4

Role of Data Mining in Bioinformatics

Vivek P. Chavda[1]*, Amit Sorathiya[2], Disha Valu[3] and Swati Marwadi[4]

[1]Department of Pharmaceutics, L.M. College of Pharmacy, Ahmedabad, India
[2]Formulation & Development, Sun Pharma Advanced Research Centre, Vadodara, India
[3]Analytical Development, Orbicular Pharmaceutical Ltd, Hyderabad, India
[4]Department of Biotechnology, Sinhgad College of Engineering, Pune, India

Abstract

In the recent days, the term which was very challenging in the discipline of informational science is data mining, which includes extraction of tremendous amount of data. Data mining is very useful and interesting in accessing the different patterns of data from the pre-existing data in the database to get knowledge based information using different software techniques. Data mining looks most suitable for bioinformatics, as bioinformatics is enrichment of data, though the evolutionary phases of human existence at molecular level are lacking.

Nevertheless, gathering information from different databases is very helpful and informative in informational science which is affected by the several different aspects of the data stored in the libraries. These libraries involves the persistent data with the relevant domain of science. Also, it consists of various factors such as diversity, count, dimension, etc. Later on, it is not that much easier to manage and access the data which is most useful in data discovery as it requires the cogent scientific skills and deep insight for the body of knowledge around that data. The biological databases compilation is also difficult. The most effective way for accessing the databases in informational science and developing new technologies for studying the biotic system at molecular level includes extracting the raw data from the other relevant databases and also its further evaluation has become more important and crucial concept in informational science.

***Keywords*:** Data mining, proteomics, DNA, RNA, gene expression, analysis, etc.

**Corresponding author*: Vivek7chavda@gmail.com

S. Balamurugan, Anand Krishnan, Dinesh Goyal, Balakumar Chandrasekaran and Boomi Pandi (eds.) Computation in Bioinformatics: Multidisciplinary Applications, (69–84)

4.1 Introduction

The Data Science is extraction of useful information from the huge amount of data which leads to identify the fresh and understandable data models and patterns out of it [1]. Bioinformatics is the computer-assisted science aiming at managing a huge volume of genomic data, which combines the power of computerized science of storing, analyzing, and utilizing information from biological data such as molecules, sequences, gene expressions, and pathways to solve multiple genetic puzzles [2]. The approaches of improvement to provide some important details about the rapidly expanding sources of biological data in data mining (DM) will play a vital role [3]. DM is the fundamental science of discovery of new interesting configurations and correlation in huge amount of data. It is defined as "the process of finding meaningful new relationships, patterns, and trends by digging into large amounts of data stored in storerooms" [4]. Knowledge Discovery in Databases (KDD) is the word sometime used for DM. These processes are blood for the future research as well as driving force for the technological paradigm. It is the need of the hour which is consistently evolving for identifying the probability of hidden knowledge that exist in the data which is already generated by the science and technology [5].

DM methodologies have paved the futuristic advancement for bioinformatics and an appropriate tool for the same which provides deep insight for the data at the molecular level for recognizing patterns and making algorithms around it. The wide-ranging databases of biological information create both challenges and opportunities for development of novel KDD methods [6, 7]. Biological data handling will be beneficial for identifying the appropriate targets and generating therapeutics around the identified target in a sort span. The entire human genome, the complete set of genetic information within each human cell, has now been determined [8]. Understanding these genetic instructions promises to permit scientists to better understand the nature of diseases and their cures, to identify the mechanisms underlying biological processes such as growth and ageing and to clearly track our evolution and its relationship with other species [9]. The main hindrance lying between investigators and the knowledge they seek is the sheer volume of data available. This is evident from the rapid increase in the number of base pairs and DNA sequences in the repository of GenBank [10].

4.2 Data Mining Methods/Techniques

Selection of DM technique is very crucial in the context of studies being undertaken and in the complexity of the identified problem and business type. Various DM techniques are summarized in Figure 4.1, and a common platform approach is adopted generally for the accuracy and cost-effectiveness of the entire process.

