Role of Bioinformatics in Analysing Big Data Using Statistical Computing and Computer Science

Kuldeep Singh Raina[1], Dr.Mushtaq Ahmad Sofi[2], Nazir Ahmad Dar [3], Dr.Owais Zargar[4]

Population Based Cancer Registry
Department of Radiation Oncology
Sher-i Kashmir Institute of Medical Sciences –Skims Srinagar
(kuldeepsingh.raina77@gmail.com)

This article discusses the links between computer science, statistics and biology education. Bioinformatics study is considered from two aspects – as one for biologists learning Information Technologies (IT) to use within their speciality, or for IT specialists learning biology so they can apply their skills to biological problems. The different computer science technologies and statistical methods in bioinformatics are considered. Currently, the concept of Big Data has become common, under big data, we understand information of a huge volume and of a diverse composition, which is often updated and located in different sources, as well as special technologies for storage, transfer, processing and analysis of this information. The multidisciplinary approach facilitates the understanding of interrelations between computer science technologies, statistical methods and bioinformatics. Bioinformatics covers a wide range of biology topics, such as genetics, evolution, biochemistry, biophysics, and cell biology. Computational biology leverages quantitative tools such as machine learning, statistical physics, and algorithm design and frequency statistics.

INTRODUCTION

The term "bioinformatics" was first used in 1970 by B. Hesper and P. Hogeweg in an article published in Dutch .There it was defined as "the study of information processes in biotic systems". The authors considered the management of

information in various forms, for example, the accumulation of information in the process of evolution, information transmission from DNA to intra- and intercellular processes, the interpretation of information at various levels of life as the defining property of life. Modern bioinformatics is a science that develops the use of computer methods for the analysis of a variety of genomic data. A huge role in the development of bioinformatics was played by the rapid development of computer technology and

DOI: 10.55041/IJSREM18829

computational methods of data processing, and the emergence of modern telecommunications technologies. Bioinformatics is

one of the science areas that are more dependent on the Internet. The very important for biology and medicine political decision, about the open accessibility of the most complex biological text – the human genome – has made this valuable source of knowledge accessible to scientists around the world and has enabled the formation of bioinformatics as a collective science, in which the achievements of separated teams are immediately made available to the entire scientific community, and where it is customary to freely distribute developed software and data.

Computational biologists are tasked with the development and application of data-analytical tools, theoretical methods, and mathematical modelling and software simulation techniques to explore biological systems. Computation is now an essential part of biological research projects. For example, protein data banks, genomic databases and brain MRI images contain massive amounts of raw data that can be translated into insightful information about all aspects of biology.Now with the strengthening of information technology facilities, it was inevitable that the subject of Bioinformatics would arise. Bioinformatics can be approached from two aspects – as one for biologists learning Information Technologies (IT) to use within their specialty, or for IT specialists learning biology so they can apply their skills to biological problems. It is the combination of mathematics, computer science, statistics and biology. In this article there are different computer science technologies and statistical methods in bioinformatics are considered. Multidisciplinary approach allows facilitating the understanding of interrelations between computer science technologies, statistical methods and bioinformatics applications . Oncology research and development departments employ computational biologists as key members of multidisciplinary teams who work to discover and develop new cancer treatment medicines. They are responsible for providing computational support to their fellow researchers by mining public and private genomic data and investigating potential treatment pathways. They strive to identify predictive biomarkers through understanding cellular mechanisms in clinical settings.These computational biologists will analyse and do processing of genomic data as well as knowledge of industry standard pathway tools and network analysis databases used in the field of cancer research, such as GSEA and DAVID. They will also need the ability to effectively work in a multidisciplinary team setting and effectively communicate in a clear and concise manner. Experience with scripting, machine learning and drug discovery and development is preferred.

Readers should note that the terms computational biology and bioinformatics are often used interchangeably, but the first usually connotes the development of algorithms and mathematical models, while bioinformatics is commonly associated with the development of software and visualization tools.

