



Intelligent mining of large-scale bio-data: Bioinformatics applications

Farahnaz Sadat Golestan Hashemi, Mohd Razi Ismail, Mohd Rafii Yusop,
Mahboobe Sadat Golestan Hashemi, Mohammad Hossein Nadimi Shahraki,
Hamid Rastegari, Gous Miah & Farzad Aslani

To cite this article: Farahnaz Sadat Golestan Hashemi, Mohd Razi Ismail, Mohd Rafii Yusop, Mahboobe Sadat Golestan Hashemi, Mohammad Hossein Nadimi Shahraki, Hamid Rastegari, Gous Miah & Farzad Aslani (2018) Intelligent mining of large-scale bio-data: Bioinformatics applications, *Biotechnology & Biotechnological Equipment*, 32:1, 10-29, DOI: [10.1080/13102818.2017.1364977](https://doi.org/10.1080/13102818.2017.1364977)

To link to this article: <https://doi.org/10.1080/13102818.2017.1364977>



© 2017 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 14 Aug 2017.



Submit your article to this journal [↗](#)



Article views: 4524



View related articles [↗](#)





View Crossmark data [↗](#)



Citing articles: 6 View citing articles [↗](#)

Intelligent mining of large-scale bio-data: Bioinformatics applications

Farahnaz Sadat Golestan Hashemi ^{a,b}, Mohd Razi Ismail^{b,c}, Mohd Rafii Yusop ^{b,c},
Mahboobe Sadat Golestan Hashemi^{d,e}, Mohammad Hossein Nadimi Shahraki^{d,e}, Hamid Rastegari^d, Gous Miah^b
and Farzad Aslani^c

^aPlant Genetics, AgroBioChem Department, Gembloux Agro-Bio Tech, University of Liege, Liege, Belgium; ^bLaboratory of Food Crops, Institute of Tropical Agriculture and Food Security, Universiti Putra Malaysia, Serdang, Selangor, Malaysia; ^cDepartment of Crop Science, Faculty of Agriculture, Universiti Putra Malaysia, Serdang, Selangor, Malaysia; ^dDepartment of Software Engineering, Faculty of Computer Engineering, Najafabad Branch, Islamic Azad University, Isfahan, Iran; ^eBig Data Research Center, Najafabad Branch, Islamic Azad University, Isfahan, Iran

ABSTRACT

Today, there is a collection of a tremendous amount of bio-data because of the computerized applications worldwide. Therefore, scholars have been encouraged to develop effective methods to extract the hidden knowledge in these data. Consequently, a challenging and valuable area for research in artificial intelligence has been created. Bioinformatics creates heuristic approaches and complex algorithms using artificial intelligence and information technology in order to solve biological problems. Intelligent implication of the data can accelerate biological knowledge discovery. Data mining, as biology intelligence, attempts to find reliable, new, useful and meaningful patterns in huge amounts of data. Hence, there is a high potential to raise the interaction between artificial intelligence and bio-data mining. The present paper argues how artificial intelligence can assist bio-data analysis and gives an up-to-date review of different applications of bio-data mining. It also highlights some future perspectives of data mining in bioinformatics that can inspire further developments of data mining instruments. Important and new techniques are critically discussed for intelligent knowledge discovery of different types of row datasets with applicable examples in human, plant and animal sciences. Finally, a broad perception of this hot topic in data science is given.

ARTICLE HISTORY

Received 13 February 2017
Accepted 4 August 2017

KEYWORDS

Bioinformatics; data mining;
artificial intelligence;
intelligent knowledge
discovery; bio-data analysis;
heuristic algorithms

Abbreviations

AUC	area under the curve
BADH	betaine aldehyde dehydrogenase
CRM	customer relationship management
GABA	4-Aminobutyric acid
GABald	γ -aminobutyraldehyde
GB	glycine betaine
HBH-AMADHs	high BADH homology aminoaldehyde dehydrogenases
MAS	marker assisted selection
MAPK	mitogen-activated protein kinase
NAD	nicotinamide adenine dinucleotide
NB	naive Bayes
OLAP	on-line analytic processing
Put	putrescine
QTL	quantitative trait loci
ROC	receiver operating characteristic
ROS	reactive oxygen species
SMG	selection marker gene

Spd	spermidine
Spm	spermine
2AP	2-acetyl-1-pyrroline

Introduction

A recent paper in the Science Policy Forum on increasing scientific exploration with Artificial Intelligence (AI) discusses that the human bottleneck in scientific discoveries could be overcome through 'systems that use encoded knowledge of scientific domains and processes in order to assist analysts with tasks that previously required human knowledge and reasoning' [1]. The Hanalyzer (high-throughput analyser) was a pioneer in supporting this knowledge-based genome-scale interpretation technique [2]. Techniques developed by computer scientists have provided the opportunity for researchers to sequence approximately 3 billion base pairs (bp) of the human genome. Currently, achievements generated from the application of next-generation DNA sequencing (NGS)

technologies have inaugurated genomics science, and facilitated critical progress in various areas such as epidemiology, biotechnology, forensics, biomedical sciences and evolutionary biology [3].

Bioinformatics as an interdisciplinary area explores new biological insights from biological data [4]. Biological databases are the heart of bioinformatics [5,6], and represent an organized set of a huge variety of biological data from past research conducted in laboratories (including *in vivo* and *in vitro*), from bioinformatics (*in silico*) analysis and scientific articles. Databases related to 'omics' (e.g. genomics, transcriptomics, proteomics and metabolomics) collect experimental data and can be browsed with designed software [7]. Recently, it has been revealed that analysis of large volumes of biological data through traditional database systems is very troublesome and challenging [8], whereas biological knowledge discovery can be accelerated by intelligent use of the data. Such action is called data mining (DM) and can include simple, complex and/or combinational queries. Consequently, numerous techniques of genomic DM have been created for experimental and computational biologists [9]. DM methods can be used in bioinformatics studies because bioinformatics is data-rich, while no comprehensive theory of life organization can be detected at the molecular level [8].

The question is how to converge the two domains, AI and DM, for successful mining of bio-data. The present paper argues how AI can assist bio-data analysis. Then, an up-to-date review of different applications of bio-data mining is presented. It also highlights some future perspectives of DM in bioinformatics that can inspire further developments of DM instruments.

Intelligent knowledge discovery in bioinformatics

A challenging and hot research area for AI was generated when the Human Genome Project and other large-scale biological studies collected a huge quantity of data [10]. Hunter's sentinel article [10] entitled 'Artificial Intelligence and Molecular Biology' appeared in AI Magazine 25 years ago. Today, bioinformatics is involved in 'big data' and encounters such challenges as sequence, expression, structure and pathway analyses [11]. For the present and future developments of bioinformatics, AI and heuristic approaches are highly essential. Today, it is widely agreed that these two potential domains are converging [12].

Bioinformatics is a highly new interdisciplinary and strategic area of study integrating and interpreting the complexity of any biological data through information technology and computer science. This area of science

attempts to develop novel algorithms and software, data storage methods and new computer architectures in order to fulfil the computational requirements [13]. Algorithm architecture is a step-by-step process (a list of well-defined instructions) for calculation, data processing and automated reasoning. In fact, an algorithm is applied to calculate a function. For instance, Hilbert et al. [14] introduced a partial formalization of the concept in order to figure out the Entscheidungsproblem. Bioinformatics basically copes with four aspects of analysis, including DNA sequence analysis, protein structure prediction, functional genomics and proteomics, and systems biology, through the development and application of innovative algorithmic methods [3].

Finding solutions to the biological issues is in the area of bioinformatics where the DM approaches could be used efficiently. Both DM and bioinformatics are fast developing fields of research [8]. The growth of information storage technology has generated a vast volume of raw data considering two aspects: algorithm development and rise of modern storage equipment. These raw data include important information. In the 1990s, researchers used knowledge discovery from data (KDD) in order to extract knowledge from databases. As Piatetsky-Shapiro and Frawley [15] argue, 'Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data.' Of course, reasonable time complexity, accuracy, comprehensibility and useful results are necessary features that should be considered for the extraction of new knowledge. Furthermore, according to Fayyad et al. [16], DM is synonymous with KDD. DM can be applied in bioinformatics for areas such as gene finding, function motif detection, protein function domain detection, protein function inference, protein and gene interaction network reconstruction, protein sub-cellular location prediction, disease diagnosis, disease treatment optimization, disease prognosis and data cleansing [17]. For instance, a novel learning algorithm (KODAMA package) can be used for knowledge discovery and DM [18].

The process of DM has three levels, including (i) data pre-processing, (ii) data modelling and (iii) data post-processing (Figure 1). In the first phase, raw data are prepared for mining. Because of the widely distributed, uncontrolled generation and utilization of numerous bio-data, data cleaning, data pre-processing and the semantic integration of such heterogeneous and highly distributed databases have become significant in systematic and coordinated analyses of bio-databases [19]. As indicated in Figure 2, the second phase discovers relationships between different data for extraction of significant new patterns [20]. In this regard, prediction and description are the primary goals of DM [17]. The

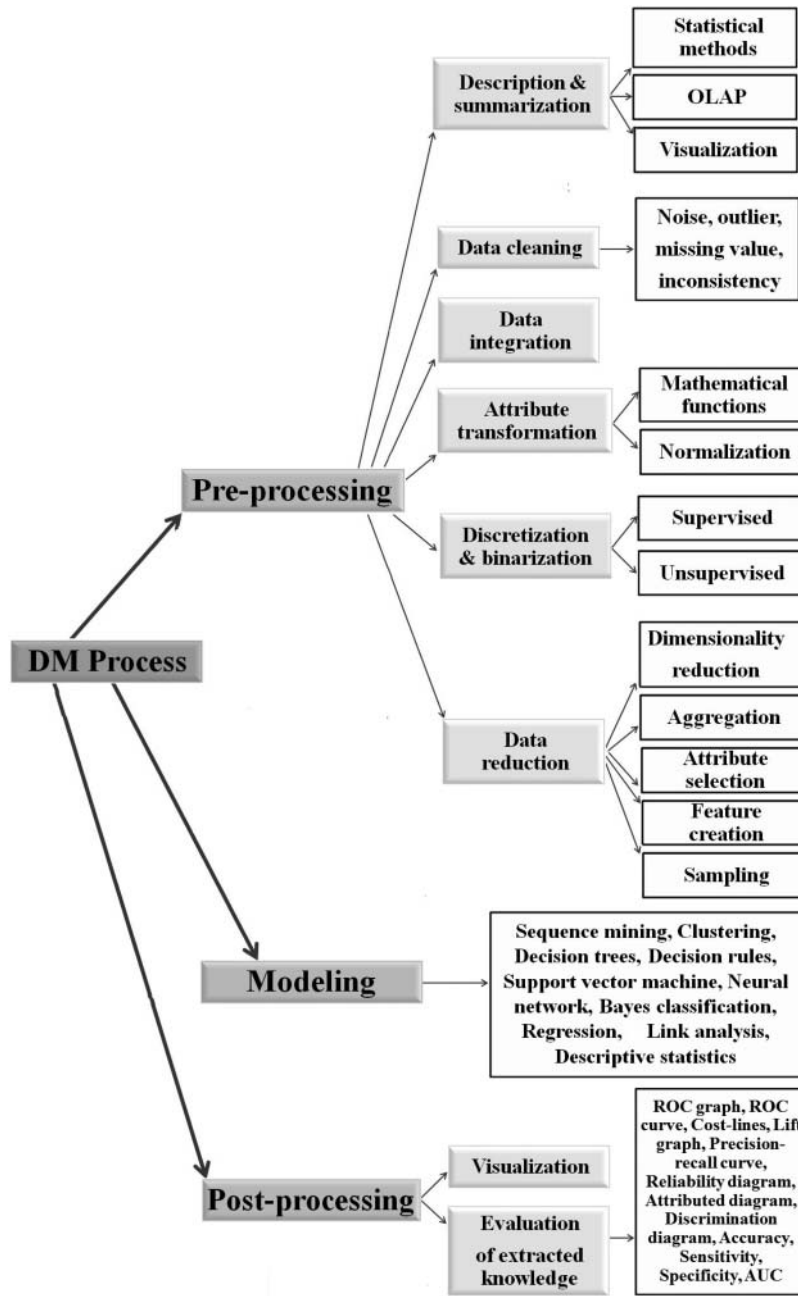


Figure 1. Basic concepts of data mining. The DM process includes three levels: (1) data pre-processing (raw data is prepared for mining), (2) data modelling (discovers relationships between different data for extraction of significant new patterns), and (3) data post-processing (extracted data and pattern are evaluated and then verified as knowledge).

predictive models (such as classification, regression, Time series analysis, prediction, etc.) can predict unknown data values using the known values. On the other hand, the descriptive models (such as clustering, sequence discovery, association rule and summarization) can detect the patterns in data and discover the properties of the data assessed [21]. In the final phase, post-processing, the extracted data and patterns are evaluated and then verified as knowledge. Background knowledge can also be used to verify the extracted knowledge [22].

