

Bioinformatics and biological data mining

Aditya Harbola¹, Deepti Negi¹, Mahesh Manchanda¹ and Rajesh Kumar Kesharwani²

¹*School of Computing, Graphic Era Hill University, Dehradun, India,* ²*Nehru Gram Bharati Deemed University, Prayagraj, India*

27.1 Biological data mining

The genomics and DNA microarrays projects have contributed toward a large set of biological data, which present a challenge for the researchers to store, filter, and analyze the information (Wegener, Rossi, Buffa, Delorenzi, & Riiping, 2011). There is an acute necessity for biological data mining after the postgenomic era when data size increased with tremendous growth and a lot of genome projects started experimentation. Bioinformatics is all about collecting, mining, analyzing, and finding valuable information from the vast available biological data (Dua & Chowriappa, 2012). In recent years the field of bioinformatics has become the driving force behind biological research advancements which has surpassed the traditional scientific procedures. Now researchers have started applying the computational resources to mine the knowledge through novel techniques and a lot of insight from biological data is gathered. This trend will continue further as many bioinformatics researchers will work in developing novel algorithms and techniques which are accurate, fast, and reliable while drawing knowledge from the biological data (Bayer, Aksogan, Celik, & Kondiloglu, 2017).

With the advancements in computing power and related data mining technologies, a lot of research is going on biological datasets and research problems, but still, there are many fundamental problems in bioinformatics that need to be addressed through data mining. Data mining is perfectly suitable for the bioinformatics processes as the term data mining started way back in 1990 when there was a need of discovering patterns from a large set of data (Mahmud, Kaiser, Hussain, & Vassanelli, 2018). With the advancement in bioinformatics, the size of data is increasing in leaps and bounds. The vast availability and easily accessible data sets have opened new research directions and also created new research challenges. So data mining is becoming an important tool for understanding and developing new solutions for bioinformatics and biological data sets and data mining will be an important tool for analyzing huge volumes of heterogeneous, distributed, semistructured, and interrelated data for knowledge discovery (Agarwal, 2014).

In this chapter first, we will study data mining concepts and techniques. Then, we will focus on different data mining algorithms for the knowledge discovery process. In the next section, we will study the evolution and different category of biological data. Then, we will focus on the online available biological databases and applications of data mining techniques in biological science.

27.2 Data mining applications

Data mining is a practice of discovering patterns in huge data sets involving methods at the meeting point of machine learning (ML), statistics, and database systems. Data mining is a subfield of computer science and statistics with a general goal to dig out information (with intellectual methods) from a data set and convert the information into an understandable arrangement for further use. Data mining is the primary step of the knowledge discovery in database process, or (Herbert & Wang, 2007). Apart from the raw scrutiny step, it also involves the following steps (Fig. 27.1):

1. database and data management aspects;
2. data preprocessing;
3. model and inference consideration;
4. interestingness metrics;
5. complexity considerations;

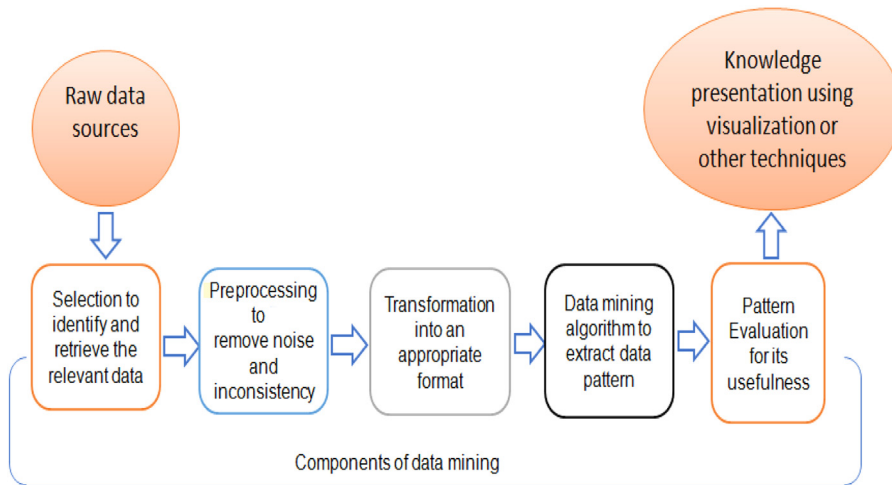


FIGURE 27.1 Data mining steps converting raw data to knowledge representation.

6. postprocessing of discovered structures;
7. visualization; and
8. online updating.

The user needs to know that data mining is an interactive and iterative process. It is due to this interactive and iterative nature that data mining finds its place as an experimental approach and that researchers can try various possibilities before discovering a single solution.

27.3 Data mining process

Data mining is an area of computer science with enormous potential, which is the process of mining information from huge biological databases or datasets. Data mining has a wide application in bioinformatics because of the fast-evolving biological data (Ocana, Silva, Oliveira, & Mattoso, 2015). The data mining process is indulged in nearly every aspect of biological data analysis, which includes application in gene analysis, gene discovery, protein function analysis, new disease analysis, disease treatment, and drug optimization. The principal tasks for any data mining process are as follows.

27.3.1 Classification

There are numerous techniques under data mining and one of them is classification. It is a process of assigning a data record to a predefined class or label so that each record of the dataset can be assigned a label. For classification, integrated algorithms in the Waikato Environment for Knowledge Analysis (WEKA) tool is used for the inbuilt data classifier algorithms, such as k-nearest neighbor (k-NN), Naïve Bayes (NB), and support vector machine (SVM) classifiers.

Classification further can be divided into two categories, supervised and unsupervised (Ang, Mirzal, Haron, & Hamed, 2016). In a supervised classification model, the method gains knowledge of a categorized dataset, on the condition that a response key which the algorithm can utilize to estimate its correctness on training data. In disparity, an unsupervised representation provides unlabeled data that the algorithm outputs by mining features and patterns. Data mining approaches are being used for health, biomedical, and clinical data analysis to find the knowledge full information. Medical practitioner extracts the information from the analysis of large data set to make a decision related to health risk, and health services. Data mining techniques can be used for the prediction of risks for cancer and other diseases to suggest an early diagnosis, preventive measures, and therapy to avoid the condition of the critical stage of the disease. For example, in a cancer dataset, a particular record of the data can be classified as benign (noncancerous) or malignant (cancerous). Using different data mining algorithms, we can know a cancer patient is a benign or malignant type. Chaurasia, Pal, and Tiwari (2018) have used NB classifier in the breast cancer dataset to achieve a high degree of certainty.

27.3.2 Estimation

Estimation is required to determine a value for unidentified continuous variables. It is of two types, point estimate, and interval estimate. This step answers how to estimate a set of parameters for quantitatively characterizing each patient

(Hirata, Morino, & Suzuki, 2016). In the earlier period, several artificial neural network (ANN) models have been developed by researchers for breast cancer risk prediction. The researchers have evaluated whether an ANN trained on a huge composed dataset of successive mammography findings can distinguish between benign and malignant disease and precisely predict the likelihood of breast cancer for individual patients. Finally, 96.5% accuracy was achieved in cancer estimation using 3-layer feedforward ANN with 1000 hidden-layer nodes (Ayer et al., 2010).