4.2.1 Classification

4.2.1.1 Statistical Techniques

Statistical techniques of DM is a branch of mathematics mainly deals with collection and description of data [11]. Despite of the above fact, many data scientists are against it for being considered as DM technique. But still, it assists to determine the patterns of data evaluated and can build a predictive models around it [12]. For this reason, data analyst should possess some knowledge about the different statistical techniques. In today's world, people have to deal with a large amount of data and have to derive important patterns from it. The application of statistical analysis is given in Figure 4.2, while the different data collection methods are summarized in Figure 4.3.

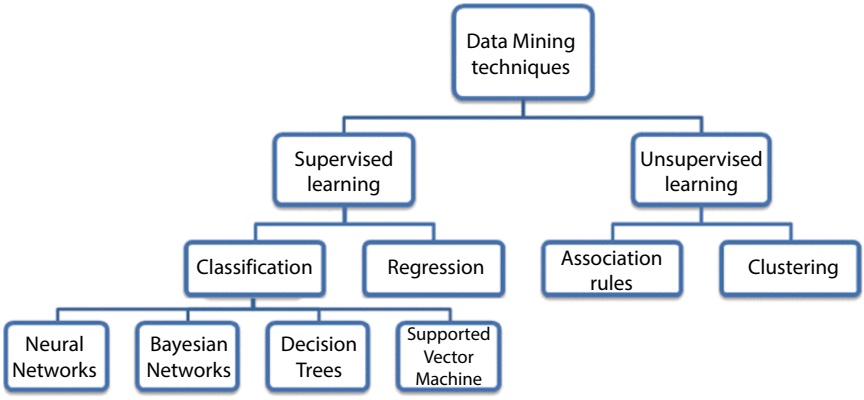

Figure 4.1 Data mining techniques.

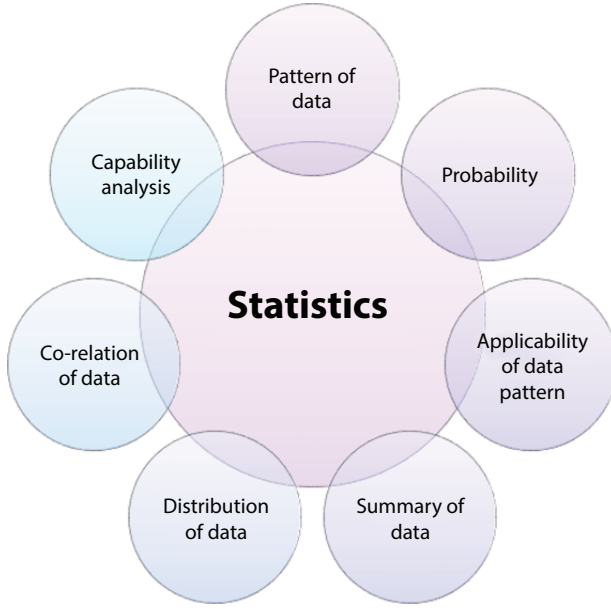

Figure 4.2 Application of statistics in data mining.

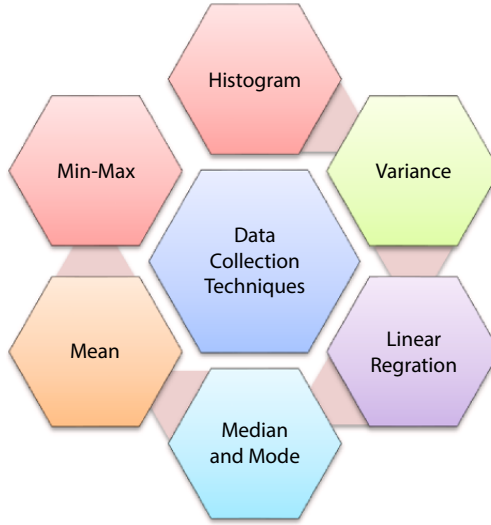

Figure 4.3 Data collection techniques.

For data analysis, statistics is the essential component to understand and evaluate the data. Statistics will help the scientist to identify the data pattern while DM process through data visualization. This will further help in detecting the noise and then in optimizing and identifying the significant finding out of the data jargon.