DM systems are classified based on criteria such as: (i) the type of data source mined (e.g. text, image, audio, video, etc), (ii) the data model (e.g. Object Model, Relational data model, Object Oriented data Model, Hierarchical data Model/W data model), (iii) mining techniques (e.g. machine learning, genetic algorithms (GA), statistics, neural networks, visualization, database oriented or data warehouse-oriented, etc.), and (iv) the kind of knowledge discovered (such as classification, clustering, association, characterization, discrimination, etc.). The classification can also consider the degree of user interaction engaged

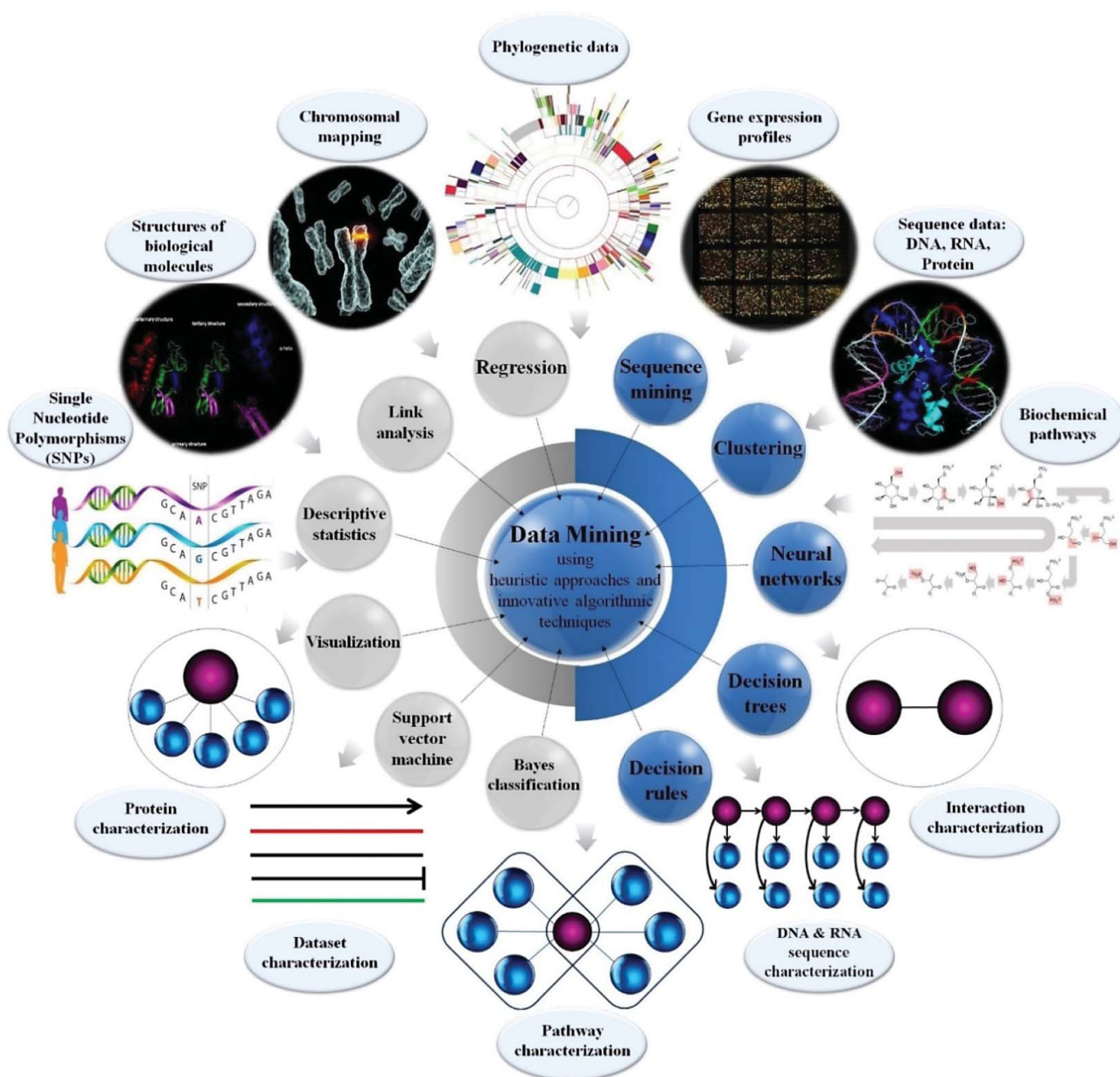


Figure 2. Schematic overview of possible inputs for DM process and subsequently possible predictions and outputs from DM algorithms leveraging many genome-scale datasets. The upper side of the circle shows different selected inputs/datasets including single nucleotide polymorphisms, structures of biological molecules, chromosomal mapping, phylogenetic data, gene expression profiles, DNA/RNA/protein sequence data and biochemical pathways. In the heart of the circle, the most popular DM algorithms and techniques are presented. On the lower side of the circle, different types of possible outputs extracted from DM approaches are displayed. These outputs include protein characterization, dataset characterization, pathway characterization, DNA and RNA sequence characterization, and interaction characterization.

in DM. A comprehensive system can provide different DM approaches be appropriate in various conditions and options, and represent various levels of user interaction [21].

DM approaches and techniques can be categorized into three key groups: (i) supervised learning techniques, (ii) unsupervised learning techniques, and (iii) other. The first group involves classification and prediction tasks. Clustering and association rules mining are in the second

category. On the other hand, some tasks are not classified either as supervised, or as unsupervised learning techniques. Hence, they are assigned into the third category. Yet, there is not a comprehensive list of DM tasks. Nevertheless, according to Piatetsky-Shapiro [23], the most common DM approaches are (a) sequence mining, (b) clustering, (c) decision trees and decision rules (classification), (d) support vector machine (SVM), (e) neural networks (classification), (f) Bayes classification,

(g) regression, (h) link analysis, (i) descriptive statistics and (j) visualization. DM tasks include the selection of suitable algorithms. Both the selection of DM approach and algorithm, and parameterization of the optimal algorithm depend on the goals of the analysis and features of the available data [24]. A couple of DM activities such as manipulation, mining of sequence data, string searching algorithms, machine learning and database theory have been considered seriously. The developed methods for such tasks have led to the extensive progress in computer science [8].

Sequence mining

DM can be used in such fields as text mining, sequential pattern mining, image mining and web mining [8]. Among these areas, sequence data mining (SDM) is the most primitive operation in computational biology [17], and helps to discover the sequential relationships and knowledge hidden in the ocean of sequence data [8]. For example, by mining of DNA sequences alone, the BiRen algorithm predicts enhancers using a deep-learning-based model [25]. Lim et al. [26] also presented an automated information extraction system (@Minter) based on Support Vector Machines for text-mining of microbial interactions. SDM has a broad range of applications such as web access patterns, the analysis of customer purchase patterns, business, security, weather observations, medical data, DNA/RNA/protein sequencing, and so on [8]. In bio-data analysis, the most critical search problems are similarity search and comparison among bio-sequences and structures [19]. In fact, the sequence analysis refers to subjecting a DNA, RNA or peptide sequence to sequence alignment, sequence databases, repeated sequence search, or other bioinformatics approaches on a computer [17].

With the reducing costs, rapid advancements in NGS and related bioinformatics computing sources, and the generation of complete genome sequences of various organisms, bioinformatics provides both conceptual bases and practical approaches for discovering systemic functional behaviours of cells and organisms [27]. In the area of DNA, RNA and protein sequence analysis, SDM approaches are utilized for sequence alignment, sequence searching and sequence classification. Protein sequence classification is the favourite area of many researchers [8].

Sequence alignment is essential in solving such issues as prediction of the secondary and tertiary structures of proteins, prediction of the ancestral sequence or tracing the common genes in two organisms [28], prediction of gene function, sequence divergence, sequence assembly, database searching and so on [29]. However,

sequence alignment is a highly complicated task because of the high number of possible combinations and searches. This complexity rises exponentially along with the size of the sequence. Therefore, sequence alignment is considered a highly computationally intensive problem [28]. Thus, both software and hardware advancements have the potential to improve the accuracy and speed. Consequently, new algorithms have emerged. These algorithms are classified as optimal and heuristic. Although optimal algorithms are efficient in alignment sensitivity, they are computationally expensive. In modern computational biology, the computational cost of all dynamic programming algorithms aforementioned is prohibitive especially for large-scale applications such as database searching. As a result, scientists have shifted their attention to heuristic algorithms. Heuristic approaches are faster algorithms that do not guarantee delivery of the optimal solutions [28].

Furthermore, pairwise sequence alignment is categorized into local and global. Local sequence alignments discover the best approximate *sub-sequence* match within two given sequences. Local sequence alignments find extremely similar areas within the two sequences. Some popular local sequence alignment algorithms include Smith–Waterman [30], FASTA [31], BLAST (Basic Local Alignment Search Tool) [32], Gapped BLAST [33], BLAT (BLAST Like Alignment Tool) [34], BLASTZ [35] and PatternHunter [36]. BLAST is the most popular bioinformatics algorithm worldwide that has been developed at the National Center for Biotechnology Information (NCBI) for fast sequence alignment [32]. The strategy utilized in BLAST for raising the speed is basically fulfilled by two shortcuts: do not bother finding the optimal alignment, and do not search all of the sequence space. Efficiently, BLAST tends to rapidly find the areas with high similarity, without checking every acceptable local alignment [29]. On the other hand, global sequence alignments detect the best alignment of both sequences in their entirety. Therefore, they look for global mapping between entire sequences. Some popular global sequence alignment algorithms include Needleman–Wunch [37], MUMmer (Maximal Unique Match-mer) [38], GLASS [39], AVID [40] and LAGAN [41] (Table 1). All pairwise algorithms are different in terms of indexing step, identifying seeds/anchors and the final step. Some algorithms seem to be more suitable to homologous sequences, whereas others target divergent sequences [28].

Besides pairwise alignments, Multiple Sequence Alignments (MSAs) have been used to align closely related sequences, distantly related sequences or both [42]. MSA algorithms are an interesting field of study since the 1980s. Traditionally, the most common method is the progressive alignment procedure, exploiting the

Table 1. Categorized pairwise alignment algorithms.

Type of pairwise alignment	Algorithm		Characteristics	References
Optimal	Local	Smith–Waterman	Dynamic programming	[30]
Heuristic	Global	Needleman–Wunch	Dynamic programming	[37]
	Local	FASTA	Disadvantages: if the sequences possess more than one area of homology (two optimal diagonals), just the area around init ¹ ^a could be found, while the area contributing to initn ^b will be discarded. Advantage: speed over optimal algorithm.	[31]
		BLAST	Disadvantages: it cannot find seeds ^c smaller than the minimum length 'l' regarded for the precise match seed (DNA alignment) and reports just local alignments. Also it can find too many seeds per sequence; therefore, decreasing speed (protein alignment) and allows no gaps in sequence.	[32]
		BLAST2	It was developed to overcome the disadvantages of BLAST.	[33]
		BLAT	Same as BLAST and FASTA. BLAT is different from BLAST in that which sequence it indexes. BLAT is confined as it does not find small homologous areas due to the small seed length.	[34]
		PatternHunter	It introduces spaced seed to increase the sensitivity. Also, its performance is higher than that of the above-mentioned algorithms regarding sensitivity. The speed is not higher than BLAST, as it is performed in Java and induces memory problems for very long sequences.	[36]
		BLASTZ	It is the fastest algorithm in the BLAST series. To speed up the algorithm, all repeats should be removed in the sequences.	[35]
		MASAA (Multiple anchor staged alignment algorithm)	MASAA employs the searching methods (suffix tree) utilized in global sequence alignment algorithms to identify long common substrings in both sequences. The simulations show that this algorithm outperforms BLASTZ when the sequences are divergent and sometimes generates an alignment when BLASTZ does not return any alignment. On homologous sequences, the performance is comparable. Overall, MASAA finds the alignment faster than BLASTZ.	[28]
	Global	MUMmer	It is one of the first global alignment algorithms that align two long genomes.	[38]
		GLASS	It aligns long genomic sequences. It aims to remove the limitations of standard dynamic programming (SDP) approaches which had running time problems and to increase the sensitivity when aligning the sequences in their entirety.	[39]
		AVID	It balances sensitivity and speed when aligning very long sequences.	[40]
		LAGAN	More sensitive than previous algorithms. An effective pairwise aligner which can be appropriate for genomic comparison of distantly related organisms. It is not faster than MUMmer and BLASTZ. It is not also sensitive in detecting transpositions.	[41]

^aFASTA refers to a diagonal, scoring the highest value, 'init1.'

