INTERNATIONAL JOURNAL OF RESEARCHES IN BIOSCIENCES, AGRICULTURE & TECHNOLOGY
© VISHWASHANTI MULTIPURPOSE SOCIETY (Global Peace Multipurpose Society) R. No. MH-659/13(N)
www.vmsindia.org

Bioinformatics: an overview

Jayshree Tirpude,
Asst. Professor, Sevadal Mahila Mahavidyalaya, Nagpur
Email: jayashreetirpude@gmail.com

Abstract:
The amount of biological information increased rapidly in accordance with the completion of several genome projects. So the huge demand for analysis and interpretation of these data is being managed by the evolving science of bioinformatics. Several sequence-based methods of analyzing individual genes or proteins have been elaborated and expanded, and methods have been developed for analyzing large numbers of genes or proteins simultaneously, such as in the identification of a drug target of any disease genes and its homologous proteins. Application of techniques of molecular modeling has a long history in the prediction of 3D structure of a protein and is now established as a cornerstone of modern structural biology. Then to predict the predominant binding modes of a drug with the target protein, principles of molecular docking is used and finally techniques of MD simulation help us to validate protein drug complex. All these techniques all-together leads us to understand the molecular mechanism of any putative protein and conduct a privilege towards structure based drug designing.

Keywords: Bioinformatics, genomics, proteomics, drug designing, HGP

Introduction

Bioinformatics has emerged as an important discipline within the biological sciences that allows scientists to decipher and manage the vast quantities of data (such as genome sequences) that are now available with the help of Information Technology [1]. Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline [2]. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics with which to assess relationships among members of large data sets; the analysis and interpretation of various types of data including nucleotide and amino acid sequences, protein domains, and protein structures; and the development and implementation of tools that enable efficient access and management of different types of information [2].

Bioinformatics has provided a strong tool for advancement of research and development in the field of biotechnology. Different research work in the field of computational Biology and Bioinformatics brings analytical results in recognition, identification and structural relationship at genome level in different organisms. Development of new tools, various algorithms and the databases helps to maintain and analyze the medical records. The i*n silico* drug development methods is very much useful to prepare a candidate drug for incurable diseases.

Aims of Bioinformatics

The aims of bioinformatics are threefold [3]. First, at its simplest bioinformatics organizes data in a way that allows researchers to access existing information and to submit new entries as they are produced, e.g. the Protein Data Bank for 3D macromolecular structures [4,5]. While data-curation is an essential task, the information stored in these databases is essentially useless until analyzed. Thus the purpose of bioinformatics extends much further. The second aim is to develop tools and resources that aid in the analysis of data. For example, having sequenced a particular protein, it is of interest to compare it with previously characterized sequences. This needs more than just a simple text-based search and programs such as FASTA [6] and PSI-BLAST [7] must consider what comprises a biologically significant match. Development of such resources dictates expertise in computational theory as well as a thorough understanding of biology. The third aim is to use these tools to analyse the data and interpret the results in a biologically meaningful manner. Traditionally, biological studies examined individual systems in detail, and frequently compared them with a few that are related. In bioinformatics, we can now conduct global analyses of all the available data with the aim of uncovering common principles that apply across many systems and highlight novel features.

Different Areas of Bioinformatics [8]

- **Genomics**

Genome is defined as the total genetic information present in an organism. So Genomics is the study of an organism's entire genome that means the study of all the genes of a cell or tissue, at the DNA (genotype), mRNA (transcriptome), or protein (proteome) levels.

- **Proteomics**
  Proteomics is the study of proteins - their location, structure and function. It is the identification, characterization and quantification of all proteins involved in a particular pathway, organelle, cell, tissue, organ or organism that can be studied in concert to provide accurate and comprehensive data about that system [9].
- **Sequence analysis**
  Sequence analysis is the application of Information Technologies to Molecular Biology. It deals with biological sequences, and processes them to extract significant information that may yield new insights and guidelines in the understanding of biological organisms.
- **Comparative Genomics**
  Comparative genomics is the study of the relationship of genome structure and function across different biological species or strains. Comparative genomics is an attempt to take advantage of the information provided by the signatures of selection to understand the function and evolutionary processes that act on genomes. While it is still a young field, it holds great promise to yield insights into many aspects of the evolution of modern species. Gene finding is an important application of comparative genomics, as is discovery of new, non-coding functional elements of the genome.
- **Protein Structure Prediction**
  Protein structure prediction is one of the most important goals pursued by bioinformatics and theoretical chemistry. Its aim is the prediction of the three-dimensional structure of proteins from their amino acid sequences, sometimes including additional relevant information such as the structures of related proteins. In other words, it deals with the prediction of a protein's tertiary structure from its primary structure. Protein structure prediction is of high importance in medicine (for example, in drug design) and biotechnology (for example, in the design of novel enzymes).
- **Drug Designing**
  Drug design is the approach of finding drugs by design, based on their biological targets. Typically a drug target is a key molecule involved in a particular metabolic or signaling pathway that is specific to a disease condition or pathology, or to the infectivity or survival of a microbial pathogen.

