Bioinformatics Computing: Ongoing Research Directions and Future Scope

Deepak Garg[1], S C Saxena[2] and L M Bhardwaj[3]

[1,2]*Thapar Institute of Engg. & Tech., Patiala*

[3]*Biomolecular Electronics & Nanotechnology, CSIO, Chandigarh*

Abstract

Bioinformatics Computing has become the buzzword throughout the world. The people from diverse backgrounds such as mathematics, computer science, biology, information technology, chemistry, medical science, agriculture engineering and life sciences are jumping on the bioinformatics track to get some pie of the benefits that may come. The students who want to get lucrative jobs, the academicians who want to remain in the forefront, the scientists who want to do something exciting and the business people who are looking for a profitable venture are all now looking towards bioinformatics. In this paper we had given some research directions and the future scope of bioinformatics computing so that the students, scientists, teachers and the agencies who are involved in bioinformatics can get some ideas out of it and work out something new.

1. Introduction

The bioinformatics computing is the area that involves applying computational powers of the computing tools & machines on the biological data so as to help the biologists and the life scientists in their work.

The subject of biology is not new. It is as old as the birth of any living species on earth. We have tons of information available on various biological entities and processes that has been written in different languages in different parts of the world. Much of the material is lying in various libraries across geographical locations, in hospitals as the files of individual patients, in the agricultural labs as the project data and in the research centers as the results of various experiments done. Apart from this the information is scattered so much that there is no standardization, classification, reusability and applicability of the data.

The discovery of a new drug used to require a lot of work and will take many years for experimentation and for final production. The complexity of the various metabolisms was little understood. There were no solutions for many horrific diseases that created havoc in many parts of the world.

The evolution of life is still a debatable topic. Everything not understood is attributed to some supreme force called GOD. We are not able to find out why the life is limited to 100 years or so? Why the people get old? Why the people cannot live a disease free life? Why we cannot decide about the qualities we want in our newborn child? Why we cannot grow plants that are very rich in our requirements? How we can engineer the animals so that supplement us for various needs?

There are so many questions like this that still look for an answer. Bioinformatics is the area that is capable of solving all these queries and that is why people are having a lot of expectations from the people involved in this field. Billions of dollars are being invested to get something out of the bioinformatics research.
The fictional stories and movies related to bioinformatics predict that in the coming years we will be able to decipher the secrets written in the genetic codes of various species and the life will be different on this planet.

2. Classification

1. The information that is available in the books or at any other place has to be made available online so that it is available at a click.
   The field of medical transcription is a step in this direction where the material is being standardized & written as per some format. The classification is also being done. Pub Med can be taken as one example that is already functioning and has become very popular in the bioinformatics community. Still lots of efforts are required for this task.
   Companies are building biomarts (Biological Data Markets) that will supply biological data as per your requirements. This is not as simple task as said. Because the data is growing exponentially, lot of junk, dirty, duplicate and hyper data is also there that has to be filtered out.
2. The biologists need to understand a little about the tools and the computational techniques that are being developed for their benefit. They have to come out of their setups and face the new things.
   How quickly can these methods be integrated with bioscience research to meet demands of rapidly growing body of data and to speed the flow of knowledge?
3. The computer scientists have to work vigorously in the direction of mining the expanding data so as to extract relevant information and patterns out. The researchers are working very hard but relatively the speed and sensitivity of the information extracted remains the same because of the exponential growth of the data. Some of the tools like BLAST, FASTA, ClustalW etc. are very good examples of this. The Genomic sequences of various species are now available online. NCBI databases are very good examples in this direction. Human Genome project has further strengthened the beliefs of the scientific community.

The specific problems and research directions are given below:

1. An ideal architecture for building the bioinformatics data warehouse for handling the terabytes of heterogeneous data has still eluded the researchers. This database should have the capability of queries that can be functionality specific or classification specific. Another important aspect of the data warehouse should be to accommodate future data as part of the warehouse and to integrate it with the classic data. The baseline can be existing data warehouse architecture for the business data and the associated functionalities. The Business data warehouses are already in use & there utility is well established.
2. Data mining algorithms in techniques such as classification, aggregation, generalization, cleaning and optimization can be taken and applied to bioinformatics data with minor or major improvements. New algorithms can also be developed. The need for specific implementation of these algorithms for bioinformatics is required because of the different nature of bioinformatics data. The direction most of the people are taking is that they are applying soft computing techniques upon the existing algorithms to improve their speed & also inculcate the scalability in them to handle vast amounts of data.

