuvant used and the route of immunization (information not included in the database) might also play important roles in the induction of protective anti-peptide antibody response against the pathogen [29-31].

Authors' contributions

SS collected and compiled the data as well as developed the web server. MB helped in designing website and stored data in PostgreSQL. GPSR conceived the idea and supervised the work.

Acknowledgements

Authors are thankful to Dr Grish C. Varshney for critically reading the manuscript. We are also thankful to Council of Scientific and Industrial Research (CSIR) and Department of Biotechnology (DBT), Govt. of India, for financial assistance.

References

1. Van Regenmortel MH: **Synthetic peptides versus natural antigens in immunoassaya.** *Ann Biol Clin (Paris)* 1993, **51:**39-41.
2. Barlow DJ, Edwards MS, Thornton JM: **Continuous and discontinuous protein antigenic determinants.** *Nature* 1986, **322:**747-748.
3. Langeveld JP, martinez-Torrecuadrada J, boshuizen RS, Meloen RH, Ignacio CJ: **Characterisation of a protective linear B cell epitope against feline parvoviruses.** *Vaccine* 2001, **19:**2352-2360.
4. Walter G: **Production and use of antibodies against synthetic peptides.** *J Immunol Methods* 1986, **88:**149-61.
5. El Kasmi KC, Muller CP: **New strategies for closing the gap of measles susceptibility in infants: towards vaccines compatible with current vaccination schedules.** *Vaccine* 2001, **19:**2238-2244.
6. Sabhanini L, Manocha M, Sridevi K, Shashikiran D, Rayanade R, Rao DN: **Developing subunit immunogens using B and T cell epitopes and their constructs derived from F1 antigen of *Yersinia pestis* using novel delivery vehicles.** *FEMS Immunol Med Microbiol* 2003, **1579:**1-15.
7. Kieber-Emmons T, Luo P, Qiu J, Chang TY, Insung O, Blaszczyk-Thurin M, Steplewski Z: **Vaccination with carbohydrate peptide mimotopes promotes anti-tumor responses.** *Nat Biotechnol* 1999, **17:**660-665.
8. Selo I, Clement G, Bernard H, Chatel J, Creminon C, Peltre G, Wal J: **Allergy to bovine beta-lactoglobulin:specificity of human IgE to tryptic peptides.** *Clin Exp Allergy* 1999, **29:**1055-1063.
9. Odorico M, Pellequer JL: **BEPITOPE: predicting the location of continuous epitope and patterns in proteins.** *J Mol Recognit* **16:**20-22.
10. Alix AJ: **Predictive estimation of protein linear epitopes by using the program PEOPLE.** *Vaccine* 1999, **18:**311-314.
11. Saha S, Raghava GPS: **BcePred: Prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties.** In *ICARIS, LNCS Volume 3239*. Edited by: Nicosia G, Cutello V, Bentley PJ, Timmis J. Springer; 2004:197-204.
12. Blythe MJ, Flower DR: **Benchmarking B cell epitope prediction: Underperformance of existing methods.** *Protein Science* 2005, **14:**246-248.
13. Brusic V, Rudy G, Harrison LC: **MHCPEP, a database of MHC-binding peptides: update 1997.** *Nucleic Acids Res* 1998, **26:**368-71.
14. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S: **SYFPEITHI: database for MHC ligands and Peptide Motifs.** *Immunogenetics* 1999, **50:**213-219.
15. Bhasin M, Singh H, Raghava GPS: **MHCBN: A comprehensive database of MHC binding and non-binding peptides.** *Bioinformatics* 2003, **19:**666-667.
16. Blythe MJ, Doytchinova IA, Flower DR: **JenPep: a database of quantitative functional peptide data for immunology.** *Bioinformatics* 2002, **18:**434-439.
17. McSparron H, Blythe MJ, Zygouri C, Doytchinova IA, Flower DR: **JenPep: a novel computational information resource for immunobiology and vaccinology.** *J Chem Inf Comput Sci* 2003, **43:**1276-87.
18. Korber B, Brander C, Haynes B, Koup R, Kuiken C, Moore J, Walker B, Watkins D: **HIV Monoclonal Antibodies.** In *"HIV Molecular Immunology 2001"* Theoretical Biology and Biophysics group T-10, Mail Stop K710 Los Alamos national Laboratory, Los Alamus, New Mexico 87545 U.S.A; 2002:IV-B-1-IV-B-278.
19. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000.** *Nucleic Acids Res* 2000, **28:**45-48.
20. Westbrook J, Feng Z, Jain S, Bhat TN, Thanki N, Ravichandran V, Gilliland GL, Bluhm WF, Weissig H, Greer DS, Bourne PE, Berman HM: **The Protein Data Bank: unifying the archive.** *Nucleic Acids Res* 2002, **30:**245-248.
21. Wheller DL, Church DM, Lash AE, Leipe DD, Madden TL, Pontius JU, Schuler GD, Schriml LM, Tatusova TA, Wagner L, Rapp BA: **Database resources of National Center for Biotechnology Information: 2002 update.** *Nucleic Acids Res* 2002, **30:**13-16.
22. Kleter GA, Peijnenburg AA: **Screening of transgenic proteins expressed in transgenic food crops for the presence of short amino acid sequences identical to potential, IgE – binding linear epitopes of allergens.** *BMC Structural Biology* 2002 [http://www.biomedcentral.com/1472-6807/2/8].
23. Xiao Y, Lu Y, Chen YH: **Epitope-vaccine as a new strategy against HIV-1 mutation.** *Immunol Lett* 2001, **77:**3-6.
24. Wu HC, Yeh CT, Huang YL, Tarn LJ, Lung CC: **Characterization of neutralizing antibodies and identification of neutralizing epitope mimics on the Clostridium botulinum neurotoxin type A.** *Appl Environ Microbiol* 2001, **67:**3201-3207.
25. Meng J, Dai X, Chang JC, Lopareva E, Pillot J, Fields HA, Khudyakov YE: **Identification and characterization of the neutralization epitope(s) of the hepatitis E virus.** *Virology* 2001, **288:**203-211.
26. Ulbrandt ND, Cassatt DR, Patel NK, Roberts WC, Bachy CM, Fazenbaker CA, Hanson MS: **Conformational nature of the Borrelia burgdorferi decorin binding protein A epitopes that elicit protective antibodies.** *Infect Immun* 2001, **69:**4799-4807.
27. Corcoran A, Mahon BP, Doyle S: **B cell memory is directed toward conformational epitopes of parvovirus B19 capsid proteins and the unique region of VP1.** *J Infect Dis* 2004, **189:**1873-1880.
28. An LL, Whitton JL: **A multivalent minigene vaccine, containing B-cell, cytotoxic T-lymphocyte, and Th epitopes from several microbes, induces appropriate responses in vivo and confers protection against more than one pathogen.** *J Virol* 1997, **71:**2292-2302.
29. Obeid OE, Stanley CM, Steward MW: **Immunological analysis of the protective responses to the chimeric synthetic peptide representing T- and B-cell epitopes from the fusion protein of measles virus.** *Virus Res* 1996, **42:**173-180.
30. Fernandez IM, Snijders A, Benaissa-Trouw BJ, Harmsen M, Snippe H, Kraaijeveld CA: **Influence of epitope polarity and adjuvants on the immunogenicity and efficacy of a synthetic peptide vaccine against Semliki Forest virus.** *J Virol* 1993, **67:**5843-8.
31. Todryk SM, Kelly CG, Lehner T: **Effect of route of immunisation and adjuvant on T and B cell epitope recognition within a streptococcal antigen.** *Vaccine* 1998, **16:**174-180.

Nucleic Acids Research, 2006, Vol. 34, Database issue ***D777–D780***
doi:10.1093/nar/gkj053

Epitome: database of structure-inferred antigenic epitopes

Avner Schlessinger[1,2,3,*], Yanay Ofran[1,2], Guy Yachdav[1,2,3] and Burkhard Rost[1,2,3]

[1]CUBIC, Department of Biochemistry and Molecular Biophysics, Columbia University, 1130 St Nicholas Avenue room 804, New York, NY 10032, USA, [2]Columbia University Center for Computational Biology and Bioinformatics (C2B2), 1130 St Nicholas Avenue room 804, New York, NY 10032, USA and [3]NorthEast Structural Genomics Consortium (NESG), Department of Biochemistry and Molecular Biophysics, Columbia University, 1130 St. Nicholas Avenue room 804, New York, NY 10032, USA

Received August 15, 2005; Revised September 23, 2005; Accepted October 4, 2005

ABSTRACT

Immunoglobulin molecules specifically recognize particular areas on the surface of proteins. These areas are commonly dubbed B-cell epitopes. The identification of epitopes in proteins is important both for the design of experiments and vaccines. Additionally, the interactions between epitopes and antibodies have often served as a model for protein–protein interactions. One of the main obstacles in creating a database of antigen–antibody interactions is the difficulty in distinguishing between antigenic and non-antigenic interactions. Antigenic interactions involve specific recognition sites on the antibody's surface, while non-antigenic interactions are between a protein and any other site on the antibody. To solve this problem, we performed a comparative analysis of all protein–antibody complexes for which structures have been experimentally determined. Additionally, we developed a semi-automated tool that identified the antigenic interactions within the known antigen–antibody complex structures. We compiled those interactions into Epitome, a database of structure-inferred antigenic residues in proteins. Epitome consists of all known antigen/antibody complex structures, a detailed description of the residues that are involved in the interactions, and their sequence/structure environments. Interactions can be visualized using an interface to Jmol. The database is available at http://www.rostlab.org/services/epitome/.

BACKGROUND

Protein–antigen structures

Antigen–antibody complexes have long been used as a model for understanding the general phenomenon of molecular recognition (1–5). The number of experimental high-resolution 3D structures of antibody–antigen complexes in the PDB (6) has significantly increased over the last years. Several groups have used these data to analyze and characterize antigenic interactions, i.e. interactions between the protein (the antigen) and the Complementarity Determining Regions (CDRs) of the antibody (7,8). An important first step in studying antigenic interactions is the characterization of CDRs. MacCallum *et al*. (8) observed that the hypervariable loops of CDRs adopt only a limited number of backbone conformations that are determined by a few key residues. Two recent studies have suggested that the amino acid composition and the length of CDRs determine the type of antigen that can be bound (9,10). Several studies have attempted to differentiate the residues on the antigen surface that are involved in the antigenic interaction from all others (5,7,11). The results of these studies were rather inconsistent. Differences in the data sets chosen (some of which were very small) and in the methodologies may explain some of those inconsistencies. Most importantly, however, the definitions of the CDRs often differed greatly, i.e. if two studies investigate the same PDB complex and use the same methodology, they might disagree on which of the interactions are antigenic (7). An important ramification of this problem was unveiled by Blythe and Flower (12), who showed that most existing B-cell epitope prediction methods do not work adequately. One explanation for this observation could be that most methods rely on inaccurate identifications of epitopes.

*To whom correspondence should be addressed. Tel: +1 212 851 4669; Fax: +1 212 305 7932; Email: as2067@columbia.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

Definition of the CDRs

Antibodies are composed of a skeleton of beta-sheets. Most of the amazing variety of antibodies is realized by differences in six hypervariable loops of the CDRs. Therefore, the CDRs have previously been defined through these six loops. The first definition of CDRs was as regions in the Kabat sequence variability plot (13,14). The residues in these regions are identified through an alignment between the query sequence and a consensus motif for antibodies. Although widely used, the Kabat CDR-definitions can be problematic because CDRs that are in structural loops often have very unusual sequences that are not captured by regular sequence motifs (15). In fact, any method based only on sequence information is prone to misaligning and therefore mis-assigning loopy CDRs. Chothia and co-workers (16) therefore based their CDR identification on structural information. Initially, hypervariable loops were defined according to a few structures. Later, the numbering of the residues that was used to locate the CDRs was changed to account for structures that became available subsequently (17). Studies also differ in their definition of secondary structures, thereby increasing the inconsistency in defining hypervariable loops. Additional disadvantages of both the Kabat and Chothia *et al*. method are described elsewhere (http://www.bioinf.org.uk/abs/).

Here, we address these problems through a comprehensive study of all known antigen–antibody complexes in the PDB. Analyzing the structures, we identified the consensus residues on the antibodies and thereby identified the CDRs on all known protein–antibody complexes (details below). This initial set of CDRs facilitated the automatic generation of a database with all known antigenic residues in the PDB; we also included the sequence environment and a detailed description of the CDR with which they interact. Several databases of antibody–antigen complex structures are available (15,18,19). Some of these databases focus on the structural aspects of the interaction (19,20). There are also databases that compile B-cell epitopes without their corresponding antibodies (12,21). However, none of these databases explicitly locates the CDRs or identifies the antigenic residues semi-automatically. In this sense, our resource is more comprehensive and easily adjustable to growing data, as more 3D structures of antigen–antibody complexes become available. Thus, the databases mentioned above, particularly the ones that are not structure based, are complementary to Epitome.

DATABASE

Extraction of 3D structures and identification of CDRs

In order to identify all structures in the PDB that contain at least one antibody–antigen complex, we searched with BLAST (22) for a consensus sequence of an antibody against the PDB. The rationale for using BLAST rather than PSI-BLAST was to avoid capturing molecules such as T-cell receptors which, despite their similarity to antibodies, participate in cell-mediated immune response, and therefore represent a different type of antigenic interaction. We then added PDB structures that contain an immunoglobulin fold from the Structural Classification of Proteins database (SCOP) (23) and PDB entries that are identified as antibody–antigen complexes through keywords (e.g. 'antibody' and 'antigen'). We discarded all complexes with T-cell receptors or MHC molecules, since these are formed during cell-mediated immune response. We labeled residues as interacting if any of their respective atoms were within a sphere of $\leqslant$6Å (24). This resulted in our final list of interactions between antibodies and antigens. Thus, we define antibody–antigen interaction as spatial proximity between a residue within the CDRs and a residue on the surface of the antigenic protein.

We located the CDRs in the known protein–antibody complexes through the following knowledge-based approach. We began by creating multiple structure alignments of antibody structures using SKA (25,26). Since the light and heavy chains have different CDRs, two different multiple structure alignments were performed corresponding to each type of antibody chain. Additionally, due to the fact that our database included several redundant sequences, we ran the structural alignment program on a sequence-unique subset of all protein–antibody complexes. As antibody sequences are highly similar to each other, the criteria for the redundancy of the complex set was determined by the antigen sequences; sequence redundancy was reduced at HSSP-values of 0 (corresponding to <33% pairwise sequence identity for long alignments) (27–30). Then, we identified structurally aligned positions that interact with a protein in more than 10% of the complexes of the alignment. We defined the borders of the CDRs through those highly populated positions. Given the CDRs in the aligned antibodies, we transferred their location to the antibody chains of the corresponding sequence–structure family that they represent by structural pairwise alignments using Combinatorial Extension (CE) (31) (Figure 1). Finally, we defined all the residues on the protein surface that are in contact with the residues on the antibody CDRs as antigenic residues.

Content statistics

Epitome currently contains 142 antigens from protein–antibody complex structures with a current total of 10 180 antigenic interactions. A total of 63 of the complexes consist of antigens that are sequence-unique, i.e. 63 are such that no other antigen in the database has a level of sequence similarity

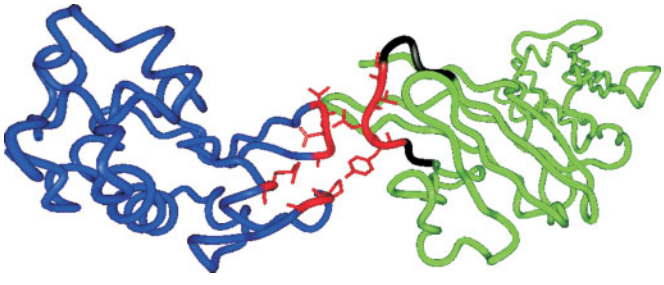


Figure 1. Antigenic residues according to Epitome. Complex structure of quail lysozyme (in blue) and the light chain of an antibody (in green), as taken from PDB ID 1bql (33). The residues that are defined to be in CDR 1 of the light chain according to Kabat definition (13) are colored in black. Residues in red are all the residues that are involved in the interaction according to Epitome. Note that not all of the residues on the antibody surface that are located on 'Kabat' CDR are involved in the antigenic reaction. Additionally, although 1bql antibody chains did not participate in the multiple structure alignment, i.e. the information about the location of the CDR was transferred from a homologous structure, the interaction was correctly identified.

PDB ID	Antigen chain	Antigen residue type	secondary structure	solvent accessibility	Antigen position	Antibody chain type	Antibody Chain	Antibody residue	Antibody position	CDR number
1eg1 Jmol	A	MRYEHID	T	178	366	heavy	H	D	31	1
1eg1 Jmol	A	MRYEHID	T	178	366	heavy	H	Y	32	1
1eg1 Jmol	A	YEHIDHT	S	29	368	heavy	H	Y	33	1
1eg1 Jmol	A	MRYEHID	T	178	366	heavy	H	Y	33	1
1eg1 Jmol	A	RYEHIDH	T	104	367	heavy	H	Y	33	1
1eg1 Jmol	A	MKMRYEH	L	80	364	heavy	H	Y	33	1
1eg1 Jmol	A	TMKMRYE	L	150	363	heavy	H	Y	33	1
1eg1 Jmol	A	KMRYEHI	L	125	365	heavy	H	Y	33	1
1eg1 Jmol	A	MRYEHID	T	178	366	heavy	H	M	34	1
1eg1 Jmol	A	MRYEHID	T	178	366	heavy	H	K	35	1
1eg1 Jmol	A	TMKMRYE	L	150	363	heavy	H	D	50	2
1eg1 Jmol	A	TMKMRYE	L	150	363	heavy	H	N	51	2
1eg1 Jmol	A	MKMRYEH	L	80	364	heavy	H	N	51	2
1eg1 Jmol	A	TMKMRYE	L	150	363	heavy	H	N	53	2
1eg1 Jmol	A	QNAHSMA	S	84	395	heavy	H	N	53	2

Figure 2. Screenshot of a database entry. Each line of the table represents different antigenic interaction, i.e. interaction of a protein surface residue with an antibody surface residue that is located on one of the antibody's 6 CDRs. Note that the search could be performed using any of the table fields and that there is additional link to visualize the interaction using Jmol (http://jmol.sourceforge.net/).

to any other of the 63 that would enable coarse-grained homology modeling.

