SJIF IMPACT FACTOR (2015): 5.79 CRDEEP Journals
International Journal of Life Sciences Bisht et. al., Vol. 5 No. 4 ISSN: 2277-193x
International Journal of Life Sciences. Vol. 5 No. 4. 2016. Pp. 222-227

Review Paper

Computational Tools of Bioinformatics and Data Repository: A Scientific Review

Deepika Bisht*[1], R.K. Singh[2] & R.C. Prasad[3]

*[*1-]Junior Project Fellow, Govind Ballabh Pant National Institute of Himalayan Environment & Sustainable Development, Almora-263643, Uttarakhand, India.*

[2-]Scientist-D, Govind Ballabh Pant National Institute of Himalayan Environment & Sustainable Development, Almora-263643, Uttarakhand, India.

[3-]Scientist-F, Govind Ballabh Pant National Institute of Himalayan Environment & Sustainable Development, Almora-263643, Uttarakhand, India.

Article history
Received: 24-12-2016
Revised: 26-12-2016
Accepted: 28-12-2016

Corresponding Author:
Deepika Bisht
Junior Project Fellow, Govind Ballabh Pant National Institute of Himalayan Environment & Sustainable Development, Almora-263643, Uttarakhand, India.

Abstract
Bioinformatics is a field of science which constitutes computer science, biology and statistics. It is the application of information technology in managing biological data, helps in decoding plant genomes. Earlier biological researches used to start in laboratories, fields and plant clinics but now, it starts by using computers for experiment planning, analysis of the data, and hypothesis development. It is used for development of algorithms and suitable data analysis tools to infer the information and make discoveries. Application of bioinformatics tools in biological research enables storage, analysis, retrieval, annotation and visualization of results and helps in better understanding of biological system. This helps in plant health care to improve the quality and quantity of production of plants. The computational tools and analysis techniques discussed here can be used for researches bioinformatics.

Keywords: *Genes, Bioinformatics, Genomics, Proteome, Agriculture, etc.*

Introduction

The term *"Bioinformatics"* was invented by *Paulien Hoyeweg* and *Ben Hesper in 1970* as *"the study of informatics processes in biotic systems"*. It is the application of *"computational biology"* to the management and analysis of biological data. Bioinformatics is very rapidly progressing in every field of biotechnology. Bioinformatics is an interdisciplinary field that combines computer science, biology, mathematics, statistics and engineering to develop methods and software tools for analysis and interpretation of biological data. Bioinformatics would not be possible without advances in computer hardware and software: Analysis of algorithms, data structures and software engineering. To elaborate algorithms on computers increased the awareness of more recent statistical methods. Statistical analysis for differently expressed genes are best carried out via hypothesis test. More complex data may require analysis via ANOVA or general linear models. [10]

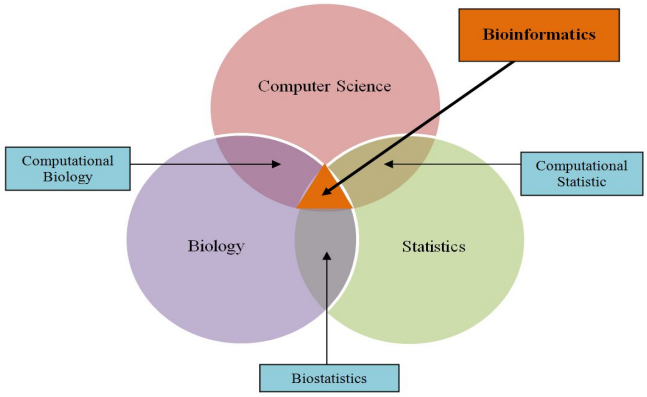

Fig.1. Taxonomy of Bioinformatics

Bioinformatics is used for biological studies that use computer programming as part of their methodology on one hand and references on the other hand, that are used repeatedly. Bioinformatics is commonly used for the identification of genes and nucleotides, which aim better understanding of the genetic basis of disease, unique adaptations, desirable features etc. It comprises

the creation and advancement of database, algorithms, computational, statistical techniques, data mining, automation, visualization tool to solve theoretical and practical problems while analysis and interpretation of biological data faster and to enhance the accuracy of results, which will reduce the time and cost of the process. The basic goal of bioinformatics is to increase the understanding of biological processes. It focuses on developing and applying computationally intensive techniques for achieving this goal.

