ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

IARJSET

International Advanced Research Journal in Science, Engineering and Technology

ICRAESIT - 2015

B V Raju Institute of Technology, Narsapur

Vol. 2, Special Issue 2, December 2015

Bioinformatics - Its Challenges

B. Pradeep Kumar Reddy[1], **S. Srinuvasarao**[2]

Assistant Professor, Computer Science and Engineering, B V Raju Institute of Technology, Narsapur, India [1, 2]

Abstract: Bioinformatics is a new science that is glowing out in the recent years. It is a multidisciplinary science that is made out of different kinds of other scientific fields like biology, computer science, chemistry, statistics, mathematics and others. Bioinformatics as it applied to medicine has changed over the years from its origins in sequence analysis and data management. It has moved from its computer science roots to interdisciplinary applications. It was a big challenge for researchers to describe this new field in a systematic scientific way and bring out the attention of its applications and services; one of these important services that Bioinformatics can be applied in, is the cancer studies, research and therapies for many beneficial reasons. This field has experienced tremendous growth since the early 2000's at the start of The Human Genome Project and has been used heavily in proteomics and genomics. The use of FPGA's is a promising way to increase computing speed, and computer tools such as BLAST and iBIRA make the ever-growing amount of biological data more manageable. Bioinformatics has a profound impact in medical sciences. The biological databases are helping physicians to diagnose the disease and develop strategies for its therapy. This paper will give a clear glance overview of bioinformatics, its definition, aims, applications, technologies, the large amount of data produced in the biological field and how bioinformatics can organize, analyse and store them, discuss some algorithms that can be implemented over bioinformatics data.

Keywords: Bioinformatics, Applications, Technologies, Data, Algorithms.

I. CURRENT CHALLENGES

We describe and classify coarsely the current challenges we envisage in bioinformatics from an algorithmic perspective, excluding biotechnological tasks that need to be overcome. We can see three main areas of interest; analyzing big data, disease analysis, and bioinformatics education as described below.

Data mining and knowledge discovery techniques have greatly progressed in the last decade. They are now able to handle larger and larger datasets, process heterogeneous information, integrate complex metadata, and extract and visualize new knowledge. Often these advances were driven by new challenges arising from real-world domains, with biology and biotechnology a prime source of diverse and hard (e.g., high volume, high throughput, high variety, and high noise) data analytics problems. The aim of this article is to show the broad spectrum of data mining tasks and challenges present in biological data, and how these challenges have driven us over the years to design new data mining and knowledge discovery procedures for bio data.

This is illustrated with the help of two kinds of case studies. The first kind is focused on the field of protein structure prediction, where we have contributed in several areas: by designing, through regression, functions that can distinguish between good and bad models of a protein's predicted structure; by creating new measures to characterize aspects of a protein's structure associated with individual positions in a protein's sequence, measures containing information that might be useful for protein structure prediction; and by creating accurate estimators of these structural aspects. The second kind of case study is focused on omics data analytics, a class of biological data characterized for having extremely high dimensionalities. Our methods were able not only to generate very accurate classification models, but also to discover new biological knowledge that was later ratified by experimentalists. Finally, we describe several strategies to tightly integrate knowledge extraction and data mining in order to create a new class of bio-data mining algorithms that can natively embrace the complexity of biological data, efficiently generate accurate information in the form of classification/regression models, and extract valuable new knowledge. Thus, a complete data-to-information-to-knowledge pipeline is presented.

II. BIG DATA ANALYSIS

The big data is a collection of large and complex data which cannot be efficiently processed using conventional database techniques. High-throughput techniques such as next-generation sequencing provided big data in bioinformatics in the last decade. This data comes in basically as sequence, network, and image data. The basic requirement in bioinformatics is to produce knowledge from the raw data. Unfortunately, our ability to process this data does not grow in proportion with the production of data. We can classify the management of big data as follows.