27.3.3 Prediction

It is required to classifying records according to estimated prospect behavior. Predicting the behavior of records is an important task. We train algorithms on medical data for example cancer data set, from the estimation sample. Prediction results can be compared with datasets of real-world diagnosis to assess the accuracy of decisions (Sidey-Gibbons & Sidey-Gibbons, 2019). The accuracy of the prediction model is a very important factor because higher accuracy guarantees the acceptance of the prediction model. Several algorithms, such as decision tree, hidden Markov model (HMM), neural networks, and regression models, are used for different types of prediction. Prediction models have been developed for different purposes, such as gene prediction, proteomics analysis, drug metabolism, and toxicity prediction, QSAR analysis, health risk analysis, clinical decision making, and many other applications. A basic concept for the development of the health risk prediction model is represented in Fig. 27.2.

27.3.4 Association

Grouping or assigning items that are collectively the same in the dataset is called association. Association rule is a data mining tool that guarantees the finding of association activities in data by processing the data at hand. The discovered patterns demonstrate the relationship between the feature values that mainly cooccur in the data set. For example, rule {fruit jam, butter} = > {bread}, according to a rule gathered from a grocery shop transaction. which is explained as if a person buys fruit jam and butter then this person will most likely buy bread, which is an analysis of the association rule. Algorithmic steps involved in finding the most frequent itemsets using the associations rule has been shown in Fig. 27.3.

27.3.5 Clustering

Working with biomedical data needed clustering algorithm implementation because dealing with huge data requires the grouping of the data in a set of categories which can help to find new domains of the data. The clustering or grouping

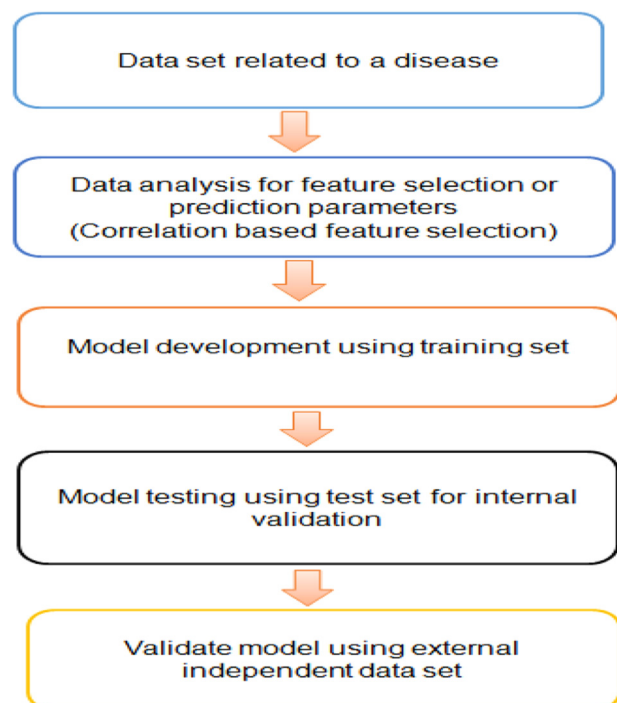


FIGURE 27.2 Basic steps in the development of health risk prediction model.

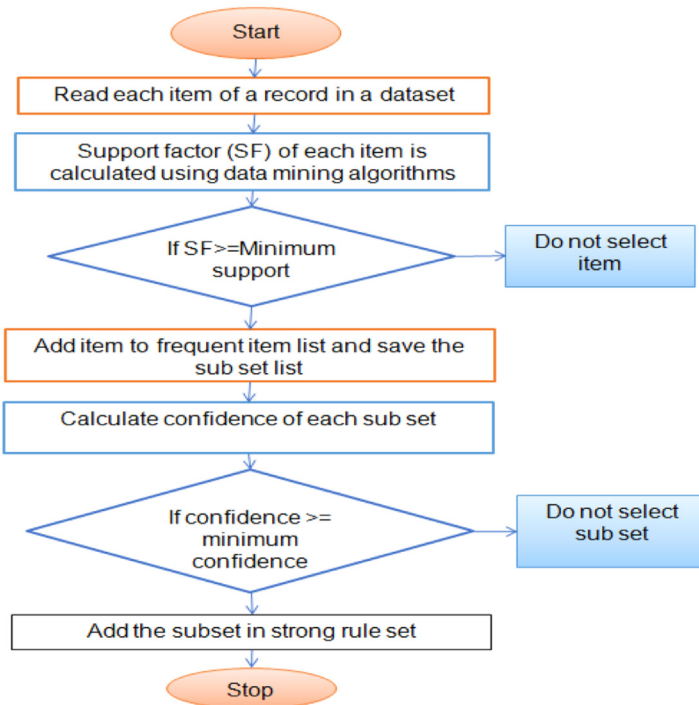


FIGURE 27.3 Steps in finding the association between data items.

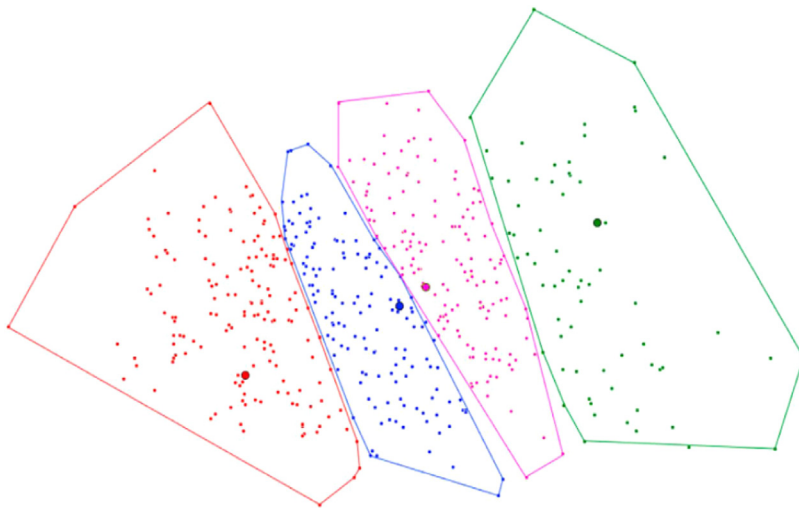


FIGURE 27.4 2D view of four subgroups of hypertensive patients using clustering.

can be based on any commonness or dissimilarity according to the clustering rules. It is an important step because it assigns a population into subgroups or clusters. Several clustering algorithms are available which are being used in bioinformatics, such as K-medoids, K-means, Fuzzy C-means, density-based spatial clustering of applications with noise, and selforganizing map clustering.

Cluster analysis of hypertensive patients' data has been done to identify the possible subgroups based on the risk of coronary artery disease (Guo et al., 2017). Four subgroups were identified in the hypertensive patients based on features, such as age, sex, blood pressure, smoking pattern, triglycerides, diabetes, cholesterol, fasting glucose, very low-density lipid, and many other criteria (Fig. 27.4).

27.3.6 Description and visualization

In the last step, the result visualization is needed so in this step data are represented in an understandable form. Each data mining tool (WEKA, MATLAB, etc.) provided visualization in form of graphs, charts, and summary features. The

visualization gives a brief representation of the results and accuracy. These steps are very general for any data mining process. If we further inspect these steps, there are a few substeps that are necessary and crucial for accurate and efficient analysis of biological data. One of the substeps is feature selection, which is an important subprocess of the classification process. In the following section, we will discuss feature selection techniques and its objectives.