Unique Problem Solving

Many fields of the life sciences, such as biology and chemistry, now rely on quantitative prediction and interpretation to address complex questions that can only be answered using advanced statistical, mathematical and computational tools. System biology uses computational and mathematical modelling to solve complex biological problem. It is now making a great deal in healthcare and medicine with the help of digital revolution. The main goal is to provide better foundation to produce and develop preventative, predictive in the field of medical science. Recently many advance computer support systems which helps us to improve protect, promote, and maintain health and well- being and to prevent disease, disability and death.

Using big data in field of preventative medicine, we can improve the health of patients and give a better diagnose while treating the disease. As the role of Big data comes, more and more information from all around the world can be balanced. As the prototypes are being made with the help of large collection of data using big data technique The amount of data available for biologists now requires substantive quantitative approaches to make accurate analyses and interpretations. The power of computers means that researchers can explore sophisticated and highly complex problems.Computational biologists focus on biology or math and computers. They take classes in probability models, inferential statistics and quantitative genetics and genomics. Three of the most common programming languages studied include C++, Python and MATLAB. Most programs require classes in statistical theories, data mining and machine learning. Other common classes include differential equations, dynamical systems and bioinformatics programming.

INTERCONNECTION BETWEEN COMPUTER SCIENCE, COMPUTATIONAL STATISTICS ANS BIOINFORMATICS

Bioinformatics is the application of computational tools and techniques to the management and analysis of biological data. It is related to such terms as “computational biology” and others. Bioinformatics would not be possible without advances in computing hardware and software analysis of algorithms, data structures and software engineering . One reason why computer scientists are attracted to molecular biology is that the way information is encoded in DNA is in some way similar to the way it is coded in computers. While computers on a basic level deal with zeros and ones (bits), DNA carries information as chains of molecules (nucleotides) that come in four different types.Besides data analysis is seen as the largest and possibly the most important area of microarray bioinformatics. Statistical analyses for differentially expressed genes are best carried out via hypothesis tests rather than using a simple fold ratio threshold. More complex data may require analysis via ANOVA or general linear models and may be also include bootstrapping. Principal Component Analysis (PCA) and Multidimensional Scaling (MDS) provide a good way to visualise data without imposing any hierarchy on them. Hierarchical clustering can be used to identify related genes or samples and portray the usage of dendrogram. There are several methods for classifying samples, each with advantages and disadvantages, including K-nearestneighbour centroid classification, linear discriminate analysis, neural network, support vector machines.

It was well understood that computing would play a vital role in the future progress of statistics. Access to elaborate algorithms on computers

Increased the awareness of more recent methodological developments in statistics. According to the definition proposed by A. Westlake "Computational statistics is related to the advance of statistical theory and methods through the use of computational methods. This includes both the use of computation to explore the impact of theories and methods, and development of algorithms to make these ideas available to users." Computation in statistics is based on algorithms which originate in numerical mathematics or in computer science. The core topics of numerical mathematics are numerical linear algebra and optimization techniques but practically all areas of modern numerical analysis may be useful. The group of algorithms highly relevant for computational statistics from computer science is machine learning, artificial intelligence (AI), and knowledge discovery in data bases or data mining. These developments have given rise to a new research area on the borderline between statistics and computer science. Besides the difficulties resulting from new problems in various research areas, for example analysis of microarrays in biology, the following three interwoven challenges for computational statistics: handling of problems stemming from new data capture techniques, from the complexity of data structures, and from the size of data The summary of the different subjects of science interrelationship is shown in Figure A