4.2.1.2 Clustering Technique

Clustering technique (segmentation) is one of the oldest DM techniques where the data which are similar in nature being are treated as a cluster. This will assist to comprehend the differences and similarities between the data and it is very accurate method of DM [13]. It seems to be pretty useful when pattern recognition is needed between the same kinds of data to draw a meaningful conclusion. For instance, an insurance firm can group its consumers based on their income, age, nature of policy, and type of claims. There are different types of clustering methods used for the DM process (Figure 4.4).

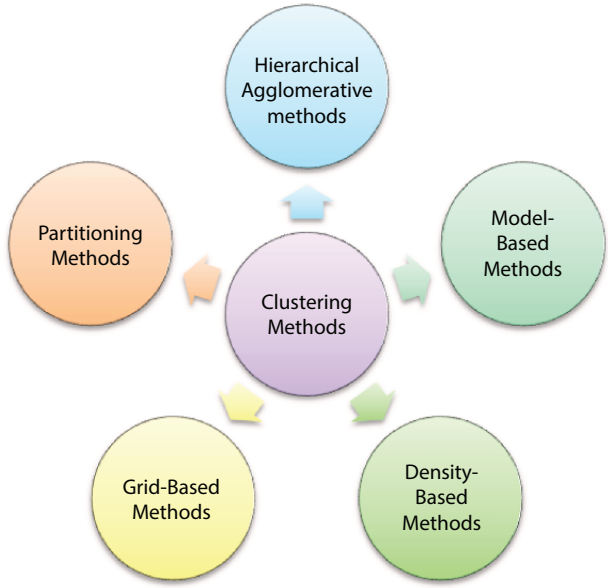

Figure 4.4 Types of clustering methods.

4.2.1.3 Visualization

Data visualization is the primary and most common method of the DM which is applied at the beginning of the DM process [14]. Many types of research are going on these days to produce an interesting projection of databases, which is called Projection Pursuit. Data visualization is very useful tool which will give the holistic overview of the analyzed data and helps the scientist to remove the data which are not suitable for the further process. It is equally useful for the recognition of the hidden patterns [12].

4.2.1.4 Induction Decision Tree Technique

As the name suggest, the decision tree technique has the tree-like structure which can generate predictive model [15]. In this technique, the data to be evaluated shall be classified as the branch of the tree and the leaves of the trees are considered as partitions of the data set related to that particular classification. This technique can be used for exploration analysis, data pre-processing, and prediction work [16]. The data are classified as the segment or the branch which possesses some similarities in their information being predicted. On a closer look, one can see that it has pretty good potential for correlate the problem statement with that of the desired outcome; hence, it is popularly used by the statistician for predictive data analysis as well as for the data pre-processing.

The first and foremost step in this technique is growing the tree. The basic of growing the tree depends on finding the best possible question to be asked at each branch of the tree. The decision tree stops growing under any one of the below circumstances.

a) If the segment contains only one record
b) All the records contain identical features
c) The growth is not enough to make any further spilt

CART which stands for Classification and Regression Trees is a data exploration and prediction algorithm which picks the questions in a more complex way. It tries them all and then selects one best question which is used to split the data into two or more segments. After deciding on the segments, it again asks questions on each of the new segment individually. Another popular decision tree technology is CHAID (Chi-Square Automatic Interaction Detector). It is similar to CART but it differs in one way. CART helps in choosing the best questions, whereas CHAID helps in choosing the splits [17].

4.2.1.5 Neural Network

Another significant approach used by individuals these days is the neural network. In the early stages of DM technology, this method is most commonly used. Neural networks are very user friendly as it is sequential process where it does not demand the scientist to be technically sound in the domain of data generated [18, 19].

A set of interconnected neurons is a neural network. A single layer or multiple layers can form. The architecture of the network is called the creation of neurons and their interconnections. There are a wide range of models of neural networks and each model has its own benefits and drawbacks. There are distinct architectures in of neural network model, and these architectures use distinct learning processes. Neural networks are a very solid method of predictive modelling. But even for specialists, it is not quite easy to comprehend. It produces very complicated models that are difficult to completely comprehend. Companies are therefore seeking new solutions to grasp the Neural Network Methodology (NNM).

Two possibilities have already been proposed [20]. The first approach is to bundle the neural network into a complete solution that will allow it to be used for a single application. The second approach is that it is related to specialist advisory services. In numerous types of applications, the neural network has been used [15].