^b In FASTA algorithm, the maximum weighted graph is chosen and the best alignment identified is marked as 'initn.'

^c A pair of highly similar areas is known as 'seed.'

idea that homologous sequences are evolutionarily related. Later, various alignment programs including global and local methods have been developed [43]. CLUSTA, an extremely common and effective heuristic algorithm for multiple alignments, was developed by Higgins and Sharp [44]. Then, it was extended into the current version, CLUSTALW, by Higgins et al. [45]. Additionally, evolutionary-based inference systems are highly crucial in such fields, as epidemiology and virulence [46], elucidation of the life tree [47], biodiversity [48], drug designs [49], human genetics [50] and cancer [51]. MSA and its subsequent analysis are the requirements for such evolutionary-based research [52–54]. Also, MSAs are very important in determining particular traits, known as 'specificity determining positions', modulating protein's function in a particular context, for instance, interaction areas, targeting signals in different cell machineries, pathways or compartments, or post-translational modification regions (cleavage, phosphorylation, etc.) [55–57].

Numerous genetic diseases are due to mutation variants of a gene or cluster of genes, or the overlapping

features of various genetic diseases mapped to near or distant loci [3]. Consequently, mutation analysis has become highly significant because of its association with different diseases [42,58]. Hence, various computational approaches are being developed to forecast the function of missense mutations and to detect residues having an important impact on maintaining wild-type function. These approaches are, sequence-based algorithms [59], structure-based algorithms [60,61] and a combination of both [62]. MSAs highlight two main trends that are particular to disease-associated mutations [42]. In addition to forecasting the function of mutant gene products, low throughput sequencing of known target genes facilitates the discovery of new mutations, thus helping scientists understand the evolving characteristics of some genetic diseases. Bioinformatics is able to predict such substitution impacts [3]. A three-phase analysis of 1514 missense substitutions in the DNA-binding domain (DBD) of TP53 (the most frequently mutated gene in human cancers) confirmed the utility of the Align-GVGD approach (<http://agvgd.iarc.fr>) for functional classification of missense mutant variants for any genes with

adequate available sequences [42]. Additionally, the discovery of single nucleotide polymorphisms (SNP) in numerous model and non-model plant species is the result of bioinformatics progress [13]. In a recent study, Huang et al. [63] offered a framework that is able to discover long, single point mutations across multiple sequences. However, this framework could not detect co-mutations involving multiple positions. Other researchers have attempted to use the translation probability matrix to evaluate the future amino-acid composition [64,65]. However, they have only considered the mutation in one position and are unable to analyse the geographical dissemination of mutations over time. Later, a different algorithm was proposed to mine co-mutations across multiple sequences [66]. However, the framework did not consider the three-dimensional (3D) structure of proteins. Recently, Wei [67] suggested an effective algorithm based on 3D-structure for discovering non-contiguous mutations in biological sequences. Furthermore, high-throughput aligners can help in mapping the sequence reads to the reference sequences. Sequence alignments have numerous functions. However, there is pressing need for highly efficient algorithms due to the large volume of the short sequence reads produced by NGS [68]. The Maq algorithm utilizes hashing methods [69]. In order to align reads, techniques based on the Burrows–Wheeler transformation can also be applied. Such techniques include BWA [70], Bowtie [71] and Soap [72]. Although these algorithms are faster than Maq [72], they are limited to split reads in order to achieve gapped alignments. Moreover, a Smith and Waterman algorithm [30] is employed in the Mosaik aligner [73] for aligning the short reads [68].

Clustering

By applying heuristic approaches, the clustering algorithm can classify objects into a default number of clusters based on the data similarity. Distance metrics which are usually utilized as a scale for similarity evaluation of the objects include Euclidean, Jacquard, Manhattan, etc. The similarity measure can be chosen based on the features of the objects [24]. Based on a machine learning view, clusters correspond to hidden patterns, the search for clusters is unsupervised learning, and the resulting system presents a data concept [74]. However, cluster analysis attempts to determine the number of clusters in a dataset. This is an open issue in cluster analysis. For example, highly utilized iterative methods, such as the *k*-means algorithm, ask the user to determine the number of clusters in the data before running the algorithm. Algorithms which can discover the number of clusters are categorized in unsupervised clustering algorithms

[75]. Hierarchical and partitional clusterings are the most popular clustering approaches (Table 2). Practically, clustering is highly important in DM applications such as information retrieval, text mining, scientific data exploration, spatial database applications, web analysis, marketing, customer relationship management (CRM), computational biology and medical diagnostics [74].

Exploring the hidden patterns in the gene expression microarray data is challenging for functional proteomics and genomics. DM methods can be used for addressing this task [75]. In gene expression data, clustering is a significant approach for deriving underlying information [20] such as biologically relevant grouping of genes and samples, gene regulation, gene function and gene expression differentiation in different circumstances [75]. For instance, Engreitz et al. [77] mined significant information from transcriptional modules in microarray data for acute myelogenous leukemia. Tasoulis et al. [75] also examined the application of the proposed *k*-windows clustering algorithm on gene expression microarray data. Besides determining the clusters present in a dataset, this algorithm can also define their number. Furthermore, the DBSCAN (density-based spatial clustering of applications with noise) clustering algorithm was used to screen colon cancer data [78]. On the other hand, a supervised fuzzy clustering approach discovered potential protein biomarkers to recognize individuals at high risk of bladder cancer [79].

Additionally, Frey and Dueck [80] proposed the Affinity Propagation (AP) algorithm, which is a state-of-the-art clustering approach. It has been used in wide fields of computer studies and bioinformatics since it has higher performance than traditional approaches such as *k*-means. In order to achieve high quality sets of clusters, real-valued messages are passed between all pairs of data points until convergence by the original AP algorithm. Like agglomerative clustering, AP is able to measure similarities between data samples.

The AP clustering algorithm is not dependent on a vector space structure, in contrast to other prototype-based techniques, and the clusters are selected from the detected data samples and not calculated as hypothetical averages of cluster samples [81]. As outlined by Bodenhofer et al. [81], AP is especially appropriate for bioinformatics purposes because: (i) numerous similarity scales applied in bioinformatics are not associated with explicit vectorial features; and (ii) detecting a small set of clusters can offer the opportunity for exploration in biological datasets. So far, AP algorithm has been demonstrated to be effective for the purpose of for microarray data analysis [80–85], Network analysis [86–88] structural biology studies [89–91], and sequence analysis [92]. For review, see [81]. Although AP has many applications,

Table 2. Most popular data-mining algorithms along with their most prominent characteristics (Modified from Li et al. [76]).

DM approaches	Algorithm	Features
Clustering	Hierarchical	Prototype-based, graph-based, bottom-up, non-parametric, less susceptible of initial value, sensitive to noise, time and space complexity.
	<i>k</i> -means	Fast, simple, popular, prototype-based, optimization problem, centre-based, partitioning problem, parametric, inappropriate for data different in density and size, sensitive to noise, susceptible initial value, different outputs in each run.
	Affinity propagation	Works for any meaningful measure of similarity between data samples; independent of a vector space structure; the clusters are chosen from the data samples observed; detects a small set of clusters, identifies clusters with minimum errors; low speed due to the necessary quadratic CPU time in the number of data points to calculate the messages.
	Fuzzy <i>c</i> -means	Determines membership of each object to the clusters; fast, simple, prototype-based, optimization problem, centre-based, parametric and unsuitable for data that are different in size and density.
	DBSCAN	Density-based, non-complete and partitioning problem; handles arbitrary size and density; resistant to noise, time and space complexity.
Association rules	DIC	Dynamic; investigates the particular distance of transactions; retrieves lost patterns through moving forward; decreases I/O complexity; sensitive to data homogeneity.
	Apriori	Popular, uses prior knowledge; simple, iterative method; searches in all variables, reviews entire database at each stage, time and I/O complexity.
	DHP	Using hash table; reducing the number of candidate patterns; collision problem in the hash table; relation between runtime and database size.
	D-CLUB	Dynamic, appropriate for parallel process and distributed database, differential optimization; reduces time and space complexity; self-adaptive; removing the empty bits.
	Eclat	Using lattice-theoretic, bottom-up method; exploring large length patterns; decreasing I/O complexity; discovering all sequential objects; inappropriate for large data; space complexity.
Classification	SVM	Eager approach; unstable; mathematical based; optimization; global minimum; diagonal separation line; appropriate for high dimensional data and little training data, parametric, black box, SVMs use kernels to learn complex functions. However they are very slow and there are multiple parameters to be chosen by the user.
	ANN	Eager approach; multi-layer network with at least one hidden layer, resistant to replication, diagonal separation line, ability to complex relation, parametric, black box, increases time by increasing hidden layers, sensitive to noise and missing values. The output of ANNs cannot be read and the training of the model is very slow.
	Decision trees	Eager approach; partitioning, stable, greedy, recursive, interpretable, non-parametric, resistant to noise and replication. The output from decision trees can be easily interpreted, but it depends on the algorithm employed and the complexity of the tree. It is also well-suited to datasets with missing values, sensitive to inconsistent data, separation line parallel to axis <i>x</i> , <i>y</i> .
	<i>k</i> -nearest neighbours (KNN)	Instance based, lazy approach, required similarity measurement, simple, prediction based on local data, parametric, flexible, arbitrary decision boundaries, sensitive to noise and replication.
	Rule-based	Eager approach, interpretable, partitioning, resistant to noise and imbalance data; produces if ... then rules; separation line parallel to axis <i>x</i> , <i>y</i> . The rules are easily readable and proper for identification of putative biomarkers. However there is a possibility of over-fitting.
	Random forest	Efficient on large datasets and can handle large numbers of attributes; not very sensitive to outliers.
	Naive Bayes	Eager approach, nondeterministic, statistical based, resistant to noise, fast and easy to implement, missing value and irrelevant features, required to determine initial probability, accuracy degraded by correlated attribute; it assumes attributes are independent of each other.
Performance evaluation	ROC graph	Independent of class distribution, comparing performance of two or more predictive models.
	ROC curve	Independent of class distribution, able to rank positive and negative samples.
	Cost-lines	Evaluating error rate based on the different costs.
	Lift graph	Identifying relations between true positive rate and profanity of positive classification.
	Precision-recall curve	Evaluation and ranking each sample based on positive class.
	Reliability diagram	Investigating probability of true calibration of model.
	Discrimination diagram	Showing discrimination between each class prediction.
	Attributed diagram	Identifying regions of model that degrade performance compared with reference models with constant performance.
	AUC	Area under the ROC curve.
	Accuracy	Rate of correct classification.
	Specificity	Rate of true negative classification.
	Sensitivity	Rate of true positive classification.

one of its most significant research problems is its speed, particularly for large-scale datasets, since it needs quadratic CPU time in the number of data points to calculate the messages [93]. In order to solve this issue, the FSAP (fast sparse affinity propagation) algorithm was suggested for AP [94]. However, the efficiency of this fast algorithm is at the expense of the clustering result accuracy. In fact, its clustering outputs are different from the

outputs of the original AP algorithm. Thus, Fujiwara et al. [93] suggested an effective AP algorithm pruning unnecessary message exchanges in the iterations and calculating the convergence values of pruned messages after the iterations to identify clusters. While it can guarantee exactness of the clustering outputs, it is quite faster than other algorithms. Furthermore, unlike FSAP, any inner-parameters are not required to be set by users. In

addition, for clustering extremely large sequencing data, Jiang et al. [95] reported a Dirichlet Process Means (DP-means) algorithm. This algorithm (DACE) follows a random projection partition approach for parallel clustering.

Association rules mining

For the first time, Piatetsky-Shapiro and Frawley [15] proposed the association rules mining technique (a market basket analysis approach), which is another area of DM. This method can detect non-trivial patterns in the data, and define the relationships among the binary variables utilized to characterize a set of objects [96] (Table 2). The most common a-priori algorithm offers two input parameters: rule support and confidence. The proportion of dataset providing the rule condition is association rule support, and the proportion of the dataset to which this rule can be applied is association rule confidence [24]. In spite of the solid nature of association analysis and its potential applications, such approach is not as popular as clustering and classification, particularly in the area of bioinformatics. However, some researchers have employed association rules techniques in their work [97–100]. For instance, Mohanty et al. [101] created a prediction model by association rules in order to discover breast cancer masses in mammograms.