27.3.7 Case studies using Waikato environment for knowledge analysis

To explain data mining processes, a breast cancer dataset is analyzed using WEKA tool. WEKA workbench contains an environment for classification, clustering, feature selection, visualization, and regression. These WEKA tools are very useful for data mining problems in bioinformatics research (Frank, Hall, Trigg, Holmes, & Witten, 2004). WEKA is open-source software under GNU license developed by the University of Waikato, New Zealand. It can be downloaded from the URL <https://www.cs.waikato.ac.nz/mL/weka/>. After downloading and installation of the open-source WEKA ML workbench, a lot of datasets are also comes as the default installed. WEKA includes data or directory, which is full of benchmark ML problems.

Example: One of the datasets is the breast cancer dataset available as breast-cancer.arff. The breast cancer dataset has 9 fields and 286 records and is available for research purposes on different dataset websites (Brownlee, 2020). Each data record represents the health details of patients and samples of their tumor tissue and the assignment is to forecast whether or not the patient has breast cancer. There are nine input variables all of which a nominal. This dataset is the binary classification where the output variable to be predicted is nominally comprised of two classes. After loading this dataset in the WEKA tool, all nine fields and the class field of the dataset can be visualized with visualizing all buttons. This visualization tool displays all the fields available in the dataset.

The study of breast cancer has been done on 10 attributes, such as clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, size of bare nuclei, bland chromatin, normal nucleoli, mitoses, and class. Clump thickness points to the radius that was calculated by averaging the length of radial line segments from the center of the nuclear mass to each of the points of the nuclear boundary. For cell size, a perimeter was calculated as the distance around the nuclear boundary which is considered to be uniform. For measuring the cell shape, the area is calculated by including the number of pixels in the interior of the nuclear border and adding one-half of the pixels on the perimeter. Trivial adhesion is measured by combining the perimeter and area to give a measure of the compactness of the cell nuclei using the formula: $\text{perimeter}^2/\text{area}$.

Classification algorithms are available in WEKA under classify button. After loading the dataset one must choose the classifier to apply. The examination has been carried on using algorithm J48 which is available under the trees classifier category. J48 predicts 75.5% accuracy in correctly classified Instances. The confusion matrix shows that the true positive rate is around 96% for no-recurrence-events or benign cancer.

Clustering algorithms are available under the clustering button. In this example, a simple K-means clustering algorithm is used. The number of clusters and distance functions for constructing clusters can be chosen by configuration. A two cluster and Euclidean distance function produce the following result, where cluster 0 has 225 (79%) records and cluster1 has 61 (21%) of records. It implies that the simple K-means algorithm has achieved an accuracy of 79% in clustering the data records in cluster0.

The association tab provides an algorithm to generate association rules in antecedent = > consequent format. Apriori algorithm for association rule can be configured in WEKA. The breast cancer dataset in WEKA is checked for association rules using the Apriori algorithm with the default parameter. The algorithm provides 10 rules. Here, are few observations of the 10 association rules.

1. Most of the rules have node-caps = no as a consequence.
2. Rule with a confidence of more than 0.9 is selected by the algorithm.
3. The long antecedent is more likely to be fragile than the short antecedent.
4. It is not always good for all datasets to generate association rules because each data set might not have association properties.

27.4 Feature selection technique in data mining

The significance of the features of the dataset is also a very important factor. The significance of a feature (f) is constantly measured by its capability to differentiate instances of the dataset concerning the target class to which the instance belongs. Significant features are categorized into two types: (1) strongly significant: those that are strongly

significant to the dataset/distribution and (2) weakly significant: those that are weakly significant to the dataset/distribution.

So the idea is to work with strong features that can contribute toward better output and knowledge discovery. There are many approaches to select strong features among various feature sets.

1. Pearson correlation coefficient;
2. information-theoretic ranking criteria;
3. feature subset generation;
4. Gram–Schmidt forward feature selection;
5. feature construction and extraction;
6. principal component analysis (PCA); and
7. independent component analysis.

The above-listed techniques can be used for the data preparation and transformation process and to extract significant features after filtering the features from the data set. These are the fundamental step in any data mining process for dataset processing.

27.4.1 Objective of feature selection

Feature selection is an important process of most bioinformatics applications. The key objectives of the feature selection are as follows.

1. to discard the problem of overfitting;
2. to improve the performance of the applied model;
3. to develop faster and efficient computational models;
4. to achieve a better understanding of the process of data generation; and
5. to achieve a set of features that performs better with certain data mining algorithm.

Among several feature selection techniques, it can be broadly classified into the following three categories.

1. Filter: The filter technique applies an autonomous test exclusive of involving any learning algorithm. Filter techniques have a small computational cost with unsatisfactory reliability in classification.
2. Wrapper: Wrapper technique necessitates a predecided learning algorithm for attribute subset assessment. Wrapper methods are likely to have better classification accuracy but involve immense computational power.
3. Hybrid: It combines both methods into an in order search method with the plan to enhance the classification performance of the features selection.

Feature selection is a vital process of biomedical data classification and strong feature selection has shown that it can enhance knowledge discovery and model analysis (Peng, Wu, & Jiang, 2010).

27.5 Major data mining algorithms applicable to biological data

Data mining methods and algorithms are being broadly used in artificial intelligence (AI) and ML techniques. Using AI and ML biological data processing and knowledge discovery techniques are being developed. Tree-based techniques, such as decision trees, are one of the accepted data mining algorithms useful in bioinformatics and statistical genetics. Some of the significant data mining algorithms applicable to biological data are as follows.

27.5.1 C4.5 algorithm

C4.5 is a decision tree based classifier algorithms and was developed by Ross Quinlan. C4.5 is used to make a classifier in the structure of a decision tree from a set of data that has previously been classified (Chen, Wang, & Zhang, 2011). For example, C4.5 can be used in the analysis of cancer patient data for the classification of different types of cancer.

27.5.2 K-means algorithm

It is one of the unsupervised clustering algorithms and works by constructing a k number of the cluster from a set of objects based on the likeness among objects. Each group member will be more alike as measure up to nongroup

members. K-means clustering has been used to cluster DNA sequences of hepatitis B virus (Bustamam, Tasman, Yuniarti, Frisca, & Mursidah, 2017)

27.5.3 Support vector machine

It is a supervised ML algorithm and can be used for classification or regression problems. Support vectors are the data points that lie nearby to the assessment surface.

Apart from the above-mentioned algorithms, a few to mention are the Apriori algorithm, which is based on association rules. The AdaBoost algorithm is based on the concept of the weak and strong learner. The k-NN based on lazy learner and naïve base is a bunch of classification algorithms. These algorithms are already available with their code for research purposes and few tools are also available. WEKA provides these algorithms in a graphical user interface for the analysis and interpretation of data sets.

27.6 Biological data evolution and related issues

In the last few decades, technological advancements in research areas have produced a sudden increase in information. The capability to differentiate and understand diseases is growing exponentially based on data obtained from genetic and proteomic studies, clinical studies, and other research accomplishments. The focus is now shifted toward data-driven research, which is searching for biological patterns and analyzing them (Xiong & Xiong, 2012). The genome project was a milestone toward biological data.