Analysis Techniques used in Bioinformatics
The following analysis techniques are used to analyze biological data with the help of bioinformatics tools:

Genome Sequencing: The genome is considered as a complete set DNA sequence of hereditary material that is passed from one generation to another. The genomics refers to sequencing and analyzing all the genome entities in an organism. Genomic information has been recorded in database and is used identifying inherited disorders, characterizing the mutations that drive cancer progression, and tracking disease outbreaks. In this process the complete DNA sequence of genome of an organism is taken at a single time. It is different form DNA profiling. The genome sequence will result in valuable shortcuts, helping scientists to find genes much more easily and quickly.

Proteome (Protein Coding): By the term "proteome" we mean the entire complement of proteins, including every modification made to a particular set of proteins. With the change of time and distinct conditions of cell or organism it undergoes variation, which is to be studied. The term "proteomics" is the study of large-scale data of specific proteome which includes information on protein, their variations and modifications. This study is done to understand cellular processes. With the help of this complete analysis of every set of proteins takes place. Here Protein sequences are compared to secondary protein databases that contain information regarding their signature, domain etc. There a number of methods available to study the proteins, set of proteins or whole proteome.

Molecular Marker Discovery: It is a most powerful tool for the analysis of genomes and allows association between heritable traits and genomic variation. It leads to the identification of optimal SNP panels for genetic purity or for the purpose of verification of varietal. After taking our fingerprinting on our parental lines and building a genetic database accordingly it helps us in finding the optimal SNP panel for your genetic purity purposes.

Transcriptomics: It is the study of complete set of RNA transcripts; mRNAs and non coding RNA present in particular cell type or a group of cells. It varies in number according to the climatic conditions. High-throughput techniques based on DNA microarray are used for this study. With the help of this Expression profiling also takes place, which examines the expression level of mRNAs in a given cell population done on the basis of DNA microarray technology. It includes two general methods for inference of transcriptomes. The first one maps sequence reads onto a reference genome while other approach is *de novo* transcriptome assembly which uses software to infer transcripts directly from short sequence reads.

Metabolomics: It can be defined as the study of metabolomes. By the term metabolome, we mean small molecules and their interaction within a biological system. Metabolomes are influenced by both genetic and environmental factors. It is the analysis of metabolites present in the present in a cell, tissue, or organism in a particular physiological or developmental state. It is the study of fingerprinting that leaves behind specific cellular processes under particular conditions. It uses the strategies for the identification and quantification of cellular metabolites with the use of different analytical techniques with the application of statistical methods for the extraction of information and data interpretation.

Candidate Gene Identification: It is an approach to conduct genetic study which aims upon the association between phenotypes and genetic variation within pre-specified genes. The candidate is most common gene selected for study. Candidate gene studies are better suited to detect genes underlying common and most complex diseases. Suitable candidate genes are generally selected based on known biological, physiological, or functional relevance to the disease in question. Genome-wide association studies and quantitative trait locus (QTL) mapping examine common variation across the entire genome, and as such can detect a new region of interest that is in or near a potential candidate gene. Candidate gene studies are relatively cost and time effective.

Taxonomy: It is the branch of science known for naming, classifying and describing organisms and includes all plants, animals and organisms. Taxonomists identify and arrange different species into classifications. Different kinds of plants, animals and microorganisms are categorized as per their features and are called different 'specie'. In the past 250 years of research, taxonomists have named about 1.78 million species of animals, plants and micro-organisms, yet the total number of species is unknown and probably between 5 and 30 million.

Non-Coding Transcripts: A non-coding transcript is a RNA molecule that is not translated in a protein. It includes highly abundant and functionally important RNAs. They produce functional RNA rather than encoding proteins which are later used in the process of study of genetics. Non-coding RNAs belong to several groups and are involved in many cellular processes. They are used to regulate gene expression at the transcriptional and post-transcriptional level. There are three major categories of short non-coding RNAs, which are microRNAs (miRNAs), short interfering RNAs (siRNAs), and piwi-interacting RNAs (piRNAs).

Computational Tools for Bioinformatics

The brief description of some of the major bioinformatics tools are given as follows:

Basic Local Alignment Search Tool (BLAST)

It is a web based search tool. It is a heuristic modification of the Smith-Waterman algorithm. In this tool, statistical methods are used to determine the possibility of a particular alignment between sequence regions (DNA).Its popular interface is available at NCBI. It is an algorithm which is used for comparing the information related to primary biological sequences. It helps in finding the areas of similar features among the biological sequences. The comparison in the algorithm is regarding nucleotide or protein sequences to databases of sequences and results to statistical significance. It is used to identify the members of similar gene families and to find out functional & evolutionary relationships among the sequences. There are a number of BLASTs available according to different query sequences. The algorithm it uses is much faster than other approaches, so it used for a huge genome database.

