IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727
PP 20-26
www.iosrjournals.org

Exploring the Applications and Potential of Bioinformatics

Mrs. Kanchan Gawande[1], Mr. Dhiraj Rane[2]
[1](kanch_vg@rediffmail.com, AP, Dept. of CS, SRPCE, Nagpur, India)
[2](dhirajrane2302@gmail.com, Dept. of CS, GHRIIT, Nagpur, India)

Abstract: *This paper has discussed basic working of Bioinformatics and its association with computation. It also summarizes the application domains of Bioinformatics. The use of computer science is also discussed with the bioinformatics. This paper has explored the potential domain of the bioinformatics in the computer science development. It has also given the technical tools discussion for the various practical implantation and conducting experiments.*

Keywords- *Bioinformatics, Bioweka, Genes, DNA*

I. INTRODUCTION

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data. As an interdisciplinary field of science, bioinformatics combines computer science, statistics, mathematics, and engineering to analyze and interpret biological data. Bioinformatics has been used for in silico analyses of biological queries using mathematical and statistical techniques.

Bioinformatics is both an umbrella term for the body of biological studies that use computer programming as part of their methodology, as well as a reference to specific analysis "pipelines" that are repeatedly used, particularly in the field of genomics. Common uses of bioinformatics include the identification of candidate genes and nucleotides (SNPs). Often, such identification is made with the aim of better understanding the genetic basis of disease, unique adaptations, desirable properties (esp. in agricultural species), or differences between populations. In a less formal way, bioinformatics also tries to understand the organisational principles within nucleic acid and protein sequences.

Bioinformatics has become an important part of many areas of biology. In experimental molecular biology, bioinformatics techniques such as image and signal processing allow extraction of useful results from large amounts of raw data. In the field of genetics and genomics, it aids in sequencing and annotating genomes and their observed mutations. It plays a role in the text mining of biological literature and the development of biological and gene ontologies to organize and query biological data. It also plays a role in the analysis of gene and protein expression and regulation. Bioinformatics tools aid in the comparison of genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology. At a more integrative level, it helps analyze and catalogue the biological pathways and networks that are an important part of systems biology. In structural biology, it aids in the simulation and modeling of DNA, RNA, and protein structures as well as molecular interactions.

Historically, the term bioinformatics did not mean what it means today. Paulien Hogeweg and Ben Hesper coined it in 1970 to refer to the study of information processes in biotic systems.[1][2][3] This definition placed bioinformatics as a field parallel to biophysics (the study of physical processes in biological systems) or biochemistry (the study of chemical processes in biological systems).[1]

II. STUDY OF BIOINFORMATICS WITH COMPUTER SCIENCE AND OTHER DISCIPLINES

1.1. SEQUENCES

Computers became essential in molecular biology when protein sequences became available after Frederick Sanger determined the sequence of insulin in the early 1950s. Comparing multiple sequences manually turned out to be impractical. A pioneer in the field was Margaret Oakley Dayhoff, who has been hailed by David Lipman, director of the National Center for Biotechnology Information, as the "mother and father of bioinformatics."[4] Dayhoff compiled one of the first protein sequence databases, initially published as books[5] and pioneered methods of sequence alignment and molecular evolution.[6] Another early contributor to bioinformatics was Elvin A. Kabat, who pioneered biological sequence analysis in 1970 with his comprehensive volumes of antibody sequences released with Tai Te Wu between 1980 and 1991.[7]

Fig 1: DNA Sequence

1.2. GOALS

To study how normal cellular activities are altered in different disease states, the biological data must be combined to form a comprehensive picture of these activities. Therefore, the field of bioinformatics has evolved such that the most pressing task now involves the analysis and interpretation of various types of data. This includes nucleotide and amino acid sequences, protein domains, and protein structures.[8] The actual process of analyzing and interpreting data is referred to as computational biology. Important sub-disciplines within bioinformatics and computational biology include:

- Development and implementation of computer programs that enable efficient access to, use and management of, various types of information
- Development of new algorithms (mathematical formulas) and statistical measures that assess relationships among members of large data sets. For example, there are methods to locate a gene within a sequence, to predict protein structure and/or function, and to cluster protein sequences into families of related sequences.

The primary goal of bioinformatics is to increase the understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques to achieve this goal. Examples include: pattern recognition, data mining, machine learning algorithms, and visualization. Major research efforts in the field include sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein–protein interactions, genome-wide association studies, the modeling of evolution and cell division/mitosis.

Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data.

Over the past few decades, rapid developments in genomic and other molecular research technologies and developments in information technologies have combined to produce a tremendous amount of information related to molecular biology. Bioinformatics is the name given to these mathematical and computing approaches used to glean understanding of biological processes.

Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, aligning DNA and protein sequences to compare them, and creating and viewing 3-D models of protein structures.

1.3. RELATION TO OTHER FIELDS

Bioinformatics is a science field that is similar to but distinct from biological computation and computational biology. Biological computation uses bioengineering and biology to build biological computers, whereas bioinformatics uses computation to better understand biology. Bioinformatics and computational biology have similar aims and approaches, but they differ in scale: bioinformatics organizes and analyzes basic biological data, whereas computational biology builds theoretical models of biological systems, just as mathematical biology does with mathematical models.

Analyzing biological data to produce meaningful information involves writing and running software programs that use algorithms from graph theory, artificial intelligence, soft computing, data mining, image processing, and computer simulation. The algorithms in turn depend on theoretical foundations such as discrete mathematics, control theory, system theory, information theory, and statistics.

III. SEQUENCE ANALYSIS

The DNA sequences of thousands of organisms have been decoded and stored in databases. This sequence information is analyzed to determine genes that encode proteins, RNA genes, regulatory sequences, structural motifs, and repetitive sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematic to construct phylogenetic trees).

For a genome as large as the human genome, it may take many days of CPU time on large-memory, multiprocessor computers to assemble the fragments, and the resulting assembly usually contains numerous

gaps that must be filled in later. Shotgun sequencing is the method of choice for virtually all genomes sequenced today, and genome assembly algorithms are a critical area of bioinformatics research.

IV. APPLICATIONS

The application domains of Bioinformatics are:

- **Molecular medicine**

The human genome will have profound effects on the fields of biomedical research and clinical medicine. Every disease has a genetic component. This may be inherited (as is the case with an estimated 3000-4000 hereditary disease including Cystic Fibrosis and Huntingtons disease) or a result of the body's response to an environmental stress which causes alterations in the genome (eg. cancers, heart disease, diabetes.).

The completion of the human genome means that we can search for the genes directly associated with different diseases and begin to understand the molecular basis of these diseases more clearly. This new knowledge of the molecular mechanisms of disease will enable better treatments, cures and even preventative tests to be developed.

- **Personalised medicine**

Clinical medicine will become more personalised with the development of the field of pharmacogenomics. This is the study of how an individual's genetic inheritence affects the body's response to drugs. At present, some drugs fail to make it to the market because a small percentage of the clinical patient population show adverse affects to a drug due to sequence variants in their DNA.

As a result, potentially lifesaving drugs never make it to the marketplace. Today, doctors have to use trial and error to find the best drug to treat a particular patient as those with the same clinical symptoms can show a wide range of responses to the same treatment. In the future, doctors will be able to analyse a patient's genetic profile and prescribe the best available drug therapy and dosage from the beginning.

- **Preventative medicine**

With the specific details of the genetic mechanisms of diseases being unravelled, the development of diagnostic tests to measure a persons susceptibility to different diseases may become a distinct reality. Preventative actions such as change of lifestyle or having treatment at the earliest possible stages when they are more likely to be successful, could result in huge advances in our struggle to conquer disease.

- **Gene therapy**

In the not too distant future, the potential for using genes themselves to treat disease may become a reality. Gene therapy is the approach used to treat, cure or even prevent disease by changing the expression of a persons genes. Currently, this field is in its infantile stage with clinical trials for many different types of cancer and other diseases ongoing.

- **Drug development**

At present all drugs on the market target only about 500 proteins. With an improved understanding of disease mechanisms and using computational tools to identify and validate new drug targets, more specific medicines that act on the cause, not merely the symptoms, of the disease can be developed. These highly specific drugs promise to have fewer side effects than many of today's medicines.