DOI 10.17148/IARJSET

- Access and management: The basic requirements for access and management of Big Data are its reliable storage, a file system, and efficient and reliable network access. Client/server systems are typically employed to access data which is distributed over a network. The file systems can be distributed, clustered, or parallel.
- The Middleware: The middleware resides between the application and the operating system which manages the hardware, the file system, and the processes. It is impossible to generate middleware that is suitable for all applications but rather, some common functionality required by the applications can be determined and the middleware can then be designed to meet these common requirements. Message Passing Interface (MPI) can be considered as one such middleware for applications that require parallel/distributed processing.
- Data mining: Data mining is the process of analysing large datasets with the aim of discovering relationships among data elements and present the user with a method to analyse data further conveniently. In practical terms, a data mining method finds patterns or models of data such as clusters and tree structures. Clusters provide valuable information about raw data as we have analysed in previously. A data mining method should specify the evaluation method and the algorithmic process, for example, the quality of clustering in a clustering method.
- Parallel and distributed computing: Given the huge size of data, the parallel/distributed computing is increasingly more required. At hardware level, one can employ clusters of tightly coupled processors, graphical processing units (GPUs), or distributed memory processor connected by an interconnection network. Our approach is using the latter as it is most versatile and can be implemented by many users more conveniently. We need distributed algorithms to exploit parallel operations on this huge data and this area has not received significant attention from the researchers of bioinformatics as we have tried to emphasize throughout this paper.

A cloud is a computation infrastructure consisting of hardware and software resources and providing computational services to the users over the Internet. Users are charged only for the services they use and the maintenance and security of the system is granted by the provider of the cloud. Clouds can be private in which the facilities are offered to an organization only, or public where the infrastructure is made available to general public for a fee. This pay as you use approach is economical as the user does not have to invest on computation. Some of the public cloud computing systems in operation are Dropbox [8], Apple iCloud [7], Google Drive [8], Microsoft Azure [5], and Amazon Cloud Drive [6].

Cloud computing is merging as a commonly used platform to handle biological data. The traditional way of managing biological data is to download it to the local computer system and use algorithms/tools in the local system to analyse data which is not efficient with big data sizes. The general idea of using cloud computing for bioinformatics applications is to store big data in the cloud and move the computation to the cloud and perform processing there. Additionally, it is of interest for researchers of bioinformatics to access data of other researchers in a suitable format and easily which can be provided by a cloud. Cloud-based services for bioinformatics applications can be classified as Data as a Service (DaaS), Software as a Service (SaaS), Platform as a Service (PaaS), and Infrastructure as a Service (IaaS). There are, however, some issues in employing cloud computing for bioinformatics applications. First, data being large necessitates the provision of efficient uploading, in some cases using hard drives should be possible. Security has to be handled by the provider and an important reason for using cloud is the scalability which requires parallel/distributed processing facilities to be provided by the cloud. The latter is possibly one of the current challenges in using cloud for bioinformatics problems as there are very few such environments such as MapReduce [3] which provides a parallel computing environment in a cloud. Hadoop [1] uses MapReduce for parallel processing of big data. An early effort of searching for SNPs using cloud computing was presented in [9] using the Crossbow system.

III. DISEASE ANALYSIS

Another grand challenge in algorithmic bioinformatics is the efficient analysis of the disease states of organisms to be able to design cures and drug therapies. In order to accomplish this formidable task, especially for complex diseases like cancer, we need to understand the causes, mechanisms, and progression of diseases at molecular level. We have the basic biochemical reactions as DNA → RNA → mRNA → protein peptides called gene expression which is the central dogma in molecular biology. Considering the disease onset follows a similar pattern but with disease genes to disease proteins, we need to look at this process closely using algorithmic techniques. The classical view of disease genes causing diseases has changed significantly to considering that a group of genes getting involved in producing a disease sub network. Network-based approaches are therefore increasingly being used to study complex diseases. A biological network can be represented by a graph and the rich theory and algorithmic techniques for graphs can be used to analyse these networks. Two important networks in the cell that are affected by the disease states of an organism are the protein–protein interaction networks and the gene regulation networks. Proteins are the fundamental molecules in the cell carrying out vital functions needed for life. They interact with each other forming protein–protein interaction (PPI) networks as we have reviewed in Part II and they also interact with DNA and RNA to perform the necessary cell processes.