EMBOSS Sequence Analysis Package

It is a freely available sequence analysis package developed for bioinformatics. It is a database that can be remotely accessible. It is a freely available Open Source software analysis package developed for helping researchers to solve problems of molecular biology. It contains more than 150 command-line tools which are used for analysis of DNA/protein sequences. It works on command-line search. It does not have comparable feature. Its applications are organized in groups according to their functions in a logical manner. It is extensively used in production environments instead of being a code only for research projects. It is a properly constructed toolkit used for creation of robust bioinformatics application. It is a comprehensive set of programs used for sequence analysis. It provides hundreds of applications covering various areas of molecular biology.

Generic Model Organism Database

This is used to develop a set of software for creating and administrating database related to organism. It is a toolkit that provides open-source software components for visualizing, managing, annotating, disseminating and storing biological data. This project was initiated in the early 2000s for the development of software tools for processing data from sequencing projects. It gives us availability of a complete set of software for creation and administration of a model organism database. This project for the development of database consists of various tools such as literature curation tools, a robust database schema, genome visualization and editing tools, biological ontology tools, and a set of standard operating procedures.

Gene Ontology

It is used to produce a vocabulary, which can be used to describe all aspects of a genetic study. It aims at the development of a common, precisely defined, structured, controlled vocabulary for describing the roles of genes and its products in all eukaryotes. The Molecular Biology Ontology Working Group is working actively to develop standards in this regard. Although nomenclature of genes itself aims to maintain vocabulary of gene and gene products, the Gene Ontology efforts by using markup language to make the data machine readable. This data consists of genes, their products and all their attributes. Gene Ontology provides a set of hierarchical controlled vocabulary which consists of 3 categories Biological process, Molecular function and Cellular component.

Gene X Gene Expression

It is software for processing and analysis of data. It offers advanced methods to provide real time data very easily to the users. It is a database system related to gene expression. It provides an integrated toolset which helps researchers and scientists to store, analyze and communicate their data. It is leading software for processing and analysis of qPCR-data. It provides user friendly interface and advanced method for analysis of real time PCR data & extraction of valuable information from the measurements.

Staden Package

It is a software program, which is used for DNA sequencing, editing and analysis. It is freely available. The Staden package provides us a fully developed set of DNA sequence assembly (Gap4 and Gap5), editing and analysis tools. It is written in C, C++, FORTRAN and TCL programming languages and is processes in UNIX, Linux, MacOs and Windows operating system. Its two main parts which are used for sequence program are PreGap and Gap. PreGap is used to process raw traces and masks all the sequences while Gap is the Genome assembly program which helps in assembling individual fragments into long contigs.

SRS Bio-database

It is a software tool used for bioinformatics and genomics for data integration, analysis and display. In recent times biological databases have developed in a large extends and became a very important part of almost every day toolbox used in biological researches. SRS (Sequence Retrieval System) has proved itself to be a valuable platform for storing, linking, and querying biological databases. It is an interface to more than 80 biological databases. It includes databases of sequences, application results (like BLAST) mappings, mutations, etc.

Table 1: SRS Databank

Library Group	Databank Name
Active Protein Sequence Databases	BCIPEP
	EPO_PRT
	IPIHISTORY
	KIPO_PRT
	NRPL1
CABRI – Bacteria	BCCM_LMG
	CABI_BACT
	CIP_BACT
	NCCB_BACT
CABRI - DNA Probes	ECACC_DNA_PROBES
EMBOSS Source Code	EDATAREL
	EFUNC

BCIPEP- It is a database of B cell epitopes; EPO_PRT- European bioinformatics institute's database; IPIHISTORY- International protein index; KIPO_PRT- Patent protein sequences; NRPLI- Non redundant patent sequences; BCCM_LMG- Bacteria collection; CAB_BACT- collection of biological products; CIP_BACT- bacteria database; NCCB_BACT- collection of bacteria; ECACC_DNA_PROBES-collection of DNA sequences; EDATAREL, EFUNC- data type and functions used in EMBOSS.