Input and fields

Epitome users can search for epitopes either by querying the database or by entering a sequence and 'BLASTing' for similar sequences that are stored in the database. The fields that can be queried include one or more of the following: PDB identifier (four-letter code used by the PDB, e.g. 1pdb); Antigen chain ID (PDB identifier for the chain of the antigen, e.g. 1pdb_C), antigen residue type (one letter code for amino acids, e.g. Y corresponds to Tyrosine), antigen residue secondary structure state as defined by DSSP (32) (1 letter code; GHI corresponds to helical structures, EB to strands and TSL to other), antigen residue solvent accessibility (the input is the accessible surface in $Å^2$ as defined by DSSP (32) and the search is on all residues with accessibility values that are bigger or equal to the input value), antigen residue position (the residue number as annotated in the PDB file), heavy/light chain (the interaction involves residues that are located either on the light or the heavy or both chains of the antibody), antibody chain identifier (similar to the antigen chain identifier), antibody residue type (one letter code for amino acids, e.g. C corresponds to Cysteine), antibody residue position in the PDB (the position of the antibody residue that is involved in the interaction as annotated by the PDB) and CDR number (possible values: 1, 2, 3).

Output

Results for database queries are presented as a table that lists all features of the result sets (Figure 2). The antigen results include the residues in the environment of the antigen (highlighted in red). If a user performs a BLAST sequence search against the Epitome database to find PDB structures containing antigens with similar sequences, the output will be all complex structures consisting of proteins with high degree of similarity to the input sequence, the corresponding *E*-value and BLAST score of the pairwise sequence alignments. Additionally, each PSI-BLAST hit contains a link that can trigger another database query.

Updates

Since most Epitome entries were identified using the SCOP database, Epitome updates will follow updates of SCOP, i.e. Epitome will be updated twice a year as soon as SCOP updates its parseable files. Additionally, all the other programs used to create the database are installed locally and can be run automatically.

ACKNOWLEDGEMENTS

Thanks to Jinfeng Liu (Columbia) for computer assistance and to Andrew Kernytsky and Henry Bigelow for helpful comments on the manuscript. Thanks also to the anonymous reviewer for an immensely supportive, helpful and enjoyable critique. This work was supported by the grants RO1-GM64633-01 from the National Institutes of Health (NIH), and RO1-LM07329-01 from the National Library of Medicine (NLM). Last, not least, thanks to Helen Berman (Rutgers), Phil Bourne (UCSD) and their crews for maintaining an excellent PDB, and to all experimentalists who enabled this analysis by making their data publicly available. Funding to pay the Open Access publication charges for this article was provided by the National Library of Medicine (NLM).

Conflict of interest statement. None declared.

REFERENCES

1. Jones,S. and Thornton,J.M. (1997) Prediction of protein-protein interaction sites using patch analysis. *J. Mol. Biol*., **272**, 133–143.
2. Lo Conte,L., Chothia,C. and Janin,J. (1999) The atomic structure of protein–protein recognition sites. *J. Mol. Biol*., **285**, 2177–2198.
3. Chen,R., Mintseris,J., Janin,J. and Weng,Z. (2003) A protein–protein docking benchmark. *Proteins*, **52**, 88–91.
4. Jones,S. and Thornton,J.M. (1997) Analysis of protein-protein interaction sites using surface patches. *J. Mol. Biol*., **272**, 121–132.
5. Jones,S. and Thornton,J.M. (1996) Principles of protein-protein interactions. *Proc. Natl Acad. Sci. USA*, **93**, 13–20.
6. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
7. Davies,D.R. and Cohen,G.H. (1996) Interactions of protein antigens with antibodies. *Proc. Natl Acad. Sci. USA*, **93**, 7–12.
8. MacCallum,R.M., Martin,A.C. and Thornton,J.M. (1996) Antibody–antigen interactions: contact analysis and binding site topography. *J. Mol. Biol*., **262**, 732–745.
9. Collis,A.V., Brouwer,A.P. and Martin,A.C. (2003) Analysis of the antigen combining site: correlations between length and sequence composition of the hypervariable loops and the nature of the antigen. *J. Mol. Biol*., **325**, 337–354.
10. Almagro,J.C. (2004) Identification of differences in the specificity-determining residues of antibodies that recognize antigens of different size: implications for the rational design of antibody repertoires. *J. Mol. Recognit*., **17**, 132–143.
11. Van Regenmortel,M.H.V. (1992) *Structure of Antigens*. CRC Press, Inc., 2000 Corporate Blvd, N.W., Boca Raton, Florida 33431.
12. Blythe,M.J. and Flower,D.R. (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci*., **14**, 246–248.
13. Wu,T.T. and Kabat,E.A. (1970) An analysis of the sequences of the variable regions of Bence Jones proteins and myeloma light chains and their implications for antibody complementarity. *J. Exp. Med*., **132**, 211–250.
14. Johnson,G. and Wu,T.T. (2000) Kabat Database and its applications: 30 years after the first variability plot. *Nucleic Acid Res*., **28**, 214–218.
15. Allcorn,L.C. and Martin,A.C. (2002) SACS—self-maintaining database of antibody crystal structure information. *Bioinformatics*, **18**, 175–181.
16. Chothia,C., Lesk,A.M., Tramontano,A., Levitt,M., Smith-Gill,S.J., Air,G., Sheriff,S., Padlan,E.A., Davies,D. and Tulip,W.R. (1989) Conformations of immunoglobulin hypervariable regions. *Nature*, **342**, 877–883.
17. Al-Lazikani,B., Lesk,A.M. and Chothia,C. (1997) Standard conformations for the canonical structures of immunoglobulins. *J. Mol. Biol*., **273**, 927–948.
18. Saha,S., Bhasin,M. and Raghava,G.P. (2005) Bcipep: a database of B-cell epitopes. *BMC Genomics*, **6**, 79.
19. Peters,B., Sidney,J., Bourne,P., Bui,H.H., Buus,S., Doh,G., Fleri,W., Kronenberg,M., Kubo,R., Lund,O. *et al*. (2005) The design and implementation of the immune epitope database and analysis resource. *Immunogenetics*, **57**, 326–336.
20. Kaas,Q., Ruiz,M. and Lefranc,M.P. (2004) IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Res*., **32**, D208–D210.
21. McSparron,H., Blythe,M.J., Zygouri,C., Doytchinova,I.A. and Flower,D.R. (2003) JenPep: a novel computational information resource for immunobiology and vaccinology. *J. Chem. Inf. Comput. Sci*., **43**, 1276–1287.
22. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., **25**, 3389–3402.
23. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol*., **247**, 536–540.
24. Ofran,Y. and Rost,B. (2003) Analysing six types of protein–protein interfaces. *J. Mol. Biol*., **325**, 377–387.
25. Petrey,D., Xiang,Z., Tang,C.L., Xie,L., Gimpelev,M., Mitros,T., Soto,C.S., Goldsmith-Fischman,S., Kernytsky,A., Schlessinger,A. *et al*. (2003) Using multiple structure alignments, fast model building, and energetic analysis in fold recognition and homology modeling. *Proteins*, **53** (Suppl. 6), 430–435.
26. Petrey,D. and Honig,B. (2003) GRASP2: visualization, surface properties, and electrostatics of macromolecular structures and sequences. *Methods Enzymol*., **374**, 492–509.
27. Mika,S. and Rost,B. (2003) UniqueProt: creating representative protein sequence sets. *Nucleic Acid Res*., **31**, 3789–3791.
28. Rost,B. (1999) Twilight zone of protein sequence alignments. *Protein Eng*., **12**, 85–94.
29. Sander,C.S.R. (1991) Database of homology-derived structures and the structural meaning of sequence alignment. *Proteins*, **9**, 56–68.
30. Schneider,R., de Daruvar,A. and Sander,C. (1997) The HSSP database of protein structure-sequence alignments. *Nucleic Acids Res*., **25**, 226–230.
31. Shindyalov,I.N. and Bourne,P.E. (1998) Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Eng*., **11**, 739–747.
32. Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **12**, 2577–2637.
33. Chacko,S., Silverton,E.W., Smith-Gill,S.J., Davies,D.R., Shick,K.A., Xavier,K.A., Willson,R.C., Jeffrey,P.D., Chang,C.Y., Sieker,L.C. *et al*. (1996) Refined structures of bobwhite quail lysozyme uncomplexed and complexed with the HyHEL-5 Fab fragment. *Proteins*, **26**, 55–65.

BMC Immunology

BioMed Central

Database

CED: a conformational epitope database

Jian Huang*[1,2] and Wataru Honda[1]

Address: [1]Bioinformatics Center, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan and [2]School of Life Science and Technology, University of Electronic Science and Technology of China, China

Email: Jian Huang* - hjian@kuicr.kyoto-u.ac.jp; Wataru Honda - honda@kuicr.kyoto-u.ac.jp

* Corresponding author

Published: 07 April 2006

BMC Immunology 2006, **7**:7 doi:10.1186/1471-2172-7-7

Received: 17 January 2006
Accepted: 07 April 2006

This article is available from: http://www.biomedcentral.com/1471-2172/7/7

Abstract

Background: Antigen epitopes provide valuable information useful for disease prevention, diagnosis, and treatment. Recently, more and more databases focusing on different types of epitopes have become available. Conformational epitopes are an important form of epitope formed by residues that are sequentially discontinuous but close together in three-dimensional space. These epitopes have implicit structural information, making them attractive for both theoretical and applied biomedical research. However, most existing databases focus on linear rather than conformational epitopes.

Description: We describe CED, a special database of well defined conformational epitopes. CED provides a collection of conformational epitopes and related information including the residue make up and location of the epitope, the immunological property of the epitope, the source antigen and corresponding antibody of the epitope. All entries in this database are manually curated from articles published in peer review journals. The database can be browsed or searched through a user-friendly web interface. Most epitopes in CED can also be viewed interactively in the context of their 3D structures. In addition, the entries are also hyperlinked to various databases such as Swiss-Prot, PDB, KEGG and PubMed, providing wide background information.

Conclusion: A conformational epitope database called CED has been developed as an information resource for investigators involved in both theoretical and applied immunology research. It complements other existing specialised epitope databases. The database is freely available at http://web.kuicr.kyoto-u.ac.jp/~ced

Background

Immune cells recognize epitopes (antigenic determinants) rather than entire antigens. Epitopes are the regions of an antigen that are bound by antigen-specific membrane receptors on lymphocytes or to secreted antibodies[1]. T-cells recognize T-cell epitopes, which are usually linear peptides derived from protein antigens and presented by MHC molecules. B-cells and antibodies recognize B-cell epitopes, which can be complete, small chemical compounds or components of larger macromolecules such as nucleotides, lipids, glycans and proteins. Epitopes from macromolecules, especially proteins, are further classified into another two categories. The first, termed linear epitopes, are segments composed of a continuous string of residues along the polymer chain. The second, termed conformational epitopes, are constituted by several sequentially discontinuous segments that are

Table 1: Contents of CED database entries

Entry Field	Description
Conformational epitope ID	Unique identification of the epitope used in CED database.
Constitution and location	Residues that constitute the epitope and their locations.
Epitope immunoproperty	Immunological property of the epitope. For example, any epitope will be considered as neutralizing if the activity of its source antigen is blocked when binding to its corresponding antibody.
Corresponding antibody	Antibody that recognizes and binds this conformational epitope.
Experimental method	Experiment techniques used to identify this epitope, e.g. NMR.
Source antigen	The name of the antigen on which the conformational epitope exists.
Sequence of source antigen	Sequence ID of the source antigen in primary sequence databases such as SwissProt and GenBank; hyperlinked.
Structure of source antigen	PDB ID of the source antigen; hyperlinked
Structure of Ag-Ab complex	PDB ID of the source antigen-antibody complex; hyperlinked.
Chemoproperty of source antigen	Chemical property of source antigen. If the source antigen is an enzyme, it is hyperlinked to KEGG enzyme.
Publication reference	PubMed ID of article that report this epitope; hyperlinked.
Comments	Miscellaneous information of the epitope. It is often about the location difference between the reference and the sequence database.

brought together by the folding of the antigen into its native structure [1,2].

As the molecular basis of immune recognition and the immune response, both kinds of epitope provide valuable information that is useful for disease prevention [3], diagnosis[4,5], and treatment [6-10]. Many databases, focusing on different kinds of epitopes, are available as a large number of epitopes have been identified in the past 20 years. MHCPEP [11], SYFPEITHI[12], FIMM[13], MHCBN[14], EPIMHC[15] are T cell oriented. Bcipep[16] and Epitome[17] are B cell oriented. AntiJen[18] is a multifaceted database with entries on both T cell and B cell epitopes. However, most existing databases focus on linear rather than conformational epitopes.

Despite this, as many as 90% of B-cell epitopes from native proteins are conformational rather than linear [19-22]. In one report, 10 monoclonal antibodies were produced against Helicobacter pylori vacuolating cytotoxin, and all of them proved to react with conformational epitopes [22]. As well as being very common, conformational epitopes may also have functional advantages over linear epitopes. For example, human antibodies directed to conformational epitopes neutralized a wider range of human immunodeficiency virus isolates than human antibodies directed to linear epitopes[23]. A better understanding of conformational epitopes can not only provide useful information for new vaccine design [3] and new diagnostic reagent development[5], but also will be of great value in disease treatment, including against virus infection and cancer [6-10]. Conformational epitopes also have implicit structural information relating to the antigen itself and the mode of binding, which make them attractive to theoretical biology research.

It is money and labour intensive to identify a B cell epitope in detail experimentally. To predict B cell epitopes accurately based on sequence profiling has been a long term goal of many groups. However, a recent report by Blythe et al shows that even the best set of scales and parameters performed only marginally better than random[24]. This suggests that structural approaches and conformational epitope prediction is vital. A conformational epitope prediction server became available recently[25] which will help in conformational epitope discovery, but to predict conformational epitopes accurately is still a very difficult task. For this reason, only several hundred conformational epitopes are currently well defined. We would expect this number to expand rapidly in the future and so it will be necessary to have a database to store and manage all the experimentally well-defined conformational epitopes. Such a specialised conformational epitope database will also be helpful for B cell epitope prediction research.

In this paper, we describe the Conformational Epitope Database (CED). It provides a curated dataset that can be used to evaluate existing epitope prediction methods as well as develop new and better algorithms for prediction. It also complements other existing epitope databases and provides a resource for applied biomedical research in disease prevention, diagnosis, and treatment.

Construction and content

The MySQL relational database management system is used in CED to store, retrieve and manage the data. The web interface between the user and CED are coded in PHP with PEAR database abstraction layer support. All entries in this database are sourced from articles published in peer reviewed journals. Initially, exhaustive queries are made to PubMed and ScienceDirect; returning more than 3000 references that are loaded into an EndNote reference

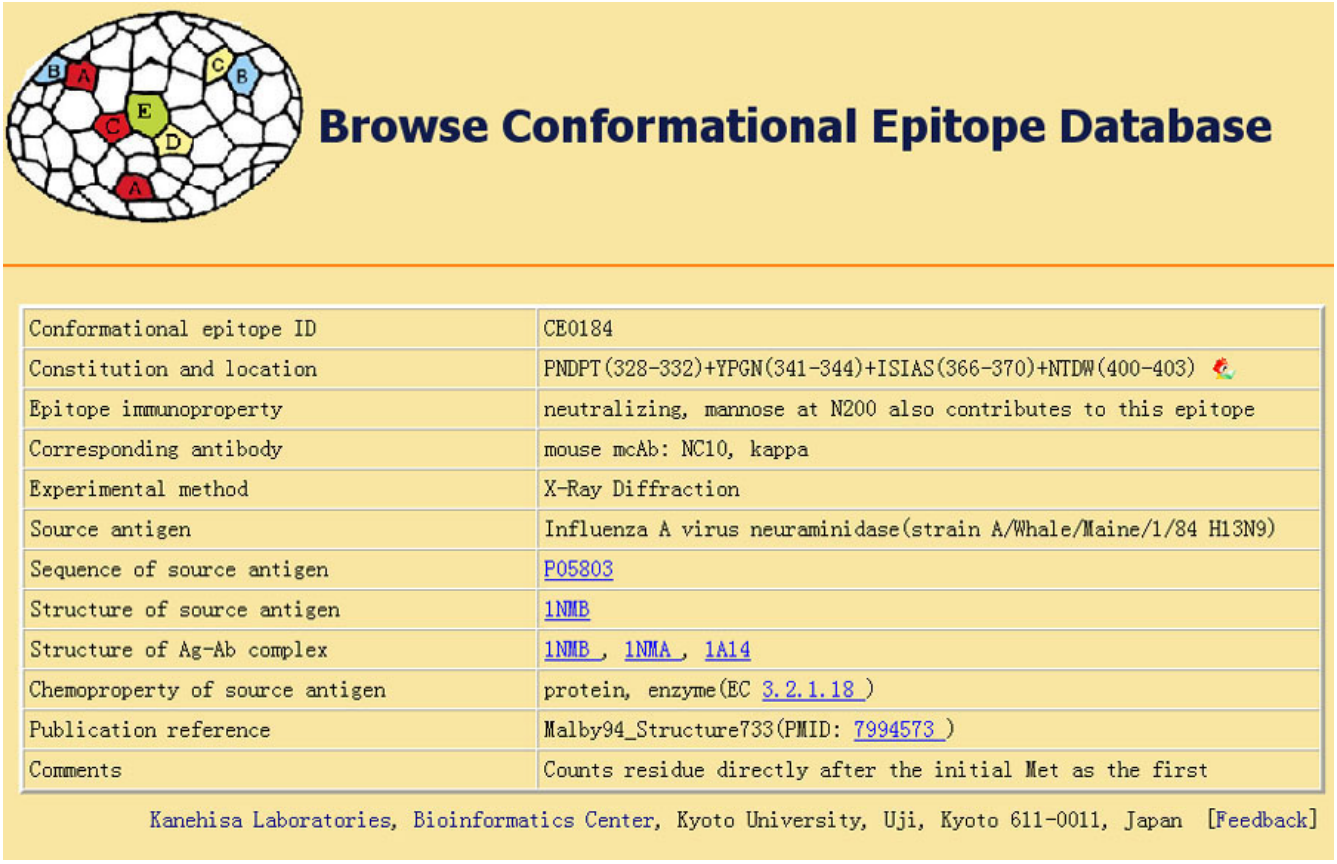


Figure 1
An example entry of CED.

database. The references are then filtered manually to exclude articles that do not define a conformational epitope or where the defined epitope is only at a very low resolution or completeness. The remaining data is checked and entered manually into CED.