4.2.1.6 Association Rule Technique

It is the different techniques of all techniques in DM which identify the hidden pattern in the data set through which one can find out the variable of interest. Even one can also find out the appearance frequency of any particular variable [21]. Different types of association rules are depicted in Figure 4.5. It will give the frequency as well as correctness of the same rule for particular problem.

This rule technique extends up to a two-step process. First, Step-Find all the data sets which are occurring regularly or repeatedly. Second, Step-Over the time it creates strong association rules from the data sets which appear constantly.

4.2.1.7 Classification

Among all DM techniques, it is the most common and widely used specially to generate model from a large data set (Figure 4.6). Also, the other advantage of this technique is to obtain important information about data

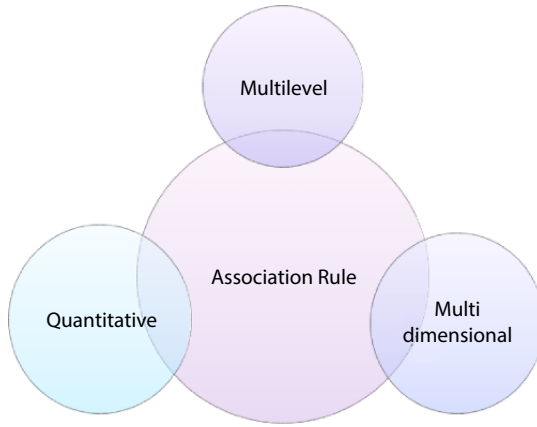

Figure 4.5 Type of association rules.

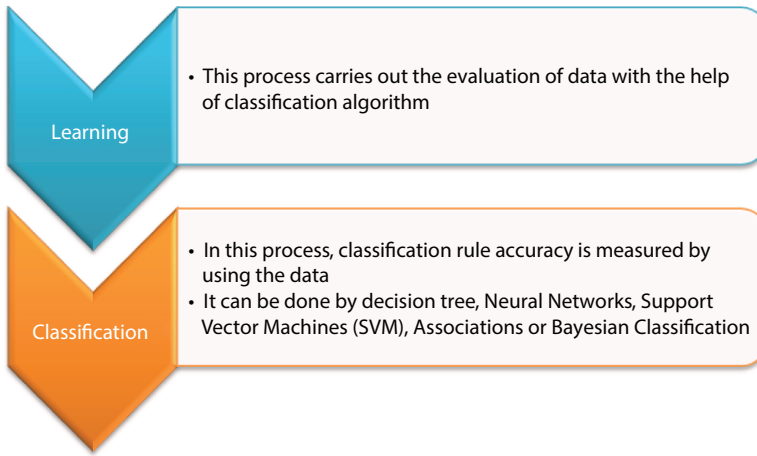

Figure 4.6 Classification technique.

and metadata (data about data). If you evaluate it closely, then it is very much similar to cluster analysis technique, but here, it utilizes decision tree or neural network system. This technique includes two main processes, i.e., Learning and Classification [16, 22].

4.3 DNA Data Analysis

DNA, RNA, and protein are three important elements of life science; they set the groundwork of all living organisms [23]. Tons of researchers are working together to explain the nature of life, and lots of study has been performed to define interactions between structures and their properties [24]. For microbiologists, information encoded in nucleic acid molecular sequences is important because not only it transmits genetic information from generation to generation and furthermore affects transcription and translation activity [25]. Without gathering and processing such DNA sequences, research on the origins of life sciences cannot be performed, which involves identifying the exact order and proportion of the four nitrogen bases in a DNA strand: adenine, thymine, guanine, and cytosine [26].

The most interesting advances in the area of life science research is the isolation of DNA fragments which are exist in large numbers and identification of biologically active using recombinant DNA technology which is the most advantageous aspect in the area of molecular biology. Large DNA molecules could be split into multiple small fragments in an organized manner using restriction endonucleases [16]. Recombinant DNA techniques also help in the purification and classification of independent mixture of restriction fragments and most significantly, at least three steps are required to do DNA sequencing: cloning, sequencing, and analyzing. In DNA sequencing, there are two main techniques: Maxam-Gilbert sequencing (also known as chemical sequencing) and the process of chain termination (also known as Sanger sequencing). The previous approach applies radioactive labels to the 50th end of DNA and produces subsequent breaks at specific bases by using chemical process [27, 28]. In the form of dark bands, which reflect radiolabeled DNA fragments, autoradiography helps to generate a sequence or chain of fragments. Sanger's mechanism, on the other hand, requires modified di-deoxynucleoside triphosphates (ddNTPs).