Repository available for Bioinformatics Data

The following data repositories are available which collect, synthesis biological data with the help of bioinformatics tools:

- ***GenBank:*** It is an open access systematic collection of all publically available sequences of nucleotides and their protein translations. It receives data sequences obtained from a large number of laboratories from all over the world for more than 1,000,000 organisms. It is growing exponentially doubling in every 18 months. It gains data from direct submission for individual laboratories as well as data submitted in bulk from large scale sequencing centers. This is a database which is produced and maintained by National Center for Biotechnology Information (NCBI). GenBank is now the most important and most used database for research in almost all biological fields. It contains more than 150 billion nucleotide bases for more than 162 million sequences. It accepts only original sequences of nucleotides.

- ***The Genome Online Database (GOLD):*** It is an online database to monitor data related to genome, meta-genome sequencing and their associated metadata. The GOLD database was developed in 1997, which was later released with different updated versions. Specialty of this database is it supports the minimum information standards metadata recommended by Genomic Standard Consortium. It enables to access information regarding complete and ongoing genomic projects. This database at present can make information available on 350 genome projects, of which 48 have been completely sequenced.

- ***The DNA Databank of Japan (DDBJ):*** It is a biological database which collects nucleotide sequence data and makes them freely available to enhance research activities in life science. It maintains and provides archival, retrieval and analytical resources for biological information. It comprises of public, open-access nucleotide sequence databases including raw sequence reads, assembly information and functional annotation. Its objective is to promote and support the sharing and use of biological data as a public resource.

- ***The European Molecular Biology Laboratory (EMBL):*** It is an intergovernmental organization which acts as a hub for bioinformatics services and develops & maintains a large number of scientific databases. It is basically a research institute for molecular biology which undergo innovations in life sciences research, transfer and technology development. It also provides trainings and services to the members of different scientific communities. The laboratory is operated from 5 sites: main in Heidelberg and outstations in Hinxton, Grenoble, Hamburg and Monterotondo.

- ***National Center for Biotechnology Information (NCBI):*** It gives advancement in the field of science and health by providing access to biomedical and genomic information. It is a database which provides access to genomic and biomedical information. It has a large number of research groups belonging to different disciplines comprising computer scientists, mathematicians, biochemists, research physicians, molecular biologists, and structural biologists concentrating on basic and applied research in computational molecular biology.

- ***Other of Bioinformatics Resources:*** There are a number of web lists of bioinformatics resources. Some of these are: Bioinformatics.net, Amazon's Alexa, Open Directory Project at Mozilla.org, The Bionetwork project at Pasteur Institute is an example of lists of resources to be searched for bioinformatics. BioHunt is a project which uses internet robot technology for searching and updating of molecular biological resources. Bioinformatics.net is an online biological resources specializing in bioinformatics tools. Bioinformatik.de provides a collection of bioinformatics and biological resources. SourceForge.net and Bioinformatics.org provides resources that help software developers and bioinformatics engineers as well as biologists.

Applications of Bioinformatics in various disciplines

Bioinformatics is growing rapidly in various fields of biotechnology. Some of its areas of application are Microbial genome application, Molecular medicine, Personal medicine, preventive medicine, Biotechnology, Climatic change studies, Crop improvement, alternative energy sources, Gene therapy, Antibiotic resistance, Waste cleanup, Forensic Analysis, Insect resistance, Drug development, Bio weapon creation, Improvement in nutritional quality, Evolutionary studies, Agriculture, Vetinary Science, Development of Drought resistant varieties, etc. In Medical field various techniques of Bioinformatics are used in medical issues of human being. It can be in the form of providing medicines after analysis of the genes of the individual. Research in medical applications focuses on developing efficient algorithms and techniques for automatic (or semi-automatic) analysis of biomedical images, medical evaluation, computer assisted diagnosis and surgery, and treatment planning. Techniques used in this interdisciplinary area often come from Computer Vision, Imaging Processing, and High Performance Computing.

Role of Bioinformatics in Agricultural Sciences

It is used in agriculture to increase the productivity of crop which is essential as per increasing population. It is used to enhance nutritional quality of the crop. At the same time is also used to develop such type of genes of crops which are resistant to different types of pesticides and adverse conditions. In Microbial Genome Application, tools of Bioinformatics are used for different genomic applications like forecasting the climatic variations taking place. It is used for waste cleanup by using bioinformatics bacteria, microbes etc. Comparative genetics of the plant genomes has shown that the organization 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. Progress has been made in developing cereal varieties that have a greater tolerance for soil alkalinity, free aluminum 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. 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 anemia 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.