Regression

The regression tree is a machine-learning approach for creating prediction models from data by recursively subsetting the data space and fitting a prediction model within each subset. Accordingly, a decision tree can be created graphically from the subsetting [102]. In fact, regression analysis is a statistical method estimating and predicting relationships between variables [20]. Regression trees are for dependent variables taking continuous or ordered discrete values, with a prediction error [102]. Regression algorithms are simple linear, multiple linear, logistic and fuzzy. In DM, regression algorithms predict hidden data based on continuous training data. In this method, the behaviour of the dependent variable (y) is estimated by independent variables (x) [20]. For example, relationships between vaccination and risk of preterm birth can be revealed by a regression algorithm [103].

Classification

Classification, as a supervised learning technique, is a very popular task in DM. It predicts the class of a user-specified goal feature based on the class of other features, known as the predictive features [104]. Therefore,

it assigns objects to the predetermined classes. The classification process has two steps, including training and testing. The training phase involves the algorithm that analyses the data meant for learning and generates a classification model (Table 2). The testing phase checks the accuracy of the model through another data set. Although Naive-Bayes Classifier, SVM, K-Nearest Neighbour (KNN) and Genetic algorithm (GA) are popular methods of classification for gene expression and protein data, decision trees, Bayes classifications and artificial neural networks (ANNs) are the most common classification approaches [24].

Supervised machine learning can be utilized for classification. For example, a group of machine learning methods is SVMs which are based on the linear separation between groups. The features determining SVMs include (i) the principal assigning the optimal linear classifier based on separation margin maximization, (ii) detection of the support vectors, and (iii) utilizing kernels to change the initial variables into a greater-order non-linear space in which the linear separation takes places. One of the most common SVM algorithms is Sequential Minimal Optimization (SMO) [105]. Furthermore, decision trees are machine-learning models structuring the knowledge utilized to differentiate between instances in a tree-like structure. Novel examples are categorized by pursuing the tree alongside the related branches, based on the features of the sample. Approaches (e.g. C4.5) begin with an empty tree and repetitively divide the data, generating branches of the tree, until they define exemplars of a branch to a leaf of the tree [106]. The Random Forest approach is based on decision trees, whereas multiple trees are based on the training data. Each tree has only access to a randomly sampled subset of the traits of the problem. Subsequently, by the class prediction of the test samples, each tree can predict a class and the majority class predicted is utilized [107]. Furthermore, Bayesian classifiers are statistical approaches based on Bayes theorem [108]. Naive Bayes [109] is the simplest one calculating the probability that each sample input belongs to each of the classes. Naive Bayes is a highly competent machine learning approach across various application domains and has perfect scalability. As reviewed in Swan et al. [105], ANNs are inspired from the function of the brain. They include a set of neurons (computational elements) interlinked via a vast diversity of interconnectivity patterns. Depending on the received signal, the connections of a neuron define its activity. Each individual neuron is a variant of a linear classifier. However, the presence of various layers and neurons can lead to the creation of elaborate nonlinear classifiers allocating their function to complicated issues [110]. Furthermore, rule-based learners involve

BioHEL [111] as well as JRip [112]. They aim to automatically produce collections of meaningful principles that determine the allocation of a particular cluster to a given class of a problem [113]. Rule learning encompasses a variety of approaches. Their distinctions are based on (i) the kind of rule sets they create and (ii) how to establish the rules and the rule sets [105].

Sequence data analysis is very important in bioinformatics. This task can be dealt with using prediction and classification methods. For example, the research goal may be to assign a protein of interest to a family in order to elucidate the evolution of this protein and to reveal its biological function [8]. Additionally, the investigation of proteins is highly beneficial in biological and medical domains. In biology, for instance, putative amino-acid sequences are often analysed for discovery of enzyme active sites, or nucleotide sequences, in order to identify coding or non-coding regions of DNA or to identify the function of particular nucleotide sequences [8,114]. Thus, it is essential to develop an intelligent system for bio-data classification and behaviour prediction (For review see [8]). To briefly outline some of the more notable techniques, the Rough Set Classifier technique [115] has been suggested as a novel model for classification of large volumes of protein data based on protein functional and structural characteristics. This model is considered an effective classification tool due to its accuracy and fast speed. Another, three-phase model for the classification of unknown proteins into known families has been reported [116], in which the noisy sequences are first omitted in order to improve the accuracy through minimizing the computational time; second, the important features are acquired and a feature ranking algorithm is used to classify the sequences; and third, neighbourhood analysis is used to classify the sequence of interest into a particular class or family. This rule can mine significant relations between a protein sequence and protein classes, subclasses and families. This kind of classification, in addition to data analysis, generates knowledge-based information [8]. Another method for classification of protein sequences is the feature hashing technique [117], which has the advantage of reducing the dimensionality on protein sequence classification tasks. Alternatively, a hybrid GA/SVM algorithm for classification of protein sequences has been proposed [118], in which the protein features that carry precise and sufficient discriminative information are selected for classifying and training the SVM classifier simultaneously. Based on experimental outputs, the hybrid GA/SVM system has been demonstrated to outperform the BLAST and HMMer (Hidden Markov Model-based sequence search) methods [8,118]. Furthermore, Leung et al. [104] used a DM framework for predicting hepatitis B virus (HBV)

positive patients and analysing key mutation sites in the HBV DNA sequences. In this approach, two new algorithms were developed based on Rule Learning (RL) and Nonlinear Integral (NI). The NI algorithm performs well using the fuzzy measure and the nonlinear integral because the non-additivity of the fuzzy measure shows the significance of the individual features and their inherent interactions. The authors also used GA for optimization providing multimodal solutions involving sets of best solutions. Moreover, a regularization approach was applied to achieve a solution with the fewest non-zero fuzzy measure values [104].

Besides, bioinformatics opens a new window for understanding cancer biology through intelligent systems. For instance, Banwait and Bastola [119] employed supervised and unsupervised techniques for precise classification of cancer types and sub-types. The supervised classifier models based on ANN, random forest and SVM have addressed the cancer sub-type classification issues [120,121]. Combining the cancer biology knowledge with influential computational and statistical tools has the potential to discover miRNAs as new biomarkers to detect cancer and cancer sub-types. Also, combining gene and miRNA expression data with computational analysis techniques could help to determine the role of miRNAs in cancer development and metastasis and their capacity in acting as therapeutic agents in cancer treatment. Additionally, a challenge in classification of cancer tissue samples based on gene expression data is to create an influential approach selecting a parsimonious set of informative genes [122]. In this regard, Wang et al. [123] introduced a novel algorithm (Chisquare-statistic-based Top Scoring Genes (Chi-TSG) classifier) for binary and multi-class cancer classification and informative genes selection based on numerical molecular data. On the other hand, classification of gene expression data is highly important in prediction of disease related genes. Thus, an influential statistical feature selection method for classification of gene expression data set was enhanced based on statistically defined efficient range of traits for every class termed as ERGS (Effective Range based Gene Selection) using naive Bayes (NB) and SVM Classifiers [120]. Furthermore, classification of RNA structure change by 'gazing' at experimental data was proposed by Woods and Laederach [124].

Neural networks

The term neural network originally refers to a circuit of biological neurons. However, its contemporary use is in the context of ANNs, which comprise programming solutions resembling the function of artificial neurons, or nodes. Electrical signalling and other types of signalling

Table 3. Neural network techniques and their applications.

Applications	References
Back-propagation artificial neural networks	[127]
Bayesian networks	[128]
Bayesian confidence propagation neural networks	[129]
Feed forward neural networks	[130]
Flow networks	[131]
Fuzzy recurrent neural networks	[132]
Gene regulatory networks	[133]
General regression neural networks	[134]
Neural classification	[135]
Neural nets	[136]
Radial basis function networks	[137]

emerge from neural transmitter diffusion. Hence, neural networks are highly complicated [125], and have become one of the vital techniques in the bioinformatics field since the development of various biological databases storing DNA/RNA sequences, protein structures and sequences, and other macromolecular structures. Prediction is the most commonly discovered ability of neural networks in bioinformatics, especially in cases of a limited volume of available raw data that can be utilized to extract the prediction model [126]. Table 3 lists a

number of applications for neural networks in bioinformatics.

Machine-learning methods can be used in different areas of bioinformatics: support vector machine for protein fold recognition, hidden Markov model (HMM) for sequence and profile alignment, Bayesian networks for gene regulatory networks [138] and ANNs for protein secondary structure prediction [138], disease classification and biomarkers identification [139] (Figure 3). Due to gene collaboration in functional molecular networks [140–142], network-based analyses have been highly used in cancer research to provide a molecular stratification of cancer patients [141], to predict disease outcome [143,144], to understand tumourigenesis [145] and the mechanism of action of tumour-inducing viruses [146], to predict the carcinogenicity of chemical compounds [147] and to prioritize the damaging effects of cancer mutations [148]. Thus, Horn et al. [149] harness the fundamental wiring of genes into functional networks to develop a powerful statistical framework complementing gene-based tests to produce new hypotheses about driver-gene candidates. Several new methods using degree-of-interest (DoI) functions

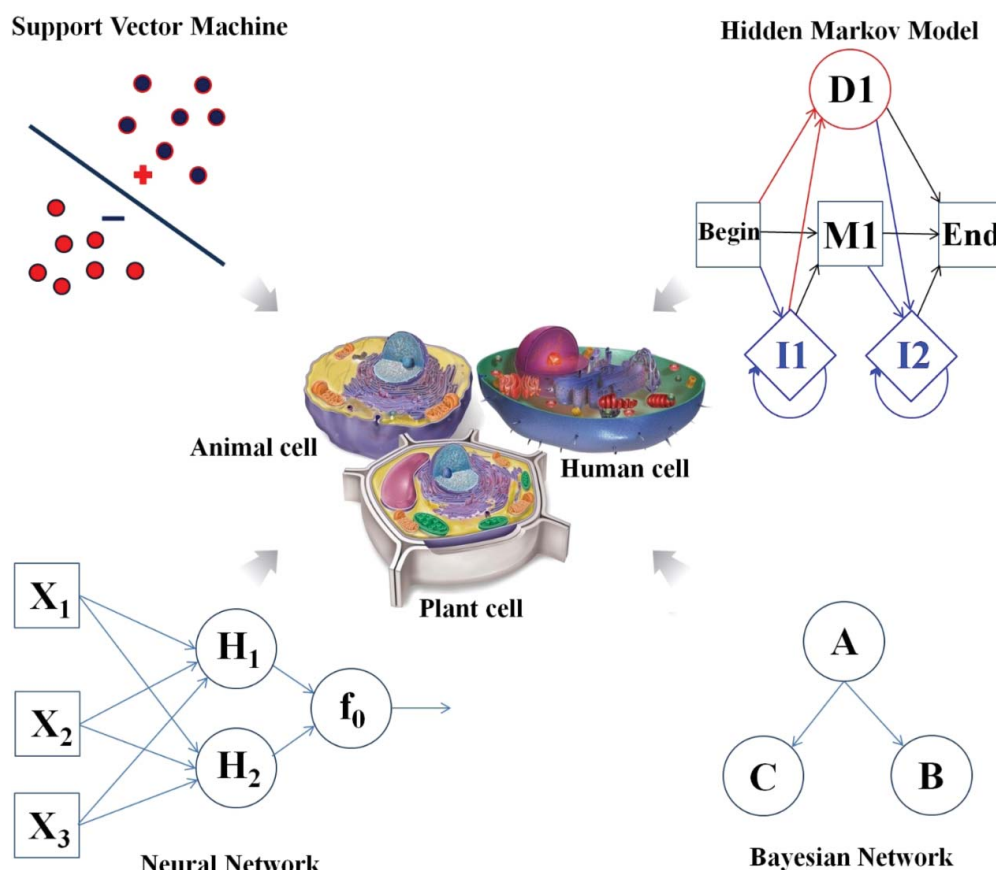


Figure 3. Schematic overview of machine-learning applications in bioinformatics. These applications include Hidden Markov Model (HMM) for sequence and profile alignment, Artificial Neural Networks (ANN) for protein secondary structure prediction, disease classification and biomarkers identification, Support Vector Machine (SVM) for protein fold recognition and Bayesian networks for gene regulatory networks.

Table 4. Strengths and weaknesses of ANNs.