Each entry in CED provides information about a given conformational epitope. The information provided includes:

(1) The residue constitution and location of the epitope

(2) The immunological property of the epitope

(3) The antibody that can bind to the epitope

(4) The experimental method used to identify the epitope

(5) The source antigen where the epitope exists

(6) The sequence of the source antigen

(7) The structure of the source antigen

(8) The structure of the antigen-antibody complex

(9) The chemical properties of the source antigen

(10) The references describing the epitope

(11) The comments and miscellaneous information

The full description of the entry fields is given Table 1. An example entry is shown in Figure 1.

Currently, CED has 225 entries. The majority (213) are epitopes from protein antigens, though the database also includes 6 from nucleic acids, 5 from glycans and 1 from lipid. 138 epitopes are from vertebrates, the majority from human (109), the rest of the epitopes are from viruses (54), bacteria (22), invertebrates (7), plants (2) and unspecified (2).

Utility and discussion

Browsing epitopes in the CED

From the homepage of CED, users can find an introduction to conformational epitopes and CED. They can

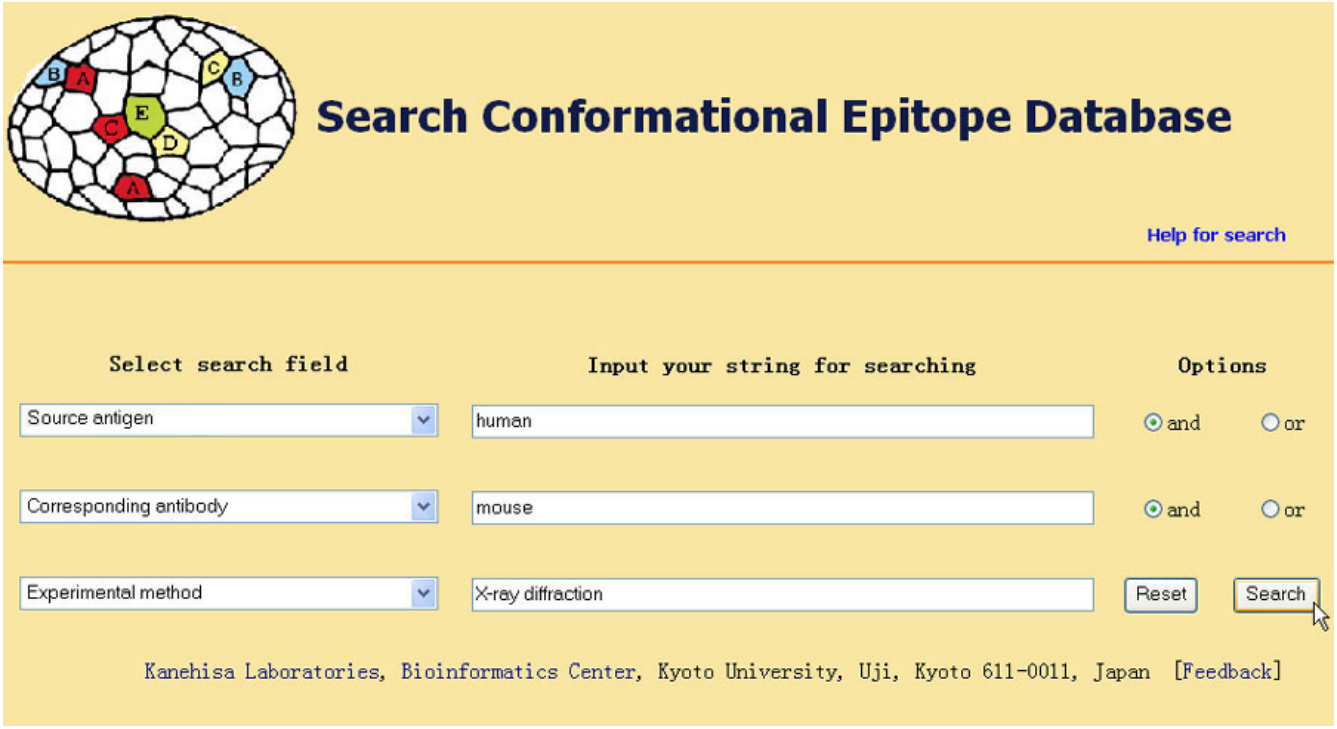


Figure 2
Interface for searching CED and an example search operation.

browse epitope records page by page and entry by entry. When browsing CED, the entries appear in a summary table at first. Only the conformational epitope ID and the constitution and location field are shown, and the data is ordered by the CED identification number (ID). Clicking on an ID opens a new window that displays the information for the selected epitope in detail. An example is shown in Figure 1.

Searching epitopes in the CED

Users can also search CED for specific conformational epitopes. A detailed help page for the search function is provided. CED can be searched by any field in the entry and by any combination of up to three of the fields. There are three pulldown listboxes and two pairs of radio buttons on the search interface. A query is formed by selecting one or more fields from the pulldown listbox and combining them logically by "and" or "or" radio buttons. Key words or strings, such as a partial protein sequence, the name of an antigen, PDB ID of a structure, clone code of a monoclonal antibody, author name or PMID of a reference are entered into the blank text forms. The HTML form parses the criteria into SQL database queries. The initial results returned are formatted into a summary table as described above. Selecting each ID in the summary table opens a new window that displays the detailed information of each epitope.

As an example, if one wanted to retrieve all well defined conformational epitopes from human antigens that are recognized by mouse antibodies and identified by X-ray diffraction, you would first select search field "Source antigen" from the pulldown listbox, input "human" into the corresponding blank text form and then select the "and" button. Then you would select the "Corresponding antibody" search field and input "mouse". Finally, you would select "Experimental method" and input "X-ray diffraction". The operations above will make the search interface appear as shown in Figure 2. After clicking on the "Search" button, records fulfilling the requirements would appear in a new window as shown in Figure 3.

Viewing epitopes in the CED

Most conformational epitopes in CED can be viewed interactively in the context of the antigen-antibody complex, antigen structure or known theoretical model, if they have corresponding PDB structures. The visualization function of CED is powered by Jmol. To view an epitope in the context of its 3D structure can help users identify important structural features and judge if entries defined by non-structural methods are reasonable. When browsing or searching the CED, entries that have PDB structures will have a conspicuous view icon in both the summary table and the entry table. Clicking the view icon will initialize the loading of the Jmol Java applet. By default,

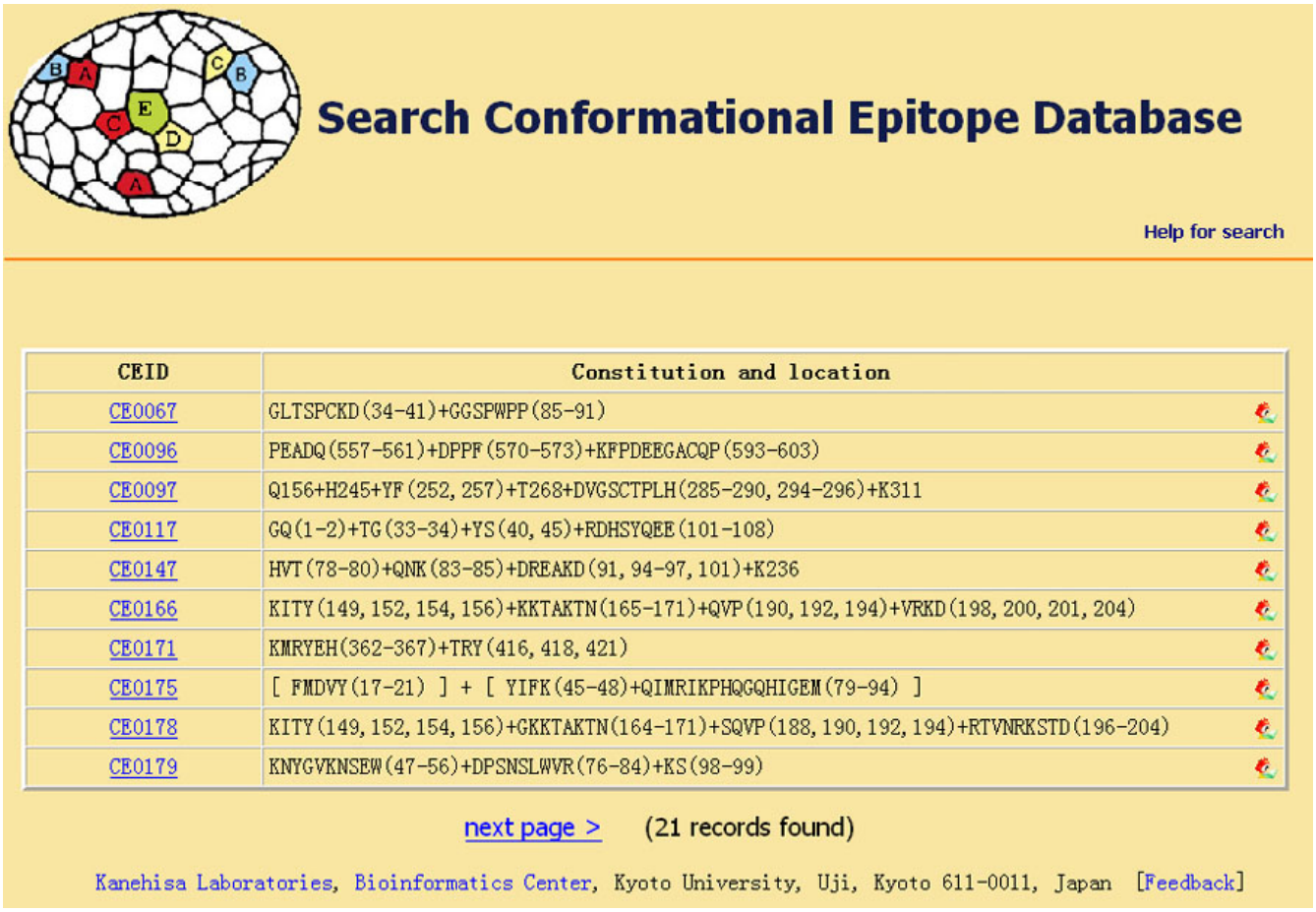


Figure 3
An example search result table.

structures in CED are displayed as backbone colored by secondary structure. After loading, users can turn on or turn off the epitope segments, antigen chains, and antibody chains. When turned on, epitope segments "blink" and are then displayed in spacefill mode colored by CPK. Users can also zoom in, zoom out, move, spin, and rotate the structure, or even measure distance, angle, and dihedral angle. A help page for viewing epitopes can be found through a link on the view page. An example visualisation of CE0096 is shown in Figure 4.

Necessity of building a specialised conformational epitope database

Conformational epitopes can not only provide useful information for new vaccine design [3] and new diagnostic reagent development [5], but also will be of great value in disease treatment, including against virus infection and cancer [6-10]. For example, two conformational epitopes in CED (CE0096 and CE0199) are targets of two FDA approved drugs (Herceptin and Erbitux), which are effective in treating some cancers [6,9]. We need a specialised database to store this kind of useful information. Conformational epitopes also provide structural information relating to the antigen itself, making them attractive to theoretical biology research. A recent research clearly shows the underperformance of existing linear B cell epitope prediction methods [24], suggesting that structural approaches and conformational epitope prediction is vital. Thus a specialised conformational epitope database will be helpful for developing new B cell epitope prediction methods.

Related databases

The existing epitope databases can be classified into four main categories: T cell oriented such as MHCPEP[11], SYFPEITHI[12], FIMM[13], MHCBN[14], EPIMHC[15]; B cell oriented such as Bcipep[16] and Epitome[17]; single pathogenic organism oriented such as the HIV Molecular Immunology Database [26] and HCV Molecular Immunology Database[27] ; multifaceted database such as AntiJen[18] and IEDB[28]. IEDB is still under construction at the time of writing.

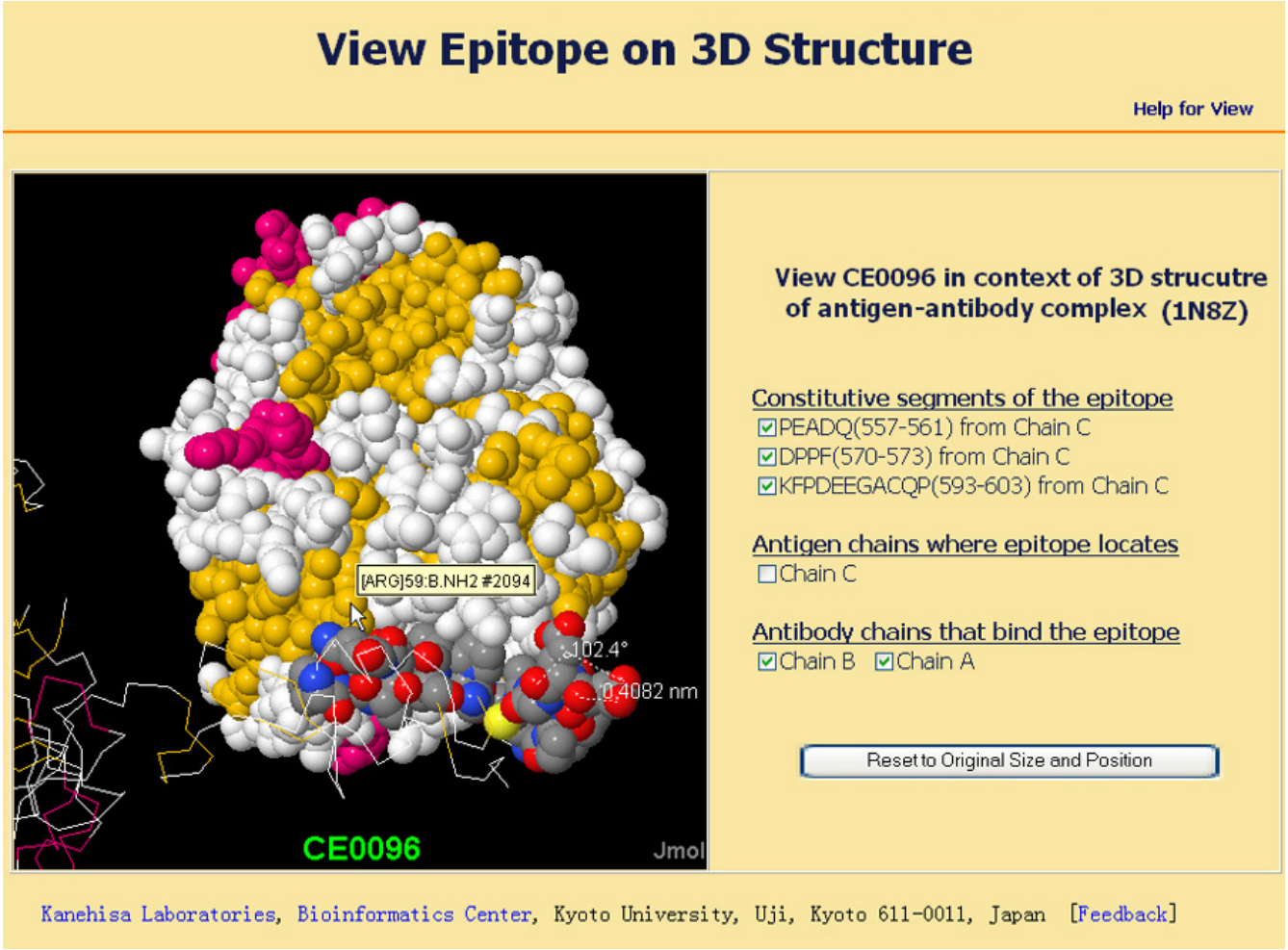


Figure 4
An example visualisation of CE0096. Epitope CE0096 has 3 segments, which are displayed as spacefill colored by CPK; the other part of antigen is visualised as backbone colored by secondary structure. The antibody chains are displayed as spacefill colored by secondary structure.