Although using computers for data analysis has obvious advantages, there still exist weaknesses (Figure 4.7). Despite of so called technological advancement in the field, it has certain disadvantages also which are summarized in Figure 4.8.

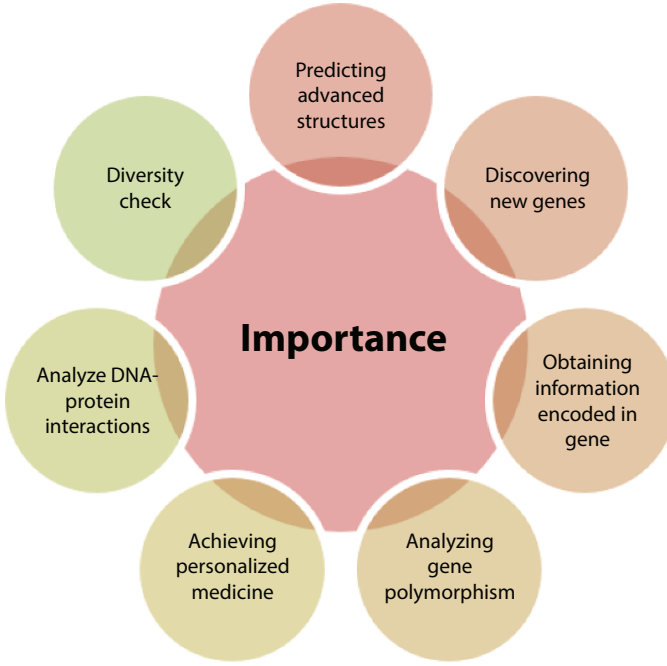

Figure 4.7 Importance of DNA sequence data analysis.

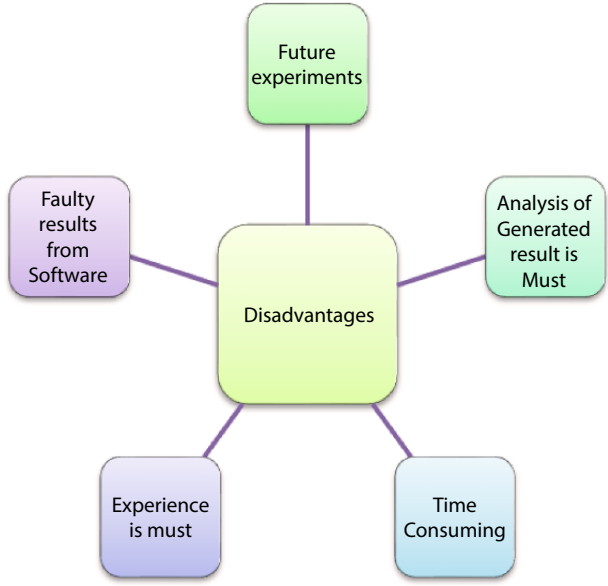

Figure 4.8 Disadvantages of next-generation sequencing data analysis.

4.4 RNA Data Analysis

When we are stating RNA sequencing (RNA-Seq), it is mainly related to transcriptome of a cell. This is mainly done with virtual screening and high-throughput screening method [29]. In comparison with previous Sanger sequencing– and microarray-based techniques, it provides good resolution and greater data coverage for RNA data. These data are very useful for identifying novel transcripts, identifying alternative spliced genes, and detecting allele-specific expression [30, 31]. All the recent advancement in this field like RNA workflow, libraries for sample preparation, and data analysis suits have enabled the scientist to get the functional transcript and transcription process [32]. RNA-Seq may be performed to estimate different RNA populations, including complete RNA, pre-mRNA, and non-coding RNA, including certain microRNA and long ncRNA, in addition to polyadenylated messenger RNA (mRNA) transcripts. A high-throughput approach is mostly used for now a days, which has some additional benefits, such as having more understanding of the complex and dynamic existence of the transcriptome [33, 34]. Elucidation of the different physiological and pathological conditions is now possible with such advanced techniques. With the aid of new mapping techniques, longer reads becomes a reality. Since prolonged readings can extend several exon-exon junctions, with the additional information encoded in longer reads, the recognition and quantification of alternative isoforms can dramatically improve. Soon, one will see the same applications as an extension in the clinical diagnosis like screening of cancer, pregnancy even for personalized medicine [35, 36].