Biological data are either sequence, structure, or functional data. Nucleic acid and protein sequences contribute to the creation of the sequence database and the structure database contains only protein structure. The first protein sequencing was done for Insulin way back in 1956 for which 51 residues were collected (Tang & Tan, 2011). Then, the nucleic acid sequencing was done around 1965 with 77 bases. The first structure of the protein was developed in a protein data bank with 10 entries in 1972 (Foulkes et al., 2017). These were the initial milestones toward biological database creation. Nucleic acid and protein sequences are stored in sequence databases and structure databases store solved structures of RNA and proteins. Functional databases provide information on the physiological role of gene products, for example, enzyme activities, mutant phenotypes, or biological pathways. Further classification of biological data types are as follows:

1. nucleotide sequences (DNA and mRNA);
2. protein sequences;
3. 3-D protein structure;
4. genome and maps;
5. gene expression; and
6. genetic variation.

Data can be classified into the following three categories:

1. Primary data: Primary data in the context of biological data are mainly the records of sequence or structure.
2. Secondary data: It contains the findings or the processed output of the primary data. It may contain the classified entries of the sequence or structure for example alpha or beta proteins.
3. Composite data: It contains a combination of different sources of primary data because there is a need to search for multiple resources. NCBI hosts several composite databases that can be freely accessed by the researchers.

The issues of handling biological data are quite different from the challenges of traditional data. The notable challenges with biological data are as follows.

27.6.1 Biological data availability

The real issue of biological data is its availability for scientists. Biological databases are results of the scientific process which are published by the researcher or scientists so that other researchers or scientist can use the data. The databases contain genome, protein, microarray related data which is collected through the large scale analysis of gene structure, cells and chromosomes, and mutation. Biological data has a variety of sources and formats. The format can be text, sequences, images, or links. Most of the databases are publicly available through online sources (Dua & Chowriappa, 2012). Few popular sources are:

1. PubMed: This search engine is available from 1996 and maintained by the National Library of Medicine, United States. It gives results from MEDLINE databases on life sciences and biomedical topics.

2. Online Mendelian Inheritance in Man (OMIM): OMIM contains the record of human genes with a special focus on gene–phenotype relationships. It is maintained by the John Hopkins University of medicine.
3. GenBank: The first release was in 1982 and it is maintained by NCBI. It contains nucleotide and protein sequences. It is an open-access database and is a collaborative work of laboratories throughout the world.
4. UniPort: It is a protein sequencing database and the major part of the database is derived from genome sequencing projects. It is maintained by a group of organizations from Europe, Switzerland, and the United States.
5. PDB: PDB is an acronym for protein data bank. It contains the 3-D structure of proteins and nucleic acids. It is a collaborative work contributed by scientists from around the world.
6. SCOP: Structural Classification of Proteins (SCOP) database contains protein structural domains. This classification is needed to find the relationship between proteins. It was released in 1994 by the laboratory of molecular biology, England. SCOP1.75 (2014), SCOPe (2018), and SCOP2 (2020) are a few versions of SCOP.
7. CATH: It is a protein classification database released in 1997. The database was created in University College London as an open-source project. The latest release CATH-Gene3D released in 2016.

27.6.2 Biological data availability in computer-readable form

As the analysis of the large-scale biological data requires computer involvement, the data must be readable by the computer. Genome, protein, nucleotide sequences data must be in a format that can be read by the computer so that results can be obtained (Tang & Tan, 2011).

27.6.3 Biological data cleaning

Data cleaning is an important issue with any kind of data analysis because it can improve the quality and knowledge discovery. Data cleaning is the process of identifying and eliminating duplicity, ambiguity, errors that can lower the quality of the results obtained after the data mining process. Biological data are loaded with data inconsistencies, data duplications, spelling mistakes, missing values that must be handled with data cleaning methodologies. Also, the inconsistent format, nomenclature, vocabulary, and annotations have contributed to a need for data cleaning (Sowmya & Suneetha, 2017). The steps for data cleaning are as follows:

1. Removal of superfluous data: In databases, superfluous observations can lead to ambiguous observations, so the very first step toward data cleaning. Duplicate and irrelevant data are categories of superfluous data. Irrelevant data consume a lot of effort but could lead to vague results.
2. Handle missing data: Some time few data records are not complete and few attributes are missing from the database. It can be handled through data validation, dropping missing values, or by inserting the missing values.
3. Fix data structure: After handling superfluous and missing data the data structure is a prime concern. There may be structural errors during data transfer or data collection.
4. Filter-out outliers: Outliers can drastically change the result. To improve the performance of the data mining models, the outliers must be removed but there must be a strong reason to remove an outlier.

27.6.4 Biological data quality

With fast-evolving biological databases, data transformation and data cleaning are two important tools for maintaining data quality and accurate results. Data transformation is very effective in mapping the format of the data and the format required by the application (Ocana et al., 2015). It also helps in multiple source database mapping while solving multiple source problems (Fig. 27.5). Data cleaning is required at the instance level where the errors and inconsistencies are visible in the actual data.

27.6.5 Biological data dimensionality

In bioinformatics, the natural outsized number of dimensions, called the curse of dimensionality, has ubiquitous effects and is a big challenge for the researchers. Let there are Y records in the database and each record has several features, X (Dua & Chowriappa, 2012). The challenges arise when there is a small Y big X case situation. Most of the data mining statistical algorithms could not handle the $X > Y$ situation, and it is very important for information mining to select, handle, and analyze the large and multidimensional data set (Fig. 27.6). For example, in many cases, $X \rightarrow \text{infinity}$ and Y can be a fixed value as in datasets of many genes describing relatively few samples of genetic diseases. Genomic and