In a graph representing a PPI network, nodes represent the proteins and the undirected edges between nodes show the interaction between the nodes. Protein interactions play a key role to sustain the healthy states of an organism from which we can deduce their dysfunction may be one of the sources of disease states. Amutation in a gene may cause unwanted new interactions in PPI networks such as with pathogens or protein misfolds to result in diseases. For example, protein misfolding due to mutations may result in lost interactions in a PPI network causing diseases [4]. Unwanted newly formed protein interactions due to mutations are considered as the main causes of certain diseases such as Huntington's disease, Alzheimer's disease, and cystic fibrosis. Moreover, some bacterial and viral infections such as Human papillomavirus may interact with the proteins of the host organism. In order to understand the mechanisms of disease in the cell, PPI networks can be used to discover pathways which are sequential biochemical reactions. Finding a sub-network of a PPI network to find the corresponding pathway can aid to understand disease progression [4]. Also, by discovering disease proteins, we can predict the disease genes. A functional module in the cell has various interacting components and can be viewed as an entity with a specific function. Gene expression is controlled by proteins via regulatory interactions, for example, transcription factor proteins bind to sites near genes in DNA to regulate them. Such formed networks which have genes, proteins, other molecules, and their interactions are called gene regulation networks (GRNs) which can be represented by directed graphs. An edge in a GRN has an orientation, for example, from the transcription factor to the gene regulated. A GRN is basically a functional network as opposed to the PPI network which is physical. An important challenge in a functional network is the identification of the sub-network associated with the disease. An effective approach to discover functional modules in the cell is the clustering process we have seen. Investigation of the physical and functional networks is needed to identify the disease-associated sub-networks. For example, discovery of a disease-affected pathway can be performed by first identifying the mutated genes, then finding the PPI sub-network associated with the mutated genes and finally searching for modules associated with the disease PPI sub-network to discover the dysfunctional pathways [2]. In conclusion, the study of PPI networks, functional networks such as GRNs in the cell using graph-theoretical analysis is needed to understand the disease states of organisms better.

IV. CONCLUSION

Looking at the current challenges, we can anticipate the future challenges and research in bioinformatics will continue to be in these topics but with possible added orientations to new subareas. Our expectation is that the bioinformatics education will be more stabilized than the current status but the main research activities in algorithmic studies in bioinformatics will probably be on the management of big data which will grow bigger, and the analysis of diseases and evolution.

REFERENCES

[1] Apache Hadoop: http://hadoop.apache.org/

[2] Cho D-Y, Kim Y-A, Przytycka TM (2012) PLOS computational biology, translational bioinformatics collection volume 8, issue 12, chapter 5: network biology approach to complex diseases.

[3] Dean J, Ghemawat S (2004) MapReduce: simplified data processing on large clusters. Proceedings of the 6th symposium on operating systems design and implementation: 6–8 Dec 2004; San Francisco, California, USA, vol 6. ACM, New York, USA, pp 137–150.

[4] Gonzalez MW, Kann MG (2012) PLOS computational biology, translational bioinformatics collection, volume 8, issue 12, chapter 4: protein interactions and disease.

[5] https://azure.microsoft.com

[6] https://www.amazon.com/clouddrive

[7] https://www.icloud.com

[8] https://www.dropbox.com

[9] https://drive.google.com/drive

[10] Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL (2009) Searching for SNPs with cloud computing. Genome Biol 10(11):R134

BIOGRAPHY

B. Pradeep Kumar Reddy, working as Assistant Professor in Department of Computer Science and Engineering, B V Raju Institute of Technology, Narsapur, Medak, Telangana, India. Completed M.Tech (CSE) in 2014 and B.Tech (CSE) in 2012.

S. Srinuvasarao, working as Assistant Professor in Department of Computer Science and Engineering, B V Raju Institute of Technology, Narsapur, Medak, Telangana, India. Completed M.Tech (CSE) in 2010 and B.Tech (CSE) in 2006.