- **Microbial genome applications**

Microorganisms are ubiquitous, that is they are found everywhere. They have been found surviving and thriving in extremes of heat, cold, radiation, salt, acidity and pressure. They are present in the environment, our bodies, the air, food and water. Traditionally, use has been made of a variety of microbial properties in the baking, brewing and food industries. The arrival of the complete genome sequences and their potential to provide a greater insight into the microbial world and its capacities could have broad and far reaching implications for environment, health, energy and industrial applications. For these reasons, in 1994, the US Department of Energy (DOE) initiated the MGP (Microbial Genome Project) to sequence genomes of bacteria useful in energy production, environmental cleanup, industrial processing and toxic waste reduction. By studying the genetic material of these organisms, scientists can begin to understand these microbes at a very fundamental level and isolate the genes that give them their unique abilities to survive under extreme conditions.

- **Waste cleanup**

Deinococcus radiodurans is known as the world's toughest bacteria and it is the most radiation resistant organism known. Scientists are interested in this organism because of its potential usefulness in cleaning up waste sites that contain radiation and toxic chemicals.

- **Climate change Studies**

Increasing levels of carbon dioxide emission, mainly through the expanding use of fossil fuels for energy, are thought to contribute to global climate change. Recently, the DOE (Department of Energy, USA) launched a program to decrease atmospheric carbon dioxide levels. One method of doing so is to study the genomes of microbes that use carbon dioxide as their sole carbon source.

- **Alternative energy sources**

Scientists are studying the genome of the microbe Chlorobium tepidum which has an unusual capacity for generating energy from light

- **Biotechnology**

The archaeon Archaeoglobus fulgidus and the bacterium Thermotoga maritima have potential for practical applications in industry and government-funded environmental remediation. These microorganisms thrive in water temperatures above the boiling point and therefore may provide the DOE, the Department of Defence, and private companies with heat-stable enzymes suitable for use in industrial processes

Other industrially useful microbes include, Corynebacterium glutamicum which is of high industrial interest as a research object because it is used by the chemical industry for the biotechnological production of the amino acid lysine. The substance is employed as a source of protein in animal nutrition. Lysine is one of the essential amino acids in animal nutrition. Biotechnologically produced lysine is added to feed concentrates as a source of protein, and is an alternative to soybeans or meat and bonemeal.

Xanthomonas campestris pv. is grown commercially to produce the exopolysaccharide xanthan gum, which is used as a viscosifying and stabilising agent in many industries.

Lactococcus lactis is one of the most important micro-organisms involved in the dairy industry, it is a non-pathogenic rod-shaped bacterium that is critical for manufacturing dairy products like buttermilk, yogurt and cheese. This bacterium, Lactococcus lactis ssp., is also used to prepare pickled vegetables, beer, wine, some breads and sausages and other fermented foods. Researchers anticipate that understanding the physiology and genetic make-up of this bacterium will prove invaluable for food manufacturers as well as the pharmaceutical industry, which is exploring the capacity of L. lactis to serve as a vehicle for delivering drugs.

- **Antibiotic resistance**

Scientists have been examining the genome of Enterococcus faecalis-a leading cause of bacterial infection among hospital patients. They have discovered a virulence region made up of a number of antibiotic-resistant genes that may contribute to the bacterium's transformation from a harmless gut bacteria to a menacing invader. The discovery of the region, known as a pathogenicity island, could provide useful markers for detecting pathogenic strains and help to establish controls to prevent the spread of infection in wards.

- **Forensic analysis of microbes**

Scientists used their genomic tools to help distinguish between the strain of Bacillus anthryacis that was used in the summer of 2001 terrorist attack in Florida with that of closely related anthrax strains.

- **The reality of bioweapon creation**

Scientists have recently built the virus poliomyelitis using entirely artificial means. They did this using genomic data available on the Internet and materials from a mail-order chemical supply. The research was financed by the US Department of Defence as part of a biowarfare response program to prove to the world the reality of bioweapons. The researchers also hope their work will discourage officials from ever relaxing programs of immunisation. This project has been met with very mixed feeelings

- **Evolutionary studies**

The sequencing of genomes from all three domains of life, eukaryota, bacteria and archaea means that evolutionary studies can be performed in a quest to determine the tree of life and the last universal common ancestor.

- **Crop improvement**

Comparative genetics of the plant genomes has shown that the organisation of their genes has remained more conserved over evolutionary time than was previously believed. These findings suggest that information obtained from the model crop systems can be used to suggest improvements to other food crops. At present the complete genomes of Arabidopsis thaliana (water cress) and Oryza sativa (rice) are available.

- **Insect resistance**

Genes from Bacillus thuringiensis that can control a number of serious pests have been successfully transferred to cotton, maize and potatoes. This new ability of the plants to resist insect attack means that the amount of insecticides being used can be reduced and hence the nutritional quality of the crops is increased.