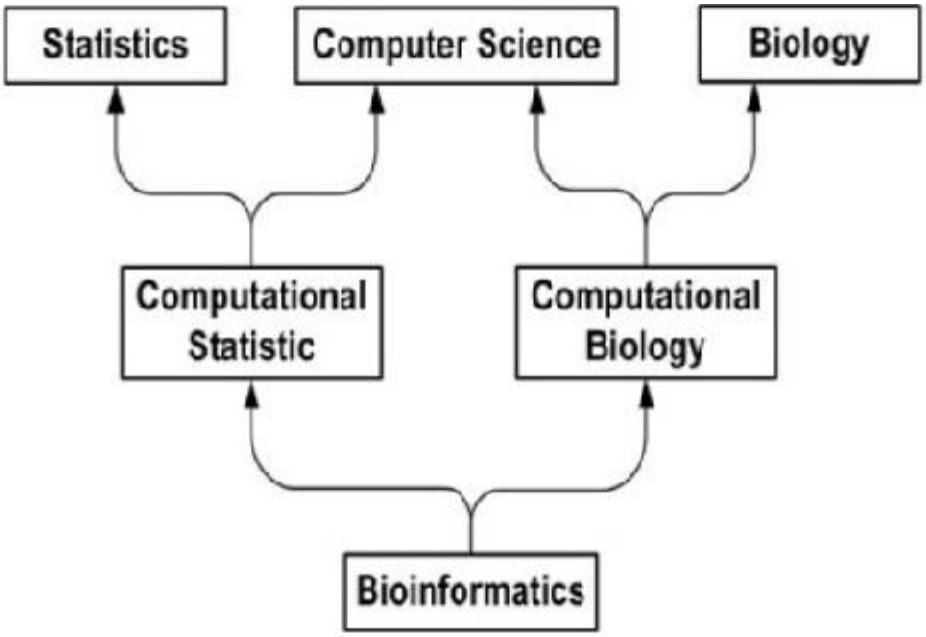

Figure A: The interrelationship of the different subjects of sciences

DATA MINING TECHNIQUES AND STATISTICAL METHODS COMPARISON

A variety of techniques have been developed over the years to explore for and extract information from large data sets. At the end of the 1980s a new discipline, named data mining, emerged. Traditional data analysis techniques often fail to process large amounts of data efficiently. Data mining is the process of discovering valid, previously unknown, and ultimately comprehensible information from large stores of data. Data Mining is the process of extracting knowledge hidden in large volumes of raw data. Data mining automates the process of finding relationships and patterns in raw data and delivers results that can be either utilized in an automated decision support system or assessed by a human analyst. Modern computer data mining systems self-learn from the previous history of the investigated system, formulating and testing hypotheses about the rules which this system obeys. In general, data mining techniques can be divided into two broad categories: predictive data mining and discovery data mining.

Predictive data mining is applied to a range of techniques that find relationship between a specific variable (called target variable) and the other variables in your data. The following are examples of predictive mining techniques:

Classification is about assigning data records into pre-defined categories. In this case the target variable is the category and the techniques discover the relationship between the other variables and the category.

Regression is about predicting the value of a continuous variable from the other variables in a data record. The most familiar value prediction techniques include linear and polynomial regression.

Discovery data mining is applied to range of techniques that find patterns inside your data without any prior knowledge of what patterns exist. The following are examples of discovery mining techniques:

Clustering is the term for range of techniques, which attempts to group data records on the basis of how similar they are.

Association and sequence analysis describes a family of techniques that determines relationship between data records.

The particularity of contemporary requirements for the data processing is the following: the data have the unlimited quantity, the data differ (quantitative, qualitative and textual), the results should be particular and comprehensible and the tools for the processing of raw data should be easy to use. The modern technology of Data Mining (discovery-driven data mining) is based on the concept of patterns, reflecting fragments of polydimensional relationships within the data. These patterns are regularities, which are characteristic to the data sub-retrievals that can be reflected compactly in the form, which is easy to comprehend for the human.

The search for the patterns is carried out by means of techniques, which are not limited by a priori proposals about the structure of retrieval and type of the value distribution of indicators to be analyzed .