Compared with most existing epitope databases, the current size of CED is limited. However, most existing epitope databases are focused on collecting linear epitopes; whereas CED only contains conformational epitopes, which are often less well defined and are harder to identify experimentally or theoretically. Many articles report new conformational epitopes, but few are defined completely and precisely enough for inclusion in CED. Taking autoimmune epitopes as an example: In many cases, the nature of the epitopes have often been successively defined and refined from the level of whole cellular organelles (using immunofluorescence methods), to identifying the macromolecule involved (immunoblot, gene expression libraries), to epitope regions (truncated cDNAs, peptide scanning), but few are identified at the contact residue level [29]. Since data quality is vital in bioinformatics research, the aim of CED is to provide high quality, well defined conformational epitope information. So conformational epitopes that are not defined clearly are discarded which limits the database size.

Epitome[17], a very recently released B cell epitope database has a similar size to CED. Epitome collects B cell epitopes only from PDB structures and includes CDR information. In contrast, CED is sourced from the literature and also has conformational epitopes defined by methods other than X-ray diffraction and NMR, such as scanning mutagenesis, overlapping peptides, and phage display. Although CED and Epitome are similar they are derived from different resources and so provide complementary information. We believe that a simple database, such as CED, specialising in conformational epitopes has value in conjunction with these more complex, less specialised, databases. Although our manual search proce-

dure may have missed a few known conformational epitopes, we believe that CED is an essentially complete database of well defined conformational epitopes.

Future work

Firstly, like all other databases, errors will have occurred during the data accumulation phase. We hope users will send their feedback to help us maintain and revise CED in future.

Secondly, we will scan newly published peer review articles for well defined or refined conformational epitopes routinely. New epitopes will be added; and newly refined epitopes will be updated. Due to the rapid progress of techniques in this field, we expect that more and more new conformational epitopes will be identified in the future. Thus, both the quantity and quality of CED entries will increase.

Lastly, we have noticed that a formal ontology of epitopes has been developed and suggested recently[30]. To represent and communicate epitope information systematically and effectively, we will make the next release of CED completely compatible with IEDB's ontology.

Conclusion

A conformational epitope database (CED) has been developed, which can be browsed or searched through a simple user friendly web interface. It is an essentially complete database of well defined conformational epitopes and provides a complement to other existing specialised epitope databases. CED is also hyperlinked to several external databases, providing wide background information for each entry. Though currently relatively small in size, the data in CED provides valuable information for disease prevention, diagnosis, and treatment. Thus, we hope it will be an important information resource for investigators involved in both theoretical and applied immunology research.

Availability and requirements

The database is available at http://web.kuicr.kyoto-u.ac.jp/~ced, suitable for most graphical web browsers support Java applets and JavaScript. We have tested on the Windows, Mac and Linux operating systems.

Abbreviations

CED: Conformational Epitope Database; CDR: Complementarity-Determining Region; HCV: Hepatitis C Virus; HIV: Human Immunodeficiency Virus; HTML: Hypertext Markup Language; IEDB: Immune Epitope Database and Analysis Resource; KEGG: Kyoto Encyclopaedia of Genes and Genomes; MHC: Major Histocompatibility Complex; NCBI: National Center for Biotechnology Information; PEAR: PHP Extension and Application Repository; PDB: Protein Data Bank; PHP: Hypertext Preprocessor; SQL: Structured Query Language.

Authors' contributions

JH and WH extracted the data, compiled the database and wrote the code for the web interface. All authors have read and approved the final manuscript.

Acknowledgements

We thank Professor Kanehisa for supervising the work, our lab colleagues Dr Alex Gutteridge for copyediting the manuscript, Dr Shanfeng Zhu for useful discussion, Mr Koichiro Tonomura and Dr Nobuya Tanaka for help with coding. We also thank Professor Mackay of Monash University for permitting us to adapt their published figure to the logo of CED. The computational resources were provided by the Bioinformatics Center, Institute for Chemical Research, Kyoto University.

References

1. Goldsby RA, Kindt TJ, Kuby J, Osborne BA: **Immunology.** Fifth edition. New York: W. H. Freeman; 2002.
2. Barlow DJ, Edwards MS, Thornton JM: **Continuous and discontinuous protein antigenic determinants.** *Nature* 1986, **322(6081):**747-748.
3. Saxena AK, Singh K, Su HP, Klein MM, Stowers AW, Saul AJ, Long CA, Garboczi DN: **The essential mosquito-stage P25 and P28 proteins from Plasmodium form tile-like triangular prisms.** *Nat Struct Mol Biol* 2005.
4. Matsuo H, Kohno K, Niihara H, Morita E: **Specific IgE Determination to Epitope Peptides of {omega}-5 Gliadin and High Molecular Weight Glutenin Subunit Is a Useful Tool for Diagnosis of Wheat-Dependent Exercise-Induced Anaphylaxis.** *J Immunol* 2005, **175(12):**8116-8122.
5. Tegoni M, Spinelli S, Verhoeyen M, Davis P, Cambillau C: **Crystal structure of a ternary complex between human chorionic gonadotropin (hCG) and two Fv fragments specific for the alpha and beta-subunits.** *J Mol Biol* 1999, **289(5):**1375-1385.
6. Li S, Schmitz KR, Jeffrey PD, Wiltzius JJ, Kussie P, Ferguson KM: **Structural basis for inhibition of the epidermal growth factor receptor by cetuximab.** *Cancer Cell* 2005, **7(4):**301-311.
7. Nybakken GE, Oliphant T, Johnson S, Burke S, Diamond MS, Fremont DH: **Structural basis of West Nile virus neutralization by a therapeutic antibody.** *Nature* 2005, **437(7059):**764-769.
8. Adams GP, Weiner LM: **Monoclonal antibody therapy of cancer.** *Nat Biotechnol* 2005, **23(9):**1147-1157.
9. Cho HS, Mason K, Ramyar KX, Stanley AM, Gabelli SB, Denney DW Jr, Leahy DJ: **Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab.** *Nature* 2003, **421(6924):**756-760.
10. Riemer AB, Kurz H, Klinger M, Scheiner O, Zielinski CC, Jensen-Jarolim E: **Vaccination with cetuximab mimotopes and biological properties of induced anti-epidermal growth factor receptor antibodies.** *J Natl Cancer Inst* 2005, **97(22):**1663-1670.
11. Brusic V, Rudy G, Harrison LC: **MHCPEP, a database of MHC-binding peptides: update 1997.** *Nucleic Acids Res* 1998, **26(1):**368-371.
12. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S: **SYFPEITHI: database for MHC ligands and peptide motifs.** *Immunogenetics* 1999, **50(3–4):**213-219.
13. Schonbach C, Koh JL, Flower DR, Brusic V: **An update on the functional molecular immunology (FIMM) database.** *Appl Bioinformatics* 2005, **4(1):**25-31.
14. Bhasin M, Singh H, Raghava GP: **MHCBN: a comprehensive database of MHC binding and non-binding peptides.** *Bioinformatics* 2003, **19(5):**665-666.
15. Reche PA, Zhang H, Glutting JP, Reinherz EL: **EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology.** *Bioinformatics* 2005, **21(9):**2140-2141.
16. Saha S, Bhasin M, Raghava GP: **Bcipep: a database of B-cell epitopes.** *BMC Genomics* 2005, **6(1):**79.

17. Schlessinger A, Ofran Y, Yachdav G, Rost B: **Epitome: database of structure-inferred antigenic epitopes.** *Nucleic Acids Res* 2006:D777-780.
18. Toseland CP, Clayton DJ, McSparron H, Hemsley SL, Blythe MJ, Paine K, Doytchinova IA, Guan P, Hattotuwagama CK, Flower DR: **AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data.** *Immunome Res* 2005, **1(1):**4.
19. Horsfall AC, Hay FC, Soltys AJ, Jones MG: **Epitope mapping.** *Immunol Today* 1991, **12(7):**211-213.
20. Padlan EA: **X-ray crystallography of antibodies.** *Adv Protein Chem* 1996, **49:**57-133.
21. Benjamin DC: **B-cell epitopes: fact and fiction.** *Adv Exp Med Biol* 1995, **386:**95-108.
22. Vinion-Dubiel AD, McClain MS, Cao P, Mernaugh RL, Cover TL: **Antigenic diversity among Helicobacter pylori vacuolating toxins.** *Infect Immun* 2001, **69(7):**4329-4336.
23. Steimer KS, Scandella CJ, Skiles PV, Haigwood NL: **Neutralization of divergent HIV-1 isolates by conformation-dependent human antibodies to Gp120.** *Science* 1991, **254(5028):**105-108.
24. Blythe MJ, Flower DR: **Benchmarking B cell epitope prediction: underperformance of existing methods.** *Protein Sci* 2005, **14(1):**246-248.
25. Kulkarni-Kale U, Bhosle S, Kolaskar AS: **CEP: a conformational epitope prediction server.** *Nucleic Acids Res* 2005:W168-171.
26. **HIV Molecular Immunology Database** [http://www.hiv.lanl.gov/content/immunology/index.html]
27. Yusim K, Richardson R, Tao N, Dalwani A, Agrawal A, Szinger J, Funkhouser R, Korber B, Kuiken C: **Los alamos hepatitis C immunology database.** *Appl Bioinformatics* 2005, **4(4):**217-225.
28. Peters B, Sidney J, Bourne P, Bui HH, Buus S, Doh G, Fleri W, Kronenberg M, Kubo R, Lund O, Nemazee D, Ponomarenko JV, Sathiamurthy M, Schoenberger SP, Stewart S, Surko P, Way S, Wilson S, Sette A: **The design and implementation of the immune epitope database and analysis resource.** *Immunogenetics* 2005, **57(5):**326-336.
29. Mackay IR, Rowley MJ: **Autoimmune epitopes: autoepitopes.** *Autoimmun Rev* 2004, **3(7–8):**487-492.
30. Sathiamurthy M, Peters B, Bui HH, Sidney J, Mokili J, Wilson SS, Fleri W, McGuinness DL, Bourne PE, Sette A: **An ontology for immune epitopes: application to the design of a broad scope database of immune reactivities.** *Immunome Res* 2005, **1(1):**2.

Hindawi
Journal of Immunology Research
Volume 2017, Article ID 2680160, 14 pages
https://doi.org/10.1155/2017/2680160

Review Article

Fundamentals and Methods for T- and B-Cell Epitope Prediction

Jose L. Sanchez-Trincado, Marta Gomez-Perosanz, and Pedro A. Reche

Laboratory of Immunomedicine, Faculty of Medicine, Complutense University of Madrid, Ave Complutense S/N, 28040 Madrid, Spain

Correspondence should be addressed to Pedro A. Reche; parecheg@med.ucm.es

Received 27 July 2017; Revised 22 November 2017; Accepted 27 November 2017; Published 28 December 2017

Academic Editor: Senthami R. Selvan

Adaptive immunity is mediated by T- and B-cells, which are immune cells capable of developing pathogen-specific memory that confers immunological protection. Memory and effector functions of B- and T-cells are predicated on the recognition through specialized receptors of specific targets (antigens) in pathogens. More specifically, B- and T-cells recognize portions within their cognate antigens known as epitopes. There is great interest in identifying epitopes in antigens for a number of practical reasons, including understanding disease etiology, immune monitoring, developing diagnosis assays, and designing epitope-based vaccines. Epitope identification is costly and time-consuming as it requires experimental screening of large arrays of potential epitope candidates. Fortunately, researchers have developed in silico prediction methods that dramatically reduce the burden associated with epitope mapping by decreasing the list of potential epitope candidates for experimental testing. Here, we analyze aspects of antigen recognition by T- and B-cells that are relevant for epitope prediction. Subsequently, we provide a systematic and inclusive review of the most relevant B- and T-cell epitope prediction methods and tools, paying particular attention to their foundations.

1. Introduction

The immune system is typically divided into two categories, innate and adaptive. Innate immunity involves nonspecific defense mechanisms that act immediately or within hours after a microbe appearance in the body. All multicellular beings exhibit some kind of innate immunity. In contrast, adaptive immunity is only present in vertebrates and it is highly specific. In fact, the adaptive immune system is able to recognize and destroy invading pathogens individually. Moreover, the adaptive immune system remembers the pathogens that fights, acquiring a pathogen-specific long-lasting protective memory that enables stronger attacks each time the pathogen is reencountered [1]. Nonetheless, innate and adaptive immune mechanisms work together and adaptive immunity elicitation is contingent on prior activation of innate immune responses [1].

Adaptive immunity is articulated by lymphocytes, more specifically by B- and T-cells, which are responsible for the humoral and cell-mediated immunity. B- and T-cells do not recognize pathogens as a whole, but molecular components known as antigens. These antigens are recognized by specific receptors present in the cell surface of B- and T-cells. Antigen recognition by these receptors is required to activate B- and T-cells but not enough, as second activation signals stemming from the activation of the innate immune system are also needed. The specificity of the recognition is determined by genetic recombination events that occur during lymphocyte development, which lead to generating millions of different variants of lymphocytes in terms of the antigen-recognizing receptors [1]. Antigen recognition by B- and T-cells differ greatly.

B-cells recognize solvent-exposed antigens through antigen receptors, named as B-cell receptors (BCR), consisting of membrane-bound immunoglobulins, as shown in Figure 1. Upon activation, B-cells differentiate and secrete soluble forms of the immunoglobulins, also known as antibodies, which mediate humoral adaptive immunity.

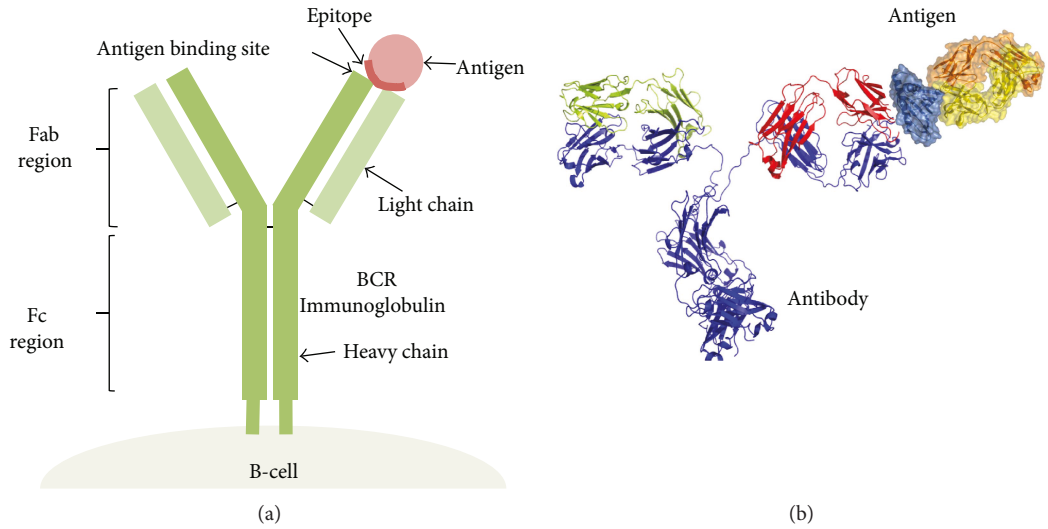


Figure 1: B-cell epitope recognition. B-cell epitopes are solvent-exposed portions of the antigen that bind to secreted and cell-bound immunoglobulins. (a) B-cell receptors encompass cell-bound immunoglobulins, consisting of two heavy chains and two light chains. The different chains and regions are annotated. (b) Molecular representation of the interaction between an antibody and the antigen. Antibodies are secreted immunoglobulins of known specificity.

Antibodies released by B-cells can have different functions that are triggered upon binding their cognate antigens. These functions include neutralizing toxins and pathogens and labeling them for destruction [1].

A B-cell epitope is the antigen portion binding to the immunoglobulin or antibody. These epitopes recognized by B-cells may constitute any exposed solvent region in the antigen and can be of different chemical nature. However, most antigens are proteins and those are the subjects for epitope prediction methods.

On the other hand, T-cells present on their surface a specific receptor known as T-cell receptor (TCR) that enables the recognition of antigens when they are displayed on the surface of antigen-presenting cells (APCs) bound to major histocompatibility complex (MHC) molecules. T-cell epitopes are presented by class I (MHC I) and II (MHC II) MHC molecules that are recognized by two distinct subsets of T-cells, CD8 and CD4 T-cells, respectively (Figure 2). Subsequently, there are CD8 and CD4 T-cell epitopes. CD8 T-cells become cytotoxic T lymphocytes (CTL) following T CD8 epitope recognition. Meanwhile, primed CD4 T-cells become helper (Th) or regulatory (Treg) T-cells [1]. Th cells amplify the immune response, and there are three main subclasses: Th1 (cell-mediated immunity against intracellular pathogens), Th2 (antibody-mediated immunity), and Th17 (inflammatory response and defense against extracellular bacteria) [2].