4.5 Protein Data Analysis

Entire units of cellular components, such as the genome, transcriptome, and, more recently, the proteome, can be analyzed now due to advancement in technology, instrumentation, molecular biology, and bioinformatics [37]. With advancement, it is now possible to monitor changes in the human tissue proteomes that are associated with differentiation, apoptosis, disease, and other important biological modifications [38]. The proteomics arm of the OMICS technology is mainly used to elucidate protein structure, its functionality, interactions, and post-translational modifications. When we compare these data with genomics, it is 100-fold complex and dynamic which contribute to the enormity of the challenge of proteomics

and the very modest progress to date [39]. Development of proteomic technologies having different sides can be depicted from relatively broad proteomics experience of the authors and the proteomics reports of others. To increase the overall number of identified and quantified proteins, multiple complimentary approaches need to be taken with any protein-profiling technology. More than one approach can be used to increase the validity of the quantitation of expressed proteins in different samples [40]. To increase the understanding of the quantitation differences arising from biological effects, a multiple-approach strategy would be preferred rather than having experimental approaches [41].

Although the reported number of identified proteins per study may range from a few hundred to a few thousand, the amount of time and material it takes to profile that many proteins is not of less importance [42]. One can see the evolution in both side means analyzing by different mode of mass spectrometry as well as proteomics approaches to be ahead of time still there is a need of better platform based approach for the same [43]. At initial stage, a platform where any raw data file format from any instrument can be deposited and converted into a common file format for preliminary spectral analysis would be extremely valuable. Adding to the same, a good reporting arm for the statistical evaluation of the data is needed [44]. Combination of both these element will surely give us the desired outcome.

4.6 Biomedical Data Analysis

Unprecedented amount of data has been generated with advancement in structural bioinformatics, molecular biology, and pharmacogenomics research in the biomedical field. In short, biomedical data can be available in varieties of sub domains. The major arm of the same is data related to gene expression, DNA sequence, and protein primary structure [27]. Apart from that, addition of the high throughput sequencing methods and cDNA microarray technology has provided efficiency in both data generation as well as its analysis. Virtual screening is also an added star to the same bucket of research tool. There are many algorithms available which can deal with bifurcation of sequences, checking similarities, and get read of weak candidates among them, separating protein coding regions from non-coding regions in DNA sequences, prediction of protein structure as well as function, and reconstructing the underlying evolutionary history [45]. There are four components of a DNA sequence, namely, adenine (A), cytosine (C), guanine (G), and thymine (T), which specifies the code of life.

Similarly, when we talk about proteins primary structure, it mainly grafted from the 20 amino acids. These amino acids will be incorporated for the gene coding of the DNA sequence. On the other side, gene expression data will be useful to understand the expression profile in terms of its regulation under specific conditions in a cell [46]. The research in this field can be accelerated if a good database is created along with a capable software.

4.7 Conclusion and Future Prospects

Bioinformatics and DM are the two side of the same coin. If we consider one side, i.e., Bioinformatics which is having huge amount of data but devoid of molecular level organizational theory to organize such data, these gaps are filled by the other side of the coin, i.e., DM. However, DM in bioinformatics is mainly affected by the heterogeneous nature of data, which make it difficult at reaching quality output some times. Not only this but data integration and level of expertise require to do the same is also the concern. It is useful for the identification of the gene, transcription process, protein function elucidation, function motif detection etc. Apart form that, it is useful for the diagnosis, prognosis and treatment optimization of a particular ailment. It also has potential to detect the interaction, sorting of the database generated and protein sub-cellular location prediction. If we talk about future of the same, the it is surely bright. There is need of an efficient scoring algorithm which can be able to execute all the data dump efficiently and in comprehensive manner is highly desirable.