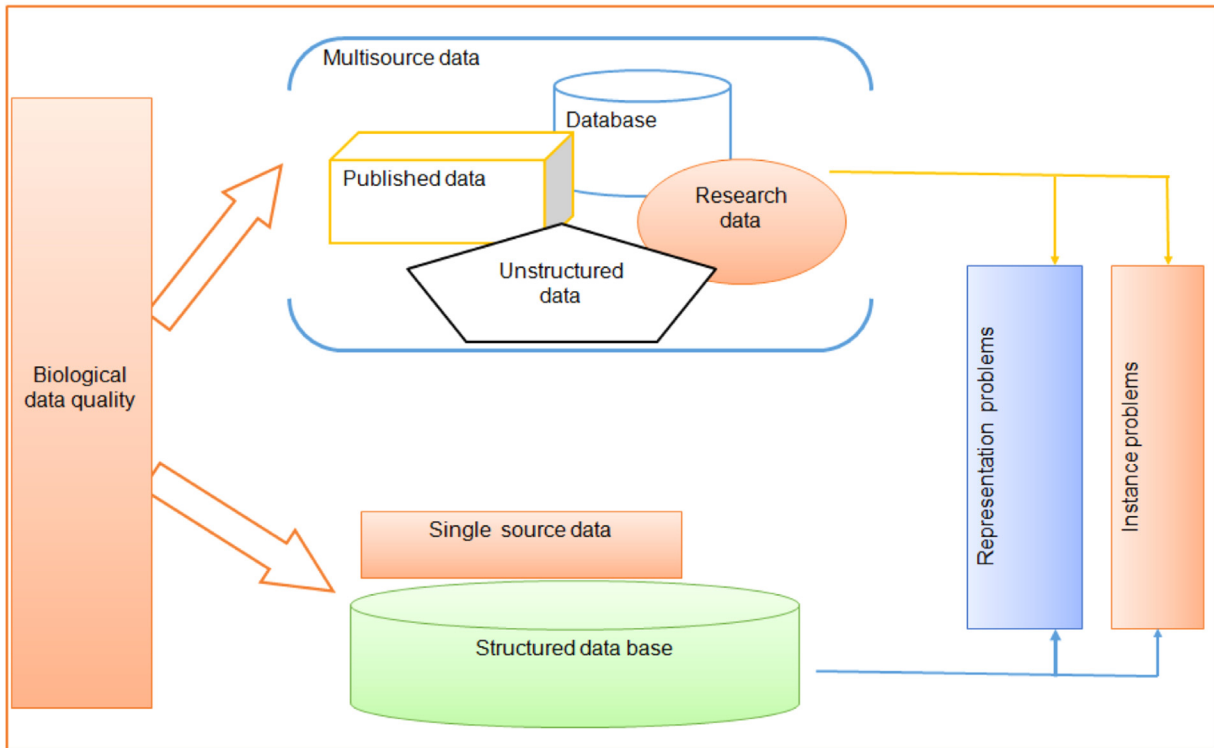


FIGURE 27.5 Quality problems in biological data.

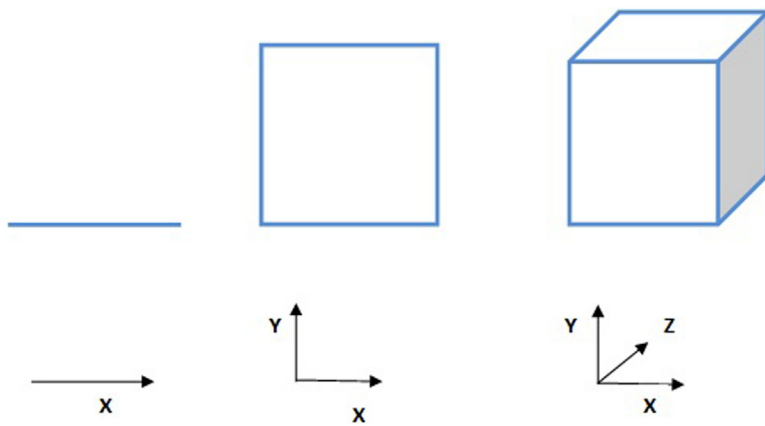


FIGURE 27.6 Data dimensionality (1D, 2D, 3D, and multidimensional data).

proteomic approaches generate high-dimensional data related to different diseases, such as cancer, diabetes, and cardiovascular diseases (Leung et al., 2011).

If required data dimensionality can be reduced by the following methods.

1. missing values ratio;
2. low variance filter;
3. high correlation filter;
4. random forests/ensemble trees;
5. PCA;
6. backward feature elimination; and
7. forward feature construction.

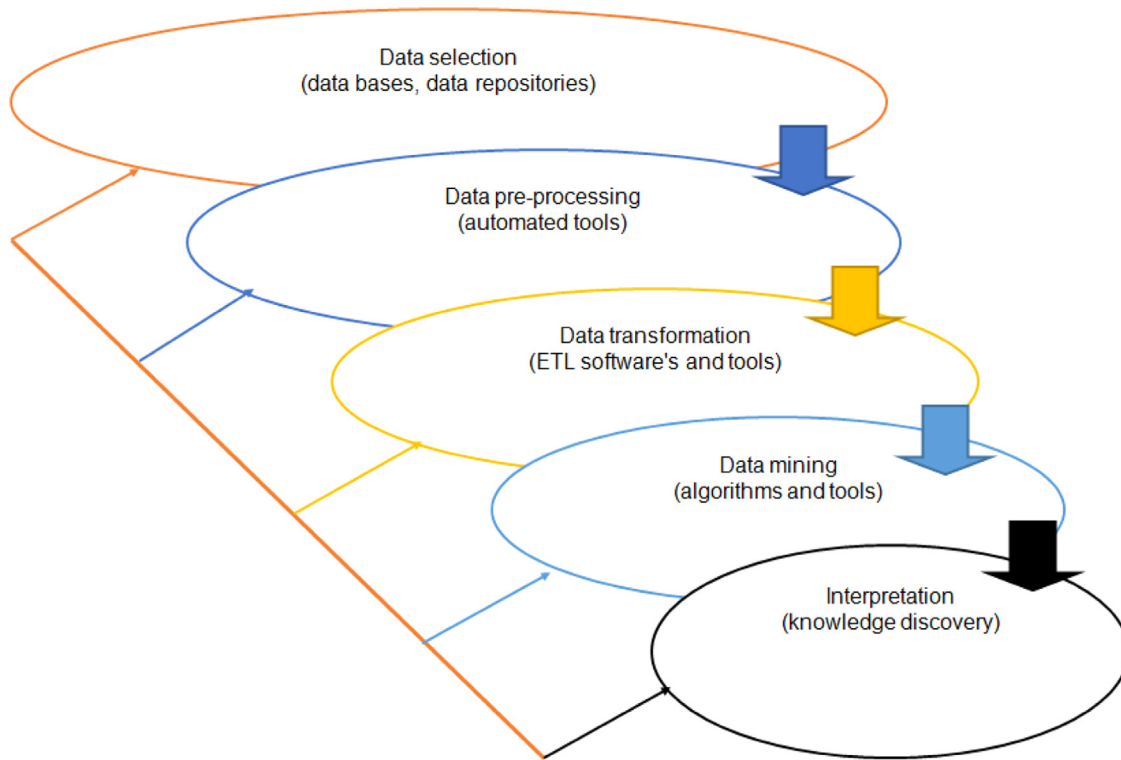


FIGURE 27.7 Biological data knowledge discovery process.

27.6.6 Biological data knowledge discovery

The idea behind the knowledge discovery process is to improve the quality of the data by handling noise, incompleteness, and other inconsistencies like missing values. In biological data knowledge discovery, a set of processes must be followed starting from data selection. The selected data must be preprocessed, which converts the data into an understandable format. Then, the data are transformed, which means that it must be in the required format for applying the data mining algorithms (Lonardi & Chen, 2010). Then, knowledge discovery can be applied to the data and some interpretations can be drawn using data mining. Fig. 27.7 shows the knowledge discovery flow model used in biological data mining.

The above-mentioned challenges are well handled by the computing resources and data mining tools. As research in data collection and data warehousing has progressed the related technologies for mining the information from the biological data has also developed.

27.7 Bioinformatics research areas and tools

In the previous section, we have discussed biological databases. A new field of science evolved which explores the issues, challenges, and new research dimensions in biological data: bioinformatics. The term bioinformatics was given by Paulin Hogeweg in 1979. It is the applications of computer technology to the management of biological data and biological information retrieval (Wegener et al., 2011). The primary goal of bioinformatics is knowledge discovery from biological processes. Bioinformatics is being used in many fields of biological sciences. Mass spectrometry is another field where a lot of data is being generated. It is the field where biologists study the mass of peptides to characterize proteins. The myriad tool is used to generate mass spectrometry data of samples. Data preprocessing of the genomic sequence data is another field where bioinformatics plays an important role. In this field, sequence data is used to identify the motifs, sites, and other key regions that are embedded in the sequence composition. Feature selection has contributed toward analyzing sequences in bioinformatics because there is a need for accurate techniques that can automatically identify genes from the sequence but only small regions of the genome sequence contain the coded information. Latest technologies, such as microarray and the high volume of data, have contributed toward a new phenomenon

called ontology, which is used to integrate the data from heterogeneous sources (Bodenreider & Stevens, 2006). Ontology techniques are used for the representation and distribution of knowledge about an entity by modeling the characteristics of the entity and the relationships among those features. Gene Ontology is an important approach used in the area of bioinformatics.