- **Improve nutritional quality**

Scientists have recently succeeded in transferring genes into rice to increase levels of Vitamin A, iron and other micronutrients. This work could have a profound impact in reducing occurrences of blindness and anaemia caused by deficiencies in Vitamin A and iron respectively. Scientists have inserted a gene from yeast into the tomato, and the result is a plant whose fruit stays longer on the vine and has an extended shelf life.

- **Development of Drought resistance varieties**

Progress has been made in developing cereal varieties that have a greater tolerance for soil alkalinity, free aluminium and iron toxicities. These varieties will allow agriculture to succeed in poorer soil areas, thus adding more land to the global production base. Research is also in progress to produce crop varieties capable of tolerating reduced water conditions.

- **Vetinary Science**

Sequencing projects of many farm animals including cows, pigs and sheep are now well under way in the hope that a better understanding of the biology of these organisms will have huge impacts for improving the production and health of livestock and ultimately have benefits for human nutrition.

- **Comparative Studies**

Analysing and comparing the genetic material of different species is an important method for studying the functions of genes, the mechanisms of inherited diseases and species evolution. Bioinformatics tools can be used to make comparisons between the numbers, locations and biochemical functions of genes in different organisms.

Organisms that are suitable for use in experimental research are termed model organisms. They have a number of properties that make them ideal for research purposes including short life spans, rapid reproduction, being easy to handle, inexpensive and they can be manipulated at the genetic level.

An example of a human model organism is the mouse. Mouse and human are very closely related (>98%) and for the most part we see a one to one correspondence between genes in the two species. Manipulation of the mouse at the molecular level and genome comparisons between the two species can and is revealing detailed information on the functions of human genes, the evolutionary relationship between the two species and the molecular mechanisms of many human diseases.

V. SOFTWARES AND TOOLS

1.1. OPEN-SOURCE BIOINFORMATICS SOFTWARE

Many free and open-source software tools have existed and continued to grow since the 1980s.[8] The combination of a continued need for new algorithms for the analysis of emerging types of biological readouts, the potential for innovative in silico experiments, and freely available open code bases have helped to create opportunities for all research groups to contribute to both bioinformatics and the range of open-source software available, regardless of their funding arrangements. The open source tools often act as incubators of ideas, or community-supported plug-ins in commercial applications. They may also provide de facto standards and shared object models for assisting with the challenge of bioinformation integration.

The range of open-source software packages includes titles such as Bioconductor, BioPerl, Biopython, BioJava, BioJS, BioRuby, Bioclipse, EMBOSS, .NET Bio, Orange with its bioinformatics add-on, Apache Taverna, UGENE and GenoCAD. To maintain this tradition and create further opportunities, the non-profit Open Bioinformatics Foundation[8] have supported the annual Bioinformatics Open Source Conference (BOSC) since 2000.[9]

An alternative method to build public bioinformatics databases is to use the MediaWiki engine with the WikiOpener extension. This system allows the database to be accessed and updated by all experts in the field[10].

1.2. WEB SERVICES IN BIOINFORMATICS

SOAP- and REST-based interfaces have been developed for a wide variety of bioinformatics applications allowing an application running on one computer in one part of the world to use algorithms, data and computing resources on servers in other parts of the world. The main advantages derive from the fact that end users do not have to deal with software and database maintenance overheads.

Basic bioinformatics services are classified by the EBI into three categories: SSS (Sequence Search Services), MSA (Multiple Sequence Alignment), and BSA (Biological Sequence Analysis).[11] The availability of these service-oriented bioinformatics resources demonstrate the applicability of web-based bioinformatics solutions, and range from a collection of standalone tools with a common data format under a single, standalone or web-based interface, to integrative, distributed and extensible bioinformatics workflow management systems.

1.3. BIOINFORMATICS WORKFLOW MANAGEMENT SYSTEMS

A Bioinformatics workflow management system is a specialized form of a workflow management system designed specifically to compose and execute a series of computational or data manipulation steps, or a workflow, in a Bioinformatics application. Such systems are designed to

- provide an easy-to-use environment for individual application scientists themselves to create their own workflows
- provide interactive tools for the scientists enabling them to execute their workflows and view their results in real-time
- simplify the process of sharing and reusing workflows between the scientists.
- enable scientists to track the provenance of the workflow execution results and the workflow creation steps.