Identifying epitopes in antigens is of great interest for a number of practical reasons, including understanding disease etiology, immune monitoring, developing diagnosis assays, and designing epitope-based vaccines. B-cell epitopes can be identified by different methods including solving the 3D structure of antigen-antibody complexes, peptide library screening of antibody binding or performing functional assays in which the antigen is mutated and the interaction antibody-antigen is evaluated [3, 4]. On the other hand, experimental determination of T-cell epitopes is carried out using MHC multimers and lymphoproliferation or ELISPOT assays, among others [5, 6]. Traditional epitope identification has depended entirely upon experimental techniques, being costly and time-consuming. Thereby, scientists have developed and implemented epitope prediction methods that facilitate epitope identification and decrease the experimental load associated with it. Here, we will first analyze aspects of antigen recognition by T- and B-cells that are relevant for a better understanding of the topic of epitope prediction. Subsequently, we will provide a systematic and inclusive review of the most important prediction methods and tools, paying particular attention to their foundations and potentials. We will also discuss epitope prediction limitations and ways to overcome them. We will start with T-cell epitopes.

2. T-Cell Epitope Prediction

T-cell epitope prediction aims to identify the shortest peptides within an antigen that are able to stimulate either CD4 or CD8 T-cells [7]. This capacity to stimulate T-cells is called immunogenicity, and it is confirmed in assays requiring synthetic peptides derived from antigens [5, 6]. There are many distinct peptides within antigens and T-cell prediction methods aim to identify those that are immunogenic. T-cell epitope immunogenicity is contingent on three basic steps: (i) antigen processing, (ii) peptide binding to MHC molecules, and (iii) recognition by a cognate TCR. Of these three events, MHC-peptide binding is the most selective one at determining T-cell epitopes [8, 9]. Therefore, prediction of peptide-MHC binding is the main basis to anticipate T-cell epitopes and we will review it next.

2.1. Prediction of Peptide-MHC Binding. MHC I and MHC II molecules have similar 3D-structures with bound peptides sitting in a groove delineated by two α-helices overlying a

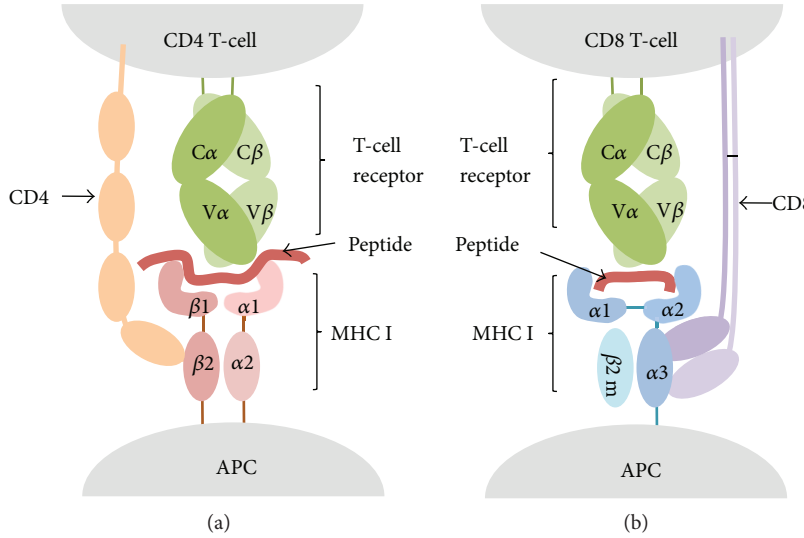


Figure 2: T-cell epitope recognition. T-cell epitopes are peptides derived from antigens and recognized by the T-cell receptor (TCR) when bound to MHC molecules displayed on the cell surface of APCs. (a) CD4 T-cells express the CD4 coreceptor, which binds to MHC II, and recognize peptides presented by MHC II molecules. (b) CD8 T-cells express the CD8 coreceptor, which binds to MHC I, and recognize peptides presented by MHC I molecules.

floor comprised of eight antiparallel β-strands. However, there are also key differences between MHC I and II binding grooves that we must highlight for they condition peptide-binding predictions (Figure 3). The peptide-binding cleft of MHC I molecules is closed as it is made by a single α chain. As a result, MHC I molecules can only bind short peptides ranging from 9 to 11 amino acids, whose N- and C-terminal ends remain pinned to conserved residues of the MHC I molecule through a network of hydrogen bonds [10, 11]. The MHC I peptide-binding groove also contains deep binding pockets with tight physicochemical preferences that facilitate binding predictions. There is a complication however. Peptides that have different sizes and bind to the same MHC I molecule often use alternative binding pockets [12]. Therefore, methods predicting peptide-MHC I binding require a fixed peptide length. However, since most MHC I peptide ligands have 9 residues, it is generally preferable to predict peptides with that size. In contrast, the peptide-binding groove of MHC II molecules is open, allowing the N- and C-terminal ends of a peptide to extend beyond the binding groove [10, 11]. As a result, MHC II-bound peptides vary widely in length (9–22 residues), although only a core of nine residues (peptide-binding core) sits into the MHC II binding groove. Therefore, peptide-MHC II binding prediction methods often target to identify these peptide-binding cores. MHC II molecule binding pockets are also shallower and less demanding than those of MHC I molecules. As a consequence, peptide-binding prediction to MHC II molecules is less accurate than that of MHC I molecules.

Given the relevance of the problem, there are numerous methods to predict peptide-MHC binding. The most relevant with free online use are collected on Table 1. They can be divided in two main categories: data-driven and structure-based methods. Structure-based approaches generally rely on modeling the peptide-MHC structure followed by evaluation of the interaction through methods such as molecular dynamic simulations [8, 13]. Structure-based methods have the great advantage of not needing experimental data. However, they are seldom used as they are computationally intensive and exhibit lower predictive performance than data-driven methods [14].

Data-driven methods for peptide-MHC binding prediction are based on peptide sequences that are known to bind to MHC molecules. These peptide sequences are generally available in specialized epitope databases such as IEDB [15], EPIMHC [16], Antijen [17, 18]. Both MHC I and II binding peptides contain frequently occurring amino acids at particular peptide positions, known as anchor residues. Thereby, prediction of peptide-MHC binding was first approached using sequence motif (SM) reflecting amino acid preferences of MHC molecules at anchor positions [19]. However, it was soon shown that nonanchor residues also contribute to the capacity of a peptide to bind to a given MHC molecule [20, 21]. Subsequently, researchers developed motif matrices (MM), which could evaluate the contribution of each and all peptide positions to the binding with the MHC molecule [22–25]. The most sophisticated form of motif matrices consists of profiles [24–26] that are similar to those used for detecting sequence homology [27]. We would like to remark that motif matrices are often mistaken with quantitative affinity matrices (QAMs) since both produce peptide scores. However, MMs are derived without taking in consideration values of binding affinities and, therefore, resulting peptide scores are not suited to address binding affinity. In contrast, QAMs are trained on peptides and corresponding binding affinities, and aim to predict binding affinity. The first method based on QAMs was developed by Parker et al. [28] (Table 1). Subsequently, various approaches were developed to obtain QAMs from peptide

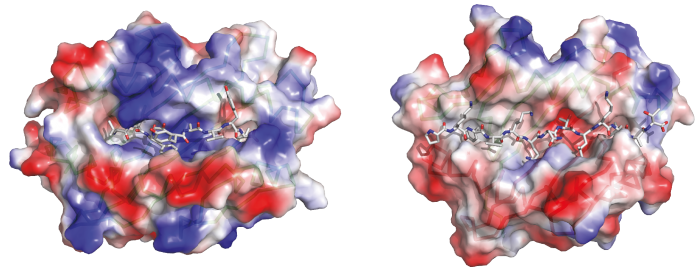


Figure 3: MHC molecule binding groove. The figure depicts the molecular surface as seen by the TCR of representative MHC I and II molecules. Note how the binding groove of the MHC I molecule is closed but that of MHC II is open. As a result, MHC I molecules bind short peptides (8–11 amino acids), while MHC II molecules bind longer peptides (9–22 amino acids). The figure was prepared from PDB files 1QRN (MHC I) and 1FYT (MHC II) using PyMol.

Table 1: Selected T-cell epitope prediction tools available online for free public use.

Tool	URL	Method[1]	MHC class	A	S	T	P	Ref.
EpiDOCK	http://epidock.ddg-pharmfac.net	SB	II	—	—	—	—	[86]
MotifScan	https://www.hiv.lanl.gov/content/immunology/motif_scan/motif_scan	SM	I and II	—	X	—	—	—
Rankpep	http://imed.med.ucm.es/Tools/rankpep.html	MM	I and II	—	—	—	X	[26]
SYFPEITHI	http://www.syfpeithi.de/	MM	I and II	—	—	—	—	[23]
MAPPP	http://www.mpiib-berlin.mpg.de/MAPPP/	MM	I	—	X	—	X	[87]
PREDIVAC	http://predivac.biosci.uq.edu.au/	MM	II	—	—	—	—	[88]
PEPVAC	http://imed.med.ucm.es/PEPVAC/	MM	I	—	X	—	X	[63]
EPISOPT	http://bio.med.ucm.es/episopt.html	MM	I	—	X	—	—	[64]
Vaxign	http://www.violinet.org/vaxign/	MM	I and II	—	—	—	—	[89]
MHCPred	http://www.ddg-pharmfac.net/mhcpred/MHCPred/	QSAR	I and II	X	—	—	—	[34]
EpiTOP	http://www.pharmfac.net/EpiTOP	QSAR	II	X	—	—	—	[90]
BIMAS	https://www-bimas.cit.nih.gov/molbio/hla_bind/	QAM	I	X				[28]
TEPITOPE	http://datamining-iip.fudan.edu.cn/service/TEPITOPEpan/TEPITOPEpan.html	QAM	II	X	—	—	—	[32]
Propred	http://www.imtech.res.in/raghava/propred/	QAM	II	X	X	—	—	[91]
Propred-1	http://www.imtech.res.in/raghava/propred1/	QAM	I	X	X	—	X	[92]
EpiJen	http://www.ddg-pharmfac.net/epijen/EpiJen/EpiJen.htm	QAM	I	X	—	X	X	[82]
IEDB-MHCI	http://tools.immuneepitope.org/mhci/	Combined	I	X	—	—	—	[93]
IEDB-MHCII	http://tools.immuneepitope.org/mhcii/	Combined	II	X	—	—	—	[93]
IL4pred	http://webs.iiitd.edu.in/raghava/il4pred/index.php	SVM	II	—	—	—	—	[67]
MULTIPRED2	http://cvc.dfci.harvard.edu/multipred2/index.php	ANN	I and II	—	X	—	—	[62]
MHC2PRED	http://www.imtech.res.in/raghava/mhc2pred/index.html	SVM	II	—	—	—	—	[38]
NetMHC	http://www.cbs.dtu.dk/services/NetMHC/	ANN	I	X	—	—	—	[49]
NetMHCII	http://www.cbs.dtu.dk/services/NetMHCII/	ANN	II	X	—	—	—	[30]
NetMHCpan	http://www.cbs.dtu.dk/services/NetMHCpan/	ANN	I	X	—	—	—	[54]
NetMHCIIpan	http://www.cbs.dtu.dk/services/NetMHCIIpan/	ANN	II	X	—	—	—	[55]
nHLApred	http://www.imtech.res.in/raghava/nhlapred/	ANN	I	—	—	—	X	[94]
SVMHC	http://abi.inf.uni-tuebingen.de/Services/SVMHC/	SVM	I and II	—	—	—	—	[95]
SVRMHC	http://us.accurascience.com/SVRMHCdb/	SVM	I and II	X	—	—	—	[46]
NetCTL	http://www.cbs.dtu.dk/services/NetCTL/	ANN	I	X	X	X	X	[83]
WAPP	https://abi.inf.uni-tuebingen.de/Services/WAPP/index_html	SVM	I	—	—	X	X	[37]

[1]Method used for prediction of peptide-MHC binding. Keys for methods: SM: sequence motif; SB: structure-based; MM: motif matrix; QAM: quantitative affinity matrix; SVM: support vector machine; ANN: artificial neural network; QSAR: quantitative structure-activity relationship model; combined: tool uses different methods including ANN and QAM, selecting the more appropriate method for each distinct MHC molecule. The table also indicates whether the tools predict quantitative binding affinity (A), supertypes (S), TAP binding (T), and proteasomal cleavage (P); marked with an X in the affirmative case.

affinity data and predict peptide binding to MHC I and II molecules [29–32].

QAMs and motif matrices assume an independent contribution of peptide side chains to the binding. This assumption is well supported by experimental data but there is also evidence that neighboring peptide residues interfere with others [33]. To account for those interferences, researchers introduced quantitative structure-activity relationship (QSAR) additive models wherein the binding affinity of peptides to MHC is computed as the sum of amino acid contributions at each position plus the contribution of adjacent side chain interactions [34]. However, machine learning (ML) is the most popular and robust approach introduced to deal with the nonlinearity of peptide-MHC binding data [8]. Researchers have used ML for two distinct problems: the discrimination of MHC binders from nonbinders and the prediction of binding affinity of peptides to MHC molecules.

For developing discrimination models, ML algorithms are trained on data sets consisting of peptides that either bind or do not bind to MHC molecules. Relevant examples of ML-based discrimination models are those based on artificial neural networks (ANNs) [35, 36], support vector machines (SVMs) [37–39], decision trees (DTs) [40, 41], and Hidden Markov models (HMMs), which can also cope with nonlinear data and have been used to discriminate peptides binding to MHC molecules. However, unlike other ML algorithms, they have to be trained only on positive data. Three types of HMMs have been used to predict MHC-peptide binding: fully connected HMMs [42], structure-optimized HMMs [43], and profile HMMs [43, 44]. Of these, only fully connected HMMs (fcHMMs) and structure-optimized HMMs (soHMMs) can recognize different patterns in the peptide binders. In fact, profile HMMs that are derived from sets of ungapped alignments (the case for peptides binding to MHC) are nearly identical to profile matrices [45] (Table 1).

With regard to predicting binding affinity, ML algorithms are trained on datasets consisting of peptides with known affinity to MHC molecules. Both SVMs and ANNs have been used for such purpose. SVMs were first applied to predict peptide-binding affinity to MHC I molecules [46] and later to MHC II molecules [47] (Table 1). Likewise, ANNs were also applied first to the prediction of peptide binding to MHC I [48, 49] and later to MHC II molecules [50] (Table 1). Benchmarking of peptide-MHC binding prediction methods appears to indicate that those based on ANNs are superior to those based on QAMs and MMs. However, the differences between the distinct methods are marginal and vary for different MHC molecules [51]. Moreover, it has been shown that the performance of peptide-MHC predictions is improved by combining several methods and providing consensus predictions [52].

A major complication for predicting T-cell epitopes through peptide-MHC binding models is MHC polymorphism. In humans, MHC molecules are known as human leukocyte antigens (HLAs), and there are hundreds of allelic variants of class I (HLA I) and class II (HLA II) molecules. These HLA allelic variants bind distinct sets of peptides [53] and require specific models for predicting peptide-MHC binding. However, peptide-binding data is only available for a minority of HLA molecules. To overcome this limitation, some researchers have developed pan-MHC-specific methods by training ANNs on input data combining MHC residues that contact the peptide with peptide-binding affinity that are capable of predicting peptide-binding affinities to uncharacterized HLA alleles [54, 55].

HLA polymorphism also hampers the development of worldwide covering T-cell epitope-based vaccines as HLA variants are expressed at vastly variable frequencies in different ethnic groups [56]. Interestingly, different HLA molecules can also bind similar sets of peptides [57, 58] and researchers have devised methods to cluster them in groups, known as HLA supertypes, consisting of HLA alleles with similar peptide-binding specificities [59–61]. The HLA-A2, HLA-A3, and HLA-B7 are relevant examples of supertypes; 88% of the population expresses at least an allele included in these supertypes [25, 57, 58]. Identification of promiscuous peptide-binding to HLA supertypes enables the development of T-cell epitope vaccines with high-population coverage using a limited number of peptides. Currently, several web-based methods allow the prediction of promiscuous peptide-binding to HLA supertypes for epitope vaccine design including MULTIPRED [62] and PEPVAC [63] (Table 1). A method to identify promiscuous peptide-binding beyond HLA supertypes was developed and implemented by Molero-Abraham et al. [64] with the name of EPISOPT. EPISOPT predicts HLA I presentation profiles of individual peptides regardless of supertypes and identifies epitope combinations providing a wider population protection coverage.

Prediction of peptide binding to MHC II molecules readily discriminate CD4 T-cell epitopes, but cannot tell their ability to activate the response of specific CD4 T-cell subsets (e.g., Th1, Th2, and Treg). However, there is evidence that some CD4 T-cell epitopes appear to stimulate specific subsets of Th cells [65, 66]. Distinguishing the ability of MHC II-restricted epitopes to elicit distinct responses is clearly relevant for epitope vaccine development and has prompted researchers' attention. A relevant example is the work by Dhanda et al. [67] who generated classifiers capable of predicting potential peptide inducers of interleukin 4 (IL-4) secretion, typical of Th2 cells, by training SVM models on experimentally validated IL4 inducing and noninducing MHC class II binders (Table 1).

2.2. Prediction of Antigen Processing and Integration with Peptide-MHC Binding Prediction. Antigen processing shapes the peptide repertoire available for MHC binding and is a limiting step determining T-cell epitope immunogenicity [68]. Subsequently, computational modeling of the antigen processing pathway provides a mean to enhance T-cell epitope predictions. Antigen presentation by MHC I and II molecules proceed by two different pathways. MHC II molecules present peptide antigens derived from endocyted antigens that are degraded and loaded onto the MHC II molecule in endosomal compartments [69]. Class II antigen degradation is poorly understood, and there is lack of good prediction algorithms yet [70]. In contrast, MHC I molecules

present peptides derived mainly from antigens degraded in the cytosol. The resulting peptide antigens are then transported to the endoplasmic reticulum by TAP where they are loaded onto nascent MHC I molecules [69] (Figure 4). Prior to loading, peptides often undergo trimming by ERAAP N-terminal amino peptidases [71].