Strengths	Definition	Weaknesses	Definition
Adaptive learning	Learns new tasks with relatively small amount of training data	Design difficulty	ANN is challenging to design and train particularly with complex issues
Self-organization	Organizes its data to obtain pattern recognition	Suboptimal speed	If the number of neurons is high, ANN will be computationally slow and challenging
Fault tolerance	Retains performance with destruction of parts of the infrastructure	Lack of transparency	Is not as easy to comprehend as other machine-learning systems
Real-time operation	Computations can be performed efficiently in parallel	Overfitting tendency	Overfits particularly if the training data are small and cannot be generalized well into the unseen data
Computationally powerful	Predicts complex biological patterns with training	Data pre-processing	ANN has better performance if there is data normalization as part of pre-processing

have been reported [150]. They use Dol-based filtering, graph layout and a network comparison method. Furthermore, the *RenoDol* framework has been developed as an application to untangle huge and dense networks through Dol function, and has been integrated in the network visualization framework Cytoscape [150]. Topological network analysis of gene–disease associations can reveal significant properties of the nature of Mendelian diseases [151]. Hence, four different bipartite networks including OMIM, CURATED, LHGDN and ALL have been employed to examine human diseases at a global scale [152]. For further exploration of the diseases and disease-related genes, gene and disease centric views of the data are produced through projecting the bipartite gene–disease networks to monopartite networks [152]. Godinez et al. [153] also reported a multi-scale convolutional neural network for phenotyping high-content cellular images. A syntax convolutional neural network (SCNN) based DDI extraction approach has been proposed for extraction of drug–drug interaction information from biomedical literature [154].

On the other hand, knowledge about protein secondary structure can help to understand human diseases and to develop therapeutic enzymes and drugs. Hence, various AI techniques are applied for prediction of protein secondary structure. Standard statistical approaches such as discriminant analysis and generalized linear models have limitations when there are highly nonlinear and complicated interactions. Currently, machine learning makes computer programming enable to increase performance with biological data sets [138]. Because of the high capability of ANN to reveal complicated patterns, categorize big data and make precise predictions in huge complex amino acid/protein data sets, ANNs have become a key technique in computational molecular biology issues such as DNA and RNA nucleotide sequence analysis, sequence correlations, sequence encoding and result interpretation, and protein structure prediction. Of course, it has its own strengths and weaknesses (Table 4). Current developments in accuracy using statistical context-based scores (SCORPION) [155] and incorporating tertiary structure information with the ROSETTA *de novo* tertiary structure prediction approach

[156] have shown continual improvements in the ANN method for protein structure prediction. Table 5 shows a comparison of ANN with other machine-learning approaches in protein structure prediction. Additionally, Uziela et al. [167] proposed a model for assessment of protein quality using deep learning neural network approach. Moreover, forecasting the errors of predicted local backbone angles and non-local solvent- accessibilities of proteins using deep neural networks are valuable for prediction, evaluation, and refinement of protein structures [168]. Zeng et al. [169] also reported a systematic exploration of CNN architectures predicting DNA–protein binding.

Performance evaluation and visualization

Because of numerous descriptive and predictive algorithms for knowledge mining, various performance assessment approaches are required (Figure 1 and Table 2). Performance assessment techniques generally include single scalar and graphical approaches [170]. Specificity, sensitivity and accuracy are in the first group. Simplicity in implementation but lower efficiency in assessment is the major feature of this group. The second group considers Receiver Operating Characteristic (ROC) Curve, Cost-Line and Lift. This group has a complicated implementation but it makes good sense. A system was suggested for fast extraction of important knowledge about cancer by summarization and visualization [171]. The model employs clinical trial registries and analyses data related to cancer vaccine trials. The system output is used as key information regarding cancer vaccine trials and can be utilized for future vaccine development [171].

After information evaluation, scientific data representation plays an important role. Different techniques of data representation can sometimes influence the explanation of the results or even change the conclusion of some experiments [172]. However, along with technological developments, data visualization is becoming a bottleneck, as in the postgenomic era, data visualization tools are necessary [173]. Consequently, Information

Table 5. Comparison of ANN with other machine-learning approaches in terms of protein structure prediction.

Machine learning methodologies	Description of the methodology	versus ANN
Bayesian networks (BN)	The first machine-learning method in protein structure prediction was partly based on Bayesian statistics [157]; BN performs well over huge databases.	Less opaque [158]
Hidden Markov models (HMM)	HMM (a probabilistic model) can provide relevant information about the sequence-structure relation [158]; its accuracy is less than that of the other machine-learning methods.	ANN is more successful [159]
Support vector machines (SVM)	A supervised learning model; associated with learning algorithms and classification and regression analysis in its construction of a hyperplane; can handle high-dimensional data; flexibility in modelling diverse types of data; high accuracy.	SVM is superior in predicting the location of turns [160]; in ubiquitin protein structure prediction, SVM is superior to both ANN and HMM [161]; SVM requires a relatively small training set to avoid overfitting of the data [162]; ANN have much better accuracy and take much less training and computation time [163]; SVM require much larger memory and powerful processor [163]; SVM outperformed ANN with an overall accuracy of 89.3% in identification of lipid-binding proteins (LPBs) from non-LPBs [164]
Other	–	Nearest-neighbour method had an overall three-state accuracy of 72%, higher than neural network [165]; nonlinear dimensional reduction in protein secondary structure prediction yielded similar results compared to ANN [166]

Visualization (IV) is highly vital in presenting experimental results in the bioinformatics area [172]. Furthermore, visualization, as an advantage for an algorithm, is very important in DM [20]. IV methods are accepted as computerized techniques such as data selection, data transformation and data representation in a visual form facilitating human interaction for discovering and understanding the data (reviewed in [174]). IV approaches are based on two main functions of the human visual system: first, a human visual system with a broad bandwidth that can process a huge amount of information at one time; second, a human visual system with the ability to distinguish trends and patterns within visual areas, such as shape, location, size, and colour of objects. Thus, IV techniques have two major objectives: first, they consider a huge amount of information at a time which would not be readily perceivable by humans otherwise; second, they retrieve useful knowledge from a huge amount of information by recognizing patterns and trends [174].

There is a wide variety of IV methods. Thus, various classifications have been developed from different angles. For instance, Shneiderman's taxonomy [175], which is based on data types and tasks, includes seven data types, namely, temporal data, tree data, multidimensional data, network data, 1-D linear data, 2-D planar or map and 3-D data, and also seven tasks, namely, zoom, history, details-on-demand, filter, overview, extract and relate [174]. On the other hand, IV approaches are categorized into six groups based on data visualization methods including pixel-oriented, geometric, hierarchical, hybrid, icon-based and graph-based techniques. Besides these dimensions of IV techniques, other aspects can also be used in IV taxonomy such as distortion, data preprocessing and dynamic/interaction

techniques [176]. Another taxonomy has been proposed based on a 'data state reference model', describing four steps of data state in IV and three transformation operators between every two adjacent steps [177]. A unified taxonomic framework in the perspective of IV system designers has also been proposed [178], including further perspectives such as display dimensions, data relationships, user's skill level and context factors [174].

Hérissou and Gherbi [179] suggested a method for the three-dimensional visualization of the DNA molecule. Their method is based on a biological 3D model predicting the complex spatial trajectory of big naked DNA. This method could help to achieve a general view of the sequence instead of the textual presentation. Thus, a novel vision and an original method emerge. This method is appropriate for conducting original bioinformatics research and for analysing the spatial architecture of the genome [172]. Moreover, a new visual method and software for analysing residue mutations has been developed. This approach can combine various biological visualizations such as one-dimensional sequence views, three-dimensional protein structure views and two-dimensional views of residue interaction networks and aggregated views [180]. A method for analysing the huge and complicated datasets is to generate integrated data-knowledge networks allowing biomedical researchers to analyse the results of an experiment in the context of existing knowledge. Hence, Vehlow et al. [181] proposed a visual analytics method integrating interactive filtering of dense networks according to degree-of-interest functions with attribute-based layouts of the resulting sub-networks. Comparing multiple sub-networks with different analysis facets was provided through an interactive super-network that could integrate brushing-and-linking

methods for highlighting components across networks [181]. Additionally, for multivariate data visualization, Kuntal and Mande [182] offered a web-based platform (Web-Igloo) which is useful for visual DM.

Future perspectives

In spite of great advances in the area of bioinformatics, various issues still remain to be addressed. High-throughput sequencing, with its increasing tools and decreasing expenses, has been widely used. Scientists have been able to sequence entire genomes, analyse DNA sequence variation, quantify transcript abundance and understand mechanisms such as alternative splicing and epigenetic regulation using the first (Sanger) and the second (next) generation sequencing technologies [183]. However, yet, NGS has important challenges, such as data processing and storage. Genome interpretation is also another major challenge, which involves not only the analysis of genomes for functional elements, but the understanding of the importance of variants in individual genomes on phenotypes and disease. On the other hand, the next generation of modern and effective sequencing technologies can determine a huge deal of elusive knowledge regarding the repetitive and noncoding elements. Developments in TGS (Third Generation Sequencing) promise synergies with NGS technologies to raise our understanding of human/animal/plant genomics and genetics. NGS made a revolution in genomics-related research, and it is believed that the NGS discoveries will be continuing in near future.

Constant developments in Pool-seq (whole-genome sequencing of pools of individuals) will raise its implications in the future. First, the availability of novel software will accelerate the analysis of Pool-seq data. Then, analyses of low-frequency variants will become typical through the use of new tools. The third development considers the haplotype phasing of Pool-seq data [184]. Although existing methods are based on sequence information of founder haplotypes, an extension relaxing this requirement to only a subset of the haplotypes in the pool will make this method more general and lead to more precise estimates. Ultimately, the availability of longer sequencing reads will accelerate the reconstruction of haplotype information from Pool-seq data. This can be achieved through technological developments (such as Nanopore and PacBio sequencing), and through new library preparation protocols (such as Illumina's Synthetic Long-Read technology), allowing haplotype sequencing for DNA fragments of up to 10 kb with the current sequencing technology. Such technological advances, along with the wide variety of biological research questions requiring huge sample sizes, mean

that Pool-seq will continue to complement the sequencing of individual genomes in future [185].

Single-cell sequencing technologies have two main weaknesses: low genome coverage and high amplification bias. Despite the existence of some bioinformatics tools, new algorithms and software should be developed in order to analyse single-cell genomics data. Particularly, tools are required to assess the function of different single-cell sequencing technologies. Additionally, technical standards are needed for evaluation of the genome coverage and amplification biases. In spite of the limitations, we expect the nucleic acid sequence analysis of single-cell genomic DNAs and RNAs will be resolved in future via novel advancements in microfluidics and NGS technologies.

Various plant genomes have been sequenced at different levels of completion and many plant genome projects are underway [186–188]. Consequently, SNP discovery has become possible even in complex genomes. However, at present, there are limited SNPs from crops. Hence, there is a wide scope for production of reference genome sequences and discovery of such SNPs using NGS technologies for further understanding of plant genetics and genomics. Moreover, other issues that should be addressed are the ascertainment bias of popular bi-parental populations and the low validation rate of some array-based genotyping platforms. On the other hand, the area of epigenetic regulation of many genome components can be understood comprehensively by achieving deeper and more accurate sequencing [13].

What is more, various studies on protein classification algorithms show that no method has been developed for the classification of the proteins based on their amino-acid sequence. Therefore, novel methods could be created for the classification of the proteins based on their sequences, rather than their functional and structural features. Moreover, new ANN-inspired approaches and strategies can be used to offer predictions for higher levels of protein structures (tertiary and quaternary). Thus, protein function can be revealed and drug/enzyme therapy could be considered in the future.

Assessing the efficiency of bioinformatics methods is very important in the future improvement of the present applications and tools. For example, a comprehensive assessment is essential for obtaining insight into the effect of mutations, how they should be best mapped onto the sequence, structure, and network presentations, and how they should be combined into the visual layout [180]. Furthermore, the aggregation of network areas is another issue that can reduce the visual complexity. In fact, identifying areas of particular interest for evaluation of the potential influence of mutations could make mutation patterns with specific functional

consequences more apparent, especially, in the analysis of multiple proteins [180]. Additionally, it is thought that improving the software integration of various applications in an automated way would involve better synchronization over linked views and automated retrieval of external data [180]. Lastly, based on the present evidence, it is our belief that the discoveries in the wide range of bioinformatics domains will continue in the next decade.

Conclusions

The developments of omics technologies have led to flourishing of high throughput genome-wide scanning data. Consequently, both bioinformatics and DM is a very fast ongoing research area. They need various skills for the gathering and storing, managing and analysing, interpreting and spreading of biological information. Furthermore, high performance computers (HPC) and innovative software are required to handle and organize tremendous quantities of genomic and proteomic data. Besides low cost and high speed, another motivating reason for wide-ranging computational screens of genomic data is the fact that the complexity and extent of biological systems might best be discovered by simultaneous consideration of a broad range of genome-scale data. Hence, it is essential to explore the hot research issues in bioinformatics and enhance innovative and intelligent data-mining techniques for effective and scalable bio-data analysis.