Conclusion

Bioinformatics is used in various fields of science. Algorithms and tools are developed accordingly to provide us accurate results. These tools made it possible to predict the function of different genes & factors affecting them. The information obtained makes it easy to produce enhanced species of plants which are resistant to pesticides, herbicides and other different adverse conditions. In addition to this, by making some changes in their genome can make the specie disease resistant. This information with appropriate technology provides predictive measures of quality and health of plants and in future can become decision management system for plant breeding.

Recommendations

As the application of bioinformatics is increasing day by day, the future of bioinformatics is very bright. The tools are easy to implement, the results obtained from are accurate, at the same time are time & cost effective. As a result majority of the Indian Bioinformatics companies are planning to enhance its use.

References

[1] Achuthsankar; S Nair (2007). Computational Biology & Bioinformatics – A gentle Overview. *Communications of Computer Society of India*.

[2] Sali A; Blundell TL (1993). Comparative protein modelling by satisfaction of spatial restraints. *Journal Mol. Biology*, 234 (3): 779–815.

[3] Berman HM; Westbrook J; Feng Z; Gilliland G; Bhat TN; Weissig H; Shindyalov IN; Bourne PE (2000). The Protein Data Bank. *Nucleic Acids Res.* 28 (1):235-242.

[4] Altschul SF; Gish W; Miller W; Myers EW; Lipman DJ (1990). Basic local alignment search tool. *Journal Mol. Biology,* 215 (3):403-410.

[5] Ji-Hong Zhang; Ling-YunWu and Xiang-Sun Zhang (2003). Reconstruction of DNA sequencing by hybridization. *Bioinformatics*, Vol.19 (1), pp.14-21.

[6] Xiang-Sun Zhang; Yong Wang; Zhong-Wei Zhan; Ling-Yun Wu and Luonan Chen (2005). Exploring protein's optimal HP configurations by self-organizing mappings. *Journal of Bioinformatics and Computational Biology*, Vol.3(2), pp.385-400.

[7] Tianshou Zhou; Luonan Chen; Yun Tang; and Xiangsun Zhang (2005). Aligning multiple protein structures by deterministic annealing. *Journal of Bioinformatics and Computational Biology*, Vol.3(4), pp.837-860.

[8] Xiang-Sun Zhang; Rui-Sheng Wang; Ling-Yun Wu; Luo-Nan Chen (2006). Models and algorithms for the haplotyping problem. *Current Bioinformatics*, Vol.1(1), pp.105-114.

[9] Yong Wang; Trupti Joshi; Xiang-Sun Zhang; Dong Xu; Luonan Chen (2006). Inferring gene regulatory networks from multiple microarray datasets. *Bioinformatics*, Vol.22(19), pp.2413-2420.

[10] http://bioinfo.mbb.yale.edu/mbb452a/intro/intro.pdf

[11] http://www.bioinfopublication.org/files/articles/3_2_5_IJBR.pd
[12] http://Bioinformatics/WHAT%20IS%20BIOINFORMATICS_%20_%20SCQ.html
[13] https://www.sciencedaily.com/terms/genbank.htm
[14] http://www.ddbj.nig.ac.jp/intro-e.html
[15] https://www.ncbi.nlm.nih.gov/
[16] http://link.springer.com/chapter/10.1007%2F978-0-387-92738-1_8
[17] http://www.omicsonline.org
[18] www.ijbs.com/v03p0420.htm
[19] http://en.wikipedia.org/wiki/
[20] www.webcitation.org/getfile?fileid=d8f2d009329eb7070621236302c3cfaba93ee2
[21] www.biomcc.com/genex-software.html
[22] https://www.its.hku.hk/news/ccnews101/ccnews-SRS.htm
[23] https://blast.ncbi.nlm.nih.gov/Blast.cgi
[24] https://en.wikipedia.org/wiki
[25] http://emboss.sourceforge.net/apps/
[26] http://manuals.bioinformatics.ucr.edu/home/emboss
[27] http://www.cs.tau.ac.il/~rshamir/algmb/98/scribe/pdf/lec04.pdf
[28] https://www.ebi.ac.uk/seqdb/confluence/display/SRS/SRS/
[29] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3037419/
[30] http://www.uniprot.org/help/gene_ontology
[31] http://www.webcitation.org/getfile?fileid=d8f2e2d009329eb7070621236302c3cfaba93ee2
[32] https://bioinf.comav.upv.es/courses/intro_bioinf/_downloads/staden_course.pdf
[33] http://www.biomcc.com/genex-software.html
[34] http://www.uniprot.org/help/proteome
[35] http://www.genomenewsnetwork.org/resources/whats_a_genome/Chp2_2.shtml