Proteasomal cleavage and peptide-binding to TAP have been studied in detail and there are computational methods that predict both processes. Proteasomal cleavage prediction models have been derived from peptide fragments generated *in vitro* by human constitutive proteasomes [72, 73] and from sets of MHC I-restricted ligands mapped onto their source proteins [74–76]. On the other hand, TAP binding prediction methods have been developed by training different algorithms on peptides of known affinity to TAP [77–80]. Combination of proteasomal cleavage and peptide-binding to TAP with peptide-MHC binding predictions increases T-cell epitope predictive rate in comparison to just peptide-binding to MHC I [37, 77, 81–83]. Subsequently, researchers have developed resources to predict CD8 T-cell epitopes through multistep approaches integrating proteasomal cleavage, TAP transport, and peptide-binding to MHC molecules [26, 37, 82–85] (Table 1).

3. Prediction of B-Cell Epitopes

B-cell epitope prediction aims to facilitate B-cell epitope identification with the practical purpose of replacing the antigen for antibody production or for carrying structure-function studies. Any solvent-exposed region in the antigen can be subject of recognition by antibodies. Nonetheless, B-cell epitopes can be divided in two main groups: linear and conformational (Figure 5). Linear B-cell epitopes consist of sequential residues, peptides, whereas conformational B-cell epitopes consist of patches of solvent-exposed atoms from residues that are not necessarily sequential (Figure 5). Therefore, linear and conformational B-cell epitopes are also known as continuous and discontinuous B-cell epitopes, respectively. Antibodies recognizing linear B-cell epitopes can recognize denatured antigens, while denaturing the antigen results in loss of recognition for conformational B-cell epitopes. Most B-cell epitopes (approximately a 90%) are conformational and, in fact, only a minority of native antigens contains linear B-cell epitopes [3]. We will review both, prediction of linear and conformational B-cell epitopes.

3.1. Prediction of Linear B-Cell Epitopes. Linear B-cell epitopes consist of peptides which can readily be used to replace antigens for immunizations and antibody production. Therefore, despite being a minority, prediction of linear B-cell epitopes have received major attention. Linear B-cell epitopes are predicted from the primary sequence of antigens using sequence-based methods. Early computational methods for the prediction of B-cell epitopes were based on simple amino acid propensity scales depicting physicochemical features of B-cellepitopes. For example, Hopp and Wood applied residue hydrophilicity calculations for B-cell epitope prediction [96, 97] on the assumption that hydrophilic regions are predominantly located on the protein surface and are potentially antigenic. We know now, however, that protein surfaces contain roughly the same number of hydrophilic and hydrophobic residues [98]. Other amino acid propensity scales introduced for B-cell epitope prediction are based on flexibility [99], surface accessibility [100], and β-turn propensity [101]. Current available bioinformatics tools to predict linear B-cell epitopes using propensity scales include PREDITOP [102] and PEOPLE [103] (Table 2). PREDITOP [102] uses a multiparametric algorithm based on hydrophilicity, accessibility, flexibility, and secondary structure properties of the amino acids. PEOPLE [103] uses the same parameters and in addition includes the assessment of β-turns. A related method to predict B-cell epitopes was introduced by Kolaskar and Tongaonkar [104], consisting on a simple antigenicity scale derived from physicochemical properties and frequencies of amino acids in experimentally determined B-cell epitopes. This index is perhaps the most popular antigenic scale for B-cell epitope prediction, and it is actually implemented by GCG [105] and EMBOSS [106] packages. Comparative evaluations of propensity scales carried out in a dataset of 85 linear B-cell epitopes showed that most propensity scales predicted between 50 and 70% of B-cell epitopes, with the β-turn scale reaching the best values [101, 107]. It has also been shown that combining the different scales does not appear to improve predictions [102, 108]. Moreover, Blythe and Flower [109] demonstrated that single-scale amino acid propensity scales are not reliable to predict epitope location.

The poor performance of amino acid scales for the prediction of linear B-cell epitopes prompted the introduction of machine learning- (ML-) based methods (Table 2). These methods are developed by training ML algorithms to distinguish experimental B-cell epitopes from non-B-cell epitopes. Prior to training, B-cell epitopes are translated into feature vectors capturing selected properties, such as those given by different propensity scales. Relevant examples of B-cell epitope prediction methods based on ML include BepiPred [110], ABCpred [111], LBtope [112], BCPREDS [113], and SVMtrip [114]. Datasets, training features, and algorithms used for developing these methods differ. BepiPred is based on random forests trained on B-cell epitopes obtained from 3D-structures of antigen-antibody complexes [110]. Both BCPREDS [113] and SVMtrip [114] are based on support vector machines (SVM) but while BCPREDS was trained using various string kernels that eliminate the need for representing the sequence into length-fixed feature vectors, SMVtrip was trained on length-fixed tripeptide composition vectors. ABCpred and LBtope methods consist on artificial neural networks (ANNs) trained on similar positive data, B-cell epitopes, but differ on negative data, non-B-cell epitopes. Negative data used for training ABCpred consisted on random peptides while negative data used for LBtope was based on experimentally validated non-B-cell epitopes form IEDB [15]. In general, B-cell epitope prediction methods employing ML-algorithm are reported to outperform those based on amino acid propensity scales. Nevertheless, some authors have reported that ML algorithms show little improvement over single-scale-based methods [115].

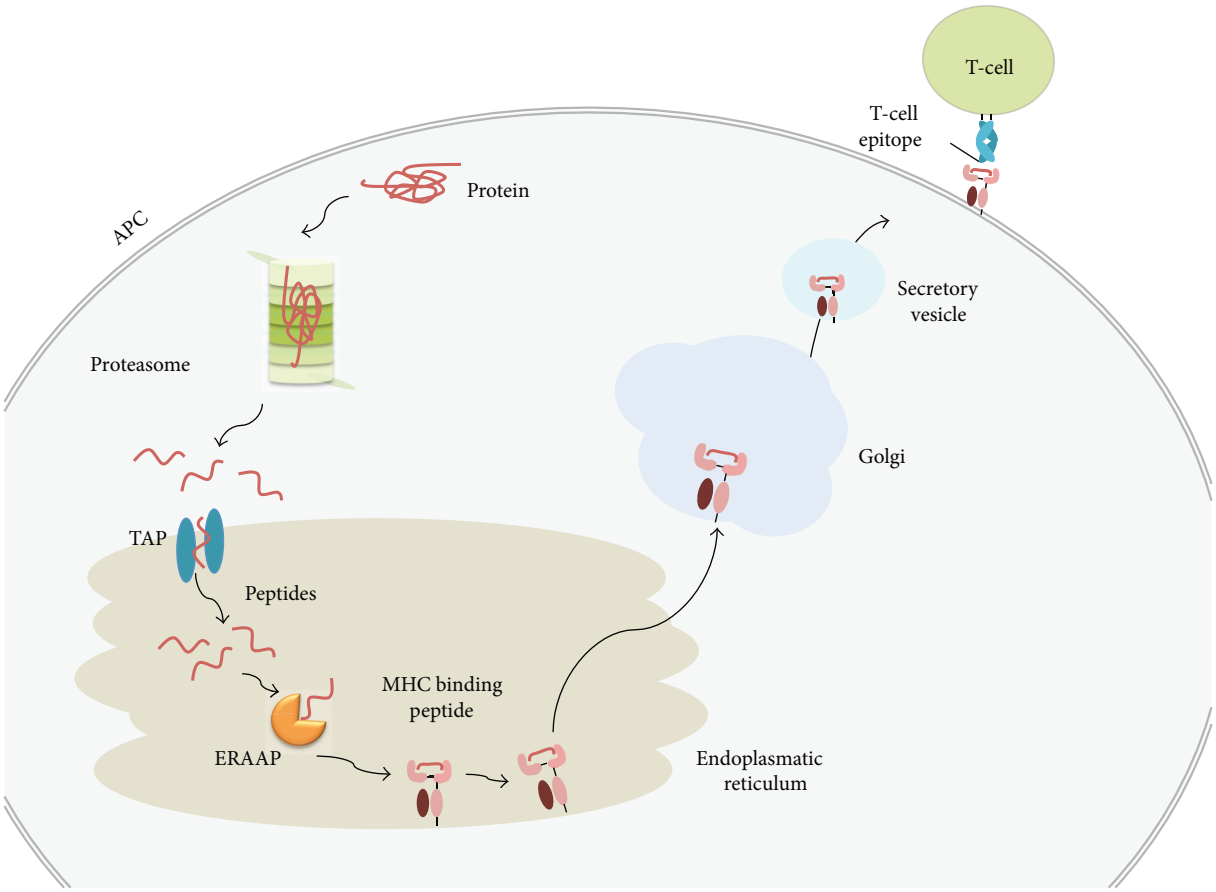


Figure 4: Class I antigen processing. The figure depicts the major steps involved in antigen presentation by MHC I molecules. Proteins are degraded by the proteasome and peptide fragments transported to the endoplasmic reticulum (ER) by TAP where they are loaded onto nascent MHC I molecules. TAP transports peptides ranging from 8 to 16 amino acids. Long peptides cannot bind MHC I molecules but often become suitable for binding after N-terminal trimming by ERAAP.

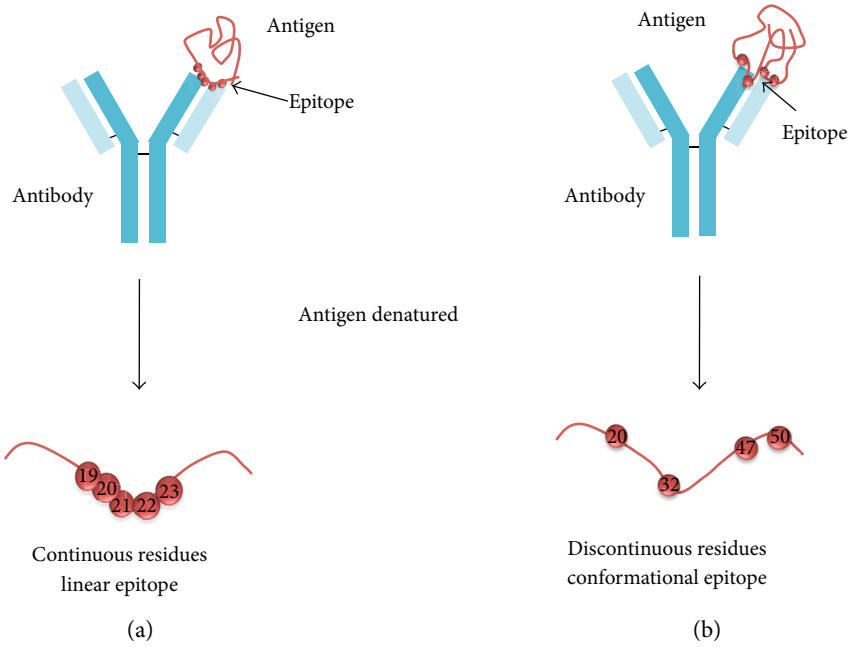


Figure 5: Linear and conformational B-cell epitopes. Linear B-cell epitopes (a) are composed of sequential/continuous residues, while conformational B-cell epitopes (b) contain scattered/discontinuous residues along the sequence.

Table 2: Selected B-cell epitope prediction methods available for free online use.

Tool	Method	Server (URL)	Ref.
Linear B cell epitope			
PEOPLE	Propensity scale method	http://www.iedb.org/	[103]
BepiPred	ML (DT)	http://www.cbs.dtu.dk/services/BepiPred/	[110]
ABCpred	ML (ANN)	http://www.imtech.res.in/raghava/abcpred/	[111]
LBtope	ML (ANN)	http://www.imtech.res.in/raghava/lbtope/	[112]
BCPREDS	ML (SVM)	http://ailab.ist.psu.edu/bcpred/	[113]
SVMtrip	ML (SVM)	http://sysbio.unl.edu/SVMTriP/prediction.php	[114]
Conformational B-cell epitope			
CEP	Structure-based method (solvent accessibility)	http://bioinfo.ernet.in/cep.htm	[118]
DiscoTope	Structure-based method (surface accessibility and propensity amino acid score)	http://tools.iedb.org/discotope/	[119]
ElliPro	Structure-based method (geometrical properties)	http://tools.iedb.org/ellipro/	[121]
PEPITO	Structure-based method (physicochemical properties and geometrical structure)	http://pepito.proteomics.ics.uci.edu/	[122]
SEPPA	Structure-based method (physicochemical properties and geometrical structure)	http://lifecenter.sgst.cn/seppa/	[123]
EPITOPIA	Structure-based method (ML-naïve Bayes)	http://epitopia.tau.ac.il/	[125]
EPSVR	Structure-based method (ML-SVR)	http://sysbio.unl.edu/EPSVR/	[126]
EPIPRED	Structure-based method (ASEP, Docking)	http://opig.stats.ox.ac.uk/webapps/sabdab-sabpred/EpiPred.php	[129]
PEASE	Structure-based method (ASEP, ML)	http://www.ofranlab.org/PEASE	[130]
MIMOX	Mimotope	http://immunet.cn/mimox/helps.html	[131]
PEPITOPE	Mimotope	http://pepitope.tau.ac.il/	[132]
EpiSearch	Mimotope	http://curie.utmb.edu/episearch.html	[133]
MIMOPRO	Mimotope	http://informatics.nenu.edu.cn/MimoPro	[134]
CBTOPE	Sequence based (SVM)	http://www.imtech.res.in/raghava/cbtope/submit.php	[136]

Antibodies elicited in the course of an immune response are generally of a given isotype that determines their biological function. A recent advance in B-cell epitope prediction is the development of a method by Gupta et al. [116] that allows the identification of B-cell epitopes capable of inducing specific class of antibodies. This method is based on SMVs trained on a dataset that includes linear B-cell epitopes known to induce IgG, IgE, and IgA antibodies.

3.2. Prediction of Conformational B-Cell Epitopes. Most B-cell epitopes are conformational and yet, prediction of conformational B-cell epitopes has lagged behind that of linear B-cell epitopes. There are two main practical reasons for that. First of all, prediction of conformational B-cell epitopes generally requires the knowledge of protein three-dimensional (3D) structure and this information is only available for a fraction of proteins [117]. Secondly, isolating conformational B-cell epitopes from their protein context for selective antibody production is a difficult task that requires suitable scaffolds for epitope grafting. Thereby, prediction of conformational B-cell prediction is currently of little relevance for epitope vaccine design and antibody-based technologies. Nonetheless, prediction of conformational B-cell epitopes is interesting for carrying structure-function studies involving antibody-antigen interactions.

There are several available methods to predict conformational B-cell epitopes (Table 2). The first to be introduced was CEP [118], which relied almost entirely on predicting patches of solvent-exposed residues. It was followed by DiscoTope [119], which, in addition to solvent accessibility, considered amino acid statistics and spatial information to predict conformational B-cell epitopes. An independent evaluation of these two methods using a benchmark dataset of 59 conformational epitopes revealed that they did not exceed a 40% of precision and a 46% of recall [120]. Subsequently, more methods were developed, like ElliPro [121] that aims to identify protruding regions in antigen surfaces and PEPITO [122] and SEPPA [123] that combine single physicochemical properties of amino acids and geometrical structure properties. The reported area under the curve (AUC) of these methods is around 0.7, which is indicative of a poor discrimination capacity yet better than random. Though, in an independent evaluation, SEPPA reached an AUC of 0.62 while all the mentioned methods had an AUC around 0.5 [124]. ML has also been applied to predict conformational B-cell epitopes in 3D-structures. Relevant examples include EPITOPIA [125] and EPSVR [126] which are based on naïve Bayes classifiers and support vector regressions, respectively, trained on feature vectors combining different scores. The reported AUC of these two methods is around 0.6.

The above methods for conformational B-cell epitope prediction identify generic antigenic regions regardless of antibodies, which are ignored [127]. However, there are also methods for antibody-specific epitope prediction. This approach was pioneered by Soga et al. [128] who defined an antibody-specific epitope propensity (ASEP) index after analyzing the interfaces of antigen-antibody 3D-structures. Using this index, they developed a novel method for predicting epitope residues in individual antibodies that worked by narrowing down candidate epitope residues predicted by conventional methods. More recently, Krawczyk et al. [129] developed EpiPred, a method that uses a docking-like approach to match up antibody and antigen structures, thus identifying epitope regions on the antigen. A similar approach is used by PEASE [130], adding that this method utilizes the sequence of the antibody and the 3D-structure of the antigen. Briefly, for each pair of antibody sequence and antigen structure, PEASE uses a machine learning model trained on properties from 120 antibody-antigen complexes to identify pair combination of residues from complementarity-determining regions (CDRs) of the antibody and the antigen that are likely to interact.

Another approach to identify conformational B-cell epitopes in a protein with a known 3D-structure is through mimotope-based methods. Mimotopes are peptides selected from randomized peptide libraries for their ability to bind to an antibody raised against a native antigen. Mimotope-based methods require to input antibody affinity-selected peptides and the 3D-structure of the selected antigen. Examples of bioinformatics tools for conformational B-cell epitope prediction using mimotopes include MIMOX [131], PEPITOPE [132], EPISEARCH [133], MIMOPRO [134], and PEPMAPPER [135] (Table 2).