Bioinformatics is seen to be vitally important in the storage, analysis, and distribution of the data generated from such analysis. With the advent of high-throughput technologies, such as microarrays and next gene sequencing, one predicted application lies in the areas of genome-wide association studies. The use of computing power and data mining methods has opened the following research areas in bioinformatics (Bayer et al., 2017). The Human Genome Project is another milestone in bioinformatics application. Chromosomes can be visualized and represented by three ways; (1) display of three mapping data: two RH (radiation hybrid G3 and GB4) and one genetic map; (2) gene density on chromosome; and (3) ideogram.

27.7.1 Sequence searching, comparison, and evolutionary analysis

Sequence study covers finding out the biological sequences, which are same and which part are different (Mathura & Kanguane, 2009). It is the procedure of studying a DNA, RNA, or peptide sequence with any of a broad collection of analytical processes to comprehend its features, purpose, arrangement, or growth. Methodologies used are comprised of sequence configuration, exploration against biological databases. The genome or protein sequence available at various databases is studied using and analyzed. Entrez is used for searching database entries at NCBI (Brzeski, 2002). The NCBI has a powerful computational facility for an easy-to-use, friendly, and web-based interface to access, visualize, and systematically represent sequence information (<http://www.ncbi.nlm.nih.gov/Entrez/>).

Basic Local Alignment Search Tool (BLAST) is used for sequence similarity searching using an input sequence of protein or nucleotide as a query. The searching algorithm aligns the nucleotide or protein sequences to sequences available in the databases and calculates the percentage identity, similarity/positives, gap (insertion/deletion), and e-values. BLAST also helps to deduce functional and evolutionary relationships among sequences as well as also identify the sequences related to a gene or protein families. This tool is developed and maintained by the NCBI, which provides the nucleotide, protein, genome, gene, mRNA, single-nucleotide polymorphism (SNP), structure, expressed sequence tag (EST), and many other information, as well as the resources for the analysis of genome and protein-related analysis (from <https://blast.ncbi.nlm.nih.gov/>).

HMMER is a tool for biological sequence analysis using profile hidden Markov models. It is used for sequence homolog's searching using databases and also for generating sequence alignments. This site has been considered to offer interactive searches for most queries, coupled with intuitive and interactive results visualizations. It can be downloaded from <https://www.ebi.ac.uk/Tools/hmmer/>. FASTA is another tool for nucleotide and protein sequence alignment. It is a collection of programs and algorithms for searching nucleotide and protein through a query sequence (<https://www.ebi.ac.uk/Tools/sss/fasta/>).

In the evolutionary analysis, different types of clustering approaches are used to deduce the interesting relationship between organisms. Clustering of a different organism based on its particular gene or protein sequence infers the knowledge about the closeness or distance among them and also provides the information about the organism that are distinct or very diverse from the rest organism.

27.7.2 Annotation of gene/protein structure and function

Genome annotation is the process of deriving the structural and functional information of a protein or gene from a raw data set using different analysis, comparison, estimation, precision, and other mining techniques. Genome annotation is essential because the sequencing of the genome or DNA generates sequence information without its functional role. After the genome is sequenced, it must be annotated to bring more logical information about its structural features and functional roles (Salzberg, 2019). It consists of three major steps:

1. recognizing pieces of the genome that do not encode for proteins;
2. recognizing essentials of the genome, a procedure called gene prediction; and
3. recognizing organic information to these elements.

The genome sequence information is stored in annotation files. Some of the file formats are FASTA, GFF3, and GENBANK. There are different file formats for the representation of sequence, structure, and pathway information related to gene and protein, and the facility to select and download a particular file is available over online databases.

Using gene annotation approaches, the genes or proteins that may be recruited by a particular genome sequence can be predicted. Functional annotation of these new genes or proteins can be done by searching their similarity with well experimentally verified sequences available in the databases. For example; if an unknown gene A shows 85% sequence similarity with another gene B whose structure, function, and related protein information is known, then the structure, function, and other information related to gene B can be assigned to gene A.

27.7.3 Gene and protein expression analysis

It is the study of the quantitative and qualitative measurement of mRNA levels expressed by a cell or tissues using different gene factors and conditions. Information about the differentially expressed genes, low or highly expressed genes, as well as genes that are not expressed or switch off can be found. Clustering, association, partial least square, correlation, and other statistical approaches are used to find the information about the gene clusters and outliers.

Protein microarrays are used to find the presence or absence of a protein in the biological sample (Guo et al., 2008). The quantitative level of different proteins expressed by a cell or tissues can also be measured. Data mining and comparative analysis of protein expression data can provide us the valuable information about the proteins that are expressing in a particular condition/stage, that is, diseased or mutant form. Association between the different proteins and required conditions can also be traced out. If a gene or protein-related data is clustered into five different subgroups using K-means clustering, and one gene or protein-related data does not belong to any cluster out of 5, then these new outlier data are also interesting from the study point of view as it represents a new case or type.

27.7.4 Mutation and disease association study

Mutations are changes, such as insertion, deletion, and substitutions, that take place in the nucleotides/genes encoding for a protein. As a result of mutations many genetic and metabolic diseases originates and mutations related to different genetic diseases, such as cancer have been identified using data mining approaches. Mutations are also associated with resistance against antibiotics, HIV, malaria, and many other diseases. SNP, a physical marker of the genome, is also related to the mutations (allelic variation) and identified by data mining from genome sequencing data.

27.7.5 Protein structure prediction, docking, and protein–protein interaction analysis

It is used for the prediction of 2D and 3D structures of a protein from its amino acid sequence information. Different computational and statistical approaches, such as the HMM, genetic algorithm, SVM, and ANN, are used for protein structure predictions. Understanding the structure of the protein is important for finding the functions of the protein (Rapaport, 1999). It is used in designing drugs and new enzymes. X-ray crystallography and nuclear magnetic resonance methods are used for 3D structure determination, and they require the protein crystals whereas, modeling approaches, such as homology modeling, threading, and ab initio methods, can build a 3D structure of a protein from the sequence composition of proteins.

In bioinformatics, docking is the computational learning of how the two molecules interact and fit mutually. It is a molecular modeling method that is used to predict how small molecules (ligands) binds and fit their particular pose in the binding site of a target protein (enzyme) (Porter, Desta, Kozakov, & Vajda, 2019).

For protein–protein interaction, ortholog-based and domain pair-based search methods are used to infer the interaction. If one protein is an ortholog of another protein or shares a conserved domain with a protein, then their interaction partner can be the same. For evolutionarily related organisms, if two proteins have the same functional linkage, then they may have a common interacting partner (Khare & Singh, 2020). Genes/proteins that show common expression patterns have chances to interact with each other. To address these issues, bioinformatics tools are available to find the mode of interaction for a protein as well as an interacting partner (Xiong & Xiong, 2012).