Acknowledgments

The authors express their acknowledgements of Universiti Putra Malaysia and the Ministry of Higher Education, Malaysia for providing the financial support through Long-Term Research Grant Scheme (LRGS 2011-2016) For Food Security—Enhancing Sustainable Rice Production and Higher Institution Centre of Excellence, HiCOE initiative to the Institute of Tropical Agriculture and Food Security, Universiti Putra Malaysia, Malaysia. Additionally, the authors extend their appreciation to Marie-Curie COFUND postdoctoral fellowship at the University of Liege, Co-funded by the European Union, Belgium. We are also thankful to Dr Ruzbeh Babaee, Faculty of Letters, University of Porto, Portugal, for his valuable time in proofreading this article.


Disclosure statement

No conflicting relationship exists for any authors.

Funding

This study was financially supported by the Long-term Research Grant Scheme (LRGS), Food Security Project, Ministry of Higher Education, Malaysia [grant number 5525001].

ORCID

Farahnaz Sadat Golestan Hashemi  <http://orcid.org/0000-0001-6269-5352>

Mohd Rafii Yusop  <http://orcid.org/0000-0003-4763-6367>

References

- [1] Gil Y, Greaves M, Hendler J, et al. Amplify scientific discovery with artificial intelligence. *Science*. 2014;346:171–172.
- [2] Leach SM, Tipney H, Feng W, et al. Biomedical discovery acceleration, with applications to craniofacial development. *PLoS Comput Biol*. 2009 [cited 2017 Feb 5];5: e1000215. DOI:10.1371/journal.pcbi.1000215
- [3] Al-Haggar M, Khair-Allaha B, Islam M, et al. Bioinformatics in high throughput sequencing: application in evolving genetic diseases. *J Data Mining Genomics Proteomics*. 2013 [cited 2017 Feb 2];4:131. DOI: 10.4172/2153-0602.1000131
- [4] He Z. Data mining for bioinformatics applications. Cambridge: Elsevier; 2015. (Woodhead Publishing Series in Biomedicine; 76).
- [5] Baxeavanis AD. The importance of biological databases in biological discovery. *Curr Protoc Bioinformatics*. 2011;34:111–116.
- [6] Kadhodaie S, Barantalab F, Taheri S, et al. BioInfoBase: a bioinformatics resourceome. Ithaca (NY): Cornell University Library; 2016. [cited 2017 July 20]. Available from: <https://arxiv.org/abs/1607.02974>
- [7] Bianco AM, Marcuzzi A, Zanin V, et al. Database tools in genetic diseases research. *Genomics*. 2013;101:75–85.
- [8] Vijayarani S, Deepa MS. Protein sequence classification in data mining – a study. *Int J Infor Technol Mod Comput*. 2014 [cited 2017 Feb 5];2. DOI:10.5121/ijitmc.2014.2201
- [9] Lee GW, Kim SS. Genome data mining for everyone. *BMB Rep*. 2008;41:757–764.
- [10] Hunter L. Artificial intelligence and molecular biology. San Jose (CA): AAAI Press; 1992.
- [11] Valentini G, Tagliaferri R, Masulli F. Computational intelligence and machine learning in bioinformatics. *Artif Intell Med*. 2009;45:91–96.
- [12] Pitrat J. Artificial intelligence and heuristic methods. *Revue Francaise De Recherche Operationnelle*. 1996;10:137–137.
- [13] Kumar S, Banks TW, Cloutier S. SNP discovery through next-generation sequencing and its applications. *Int J Plant Genom*. 2012 [cited 2017 Feb 10];2012:831460. DOI:10.1155/2012/831460
- [14] Hilbert D, Neumann JV, Nordheim L. Über die Grundlagen der quantenmechanik [On the fundamentals of quantum mechanics]. *Math Ann*. 1928;98:1–30.
- [15] Piatetsky-Shapiro G, Frawley W. Knowledge discovery in databases. San Jose (CA): AAAI/MIT Press; 1991.
- [16] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. *AI Mag*. 1996;17:37–54.
- [17] Raza K. Application of data mining in bioinformatics. *Indian J Comp Sci Eng*. 2012;1:114–118.
- [18] Cacciatore S, Tenori L, Luchinat C, et al. KODAMA: an R package for knowledge discovery and data mining. *Bioinformatics*. 2017;33:621–623.

- [19] Han J. How can data mining help bio-data analysis? [extended abstract]. Paper presented at: BIODDD02: Workshop on Data Mining in Bioinformatics (with SIGKDD02 Conference); 2002 Jul 23; Edmonton (Canada). Available from: <https://web.njit.edu/~wangj/publications/bioddd02/01-han.pdf>
- [20] Esfandiari N, Babavalian MR, Moghadam AME, et al. Knowledge discovery in medicine: current issue and future trend. *Expert Syst Appl*. 2014;41:4434–4463.
- [21] Padhy N, Mishra P, Panigrahi R. The survey of data mining applications and feature scope. *Int J Comp Sci Eng Inf Tech*. 2012 [cited 2017 Feb 10];2:2. DOI:10.5121/ijcseit.2012.2303.
- [22] Pang-Ning T, Steinbach M, Kumar V. Introduction to data mining. Boston: Pearson Education Inc; 2006.
- [23] Piatetsky-Shapiro G. CRISP-DM, still the top methodology for analytics, data mining, or data science projects [Internet]. KDNuggets. 2014 [cited 2017 Feb 10]. Available from: <http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html>
- [24] Niakšū O. Development and application of data mining methods in medical diagnostics and healthcare management. [dissertation]. Vilnius: Vilnius University; 2015.
- [25] Yang B, Liu F, Ren C, et al. BiRen: predicting enhancers with a deep-learning-based model using the DNA sequence alone. *Bioinformatics*. 2017;33(13):1930–1936.
- [26] Lim KMK, Li C, Chng KR, et al. @MInter: automated text-mining of microbial interactions. *Bioinformatics*. 2016;32:2981–2987.
- [27] Kanehisa M, Bork P. Bioinformatics in the post-sequence era. *Nat Genet*. 2003;33:305–310.
- [28] Haque W, Aravind A, Reddy B, editors. Pairwise sequence alignment algorithms: a survey. Proceedings of the Conference on Information Science, Technology and Applications; 2009 Mar 20–22; Kuwait. New York (NY): ACM; 2009. p. 96–103. Available from: <http://dl.acm.org/citation.cfm?id=1551980>
- [29] Cristianini N, Hahn MW. Introduction to computational genomics: a case studies approach. New York (NY): Cambridge University Press; 2006.
- [30] Smith TF, Waterman MS. Identification of common molecular subsequence. *J Mol Biol*. 1981;147:195–197.
- [31] Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. *Science*. 1985;227:1435–1441.
- [32] Altschul SF, Gish W, Miller W, et al. Basic local alignment search tool. *J Mol Biol*. 1990;215:403–410.
- [33] Altschul SF, Madden TL, Schaffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389–3402.
- [34] Kent WJ. BLAT—The BLAST-Like alignment tool. *Genome Res*. 2002;12:656–664.
- [35] Schwartz S, Kent WJ, Smit A, et al. Human–mouse alignments with BLASTZ. *Genome Res*. 2003;13:103–107.
- [36] Ma B, Tromp J, Li M. Pattern-hunter: faster and more sensitive homology search. *Bioinformatics*. 2002;18:440–445.
- [37] Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*. 1970;48:443–453.
- [38] Delcher AL, Kasif S, Fleischmann RD, et al. Alignment of whole genomes. *Nucleic Acids Res*. 1999;27:2369–2376.
- [39] Batzoglou L, Pachter J, Mesirov B, et al. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res*. 2000;10:950–958.
- [40] Bray N, Dubchak I, Pachter L. AVID: a global alignment program. *Genome Res*. 2003;13:97–102.
- [41] Brudno M, Morgenstern B, editors. Fast and sensitive alignment of large genomic sequences. Proceedings of IEEE Computer Science Bioinformatics Conference on Comparative Genomics; 2002 Aug 14–16; Stanford (CA): IEEE; 2002. p. 138–147. DOI:10.1109/CSB.2002.1039337
- [42] Mathe E, Olivier M, Kato S, et al. Computational approaches for predicting the biological effect of p53 missense mutations: a comparison of three sequence analysis based methods. *Nucleic Acids Res*. 2006;34:1317–1325.
- [43] Thompson JD, Linard B, Lecompte O, et al. A comprehensive benchmark study of multiple sequence alignment methods: current challenges and future perspectives. *PloS One*. 2011 [cited 2017 Feb 10]; 6:e18093. DOI:10.1371/journal.pone.0018093
- [44] Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene*. 1988;73:237–244.
- [45] Higgins DG, Thompson JD, Gibson TJ. Using CLUSTAL for multiple sequence alignments. *Methods Enzymol*. 1996;266:383–402.
- [46] Bao Y, Bolotov P, Dernovoy D, et al. The influenza virus resource at the National Center for Biotechnology Information. *J Virol*. 2008;82:596–601.
- [47] Dunn CW, Hejnal A, Matus DQ, et al. Broad phylogenomic sampling improves resolution of the animal tree of life. *Nature*. 2008;452:745–749.
- [48] Eaton MJ, Martin A, Thorbjarnarson J, et al. Species-level diversification of African dwarf crocodiles (Genus *Osteolaemus*): a geographic and phylogenetic perspective. *Mol Phylogenet Evol*. 2009;50:496–506.
- [49] Kuipers RK, Joosten HJ, van Berkel WJ, et al. 3DM: systematic analysis of heterogeneous superfamily data to discover protein functionalities. *Proteins*. 2010;78:2101–2113.
- [50] Singh S, Tokhunts R, Baubet V, et al. Sonic hedgehog mutations identified in holoprosencephaly patients can act in a dominant negative manner. *Hum Genet*. 2009;125:95–103.
- [51] Zhang J, Chen X, Kent M, et al. Establishment of a dog model for the p53 family pathway and identification of a novel isoform of p21 cyclin-dependent kinase inhibitor. *Mol Cancer Res*. 2009;7:67–78.
- [52] Levasseur A, Pontarotti P, Poch O, et al. Strategies for reliable exploitation of evolutionary concepts in high throughput biology. *Evol Bioinform*. 2008;4:121–137.
- [53] Loytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*. 2008;320:1632–1635.
- [54] Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. *Science*. 2008;319:473–476.
- [55] Brandt BW, Feenstra KA, Heringa J. Multi-harmony: detecting functional specificity from sequence alignment. *Nucleic Acids Res*. 2010;38:W35–40.
- [56] Brown DP, Krishnamurthy N, Sjolander K. Automated protein subfamily identification and classification. *PLoS*