As remarked before, methods for conformational B-cell epitope prediction generally require the 3D-structure of the antigen. Exceptionally, however, Ansari and Raghava [136] developed a method (CBTOPE) for the identification of conformational B-cell epitope from the primary sequence of the antigen. CBTOPE is based on SVM and trained on physicochemical and sequence-derived features of conformational B-cell epitopes. CBTOPE reported accuracy was 86.6% in crossvalidation experiments.

4. Concluding Remarks

Currently, T-cell epitope prediction is more advanced and reliable than that of B-cell prediction. However, while it is possible to confirm experimentally the predicted binding to MHC molecules of most peptides predicted, only ~10% of those are shown to be immunogenic (able to elicit a T-cell response) [68]. Such a low T-cell epitope discovery rate is due to the fact that we do not have adequate models for predicting antigen processing yet [68]. The economic toll of low T-cell epitope discovery rate can be overcome, at least in part, by prioritizing protein antigens for epitope prediction [137–139]. For T-cell epitope vaccine development, researchers can also resort to experimentally known T-cell epitopes, available in epitope databases, selecting through immunoinformatics those that provide maximum population protection coverage [64, 140, 141]. In any case, T-cell epitope prediction remains an integral part of T-cell epitope mapping approaches. In contrast, B-cell epitope prediction utility is currently much more limited. There are several reasons to that. First of all, prediction of B-cell epitopes is still unreliable for both linear and conformational B-cell epitopes. Secondly, linear B-cell epitopes do usually elicit antibodies that do not crossreact with native antigens. Third, the great majority of B-cell epitopes are conformational and yet predicting conformational epitopes have few applications, as they cannot be isolated from their protein context. Under this scenario, the key is not only to improve current methods for B-cell epitope prediction but also to develop novel approaches and platforms for epitope grafting onto suitable scaffolds capable of replacing the native antigen.

To conclude, we wish to make two final remarks that are relevant for epitope vaccine design. First of all, it is that epitope prediction methods can provide potential epitopes from any given protein query but not all the antigens are equally relevant for vaccine development. Therefore, researchers have also developed tools to identify vaccine candidate antigens [142, 143], those likely to induce protective immunity, which can then be targeted for epitope prediction and epitope vaccine design. Second, it should be borne in mind that epitope peptides exhibit little immunogenicity and need to be used in combination with adjuvants, which increase immunogenicity by inducing strong innate immune responses that enable adaptive immunity [144–146]. Consequently, the discovery of new adjuvants is particularly relevant for epitope-based vaccines [146] and to that end, Nagpal et al. [147] developed a pioneered method that can predict the immunomodulatory activity of RNA sequences.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Jose L. Sanchez-Trincado and Marta Gomez-Perosanz contributed equally to this work.

Acknowledgments

The authors wish to thank *Inmunotek*, *SL* and the Spanish Department of Science at MINECO for supporting the Immunomedicine group research through Grants SAF2006: 07879, SAF2009:08301, and BIO2014:54164-R to Pedro A. Reche. The authors also wish to thank Dr. Esther M. Lafuente for critical reading and corrections.

References

[1] W. E. Paul, *Fundamental Immunology*, Lippincott Williams & Wilkins, 2012.

[2] B. Sun and Y. Zhang, "Overview of orchestration of CD4+ T cell subsets in immune responses," *Advances in Experimental Medicine and Biology*, vol. 841, pp. 1–13, 2014.

[3] M. H. Van Regenmortel, "What is a B-cell epitope?," *Methods in Molecular Biology*, vol. 524, pp. 3–20, 2009.

[4] J. Ponomarenko and M. Van Regenmortel, "B-cell epitope prediction," in *Structural Bioinformatics*, pp. 849–879, John Wiley & Sons, Inc, 2009.

[5] T. A. Ahmad, A. E. Eweida, and L. H. El-Sayed, "T-cell epitope mapping for the design of powerful vaccines," *Vaccine Reports*, vol. 6, pp. 13–22, 2016.

[6] L. Malherbe, "T-cell epitope mapping," *Annals of Allergy, Asthma & Immunology*, vol. 103, no. 1, pp. 76–79, 2009.

[7] R. K. Ahmed and M. J. Maeurer, "T-cell epitope mapping," *Methods in Molecular Biology*, vol. 524, pp. 427–438, 2009.

[8] E. M. Lafuente and P. A. Reche, "Prediction of MHC-peptide binding: a systematic and comprehensive overview," *Current Pharmaceutical Design*, vol. 15, no. 28, pp. 3209–3220, 2009.

[9] P. E. Jensen, "Recent advances in antigen processing and presentation," *Nature Immunology*, vol. 8, no. 10, pp. 1041–1048, 2007.

[10] L. J. Stern and D. C. Wiley, "Antigenic peptide binding by class I and class II histocompatibility proteins," *Structure*, vol. 2, no. 4, pp. 245–251, 1994.

[11] D. R. Madden, "The three-dimensional structure of peptide-MHC complexes," *Annual Review of Immunology*, vol. 13, no. 1, pp. 587–622, 1995.

[12] D. R. Madden, D. N. Garboczi, and D. C. Wiley, "The antigenic identity of peptide-MHC complexes: a comparison of the conformations of five viral peptides presented by HLA-A2," *Cell*, vol. 75, no. 4, pp. 693–708, 1993.

[13] D. V. Desai and U. Kulkarni-Kale, "T-cell epitope prediction methods: an overview," *Methods in Molecular Biology*, vol. 1184, pp. 333–364, 2014.

[14] A. Patronov and I. Doytchinova, "T-cell epitope vaccine design by immunoinformatics," *Open Biology*, vol. 3, no. 1, article 120139, 2013.

[15] R. Vita, J. A. Overton, J. A. Greenbaum et al., "The immune epitope database (IEDB) 3.0," *Nucleic Acids Research*, vol. 43, D1, pp. D405–D412, 2015.

[16] M. Molero-Abraham, E. M. Lafuente, and P. Reche, "Customized predictions of peptide-MHC binding and T-cell epitopes using EPIMHC," *Methods in Molecular Biology*, vol. 1184, pp. 319–332, 2014.

[17] C. P. Toseland, D. J. Clayton, H. McSparron et al., "AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data," *Immunome Research*, vol. 1, no. 1, p. 4, 2005.

[18] S. P. Singh and B. N. Mishra, "Major histocompatibility complex linked databases and prediction tools for designing vaccines," *Human Immunology*, vol. 77, no. 3, pp. 295–306, 2016.

[19] J. D'Amaro, J. G. A. Houbiers, J. W. Drijfhout et al., "A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide-binding motifs," *Human Immunology*, vol. 43, no. 1, pp. 13–18, 1995.

[20] M. Bouvier and D. Wiley, "Importance of peptide amino and carboxyl termini to the stability of MHC class I molecules," *Science*, vol. 265, no. 5170, pp. 398–402, 1994.

[21] J. Ruppert, J. Sidney, E. Celis, R. T. Kubo, H. M. Grey, and A. Sette, "Prominent role of secondary anchor residues in peptide binding to HLA-A2.1 molecules," *Cell*, vol. 74, no. 5, pp. 929–937, 1993.

[22] M. Nielsen, C. Lundegaard, P. Worning et al., "Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach," *Bioinformatics*, vol. 20, no. 9, pp. 1388–1397, 2004.

[23] H. G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, and S. Stevanovic, "SYFPEITHI: database for MHC ligands and peptide motifs," *Immunogenetics*, vol. 50, no. 3-4, pp. 213–219, 1999.

[24] P. A. Reche, J. P. Glutting, and E. L. Reinherz, "Prediction of MHC class I binding peptides using profile motifs," *Human Immunology*, vol. 63, no. 9, pp. 701–709, 2002.

[25] P. A. Reche and E. L. Reinherz, "Definition of MHC supertypes through clustering of MHC peptide-binding repertoires," *Methods in Molecular Biology*, vol. 409, pp. 163–173, 2007.

[26] P. A. Reche, J. P. Glutting, H. Zhang, and E. L. Reinherz, "Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles," *Immunogenetics*, vol. 56, no. 6, pp. 405–419, 2004.

[27] M. Gribskov and S. Veretnik, "Identification of sequence pattern with profile analysis," *Methods in Enzymology*, vol. 266, pp. 198–212, 1996.

[28] K. C. Parker, M. A. Bednarek, and J. E. Coligan, "Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains," *The Journal of Immunology*, vol. 152, no. 1, pp. 163–175, 1994.

[29] H. H. Bui, J. Sidney, B. Peters et al., "Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications," *Immunogenetics*, vol. 57, no. 5, pp. 304–314, 2005.

[30] M. Nielsen, C. Lundegaard, and O. Lund, "Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method," *BMC Bioinformatics*, vol. 8, no. 1, p. 238, 2007.

[31] B. Peters and A. Sette, "Generating quantitative models describing the sequence specificity of biological processes with the stabilized matrix method," *BMC Bioinformatics*, vol. 6, no. 1, p. 132, 2005.

[32] T. Sturniolo, E. Bono, J. Ding et al., "Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices," *Nature Biotechnology*, vol. 17, no. 6, pp. 555–561, 1999.

[33] B. Peters, W. Tong, J. Sidney, A. Sette, and Z. Weng, "Examining the independent binding assumption for binding of peptide epitopes to MHC-I molecules," *Bioinformatics*, vol. 19, no. 14, pp. 1765–1772, 2003.

[34] P. Guan, I. A. Doytchinova, C. Zygouri, and D. R. Flower, "MHCPred: a server for quantitative prediction of peptide–MHC binding," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3621–3624, 2003.

[35] M. Milik, D. Sauer, A. P. Brunmark et al., "Application of an artificial neural network to predict specific class I MHC binding peptide sequences," *Nature Biotechnology*, vol. 16, no. 8, pp. 753–756, 1998.

[36] V. Brusic, G. Rudy, G. Honeyman, J. Hammer, and L. Harrison, "Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network," *Bioinformatics*, vol. 14, no. 2, pp. 121–130, 1998.

[37] P. Donnes and O. Kohlbacher, "Integrated modeling of the major events in the MHC class I antigen processing pathway," *Protein Science*, vol. 14, no. 8, pp. 2132–2140, 2005.

[38] M. Bhasin and G. P. S. Raghava, "SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence," *Bioinformatics*, vol. 20, no. 3, pp. 421–423, 2004.

[39] L. Jacob and J. P. Vert, "Efficient peptide-MHC-I binding prediction for alleles with few known binders," *Bioinformatics*, vol. 24, no. 3, pp. 358–366, 2008.

[40] S. Zhu, K. Udaka, J. Sidney, A. Sette, K. F. Aoki-Kinoshita, and H. Mamitsuka, "Improving MHC binding peptide prediction by incorporating binding data of auxiliary MHC molecules," *Bioinformatics*, vol. 22, no. 13, pp. 1648–1655, 2006.

[41] C. J. Savoie, N. Kamikawaji, T. Sasazuki, and S. Kuhara, "Use of BONSAI decision trees for the identification of potential MHC class I peptide epitope motifs," *Pacific Symposium on Biocomputing*, vol. 4, pp. 182–189, 1999.

[42] H. Mamitsuka, "Predicting peptides that bind to MHC molecules using supervised learning of hidden Markov models," *Proteins*, vol. 33, no. 4, pp. 460–474, 1998.

[43] C. Zhang, M. G. Bickis, F. X. Wu, and A. J. Kusalik, "Optimally-connected hidden Markov models for predicting MHC-binding peptides," *Journal of Bioinformatics and Computational Biology*, vol. 04, no. 05, pp. 959–980, 2006.

[44] M. Lacerda, K. Scheffler, and C. Seoighe, "Epitope discovery with phylogenetic hidden Markov models," *Molecular Biology and Evolution*, vol. 27, no. 5, pp. 1212–1220, 2010.

[45] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*, Cambridge University Press, Cambridge, 1998.

[46] W. Liu, X. Meng, Q. Xu, D. R. Flower, and T. Li, "Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models," *BMC Bioinformatics*, vol. 7, no. 1, p. 182, 2006.

[47] D. R. Jandrlic, "SVM and SVR-based MHC-binding prediction using a mathematical presentation of peptide sequences," *Computational Biology and Chemistry*, vol. 65, pp. 117–127, 2016.

[48] S. Buus, S. L. Lauemoller, P. Worning et al., "Sensitive quantitative predictions of peptide-MHC binding by a 'query by committee' artificial neural network approach," *Tissue Antigens*, vol. 62, no. 5, pp. 378–384, 2003.

[49] M. Nielsen, C. Lundegaard, P. Worning et al., "Reliable prediction of T-cell epitopes using neural networks with novel sequence representations," *Protein Science*, vol. 12, no. 5, pp. 1007–1017, 2003.

[50] M. Nielsen and O. Lund, "NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction," *BMC Bioinformatics*, vol. 10, no. 1, p. 296, 2009.

[51] K. Yu, N. Petrovsky, C. Schonbach, J. Y. Koh, and V. Brusic, "Methods for prediction of peptide binding to MHC molecules: a comparative study," *Molecular Medicine*, vol. 8, no. 3, pp. 137–148, 2002.

[52] P. Wang, J. Sidney, C. Dow, B. Mothe, A. Sette, and B. Peters, "A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach," *PLoS Computational Biology*, vol. 4, no. 4, article e1000048, 2008.

[53] P. A. Reche and E. L. Reinherz, "Sequence variability analysis of human class I and class II MHC molecules: functional and structural correlates of amino acid polymorphisms," *Journal of Molecular Biology*, vol. 331, no. 3, pp. 623–641, 2003.

[54] M. Nielsen, C. Lundegaard, T. Blicher et al., "*NetMHCpan*, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence," *PLoS One*, vol. 2, no. 8, article e796, 2007.

[55] M. Nielsen, C. Lundegaard, T. Blicher et al., "Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan," *PLoS Computational Biology*, vol. 4, no. 7, article e1000107, 2008.

[56] P. I. Terasaki, "A brief history of HLA," *Immunologic Research*, vol. 38, no. 1-3, pp. 139–148, 2007.

[57] A. Sette and J. Sidney, "HLA supertypes and supermotifs: a functional perspective on HLA polymorphism," *Current Opinion in Immunology*, vol. 10, no. 4, pp. 478–482, 1998.

[58] A. Sette and J. Sidney, "Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism," *Immunogenetics*, vol. 50, no. 3-4, pp. 201–212, 1999.

[59] I. A. Doytchinova and D. R. Flower, "In silico identification of supertypes for class II MHCs," *The Journal of Immunology*, vol. 174, no. 11, pp. 7085–7095, 2005.

[60] J. Greenbaum, J. Sidney, J. Chung, C. Brander, B. Peters, and A. Sette, "Functional classification of class II human leukocyte antigen (HLA) molecules reveals seven different supertypes and a surprising degree of repertoire sharing across supertypes," *Immunogenetics*, vol. 63, no. 6, pp. 325–335, 2011.

[61] O. Lund, M. Nielsen, C. Kesmir et al., "Definition of supertypes for HLA molecules using clustering of specificity matrices," *Immunogenetics*, vol. 55, no. 12, pp. 797–810, 2004.

[62] G. L. Zhang, D. S. DeLuca, D. B. Keskin et al., "MULTI-PRED2: a computational system for large-scale identification of peptides predicted to bind to HLA supertypes and alleles," *Journal of Immunological Methods*, vol. 374, no. 1-2, pp. 53–61, 2011.

[63] P. A. Reche and E. L. Reinherz, "PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands," *Nucleic Acids Research*, vol. 33, Supplement 2, pp. W138–W142, 2005.

[64] M. Molero-Abraham, E. M. Lafuente, D. R. Flower, and P. A. Reche, "Selection of conserved epitopes from hepatitis C virus for pan-populational stimulation of T-cell responses," *Clinical and Developmental Immunology*, vol. 2013, Article ID 601943, 10 pages, 2013.

[65] S. L. Constant and K. Bottomly, "Induction of Th1 and Th2 $CD4^+$ T cell responses: the alternative approaches," *Annual Review of Immunology*, vol. 15, no. 1, pp. 297–322, 1997.

[66] E. M. Janssen, A. J. M. van Oosterhout, A. J. M. L. van Rensen, W. van Eden, F. P. Nijkamp, and M. H. M. Wauben, "Modulation of Th2 responses by peptide analogues in a murine model of allergic asthma: amelioration or deterioration of the disease process depends on the Th1 or Th2 skewing characteristics of the therapeutic peptide," *The Journal of Immunology*, vol. 164, no. 2, pp. 580–588, 2000.

[67] S. K. Dhanda, S. Gupta, P. Vir, and G. P. Raghava, "Prediction of IL4 inducing peptides," *Clinical and Developmental Immunology*, vol. 2013, Article ID 263952, 9 pages, 2013.

[68] W. Zhong, P. A. Reche, C. C. Lai, B. Reinhold, and E. L. Reinherz, "Genome-wide characterization of a viral cytotoxic T lymphocyte epitope repertoire," *The Journal of Biological Chemistry*, vol. 278, no. 46, pp. 45135–45144, 2003.

[69] J. S. Blum, P. A. Wearsch, and P. Cresswell, "Pathways of antigen processing," *Annual Review of Immunology*, vol. 31, no. 1, pp. 443–473, 2013.