The traditional mathematical statistics, for a long time applying for the role of the basic tool of data analysis, clearly gave up in front of the problems coming into existence. The main reason – the concept of the mean of retrieval that leads to the operations on the fictitious values. The techniques of mathematical statistics proved to be useful mainly for the verification of preliminary defined hypotheses (verification-driven data mining) and for the 'rough' research analysis that forms the basis of operative analytical data processing (online analytical processing, OLAP). At the same time the strong correlation exists between data mining and statistical methods, because statistical methods can be divided into the similar categories as data mining techniques: dependence methods and interdependence methods. The objective of the dependence methods is to determine whether the set of independent variables affects the set of dependent variables individually and/or jointly. That is, statistical techniques only test for the presence or absence of relationships between the two sets of variables. At the same time there exist such data sets for which it is impossible to designate conceptually the set of variables as

dependent or independent. For these types of data sets the objectives are to identify how and why the variables are related among themselves. Statistical methods for analyzing these types of data sets are called interdependence methods. The classification of the data mining and statistical methods is the following:

▪ Data Mining

- Predictive techniques: Classification, Regression.
- Discovery techniques: Association Analysis, Sequence Analysis, Clustering.

▪ Statistical methods

- Dependence methods: Discriminant analysis, Logistic regression.
- Interdependence methods: Correlation analysis, Correspondence analysis, Cluster analysis.

Conclusion

Bio Informatics and Computer science is becoming more focused on data rather than computation, and modern statistics requires more computational sophistication to work with large data sets," Lafferty says. "Machine learning draws on and pushes forward both of these disciplines." The goal is to develop computer programs that, with little or no human input, can extract knowledge from large amounts of numbers, text, audio or video and make predictions and decisions about events that haven't been coded in its instructions. As big data becomes more common in fields including astronomy, biology, and the humanities, researchers need new statistical techniques to reveal meaningful signals amid the noise Today, when the collection of biological data has been increasing at explosive rate, the processing of these data is needed. It would be extremely valuable if the specialists who know what the data mean were also able to imagine ways to collect, process and exploit these data. Biological and agricultural scientists should note that concepts from computer science, discrete mathematics and statistics are being used increasingly to study and describe biological systems. The specialists, who know the subjects of mathematics, statistics and computer programming, are needed for solving the computational problems in biology. As we enter into the information age data are being generated by variety of sources other than people and servers such as sensors embedded into phones and wearable devices, video surveillance cameras, set-top boxes and so on. High performance technologies are used in scientific research, such as fast data capturing tools and very high resolution satellite data recording.The world of big data is changing dramatically before our eyes, from the increase in big data growth to the way in which it is structured and used. The trend of big data growth presents enormous challenges, but it also presents incredible business opportunities. Taking growth chart we can see is the rapid growth of Big Data. Keeping that in mind the annual growth of data generation may reach 44 trillion zettabyte by the year 2021.

REFERENCES

1.Eidhammer, I., Jonassen, I., and Taylor, W. R. (2004). Protein Bioinformatics: An Algorithmic

2.Approach to Sequence and Structure Analysis. New York: John Wiley and Sons, Ltd.

3.Grossmann, W., Schimek, M. G. and Sint P. P. (2014). The history of Compstat and key-steps of statistical computing during the last 30 years. COMPSTAT 2004. Proceedings in Computational Statistics, (pp. 1-35).

4.Lauro, C. (1996). Computational statistics or statistical computing, is that the question? Computational Statistics and Data Analysis, 23, 191–193.

5.LeBlanc, M. D. and Dyer, B. D. (2005). Bioinformatics and Computing Curricula 2001. Why Computer Science is well positioned in a post-genomic world,

6.Lesk, A. M. (2012). Introduction to Bioinformatics. Oxford: Oxford University Press.

7.Mamcenko, J. and Kulvietiene, R. (2004). IBM intelligent miner for data and its application.

8.Scientific Proceeding of Riga Technical University, Series-Computer Science. Applied Computer Systems, 5th Volume, (pp. 81-91).

9.Roberts, E. (Ed.) (2002). Computing Curricula 2001: Computer Science Final Report. New York:

10.Sharma, S. (1996). Applied Multivariate Techniques. New York: John Wiley and Sons, Inc.

11.Stekel, D. (2013). Microarray Bioinformatics. Cambridge: Cambridge University Press.

12.Waterman, M. S. (2000). Introduction to computational biology. Maps, sequences and genomes.