- Comput Biol. 2007 [cited 2017 Feb 10];3:e160. DOI:10.1371/journal.pcbi.0030160
- [57] Rausell A, Juan D, Pazos F, et al. Protein interactions and ligand binding: from protein subfamilies to functional specificity. *Proc Natl Acad Sci USA*. 2010;107:1995–2000.
 - [58] Stenson PD, Ball EV, Mort M, et al. Human gene mutation database (HGMD): 2003 update. *Hum Mutat*. 2003;21:577–581.
 - [59] Yang Z, Ro S, Rannala B. Likelihood models of somatic mutation and codon substitution in cancer genes. *Genetics*. 2003;165:695–705.
 - [60] Damborský J, Prokop M, Koca J. TRITON: graphic software for rational engineering of enzymes. *Trends Biochem Sci*. 2001;26:71–73.
 - [61] Sunyaev S, Ramensky V, Bork P. Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet*. 2000;16:198–200.
 - [62] Ramensky V, Bork P, Sunyaev S. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res*. 2002;30:3894–3900.
 - [63] Huang Y, Zhang L, Zhang P. A framework for mining sequential patterns from spatio-temporal event data sets. *IEEE Trans Knowl Data Eng*. 2008; 20:433–448.
 - [64] Kashyap AK, Steel J, Oner AF, et al. Combinatorial antibody libraries from survivors of the Turkish H5N1 avian influenza outbreak reveal virus neutralization strategies. *Proc Natl Acad Sci USA*. 2008;105:5986–5991.
 - [65] Wu G. Prediction of mutations in H5N1 hemagglutinins from influenza A virus. *Protein Peptide Lett*. 2006;13:971–976.
 - [66] Sheng C, Hsu W, Lee ML, et al. editors. Mining mutation chains in biological sequences. *Proceedings of the 26th International conference on Data Engineering*. 2010 Mar 1–6; Long Beach (CA): IEEE; 2010. p. 473–484. DOI:10.1109/ICDE.2010.5447869
 - [67] Wei H. Mining non-contiguous mutation chain in biological sequences based on 3D-structure [dissertation]. Singapore: National University of Singapore; 2011.
 - [68] Goya R, Meyer IM, Marra MA, et al. Applications of high-throughput sequencing. In: Rodríguez-Ezpeleta N, Hackenberg M, Aransay AM, editors. *Bioinformatics for high throughput sequencing*. New York (NY): Springer; 2012. p. 27–52.
 - [69] Li R, Zhu H, Ruan J, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res*. 2010;20:265–272.
 - [70] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009;25:1754–1760.
 - [71] Langmead B, Trapnell C, Pop M, et al. Ultrafast and memory efficient alignment of short DNA sequences to the human genome. *Genome Biol*. 2009 [cited 2017 Feb 10];10:R25. DOI:10.1186/gb-2009-10-3-r25
 - [72] Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res*. 2008;18(11):1851–1858.
 - [73] Hillier LW, Marth GT, Quinlan AR, et al. Whole genome sequencing and variant discovery in *C. elegans*. *Nat Methods*. 2008;5:183–188.
 - [74] Berkhin P. A survey of clustering data mining techniques. In: Kogan J, Nicholas C, Teboulle M, editors. *Grouping multidimensional data*. Berlin, Heidelberg: Springer; 2006. p. 25–71.
 - [75] Tasoulis D, Plagianakos V, Vrahatis M. Unsupervised clustering of bioinformatics data. Paper presented at: The European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems, Eunite; 2004 Jun 10–12; Aachen (Germany). Available from: <http://www.eunite.org/eunite/events/eunite2004/eunite2004.htm>
 - [76] Li J, Zhang Y, Tian Y. Medical big data analysis in hospital information system. In: Ventura Soto S, Luna JM, Cano A, editors. *Big data on real-world applications*. Rijeka (Croatia): InTech; 2016. [cited 2017 Jul 21]; p. 65–96. DOI: <https://doi.org/10.5772/63754>
 - [77] Engreitz JM, Daigle Jr BJ, Marshall JJ, et al. Independent component analysis: mining microarray data for fundamental human gene expression modules. *J Biomed Inform*. 2010;43:932–944.
 - [78] Antonelli D, Baralis E, Bruno G, et al. Analysis of diagnostic pathways for colon cancer. *Flex Serv Manuf J*. 2012;24:379–399.
 - [79] Mueller J, Von Eggeling F, Driesch D, et al. ProteinChip technology reveals distinctive protein expression profiles in the urine of bladder cancer patients. *Eur Urol*. 2005;47:885–894.
 - [80] Frey BJ, Dueck D. Clustering by passing messages between data points. *Science*. 2007;315:972–976.
 - [81] Bodenhofer U, Kothmeier A, Hochreiter S. APCluster: an R package for affinity propagation clustering. *Bioinformatics*. 2011;27:2463–2464.
 - [82] Kiddle SJ, Windram OP, McHattie S, et al. Temporal clustering by affinity propagation reveals transcriptional modules in *Arabidopsis thaliana*. *Bioinformatics*. 2010;26:355–362.
 - [83] Leone M, Weigt M. Clustering by soft-constraint affinity propagation: applications to gene-expression data. *Bioinformatics*. 2007;23:2708–2715.
 - [84] Liu H, Zhou S, Guan J. Detecting microarray data supported microRNA-mRNA interactions. *Int J Data Min Bioinform*. 2010;4:639–655.
 - [85] Tang D, Zhu Q, Yang F. A Poisson-based adaptive affinity propagation clustering for SAGE data. *Comput Biol Chem*. 2010;34:63–70.
 - [86] Pavlopoulos GA, O'Donoghue SI, Satagopam VP, et al. Arena3D: visualization of biological networks in 3D. *BMC Syst Biol*. 2008 [cited 2017 Feb 5];2:104. DOI:10.1186/1752-0509-2-104
 - [87] Vlasblom J, Wodak SJ. Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics*. 2009 [cited 2017 Feb 3];10:99. DOI:10.1186/1471-2105-10-99
 - [88] Wozniak M, Tiuryn J, Dutkowski J. MODEVO: exploring modularity and evolution of protein interaction networks. *Bioinformatics*. 2010;26:1790–1791.
 - [89] North B, Lehmann A, Dunbrack RL. A new clustering of antibody CDR loop conformations. *J Mol Biol*. 2011;406:228–256.
 - [90] Pandit SB, Skolnick J. TASSER_low-zsc an approach to improve structure prediction using low z-score-ranked templates. *Proteins Struct Funct Bioinform*. 2010;78:2769–2780.

- [91] Wang CW, Chen KT, Lu CL. iPARTS: an improved tool of pairwise alignment of RNA tertiary structures. *Nucleic Acids Res.* **2010**;38:W340–W347.
- [92] Yang F, Zhu Q, Tang D, et al. Using affinity propagation combined post-processing to cluster protein sequences. *Protein Peptide Lett.* **2010**;17:681–689.
- [93] Fujiwara Y, Irie G, Kitahara T. Fast algorithm for affinity propagation. In: Walsh T, editor. *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*. Vol. 3; 2011 Jul 16–22; Barcelona (Spain). Menlo Park (CA): AAAI Press; **2011**. p. 2238–2243. Available from: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.208.3617&rep=rep1&type=pdf>
- [94] Jia Y, Wang J, Zhang C, et al. editors. Finding image exemplars using fast sparse affinity propagation. *Proceedings of the ACM International Conference on Multimedia (ACM MM)*; 2008 Oct 26–31; Vancouver (Canada): ACM; **2008**. p. 639–642. DOI:10.1145/1459359.1459448
- [95] Jiang L, Dong Y, Chen N, et al. DACE: a scalable DP-means algorithm for clustering extremely large sequence data. *Bioinformatics.* **2017**;33:834–842.
- [96] Atluri G, Gupta R, Fang G, et al. Association analysis techniques for bioinformatics problems. In: Rajasekara S, editor. *Bioinformatics and computational biology*. Berlin, Heidelberg (Germany): Springer; **2009**. p. 1–13.
- [97] Becquet C, Blachon S, Jeudy B, et al. Strong-association-rule mining for large-scale gene-expression data analysis: a case study on human sage data. *Genome Biol.* **2002** [cited 2017 Feb 5];3:research0067. DOI:10.1186/gb-2002-3-12-research0067
- [98] Creighton C, Hanash S. Mining gene expression databases for association rules. *Bioinformatics.* **2003**;19:79–86.
- [99] Martinez R, Pasquier N, Pasquier C. GenMiner: mining non-redundant association rules from integrated gene expression data and annotations. *Bioinformatics.* **2008**;24:2643–2644.
- [100] McIntosh T, Chawla S. High confidence rule mining for microarray analysis. *IEEE/ACM Trans Comput Biol Bioinform.* **2007**;4:611–623.
- [101] Mohanty A, Senapati M, Lenka S. An improved data mining technique for classification and detection of breast cancer from mammograms. *Neural Comput Appl.* **2013**;22:303–310.
- [102] Loh WY. Classification and regression trees. *Wiley Interdiscip Rev Data Min Knowl Discov.* **2011**;1:14–23.
- [103] Orozova-Bekkevold I, Jensen H, Stensballe L, et al. Maternal vaccination and preterm birth: using data mining as a screening tool. *Pharm World Sci.* **2007**;29:205–212.
- [104] Leung KS, Lee KH, Wang JF, et al. Data mining on DNA sequences of hepatitis B virus. *IEEE/ACM Trans Comput Biol Bioinform.* **2011**;8:428–440.
- [105] Swan AL, Mobasheri A, Allaway D, et al. Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. *Omics: J Integr Biol.* **2013**;17:595–610.
- [106] Židek R, Sidlova V, Kasarda R, et al. Methods for distinction of cattle using supervised learning. *Int J Biol Vet Agri Food Eng.* **2014**;8:500–502.
- [107] Breiman L. Random forests. *Mach Learn.* **2001**;45:5–32.
- [108] Casella G, Berger RL. *Statistical inference*. 2nd ed. Pacific Grove (CA): Duxbury/Thomson Learning; **2002**.
- [109] John G, Langley P. Estimating continuous distributions in Bayesian classifiers. In: Besnard P, Hanks S, editors. *Proceedings of the 11th Conference on Uncertainty in Artificial Intelligence*; 1995 Aug 18–20; Montréal (Canada). San Francisco (CA): Morgan Kaufmann Publishers; **1995**. p. 338–345. Available from: <http://dl.acm.org/citation.cfm?id=2074196>
- [110] Dayhoff JE, Deleo JM. Artificial neural networks: opening the black box. *Cancer.* **2001**;91:1615–1635.
- [111] Bacardit J, Burke E, Krasnogor N. Improving the scalability of rule-based evolutionary learning. *Memetic Comput.* **2009**;1:55–67.
- [112] Cohen WW. Fast effective rule induction. In: Prieditis A, Russell SJ, editors. *Proceedings of the 12th International Conference on Machine Learning*; 1995 Jul 9–12; Tahoe City (CA). San Francisco (CA): Morgan Kaufmann Publishers; **1995**. p. 115–123. Available from: <http://dl.acm.org/citation.cfm?id=3091637>
- [113] Fürnkranz J. Separate-and-conquer rule learning. *Artif Intell Rev.* **1999**;13:3–54.
- [114] Schaefer C, Bromberg Y, Achten D, Rost B. Disease-related mutations predicted to impact protein function. *BMC Genomics.* **2012** [cited 2017 Feb 5];13(Suppl. 4):S11. DOI:10.1186/1471-2164-13-S4-S11
- [115] Yellasiri R, Rao CR. Rough set protein classifier. *J Theor Appl Inform Technol.* **2009**;5(1):1–7.
- [116] Saha S, Chaki R. A brief review of data mining application involving protein sequence classification. In: Meghanathan N, Nagamalai D, Chaki N, editors. *Advances in computing and information technology*; **2013**; Berlin, Heidelberg (Germany): Springer; **2013**. p. 469–477. Available from: https://doi.org/10.1007/978-3-642-31552-7_48
- [117] Caragea C, Silvescu A, Mitra P. Protein sequence classification using feature hashing. *Proteome Sci.* **2012** [cited 2017 Feb 5];10:S14. DOI:10.1186/1477-5956-10-S1-S14
- [118] Zhao XM, Huang DS, Cheung YM, et al. A Novel Hybrid GA/SVM system for protein sequences classification. In: Yang ZR, Yin H, Everson RM, editors. *Intelligent data engineering and automated learning–IDEAL*. Berlin, Heidelberg (Germany): Springer; **2004**. p. 11–16.
- [119] Banwait JK, Bastola DR. Contribution of bioinformatics prediction in microRNA-based cancer therapeutics. *Adv Drug Deliver Rev.* **2015**;81:94–103.
- [120] Chandra B, Gupta M. An efficient statistical feature selection approach for classification of gene expression data. *J Biomed Inform.* **2011**;44:529–535.
- [121] Maulik U, Mukhopadhyay A, Chakraborty D. Gene-expression-based cancer subtypes prediction through feature selection and transductive SVM. *IEEE Trans Biomed Eng.* **2013**;60:1111–1117.
- [122] Chen Y, Wang L, Li L, et al. Informative gene selection and the direct classification of tumors based on relative simplicity. *BMC Bioinformatics.* **2016** [cited 2017 Jul 23];17:44. DOI:10.1186/s12859-016-0893-0
- [123] Wang H, Zhang H, Dai Z, et al. TSG: a new algorithm for binary and multi-class cancer classification and informative genes selection. *BMC Med Genomics.* **2013** [cited 2017 Feb 5];6:S3. DOI:10.1186/1755-8794-6-S1-S3
- [124] Woods CT, Laederach A. Classification of RNA structure change by ‘gazing’ at experimental data. *Bioinformatics.* **2017**;33(11):1647–1655.