[70] E. Hoze, L. Tsaban, Y. Maman, and Y. Louzoun, "Predictor for the effect of amino acid composition on CD4 + T cell

epitopes preprocessing," *Journal of Immunological Methods*, vol. 391, no. 1-2, pp. 163–173, 2013.

[71] G. E. Hammer, F. Gonzalez, M. Champsaur, D. Cado, and N. Shastri, "The aminopeptidase ERAAP shapes the peptide repertoire displayed by major histocompatibility complex class I molecules," *Nature Immunology*, vol. 7, no. 1, pp. 103–112, 2006.

[72] A. K. Nussbaum, C. Kuttler, K. P. Hadeler, H. G. Rammensee, and H. Schild, "PAProC: a prediction algorithm for proteasomal cleavages available on the WWW," *Immunogenetics*, vol. 53, no. 2, pp. 87–94, 2001.

[73] H. G. Holzhutter, C. Frommel, and P. M. Kloetzel, "A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20s proteasome," *Journal of Molecular Biology*, vol. 286, no. 4, pp. 1251–1265, 1999.

[74] M. Nielsen, C. Lundegaard, O. Lund, and C. Kesmir, "The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage," *Immunogenetics*, vol. 57, no. 1-2, pp. 33–41, 2005.

[75] C. M. Diez-Rivero, E. M. Lafuente, and P. A. Reche, "Computational analysis and modeling of cleavage by the immunoproteasome and the constitutive proteasome," *BMC Bioinformatics*, vol. 11, no. 1, p. 479, 2010.

[76] M. Bhasin and G. P. S. Raghava, "Pcleavage: an SVM based method for prediction of constitutive proteasome and immunoproteasome cleavage sites in antigenic sequences," *Nucleic Acids Research*, vol. 33, Supplement 1, pp. W202–W207, 2005.

[77] M. Bhasin and G. P. Raghava, "Analysis and prediction of affinity of TAP binding peptides using cascade SVM," *Protein Science*, vol. 13, no. 3, pp. 596–607, 2004.

[78] C. M. Diez-Rivero, B. Chenlo, P. Zuluaga, and P. A. Reche, "Quantitative modeling of peptide binding to TAP using support vector machine," *Proteins*, vol. 78, no. 1, pp. 63–72, 2010.

[79] S. Daniel, V. Brusic, S. Caillat-Zucman et al., "Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules," *The Journal of Immunology*, vol. 161, no. 2, pp. 617–624, 1998.

[80] V. Brusic, P. van Endert, J. Zeleznikow, S. Daniel, J. Hammer, and N. Petrovsky, "A neural network model approach to the study of human TAP transporter," *In Silico Biology*, vol. 1, no. 2, pp. 109–121, 1999.

[81] S. Tenzer, B. Peters, S. Bulik et al., "Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding," *Cellular and Molecular Life Sciences*, vol. 62, no. 9, pp. 1025–1037, 2005.

[82] I. A. Doytchinova, P. Guan, and D. R. Flower, "EpiJen: a server for multistep T cell epitope prediction," *BMC Bioinformatics*, vol. 7, no. 1, p. 131, 2006.

[83] M. V. Larsen, C. Lundegaard, K. Lamberth et al., "An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions," *European Journal of Immunology*, vol. 35, no. 8, pp. 2295–2303, 2005.

[84] B. Schubert, H. P. Brachvogel, C. Jurges, and O. Kohlbacher, "EpiToolKit—a web-based workbench for vaccine design," *Bioinformatics*, vol. 31, no. 13, pp. 2211–2213, 2015.

[85] B. Schubert, M. Walzer, H. P. Brachvogel, A. Szolek, C. Mohr, and O. Kohlbacher, "FRED 2: an immunoinformatics framework for Python," *Bioinformatics*, vol. 32, no. 13, pp. 2044–2046, 2016.

[86] M. Atanasova, A. Patronov, I. Dimitrov, D. R. Flower, and I. Doytchinova, "EpiDOCK: a molecular docking-based tool for MHC class II binding prediction," *Protein Engineering, Design and Selection*, vol. 26, no. 10, pp. 631–634, 2013.

[87] J. Hakenberg, A. K. Nussbaum, H. Schild et al., "MAPPP: MHC class I antigenic peptide processing prediction," *Applied Bioinformatics*, vol. 2, no. 3, pp. 155–158, 2003.

[88] P. Oyarzun, J. J. Ellis, M. Boden, and B. Kobe, "PREDIVAC: CD4+ T-cell epitope prediction for vaccine design that covers 95% of HLA class II DR protein diversity," *BMC Bioinformatics*, vol. 14, no. 1, p. 52, 2013.

[89] Y. He, Z. Xiang, and H. L. T. Mobley, "Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development," *Journal of Biomedicine and Biotechnology*, vol. 2010, Article ID 297505, 15 pages, 2010.

[90] I. Dimitrov, P. Garnev, D. R. Flower, and I. Doytchinova, "EpiTOP—a proteochemometric tool for MHC class II binding prediction," *Bioinformatics*, vol. 26, no. 16, pp. 2066–2068, 2010.

[91] H. Singh and G. P. S. Raghava, "ProPred: prediction of HLA-DR binding sites," *Bioinformatics*, vol. 17, no. 12, pp. 1236-1237, 2001.

[92] H. Singh and G. P. Raghava, "ProPred1: prediction of promiscuous MHC class-I binding sites," *Bioinformatics*, vol. 19, no. 8, pp. 1009–1014, 2003.

[93] Q. Zhang, P. Wang, Y. Kim et al., "Immune epitope database analysis resource (IEDB-AR)," *Nucleic Acids Research*, vol. 36, Web Server issue, pp. W513–W518, 2008.

[94] M. Bhasin and G. P. S. Raghava, "A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes," *Journal of Biosciences*, vol. 32, no. 1, pp. 31–42, 2007.

[95] P. Donnes and A. Elofsson, "Prediction of MHC class I binding peptides, using SVMHC," *BMC Bioinformatics*, vol. 3, no. 1, p. 25, 2002.

[96] T. P. Hopp and K. R. Woods, "Prediction of protein antigenic determinants from amino acid sequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 78, no. 6, pp. 3824–3828, 1981.

[97] T. P. Hopp and K. R. Woods, "A computer program for predicting protein antigenic determinants," *Molecular Immunology*, vol. 20, no. 4, pp. 483–489, 1983.

[98] L. Lins, A. Thomas, and R. Brasseur, "Analysis of accessible surface of residues in proteins," *Protein Science*, vol. 12, no. 7, pp. 1406–1417, 2003.

[99] P. A. Karplus and G. E. Schulz, "Prediction of chain flexibility in proteins: a tool for the selection of peptide antigen," *Naturwissenschaften*, vol. 72, no. 4, pp. 212-213, 1985.

[100] E. A. Emini, J. V. Hughes, D. S. Perlow, and J. Boger, "Induction of hepatitis A virus-neutralizing antibody by a virus-specific synthetic peptide," *Journal of Virology*, vol. 55, no. 3, pp. 836–839, 1985.

[101] J. L. Pellequer, E. Westhof, and M. H. V. Van Regenmortel, "Correlation between the location of antigenic sites and the prediction of turns in proteins," *Immunology Letters*, vol. 36, no. 1, pp. 83–99, 1993.

[102] J. L. Pellequer and E. Westhof, "PREDITOP: a program for antigenicity prediction," *Journal of Molecular Graphics*, vol. 11, no. 3, pp. 204–210, 1993, 191–2.

[103] A. J. P. Alix, "Predictive estimation of protein linear epitopes by using the program PEOPLE," *Vaccine*, vol. 18, no. 3-4, pp. 311–314, 1999.

[104] A. S. Kolaskar and P. C. Tongaonkar, "A semi-empirical method for prediction of antigenic determinants on protein antigens," *FEBS Letters*, vol. 276, no. 1-2, pp. 172–174, 1990.

[105] D. D. Womble, "GCG: the Wisconsin package of sequence analysis programs," *Methods in Molecular Biology*, vol. 132, pp. 3–22, 2000.

[106] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: the European molecular biology open software suite," *Trends in Genetics*, vol. 16, no. 6, pp. 276-277, 2000.

[107] J. L. Pellequer, E. Westhof, and M. H. V. Van Regenmortel, "Predicting location of continuous epitopes in proteins from their primary structures," *Methods in Enzymology*, vol. 203, pp. 176–201, 1991.

[108] M. Odorico and J. L. Pellequer, "BEPITOPE: predicting the location of continuous epitopes and patterns in proteins," *Journal of Molecular Recognition*, vol. 16, no. 1, pp. 20–22, 2003.

[109] M. J. Blythe and D. R. Flower, "Benchmarking B cell epitope prediction: underperformance of existing methods," *Protein Science*, vol. 14, no. 1, pp. 246–248, 2005.

[110] M. C. Jespersen, B. Peters, M. Nielsen, and P. Marcatili, "BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes," *Nucleic Acids Research*, vol. 45, Web Server issue, pp. W24–W29, 2017.

[111] S. Saha and G. P. S. Raghava, "Prediction of continuous B-cell epitopes in an antigen using recurrent neural network," *Proteins*, vol. 65, no. 1, pp. 40–48, 2006.

[112] H. Singh, H. R. Ansari, and G. P. S. Raghava, "Improved method for linear B-cell epitope prediction using antigen's primary sequence," *PLoS One*, vol. 8, no. 5, article e62216, 2013.

[113] Y. El-Manzalawy, D. Dobbs, and V. Honavar, "Predicting linear B-cell epitopes using string kernels," *Journal of Molecular Recognition*, vol. 21, no. 4, pp. 243–255, 2008.

[114] B. Yao, L. Zhang, S. Liang, and C. Zhang, "SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity," *PLoS One*, vol. 7, no. 9, article e45152, 2012.

[115] J. A. Greenbaum, P. H. Andersen, M. Blythe et al., "Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools," *Journal of Molecular Recognition*, vol. 20, no. 2, pp. 75–82, 2007.

[116] S. Gupta, H. R. Ansari, A. Gautam, Open Source Drug Discovery Consortium, and G. P. Raghava, "Identification of B-cell epitopes in an antigen for inducing specific class of antibodies," *Biology Direct*, vol. 8, no. 1, p. 27, 2013.

[117] M. Levitt, "Nature of the protein universe," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 27, pp. 11079–11084, 2009.

[118] U. Kulkarni-Kale, S. Bhosle, and A. S. Kolaskar, "CEP: a conformational epitope prediction server," *Nucleic Acids Research*, vol. 33, Web Server issue, pp. W168–W171, 2005.

[119] P. Haste Andersen, M. Nielsen, and O. Lund, "Prediction of residues in discontinuous B-cell epitopes using protein 3D structures," *Protein Science*, vol. 15, no. 11, pp. 2558–2567, 2006.

[120] J. V. Ponomarenko and P. E. Bourne, "Antibody-protein interactions: benchmark datasets and prediction tools evaluation," *BMC Structural Biology*, vol. 7, no. 1, p. 64, 2007.

[121] J. Ponomarenko, H. H. Bui, W. Li et al., "ElliPro: a new structure-based tool for the prediction of antibody epitopes," *BMC Bioinformatics*, vol. 9, no. 1, p. 514, 2008.

[122] M. J. Sweredoski and P. Baldi, "PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure," *Bioinformatics*, vol. 24, no. 12, pp. 1459-1460, 2008.

[123] J. Sun, D. Wu, T. Xu et al., "SEPPA: a computational server for spatial epitope prediction of protein antigens," *Nucleic Acids Research*, vol. 37, no. suppl_2, Web Server issue, pp. W612–W616, 2009.

[124] X. Xu, J. Sun, Q. Liu et al., "Evaluation of spatial epitope computational tools based on experimentally-confirmed dataset for protein antigens," *Chinese Science Bulletin*, vol. 55, no. 20, pp. 2169–2174, 2010.

[125] N. D. Rubinstein, I. Mayrose, E. Martz, and T. Pupko, "Epitopia: a web-server for predicting B-cell epitopes," *BMC Bioinformatics*, vol. 10, no. 1, p. 287, 2009.

[126] S. Liang, D. Zheng, D. M. Standley, B. Yao, M. Zacharias, and C. Zhang, "EPSVR and EPMeta: prediction of antigenic epitopes using support vector regression and multiple server results," *BMC Bioinformatics*, vol. 11, no. 1, p. 381, 2010.

[127] I. Sela-Culang, Y. Ofran, and B. Peters, "Antibody specific epitope prediction — emergence of a new paradigm," *Current Opinion in Virology*, vol. 11, pp. 98–102, 2015.

[128] S. Soga, D. Kuroda, H. Shirai, M. Kobori, and N. Hirayama, "Use of amino acid composition to predict epitope residues of individual antibodies," *Protein Engineering, Design and Selection*, vol. 23, no. 6, pp. 441–448, 2010.

[129] K. Krawczyk, X. Liu, T. Baker, J. Shi, and C. M. Deane, "Improving B-cell epitope prediction and its application to global antibody-antigen docking," *Bioinformatics*, vol. 30, no. 16, pp. 2288–2294, 2014.

[130] I. Sela-Culang, S. Ashkenazi, B. Peters, and Y. Ofran, "PEASE: predicting B-cell epitopes utilizing antibody sequence," *Bioinformatics*, vol. 31, no. 8, pp. 1313–1315, 2015.

[131] J. Huang, A. Gutteridge, W. Honda, and M. Kanehisa, "MIMOX: a web tool for phage display based epitope mapping," *BMC Bioinformatics*, vol. 7, no. 1, p. 451, 2006.

[132] I. Mayrose, O. Penn, E. Erez et al., "Pepitope: epitope mapping from affinity-selected peptides," *Bioinformatics*, vol. 23, no. 23, pp. 3244–3246, 2007.

[133] S. S. Negi and W. Braun, "Automated detection of conformational epitopes using phage display peptide sequences," *Bioinformatics and Biology Insights*, vol. 3, pp. 71–81, 2009.

[134] W. Chen, P. Sun, Y. Lu, W. W. Guo, Y. Huang, and Z. Ma, "MimoPro: a more efficient web-based tool for epitope prediction using phage display libraries," *BMC Bioinformatics*, vol. 12, no. 1, p. 199, 2011.

[135] W. Chen, W. W. Guo, Y. Huang, and Z. Ma, "PepMapper: a collaborative web tool for mapping epitopes from affinity-selected peptides," *PLoS One*, vol. 7, no. 5, article e37869, 2012.

[136] H. Ansari and G. P. S. Raghava, "Identification of conformational B-cell epitopes in an antigen from its primary sequence," *Immunome Research*, vol. 6, no. 1, p. 6, 2010.

[137] C. M. Diez-Rivero and P. A. Reche, "CD8 T cell epitope distribution in viruses reveals patterns of protein biosynthesis," *PLoS One*, vol. 7, no. 8, article e43674, 2012.

[138] D. R. Flower, I. K. Macdonald, K. Ramakrishnan, M. N. Davies, and I. A. Doytchinova, "Computer aided selection of candidate vaccine antigens," *Immunome Research*, vol. 6, Supplement 2, p. S1, 2010.

[139] M. Molero-Abraham, J. P. Glutting, D. R. Flower, E. M. Lafuente, and P. A. Reche, "EPIPOX: immunoinformatic characterization of the shared T-cell epitome between variola virus and related pathogenic orthopoxviruses," *Journal of Immunology Research*, vol. 2015, Article ID 738020, 11 pages, 2015.

[140] P. A. Reche, D. B. Keskin, R. E. Hussey, P. Ancuta, D. Gabuzda, and E. L. Reinherz, "Elicitation from virus-naive individuals of cytotoxic T lymphocytes directed against conserved HIV-1 epitopes," *Medical Immunology*, vol. 5, no. 1, p. 1, 2006.

[141] Q. M. Sheikh, D. Gatherer, P. A. Reche, and D. R. Flower, "Towards the knowledge-based design of universal influenza epitope ensemble vaccines," *Bioinformatics*, vol. 32, no. 21, pp. 3233–3239, 2016.

[142] S. J. Goodswen, P. J. Kennedy, and J. T. Ellis, "*Vacceed*: a high-throughput *in silico* vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology," *Bioinformatics*, vol. 30, no. 16, pp. 2381–2383, 2014.

[143] M. Rizwan, A. Naz, J. Ahmad et al., "VacSol: a high throughput in silico pipeline to predict potential therapeutic targets in prokaryotic pathogens using subtractive reverse vaccinology," *BMC Bioinformatics*, vol. 18, no. 1, p. 106, 2017.

[144] H. Yang and D. S. Kim, "Peptide immunotherapy in vaccine development: from epitope to adjuvant," *Advances in Protein Chemistry and Structural Biology*, vol. 99, pp. 1–14, 2015.

[145] A. Di Pasquale, S. Preiss, F. Tavares Da Silva, and N. Garcon, "Vaccine adjuvants: from 1920 to 2015 and beyond," *Vaccines*, vol. 3, no. 2, pp. 320–343, 2015.

[146] F. Azmi, A. A. Ahmad Fuaad, M. Skwarczynski, and I. Toth, "Recent progress in adjuvant discovery for peptide-based subunit vaccines," *Human Vaccines & Immunotherapeutics*, vol. 10, no. 3, pp. 778–796, 2014.

[147] G. Nagpal, K. Chaudhary, S. K. Dhanda, and G. P. S. Raghava, "Computational prediction of the immunomodulatory potential of RNA sequences," *Methods in Molecular Biology*, vol. 1632, pp. 75–90, 2017.