References

1. Hayashi, C., What is Data Science? Fundamental Concepts and a Heuristic Example. In: *Data Science, Classification, and Related Methods*. Hayashi, C., Yajima, K., Bock, H.-H., Ohsumi, N., Tanaka, Y., Baba, Y. (eds.). Studies in Classification, Data Analysis, and Knowledge Organization, pp. 40–51, Springer, Japan, 1998.
2. Grothaus, G.A., Mufti, A., Murali, T., Automatic layout and visualization of biclusters. *Algorithms Mol. Biol.*, 1, 15, 2006.
3. Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, 247, 536–540, 1995.
4. Greiner, L., *What is Data Analysis and Data Mining?*, http://www.dbta.com/Editorial/Trends-and-Applications/What-is-Data-Analysis-and-Data-Mining-73503.aspx, 2019. Jan 7, 2011.

5. Hwang, W., Cho, Y.R., Zhang, A., Ramanathan, M., A novel functional module detection algorithm for protein-protein interaction networks. *Algorithms Mol. Biol.*, 1, 24, 2006.
6. Fayyad, U., Haussler, D., Stolorz, P., Mining scientific data. *Commun. ACM*, 39, 51–57, 1996.
7. Han, J. and Kamber, M., *Data Mining: Concepts and Techniques*, 2nd Ed., Morgan Kaufmann, 2006.
8. Mount, D.W., *Bioinformatics – Sequence and Genome Analysis*, Cold Spring Harbor Laboratory Press, New York, 2001.
9. Li, J., Wong, L., Yang, Q., *Data Mining in Bioinformatics*, IEEE Intelligent System, IEEE Computer Society, 2005.
10. Edwards, D., Hansen, D., Stajich, J.E., DNA Sequence Databases. In: *Bioinformatics*, Edwards, D., Stajich, J., Hansen, D. (eds). Springer, New York, NY, 2009. https://doi.org/10.1007/978-0-387-92738-1_1.
11. Tang, H. and Kim, S., *Bioinformatics: mining the massive data from high throughput genomics experiments, analysis of biological data: a soft computing approach*, S. Bandyopadhyay (Ed.), Indian Statistical Institute, India, 2007.
12. Guillet, F., *Quality measures in data mining*, 1st ed., Springer, Berlin, 2007.
13. Tramontano, A., *Introduction to bioinformatics*, 1st ed., Chapman & Hall/CRC, London, 2007.
14. Larose, D. and Larose, C., *Discovering Knowledge in Data: An Introduction to Data Mining*, 1st ed., Wiley Publication, 2014.
15. Zhang, Y. and Rajapakse, C.J., *Machine Learning in Bioinformatics*, Wiley, 456, 2008.
16. Han, and Kamber, *Data Mining concepts and techniques*, Morgan Kaufmann Publishers, 2006.
17. Lever, J., Krzywinski, M., Altman, Model selection and overfitting. *N. Nat. Methods*, 13, 703–704, 2016.
18. Lancashire, L.J., Lemetre, C., Ball, G.R., An introduction to artificial neural networks in bioinformatics–application to complex microarray and mass spectrometry datasets in cancer studies. *Brief. Bioinform.*, 10, 3, 315–29, 2009.
19. Yang, Q., Data Mining and Bioinformatics: Some Challenges, 2012, http://www.cse.ust.hk/~qyang (Date of access: 25 Nov 2019).
20. https://www.educba.com/data-mining-techniques/.
21. Baxevanis, A.D., Petsko, G.A., Stein, L.D., Stormo, G.D. (Eds.), *Current Protocols in Bioinformatics*, Wiley, 2007.
22. Hand, D.J., Mannila, H., Smyth, P., *Principles of Data Mining*, MIT Press, p. 578, 2001.
23. Adams, M.D., Kerlavage, A.R., Fleischmann, R.D., Fuldner, R.A., Bult, C.J., Lee, N.H., Kirkness, E.F., Weinstock, K.G., Gocayne, J.D., White, O. *et al.*, Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, 377, 3–174, 1995.