- [125] Liao SH, Chu PH, Hsiao PY. Data mining techniques and applications—a decade review from 2000 to 2011. *Expert Syst Appl.* **2012**;39:11303–11311.
- [126] Chen K, Kurgan LA. Neural networks in bioinformatics. In: Bianchini M, Maggini M, Jain LC, editors. *Handbook of natural computing*. Berlin, Heidelberg (Germany): Springer;2012. p. 565–583.
- [127] Lin WT, Wang SJ, Wu YC, et al. An empirical analysis on auto corporation training program planning by data mining techniques. *Expert Syst Appl.* **2011**;38:5841–5850.
- [128] Rivas T, Paz M, Martín J, et al. Explaining and predicting workplace accidents using data-mining techniques. *Reliab Eng Syst Safe.* **2011**;96:739–747.
- [129] Cesana M, Cerutti R, Grossi E, et al. Bayesian data mining techniques: the evidence provided by signals detected in single-company spontaneous reports databases. *Drug Inf J.* **2007**;41:16–28.
- [130] Trafalis TB, White A. Data mining techniques for pattern recognition: tornado signatures in doppler weather radar data. *Int J Smart Eng Syst Des.* **2003**;5:347–359.
- [131] Zhang C, Ramirez-Marquez JE. Approximation of minimal cut sets for a flow network via evolutionary optimization and data mining techniques. *Int J Performability Eng.* **2011**;7:21–31.
- [132] Aliev RA, Aliev RR, Guirimov B, et al. Dynamic data mining technique for rules extraction in a process of battery charging. *Appl Soft Comput.* **2008**;8:1252–1258.
- [133] Ma PCH, Chan KCC. An effective data mining technique for reconstructing gene regulatory networks from time series expression data. *J Bioinform Comput Biol.* **2007**;5:651–668.
- [134] Tu C, Chang C, Chen K, et al. Application of data mining technique in the performance analysis of shipping and freight enterprise and the construction of stock forecast model. *J Convergen Infor Technol.* **2011**;6:331–342.
- [135] Dutta M, Mukhopadhyay A, Chakrabarti S. Effect of galvanizing parameters on spangle size investigated by data mining technique. *ISIJ Int.* **2004**;44:129–138.
- [136] Tsai C, Chen M. Using adaptive resonance theory and data-mining techniques for materials recommendation based on the e-library environment. *Electron Libr.* **2008**;26:287–302.
- [137] Srivastava AN, Oza NC, Stroeve J. Virtual sensors: using data mining techniques to efficiently estimate remote sensing spectra. *IEEE Trans Geosci Remote Sensing.* **2005**;43:590–600.
- [138] Brutlag D, Davison D, Chang AC. BIOMEDIN 231: computational molecular biology [Internet]. **2014** [cited 2017 Feb 5]; Available from: <http://cmgm3.stanford.edu/biochem/biochem218/Projects%202014/Chang.pdf>
- [139] Lancashire LJ, Lemetre C, Ball GR. An introduction to artificial neural networks in bioinformatics—application to complex microarray and mass spectrometry datasets in cancer studies. *Brief Bioinform.* **2009** [cited 2017 Feb 5]; bbbp012:1–15. DOI:10.1093/bib/bbp012
- [140] Garraway LA, Lander ES. Lessons from the cancer genome. *Cell.* **2013**;153:17–37.
- [141] Hofree M, Shen JP, Carter H, et al. Network-based stratification of tumor mutations. *Nat Methods.* **2013**;10:1108–1115.
- [142] Chang JN, Collisson EA, Mills GB, et al. The cancer genome atlas pan-cancer analysis project. *Nat Genet.* **2013**;45:1113–1120.
- [143] Winter C, Kristiansen G, Kersting S, et al. Google goes cancer: improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLOS Comput Biol.* **2012** [cited 2017 Feb 5];8: 002511. DOI:10.1371/journal.pcbi.1002511
- [144] Xiang Y, Zhang CQ, Huang K. Predicting glioblastoma prognosis networks using weighted gene co-expression network analysis on TCGA data. *BMC Bioinform.* **2012** [cited 2017 Feb 5];13:S12. DOI:10.1186/1471-2105-13-S2-S12
- [145] March HN, Rust AG, Wright NA, et al. Insertional mutagenesis identifies multiple networks of cooperating genes driving intestinal tumorigenesis. *Nat Genet.* **2011**;43:1202–1209.
- [146] Rozenblatt-Rosen O, Deo RC, Padi M, et al. Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature.* **2012**;487:491–495.
- [147] Thomas R, Thomas RS, Auerbach SS, et al. Biological networks for predicting chemical hepatocarcinogenicity using gene expression data from treated mice and relevance across human and rat species. *PLoS One.* **2013** [cited 2017 Feb 5];8:e63308. DOI:<https://doi.org/10.1371/journal.pone.0063308>
- [148] Won HH, Kim JW, Lee DA. Bayesian ensemble approach with a disease gene network predicts damaging effects of missense variants of human cancers. *Hum Genet.* **2013**;132:15–27.
- [149] Horn H, Lawrence MS, Hu JX, et al. A comparative analysis of network mutation burdens across 21 tumor types augments discovery from cancer genomes. *BioRxiv.* **2015** [cited 2017 Feb 5];025445. DOI:<https://doi.org/10.1101/025445>
- [150] Shannon P, Markiel A, Ozier O, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **2003**;13:2498–504.
- [151] Goh KI, Cusick ME, Valle D, et al. The human disease network. *Proc Natl Acad Sci USA.* **2007**;104:8685–8690.
- [152] Bauer-Mehren A, Bunschus M, Rautschka M, et al. Gene-disease network analysis reveals functional modules in mendelian, complex and environmental diseases. *PLoS One.* **2011** [cited 2017 Feb 5];6:e20284. DOI:10.1371/journal.pone.0020284
- [153] Godinez WJ, Hossain I, Lazic SE, et al. A multi-scale convolutional neural network for phenotyping high-content cellular images. *Bioinformatics.* **2017**;33(13):2010–2019.
- [154] Zhao Z, Yang Z, Lin H, et al. Drug drug interaction extraction from biomedical literature using syntax convolutional neural network. *Bioinformatics.* **2016**;32:3444–3453.
- [155] Yaseen A, Li Y. Context-based features enhance protein secondary structure prediction accuracy. *J Chem Inf Model.* **2014**;54:992–1002.
- [156] Meiler J, Baker D. Coupled prediction of protein secondary and tertiary structure. *Proc Natl Acad Sci USA.* **2003**;100:12105–12110.
- [157] Garnier J, Osguthorpe DJ, Robson J. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J Mol Biol.* **1978**;120:97–120.
- [158] Thompson MJ, Goldstein RA. Predicting protein secondary structure with probabilistic schemata of evolutionary derived information. *Protein Sci.* **1997**;6:1963–1975.
- [159] Bordoloi H, Sarma KK. Protein structure prediction using artificial neural network. *Int J Comput Appl Electron Inf Commun Eng.* **2011**;3:22–25.

- [160] Pham TH, Satou K, Ho TB. Support vector machines for prediction and analysis of beta and gamma-turns in proteins. *J Bioinform Comput Biol.* **2005**;03:343–358.
- [161] Jaiswal K. Prediction of ubiquitin proteins using artificial neural networks, hidden Markov models, and support vector machines. In *Slico Biol.* **2007**;7:559–568.
- [162] Zhang Q, Yoon S, Welsh WJ. Improved method for predicting beta-turn using support vector machine. *Bioinformatics.* **2005**;21:2370–2374.
- [163] Johal AK, Singh R. Protein secondary structure prediction using improved support vector machine and neural networks. *Int J Eng Comp Sci.* **2014**;3:3593–3597.
- [164] Bakhtiarzadeh MR, Moradi-Shahrbabak M, Ebrahimi M, et al. Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology. *J Theor Biol.* **2014**;356:213–222.
- [165] Salamov AA, Solovyev VV. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J Mol Biol.* **1995**;247:11–15.
- [166] Simas GM, Botelho SSC, Grando N, et al. Dimensional reduction in protein secondary structure prediction- non-linear method improvements in innovations in hybrid intelligent systems. In: Corchado E, Corchado J, Abraham A, editors. *Innovations in hybrid intelligent systems.* Berlin, Heidelberg (Germany): Springer; **2007**. p. 425–432.
- [167] Uziela K, Hurtado DM, Shu N, et al. ProQ3D: improved model quality assessments using deep learning. *Bioinformatics.* **2017**;33(10):1578–1580.
- [168] Gao J, Yang Y, Zhou Y. Predicting the errors of predicted local backbone angles and non-local solvent- accessibilities of proteins by deep neural networks. *Bioinformatics.* **2016**;32:3768–3773.
- [169] Zeng H, Edwards MD, Liu G, et al. Convolutional neural network architectures for predicting DNA-protein binding. *Bioinformatics.* **2016**;32:i121–i127.
- [170] Prati RC, Batista GE, Monard MC. A survey on graphical methods for classification predictive performance evaluation. *IEEE Trans Knowl Data Eng.* **2011**;23:1601–1618.
- [171] Cao X, Maloney K, Brusica V. Data mining of cancer vaccine trials: a bird's-eye view. *Immunome Res.* **2008**;4:7.
- [172] Ren J, Lu J, Wang L, et al. Data visualization in bioinformatics. *Adv Inf Sci Serv Sci.* **2012**;4:157–165.
- [173] Amir ED, Davis KL, Tadmor MD, et al. viSNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. *Nat Biotechnol.* **2013**;31:545–552.
- [174] Tao Y, Liu Y, Friedman C, et al. Information visualization techniques in bioinformatics during the postgenomic era. *Drug Discov Today.* **2004**;2:237–245.
- [175] Shneiderman B, editor. The eyes have it: a task by data type taxonomy for information visualizations. *Proceedings of the IEEE Symposium on Visual Languages*; 1996 Sep 3–6; Boulder (CO): IEEE Xplore; **1996**. p. 336–343. DOI:10.1109/VL.1996.545307
- [176] Keim DA, Ankerst M. Visual data mining and exploration of large databases. Paper presented at: Tutorial at the 5th European Conference on Principles and Practice of Knowledge Discovery in Databases; 2001 Sep 3–5; Freiburg (Germany).
- [177] Chi EH, editor. A taxonomy of visualization techniques using the data state reference model. *Proceedings of the IEEE Symposium on Information Visualization*; 2000 Oct 9–10; Salt Lake City (UT): IEEE Xplore; **2000**. p. 69–75. DOI:10.1109/INFVIS.2000.885092
- [178] Pfitzner D, Hobbs V, Powers D. A unified taxonomic framework for information visualization. In: Pattison T, Thomas B, editors. *Proceedings of the Asia-Pacific Symposium on Information Visualisation.* Vol. 24. 2003 Feb 3–4; Adelaide (Australia). Darlinghurst (Australia): Australian Computer Society; **2003**. p. 57–66. Available from: <http://dl.acm.org/citation.cfm?id=857087>
- [179] Hérisson J, Gherbi R, editors. Model-based prediction of the 3D trajectory of huge DNA sequences interactive visualization and exploration. *Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering (BIBE'01)*; 2001 Nov 4–6; Bethesda (MD). Washington (DC): IEEE Computer Society; **2001**. p. 263–270. Available from: <http://www.computer.org/csdl/proceedings/bibe/2001/1423/00/14230263.pdf>
- [180] Doncheva NT, Klein K, Morris JH, et al. Integrative visual analysis of protein sequence mutations. *BMC Proc.* **2014**;8(Suppl. 2):S2. DOI:10.1186/1753-6561-8-S2-S2
- [181] Vehlow C, Kao DP, Bristow MR, et al. Visual analysis of biological data-knowledge networks. *BMC Bioinformatics.* **2015** [cited 2017 Jan 28];16:135. DOI:10.1186/s12859-015-0550-z
- [182] Kuntal BK, Mande SS. Web-igloo: a web based platform for multivariate data visualization. *Bioinformatics.* **2017**;33:615–617.
- [183] Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. *Hum Mol Genet.* **2010**;19:R227–R240.
- [184] Schlötterer C, Tobler R, Kofler R, et al. Sequencing pools of individuals – mining genome-wide polymorphism data without big funding. *Nat Rev Genet.* **2014**;15:749–763.
- [185] Cao C, Sun X. Combinatorial pooled sequencing: experiment design and decoding. *Quant Biol.* **2016**;4:36–46.
- [186] Feuillet C, Leach JE, Rogers J, et al. Crop genome sequencing: lessons and rationales. *Trends Plant Sci.* **2011**;16:77–88.
- [187] Golestan Hashemi FS, Rafii MY, Ismail MR, et al. Biochemical, genetic and molecular advances of fragrance characteristics in rice. *Crit Rev Plant Sci.* **2013**;32:445–457.
- [188] Golestan Hashemi FS, Rafii MY, Ismail MR, et al. The genetic and molecular origin of natural variation for the fragrance trait in an elite Malaysian aromatic rice through quantitative trait loci mapping using SSR and gene-based markers. *Gene.* **2015**;555:101–107.