Graph mining techniques use a small subgraph to find this pattern in a given dataset. Graph mining can be used to classify the chemical structures and identify the motifs pattern in a protein family.

27.7.6 Biological systems modeling and network analysis

Modeling of biological systems is a significant task of systems biology and mathematical modeling (Motta & Pappalardo, 2013). Computational systems biology aims to develop and use efficient algorithms, data structures, visualization, and communication tools for the integration of large quantities of biological data to build a model that can be

used to understand the dynamics and interaction of different components in a system. Investigational data on any given chemical entity are used to develop a mathematical model that leads to hypotheses to test its effect on the dynamic system. Then, these investigational data are compared with computational model output to test if the hypothesis is true or false. Computational findings can be compared with the known experimental results to decide their applicability for future predictions. This modeling technique can quickly examine diverse investigational conditions for the biological system, and simply the most significant cases can be considered subsequently in the laboratory conditions. This modeling technique permits researchers to examine original development and to build up hypotheses to conduct the plan for novel experiments in the future.

Biological network analysis is performed by mining approaches for interpreting the interactions among different components. Feature-based indexing, tree-based indexing, and reference-based indexing approaches are used to processing query access and accessing the information from the complex biological networks.

27.7.7 Expressed sequence tag analysis

Identifying the coding or functional part of the genome that has the potential to encode a protein is an important aspect of genome analysis (Leung et al., 2011). These proteins play important role in the different reactions of pathways in the form of enzymes. For this purpose, mRNA is isolated from a cell and reverse transcribed into complementary DNA (cDNA) sequences, and these partial cDNA sequences are known as EST. Genome sequencing has generated a vast amount of information that can be used for structure and function analysis. A large number of ESTs can be generated from mRNA, and then, these ESTs can be compared with already known sequences in the databases to determine their function.

27.7.8 MicroRNA and target prediction

MicroRNAs (miRNAs) are short length, noncoding RNA present in plants, viruses, and animals that regulates gene expression by base pairing with mRNA. Information about various miRNAs and their target are available in miRBase. miRNA and its precursor sequence contain many important features that have been used for developing miRNA prediction tools. Using precursor information and processing features as well as characteristics of known miRNA, many miRNAs predicted in the HIV genome using the ANN approach (Gupta, Agarwal, & Prakash, 2012). In another study, precursor miRNAs predicted from the EST database using BLAST, and then miRNAs predicted from precursor using secondary structural features essential for a miRNA. Functional annotations, gene ontology, and target for predicted miRNAs were also predicted using computational resources and tools (Sahu, Singh, & Yadav, 2013).

27.7.9 Medical and health data analysis and clinical decision support system

Medical imaging is used to analyze the interior part of the body to reveal the changes in the diseased condition. Segmentation of images is done by the use of clustering techniques based on the external design and view, depth, motion, etc. Clustering is used to group pixels of an image into predefined classes or clusters based on their similarities. Medical imaging pattern analysis can be used to differentiate between the normal and diseased or altered conditions, such as normal versus dead tissues, or malignant cells. The image segmentation approach has been used to detect the possibility of lung cancer using K-means clustering, and many other optimization techniques (Senthil Kumar, Venkata Lakshmi, & Karthikeyan, 2019). Many other approaches have been used for the prediction and detection of different types of disease based on feature-based methods, decision trees, genetic algorithms, and classification.

A decision support system for medical and health-related issues can be build using feature selection, prediction, clustering and outlier detection, classification, and neural network approaches (Jacob & Ramani, 2012). It can provide support to the health practitioner for predicting disease risk based on disease characteristics, recommending drug and dose for a patient, assessing the severity and future outcome of therapy, analysis, and interpreting drug side effects, and recommending the most effective drug or therapy for a subpopulation based on genetic profile or biomarker (precision medicine).

27.8 Limitations

In recent years a large amount of biomedical data has been generated as a result of advancements in genomics, proteomics, metabolomics, transcriptomics, pharmacogenomics, metagenomics, health, and clinical science. These data must

be analyzed and mined to extract meaningful full and significant information that can be used for decision making in clinical and healthcare applications. There are many problems associated with the generated raw biomedical data, such as noise, incompleteness, and high false-positive/negative expression data. In the case of the large biological dataset, indexing, searching, and mining are a computationally complex and time-consuming task. For a better understanding of the complex human system, data available in the gene, RNA, protein, expression, pathway, structure, enzyme, function, disease, and clinical databases or resources need to be linked.

27.9 Conclusion

Both data mining and bioinformatics are emerging and strongly related research fields. It is very vital to inspect the significant research issues in bioinformatics and build up new data mining methods for accurate, fast, and effective biological data analysis. In this chapter, the basics concept of data mining techniques and their application in the field of biological sciences have been discussed. However, a review of all kinds of data mining methods and their usability in biological data analysis are provided and the selective data offered here may give readers a sense that a lot of interesting work has been done and still more can be.

Conflict of interest

The authors declare that they have no conflict of interest.

References

- Agarwal, S. (2014). Data mining: Data mining concepts and techniques. In *Proceedings of the 2013 international conference on machine intelligence research and advancement, ICMIRA 2013*. doi:10.1109/ICMIRA.2013.45.
- Ang, J. C., Mirzal, A., Haron, H., & Hamed, H. N. A. (2016). Supervised, unsupervised, and semi-supervised feature selection: A review on gene selection. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 13(5), 971–989. Available from <https://doi.org/10.1109/TCBB.2015.2478454>.
- Ayer, T., Alagoz, O., Chhatwal, J., Shavlik, J. W., Kahn, C. E., Jr., & Burnside, E. S. (2010). Breast cancer risk estimation with artificial neural networks revisited: Discrimination and calibration. *Cancer*, 116(14), 3310–3321. Available from <https://doi.org/10.1002/cncr.25081>.
- Bayer, H., Aksogan, M., Celik, E., & Kondiloglu, A. (2017). Big data mining and business intelligence trends. *Journal of Asian Business Strategy*, 7(1), 23–33. Available from <https://doi.org/10.18488/journal.1006/2017.7.1/1006.1.23.33>.
- Bodenreider, O., & Stevens, R. (2006). Bio-ontologies: Current trends and future directions. *Briefings in Bioinformatics*, 7(3), 256–274. Available from <https://doi.org/10.1093/bib/bbl027>.
- Brownlee, J. (2020). *Standard machine learning datasets to practice in Weka, Janson Brownlee*. <<https://machinelearningmastery.com/standard-machine-learning-datasets-used-practice-weka/>>. Accessed: February 04.02.21.
- Brzeski, H. (2002). An introduction to bioinformatics. *Methods in Molecular Biology*, 187, 193–208. Available from <https://doi.org/10.1385/1-59259-273-2:193>.
- Bustamam, A., Tasman, H., Yuniarti, N., Frisca, & Mursidah, I. (2017). Application of k-means clustering algorithm in grouping the DNA sequences of hepatitis B virus (HBV). In *AIP Conference Proceedings*, 1862. <<https://doi.org/10.1063/1.4991238>>.
- Chaurasia, V., Pal, S., & Tiwari, B. B. (2018). Prediction of benign and malignant breast cancer using data mining techniques. *Journal of Algorithms and Computational Technology*, 12(2), 119–126. Available from <https://doi.org/10.1177/1748301818756225>.
- Chen, X., Wang, M., & Zhang, H. (2011). The use of classification trees for bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 1(1), 55–63. Available from <https://doi.org/10.1002/widm.14>.
- Dua, S., & Chowriappa, P. (2012). Data mining for bioinformatics. *Data Mining for Bioinformatics*. Available from <https://doi.org/10.1201/b13091>.
- Foulkes, A. C., Watson, D. S., Griffiths, C. E. M., Warren, R. B., Huber, W., & Barnes, M. R. (2017). Research techniques made simple: Bioinformatics for genome-scale biology. *The Journal of Investigative Dermatology*, 137(9), e163–e168. Available from <https://doi.org/10.1016/j.jid.2017.07.095>.
- Frank, E., Hall, M., Trigg, L., Holmes, G., & Witten, I. H. (2004). Data mining in bioinformatics using Weka. *Bioinformatics (Oxford, England)*, 20(15), 2479–2481. Available from <https://doi.org/10.1093/bioinformatics/bth261>.
- Guo, Q., Lu, X., Gao, Y., Zhang, J., Yan, B., Su, D., . . . Wang, G. (2017). Cluster analysis: A new approach for identification of underlying risk factors for coronary artery disease in essential hypertensive patients. *Scientific Reports*, 7, 43965. Available from <https://doi.org/10.1038/srep43965>.
- Guo, Y., Xiao, P., Lei, S., Deng, F., Xiao, G. G., Liu, Y., . . . Deng, H. (2008). How is mRNA expression predictive for protein expression? A correlation study on human circulating monocytes. *Acta Biochimica et Biophysica Sinica*, 40(5), 426–436. Available from <https://doi.org/10.1111/j.1745-7270.2008.00418.x>.
- Gupta, M. K., Agarwal, K., Prakash, N., et al. (2012). Prediction of miRNA in HIV-1 genome and its targets through artificial neural network: A bioinformatics approach. *Network Modeling and Analysis in Health Informatics and Bioinformatics*, 1, 141–151.