24. Cristianini, N. and Hahn, M., *Introduction to Computational Genomics*, Cambridge University Press, 2006.
25. Abbott, A., Genome sequence of the nematode C. elegans: a platform for investigating biology. The C. elegans Sequencing Consortium. *Science*, 282, 2012–2018, 1998.
26. Velculescu, V.E., Zhang, L., Vogelstein, B., Kinzler, K.W., Serial analysis of gene expression. *Science*, 270, 484–487, 1995.
27. Abecasis, G.R., Cherny, S.S., Cookson, W.O., Cardon, L.R., Merlin—Rapid analysis of dense genetic maps using sparse gene flow trees. *Nat. Genet.*, 30, 97–101, 2002.
28. Adams, M.D., Kelley, J.M., Gocayne, J.D., Dubnick, M., Polymeropoulos, M.H., Xiao, H., Merril, C.R., Wu, A., Olde, B., Moreno, R.F. *et al.*, Complementary DNA sequencing: Expressed sequence tags and human genome project. *Science*, 252, 1651–1656, 1991.
29. Auer, P.L. and Doerge, R.W., Statistical design and analysis of RNA sequencing data. *Genetics*, 185, 405–416, 2010.
30. Battle, A., Mostafavi, S., Zhu, X., Potash, J.B., Weissman, M.M., McCormick, C., Haudenschild, C.D., Beckman, K.B., Shi, J., Mei, R. *et al.*, Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Res.*, 24, 14–24, 2013.
31. Blencowe, B.J., Ahmad, S., Lee, L.J., Current-generation high-throughput sequencing: Deepening insights into mammalian transcriptomes. *Genes Dev.*, 23, 1379–1386, 2009.
32. Brennecke, P., Anders, S., Kim, J.K., Kolodziejczyk, A.A., Zhang, X., Proserpio, V., Baying, B., Benes, V., Teichmann, S.A., Marioni, J.C. *et al.*, Accounting for technical noise in single-cell RNA-seq experiments. *Nat. Methods*, 10, 1093–1095, 2013.
33. Crick, F., Central dogma of molecular biology. *Nature*, 227, 561–563, 1970.
34. Zeng, W. and Mortazavi, A., Technical considerations for functional sequencing assays. *Nat. Immunol.*, 13, 802–807, 2012.
35. Wang, Z., Gerstein, M., Snyder, M., RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10, 57–63, 2009.
36. Rudloff, U., Bhanot, U., Gerald, W., Klimstra, D.S., Jarnagin, W.R., Brennan, M.F., Allen, P.J., Biobanking of human pancreas cancer tissue: Impact of *ex-vivo* procurement times on RNA quality. *Ann. Surg. Oncol.*, 17, 2229–2236, 2010.
37. Wodak, S.J. and Janin, J., Computer Analysis of Protein-Protein Interactions. *J. Mol. Biol.*, 124, 2, 323–42, 1978.
38. Lee, K., Computational Study for Protein-Protein Docking Using Global Optimization and Empirical Potentials. *Int. J. Mol. Sci.*, 9, 65–77, 2008.
39. Abbott, A., A post-genomic challenge: learning to read patterns of protein synthesis. *Nature*, 402, 715–720, 1999.
40. Aebersold, R., Rist, B., Gygi, S.P., Quantitative proteome analysis: methods and applications. *Ann. N. Y. Acad. Sci.*, 919, 33–47, 2000.

41. Yates, J.R., III, Protein structure analysis by mass spectrometry. *Methods Enzymol.*, 271, 351–377, 1996.
42. Aebersold, R., A mass spectrometric journey into protein and proteome research. *J. Am. Soc. Mass. Spectrom.*, 14, 685–695, 2003b.
43. Bischoff, R. and Luider, T.M., Methodological advances in the discovery of protein and peptide disease markers. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.*, 803, 27–40, 2004.
44. MacBeath, G. and Schreiber, S.L., Printing proteins as microarrays for high-throughput function determination. *Science*, 289, 1760–1763, 2000.
45. Watson, J.D., Hopkins, N.H., Roberts, J.W. *et al.*, *Molecular Biology of the Gene*, 4th edn., Benjamin-Cummings, Menlo Park, CA, 1987.
46. Aluru, S. (Ed.), *Handbook of Computational Molecular Biology*, Chapman & Hall/CRC Press, 2006.