3. The power of the computational machines is increasing day by day. Still due to the sheer size & nature of the bioinformatics data we need more powerful tools & machines. The experiments are going on for distributed and parallel computing environments to solve this problem. So to make the algorithms capable of running in a distributed & parallel architecture is again a challenge. It is an active area of research. People are trying to make grids to use the unutilized processor power.
   Some sample experiments has already been done & are giving good results. Still much has to be done to make the idea viable to be used in general.
4. The big question in the minds of professionals working on biological sequences is about the classification. How the sequences can be classified so as to divide the sequences into categories based on their functions or qualities. If we are able to do that then the amount of work will reduce by manifold. This work is only possible if we put a computer scientist & biologist in the same shoe.
5. Evolution of life is still the biggest question in mind of the life scientists & biologists. Currently we are able to handle single mutations or local mutations at some particular points. But how it is happening globally at the sequence or genome level. How it has traversed from its first existence to the current form. That will solve many fundamental questions. We will be able to predict the global level mutations that are happening in a sequence or a genome. That will help us discover the cause of many diseases & many transformations that are taking place in the various species. The single level mutations are not considered to be of much importance but sometime a single change in the human genome can be the cause of cancer or some other disease. So nothing can be taken as a proven fact. That is why we need more specific patterns & proofs that give credence to our theory or we invent a new theory.
6. Protein modeling is an active area of research. How you can model a protein from a sequence or vice versa. How better it can be displayed in a 3-D environment. How we can come to know about the functionality & relation in context to other proteins from the protein modeling. The better software has to be developed for good interface & application for the users & the professionals who are working in the area of protein modeling. Many research labs are already working in this direction.
7. The area of multiple sequence alignment is still unexplored. Some introductory tools & algorithms are available but much is left to be done. We cannot get as much information from comparing two sequences as we can get from multiple sequences. The large the number, more the information. Apart from giving the similarity information, it will also be important to know the dissimilarity in them. Multiple sequence alignments can also help in classifying the sequences & predicting the evolution of the sequences. More robust string comparison algorithms need to be developed. The area of dynamic programming has to be taken further to help sort out the problems in multiple sequence alignment.
8. Statistical significance of the results gained after sequence matches through various tools is very important. We had succeeded in comparing the two sequences & getting the results. Like a search engine we get a lot of results. To make something out of those results is much more important then getting those results. The people are working on various theories with the help of statistics & taking clues from the biologists who can provide some additional information based on the content of those sequences. This is also a good area of research.
9. User friendly, visually powerful and knowledge based tools for making the pattern discovery in biological data is need of the hour. Many tools are already available and are also doing well. The biologists are now discovering new things relatively fast if we see from

a historic point of view. Still much more can be done. There is scope in improving these tools based on some parameters.

10. Bioinformatics databases are growing at an unimaginable speed because of the faster methods available to sequence the genomes for various species. They are breaking the records of the previously considered largest databases. So the techniques to store, access, modify, catalogue these databases has to be changed with the requirements of the coming time. The data is not only the text, it is heterogeneous data having different formats, files, language, size and style. What are the best computing methods to be applied to collect, organize, analyze and distribute this information to advance this research? What are the critical sources of information for advancing fundamental knowledge in biology, medicine, biotechnology, pharmacology, agriculture and related biosciences? The people who are interesting in data mining & warehousing can also pick up this field.

3. Conclusion

You have the energy & the resources to do the research work; there is enough to be done in bioinformatics. The work done in this field will be satisfying also because that is for the welfare of every living being on the universe. These are many open-ended questions & problems. Some of them have been listed above for your reference.

Acknowledgements

We are thankful to All India Council for Technical Education, New Delhi, Department of Biotechnology, Government of India, New Delhi, Thapar Institute of Engineering & Technology, Patiala, Computer Science & Engineering Department, TIET, Patiala, CSIO, Chandigarh for providing the resources & infrastructure to carry the work.

References

[1] Jason, Bruce, Dennis, “Pattern Discovery in Bimolecular Data”, Oxford University Press, New York 1999.
[2] Jean Michel Claverie and Cedric Notredame " Bioinformatics A beginner’s Guide". Wiley Publishing, Inc. 2003.
[3] Jiawei Han, Micheline Kamber amd Simon Fraser University “Data Mining Concepts and Techniques” Morgan Kaufmann Publishers, USA 2001.
[4] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy “Advances in Knowledge Discovery and Data Mining.” AAAI/MIT Press, 1996.
[5] J. Han and M. Kamber “Data Mining: Concepts and Techniques”. Morgan Kaufmann, 2000.
[6] T. Imielinski and H. Mannila “A database perspective on knowledge discovery. Communications of ACM”, 39:58-64, 1996.
[7] G. Piatetsky-Shapiro, U. Fayyad, and P. Smith “Data mining to knowledge discovery: An overview”. In U.M. Fayyad, et al. (eds.), Advances in Knowledge Discovery and Data Mining, 1-35. AAAI/MIT Press, 1996.
[8] G. Piatetsky-Shapiro and W. J. “ Frawley Knowledge Discovery in Databases.” AAAI/MIT Press, 1991.
[9] Jagadish et al., “Special Issue on Data Reduction Techniques. Bulletin of the Technical Committee on Data Engineering, 20(4)”, December 1997.
[10] Dan E. Krane, Michael L. Raymer “ Fundamental concepts of Bioinformatics” Pearson Education, 2003.
[11] Scott Markel, Darryl Leon “ Sequence Analysis In a Nutshell” O’reilly 2003.
[12] Kent, W. James “BLAT- The BLAST- like Alignment Tool” Genome Research 12 (4):656-664” 2002.
[13] Higgins, D.J. Thompson “ClustalW: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice” Nucleic Acid Research 22:4673-4680, 1994.
[14] Gilbert, D. G. “ReadSeq version 2, an improved biosequence conversion tool” Bionet.Software(Aug), 1999.