- Herbert, K. G., & Wang, J. T. L. (2007). Biological data cleaning: A case study. *International Journal of Information Quality*, 1(1), 60–82. Available from <https://doi.org/10.1504/IJIQ.2007.013376>.
- Hirata, Y., Morino, K., Suzuki, T., et al. (2016). System identification and parameter estimation in mathematical medicine: Examples demonstrated for prostate cancer. *Quantitative Biology*, 4(1), 13–19. Available from <https://doi.org/10.1007/s40484-016-0059-0>.
- Jacob, S. G., & Ramani, R. G. (2012). Data mining in clinical data sets: A review. *International Journal of Applied Information Systems*, 4, 15–26.
- Khare, E., & Singh, D. B. (2020). Protein-protein interactions modeling: From dry to wet lab. In D. Singh, & T. Tripathi (Eds.), *Frontiers in protein structure, function, and dynamics* (pp. 119–143). Singapore: Springer.
- Leung, K. S., Lee, K. H., Wang, J. F., Ng, E. Y., Chan, H. L., Tsui, S. K., . . . Sung, J. J. (2011). Data mining on DNA sequences of hepatitis B virus. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2), 428–440. Available from <https://doi.org/10.1109/TCBB.2009.6>.
- Lonardi, S., & Chen, J. (2010). Data mining in bioinformatics: Selected papers from BIODDD. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(2), 195–196. Available from <https://doi.org/10.1109/TCBB.2010.28>.
- Mahmud, M., Kaiser, M. S., Hussain, A., & Vassanelli, S. (2018). Applications of deep learning and reinforcement learning to biological data. *IEEE Transactions on Neural Networks and Learning Systems*, 29(6), 2063–2079. Available from <https://doi.org/10.1109/TNNLS.2018.2790388>.
- Mathura, V. S., & Kanguane, P. (2009). *Bioinformatics: A concept-based introduction*. Boston, MA: Springer. Available from <http://doi.org/10.1007/978-0-387-84870-9>.
- Motta, S., & Pappalardo, F. (2013). Mathematical modeling of biological systems. *Briefings in Bioinformatics*, 14(4), 411–422. Available from <https://doi.org/10.1093/bib/bbs061>.
- Ocana, K. A. C. S., Silva, V., Oliveira, D. D., & Mattoso, M. (2015). Data analytics in bioinformatics: Data science in practice for genomics analysis workflows. In *Proceedings of the eleventh IEEE international conference on eScience, eScience 2015* (pp. 322–331). doi:10.1109/eScience.2015.50.
- Peng, Y., Wu, Z., & Jiang, J. (2010). A novel feature selection approach for biomedical data classification. *Journal of Biomedical Informatics*, 43(1), 15–23. Available from <https://doi.org/10.1016/j.jbi.2009.07.008>.
- Porter, K. A., Desta, I., Kozakov, D., & Vajda, S. (2019). What method to use for protein–protein docking? *Current Opinion in Structural Biology*, 55, 1–7. Available from <https://doi.org/10.1016/j.sbi.2018.12.010>.
- Rapaport, D. C. (1999). Molecular dynamics simulation. *Computing in Science and Engineering*, 1(1), 70–71. Available from <https://doi.org/10.1109/5992.743625>.
- Sahu, S., Singh, D. B., Yadav, K. K., et al. (2013). Computational identification and functional annotation of miRNAs in medicinal plant *Helianthus petiolaris*. *Network Modeling and Analysis in Health Informatics and Bioinformatics*, 2, 277–284. Available from <https://doi.org/10.1007/s13721-013-0044-8>.
- Salzberg, S. L. (2019). Next-generation genome annotation: We still struggle to get it right. *Genome Biology*, 20(1), 92. Available from <https://doi.org/10.1186/s13059-019-1715-2>.
- Senthil Kumar, K., Venkata Lakshmi, K., & Karthikeyan, K. (2019). Lung cancer detection using image segmentation by means of various evolutionary algorithms. *Computational and Mathematical Methods in Medicine*. Available from <https://doi.org/10.1155/2019/4909846>.
- Sidey-Gibbons, J. A. M., & Sidey-Gibbons, C. J. (2019). Machine learning in medicine: A practical introduction. *BMC Medical Research Methodology*, 19(1), 1–18. Available from <https://doi.org/10.1186/s12874-019-0681-4>.
- Sowmya, R., & Suneetha, K. R. (2017). Data mining with big data. In *Proceedings of the eleventh international conference on intelligent systems and control (ISCO)* (pp. 246–250). Coimbatore. doi:10.1109/ISCO.2017.7855990.
- Tang, D. Q. & Tan, Y. (2011). Graph-based bioinformatics mining research and application. In *Proceedings of the fourth international symposium on knowledge acquisition and modeling, KAM* (pp. 286–290). doi:10.1109/KAM.2011.83.
- Wegener, D., Rossi, S., Buffa, F., Delorenzi, M., & Riiping, S. (2011). Towards an environment for data mining based analysis processes in bioinformatics & personalized medicine. In *Proceedings of the IEEE international conference on bioinformatics and biomedicine workshops, BIBMW 2011* (pp. 570–577). doi:10.1109/BIBMW.2011.6112431.
- Xiong, J., & Xiong, J. (2012). Introduction to biological databases. *Essential Bioinformatics*, 10–28. Available from <https://doi.org/10.1017/cbo9780511806087.003>.