Some of the platforms giving this service: Galaxy, Kepler, Taverna, UGENE, Anduril.

1.4. BIOWEKA

One has to download both the Weka and the BioWeka distribution and include the Weka JAR in the CLASSPATH variable for BioWeka. The BioWeka startup script provides access to Weka as well as BioWeka. For the BLAST and PSI- BLAST classifiers, a BLAST installation is necessary. In the Explorer GUI, users can import the new data formats listed above using BioWeka's converters and apply BioWeka's filters and classifiers.

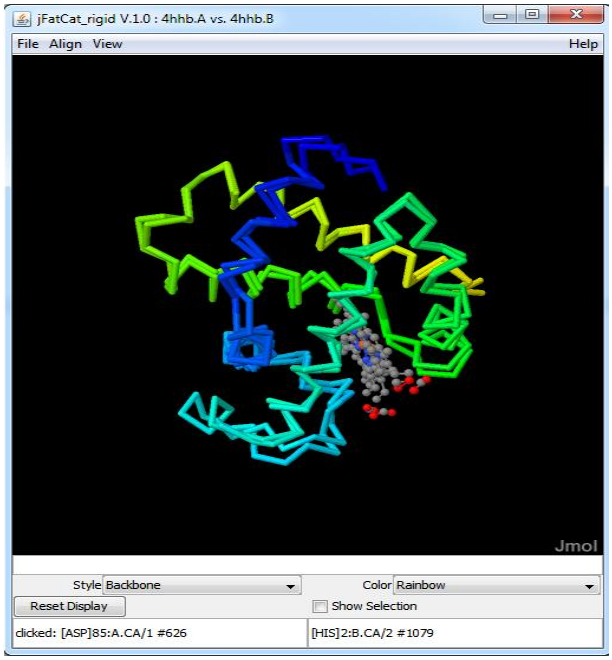

Figure 2: BioWeka

VI. CONCLUSION

Bioinformatics provide high potential to think towards the real life and mathematical problems with new human like approach. It also enhances the study of the living organism development and evolutions, which help us to open new computational areas. Bioinformatics also help us to search the solutions for the various diseases which is not possible by the computational science.

REFERENCES

[1] Hogeweg P (2011). Searls, David B., ed. "The Roots of Bioinformatics in Theoretical Biology". PLoS Computational Biology 7 (3): e1002021.
[2] Hesper B, Hogeweg P (1970). "Bioinformatica: een werkconcept" 1 (6). Kameleon: 28–29.
[3] Hogeweg P (1978). "Simulating the growth of cellular forms". Simulation 31 (3): 90–96.
[4] Moody, Glyn (2004). Digital Code of Life: How Bioinformatics is Revolutionizing Science, Medicine, and Business. ISBN 978-0-471-32788-2.
[5] Dayhoff, M.O. (1966) Atlas of protein sequence and structure. National Biomedical Research Foundation, 215 pp.
[6] Eck RV, Dayhoff MO (1966). "Evolution of the structure of ferredoxin based on living relics of primitive amino Acid sequences". Science 152 (3720): 363–6. Bibcode:1966Sci...152..363E. doi:10.1126/science.152.3720.363. PMID 17775169.
[7] Johnson G, Wu TT (January 2000). "Kabat Database and its applications: 30 years after the first variability plot"
[8] "Open Bioinformatics Foundation: About us". Official website. Open Bioinformatics Foundation. Retrieved 10 May 2011.
[9] "Open Bioinformatics Foundation: BOSC". Official website. Open Bioinformatics Foundation. Retrieved 10 May 2011.
[10] Brohée, Sylvain; Barriot, Roland; Moreau, Yves. "Biological knowledge bases using Wikis: combining the flexibility of Wikis with the structure of databases". Bioinformatics. Oxford Journals. Retrieved 5 May 2015.
[11] Nisbet, Robert (14 May 2009). "BIOINFORMATICS". Handbook of Statistical Analysis and Data Mining Applications. John Elder IV, Gary Miner. Academic Press. p. 328. Retrieved 9 May 2014
[12] http://bioinformaticsweb.net/applications.html
[13] http://www.biotecharticles.com/Bioinformatics-Article/Applications-of-Bioinformatics-3270.html
[14] https://en.wikipedia.org/wiki/Bioinformatics