Public Domain Resources / Databases in Biology

National Center for Biotechnology Information (NCBI)
http://www.ncbi.nlm.nih.gov/

The National Center for Biotechnology Information (NCBI) is part of the United States National Library of Medicine (NLM), a branch of the National Institutes of Health. The NCBI houses a series of databases relevant to biotechnology and biomedicine. Major databases include GenBank for DNA sequences and PubMed, a bibliographic database for the biomedical literature. All these databases are available online through the Entrez search engine. NCBI is directed by David Lipman, one of the original authors of the BLAST sequence alignment program and a widely respected figure in Bioinformatics.

European Bioinformatics Institute (EBI)
http://www.ebi.ac.uk/

The European Bioinformatics Institute (EBI) is a centre for research and services in bioinformatics, and is part of European Molecular Biology Laboratory (EMBL). EMBL-EBI provides freely available data from life science experiments covering the full spectrum of molecular biology. The roots of the EMBL-EBI lie in the EMBL Nucleotide Sequence Data Library (now known as EMBL-Bank), which was established in 1980 at the EMBL laboratories in Heidelberg, Germany and was the world's first nucleotide sequence database.

Biological Databases:

A database is a collection of information stored in a computer in a systematic way. In a broad sense, a database is a collection of simple facts, or data. Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high throughput experiment technology, and computational analyses. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

The history of sequence databases began in the early 1960s, when Margaret Dayhoff and colleagues at the Protein Information Resource (PIR) collected all the protein sequences known at that time. Her group published this collection as a printed work called the “Atlas of Protein sequence and Structure” in 1978.

Types of Biological Databases:

Types of Databases	Information they contain
Bibliographic databases	Literature
Taxonomic databases	Classification
Nucleic acid databases	DNA information
Genomic databases	Gene level information
Protein Databases	Protein Information
Protein families, domains and functional sites	Classification of proteins and identifying domains
Enzymes/ metabolic pathways	Metabolic pathways

There are many different types of database but for routine sequence analysis, the following are initially the most important

- Primary databases (Contains sequence data such as nucleic acid or protein)
  - ∗ Nucleic Acid Databases: EMBL, Genbank, DDBJ
  - ∗ Protein Databases: SWISS-PROT, TrEMBL, PIR
- Secondary databases (Contains results from the analysis of the sequences in the primary databases)
  e.g: PROSITE, Pfam, BLOCKS, PRINTS etc.

1. **EMBL (http://www.ebi.ac.uk/)**
   The European Molecular Biology Laboratory (EMBL) is a molecular biology research institution supported by 20 European countries & Australia as associate member state and is maintained by EBI (European Bioinformatics Institute). The EMBL Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource. Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications.
2. **GenBank**
   GenBank is the NIH (National Institute of Health) genetic sequence database, an annotated collection of all publicly available DNA sequences. There are approximately 85,759,586,764 bases in 82,853,685 sequence records in the traditional GenBank divisions and 108,635,736,141 bases in 27,439,206 sequence records in the WGS division as of February 2008. It is maintained by **National Center for Biotechnology Information (**http://www.ncbi.nlm.nih.gov/). NCBI is established in 1988 as a national resource for molecular biology information, NCBI creates public databases, conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease
3. **DDBJ (http://www.ddbj.nig.ac.jp/)**
   The **DNA Data Bank of Japan** is a DNA data bank. It is located at the National Institute of Genetics of Japan. It is also a member of the International Nucleotide Sequence Database Collaboration or **INSDC**. DDBJ is the sole DNA data bank in Japan, which is officially certified to collect DNA sequences from researchers and to issue the internationally recognized accession number to data submitters.
   INSDC (http://www.insdc.org/)
   The International Nucleotide Sequence Database Collaboration consists of a joint effort to collect and disseminate databases containing DNA and RNA sequences. It involves the following computerized databases: DNA Data Bank of Japan (Japan), GenBank (USA) and the EMBL (European Molecular Biology Laboratory, Germany).
   These three databases have collaborated since 1982. Each database collects and processes new sequence data and relevant biological information from scientists in their region e.g. EMBL collects from Europe, GenBank from the USA. These databases automatically update each other with the new sequences collected from each region, every 24 hours. The result is that they contain exactly the same information, except for any sequences that have been added in the last 24 hours.
4. **Swiss-Prot (http://www.expasy.ch/sprot/)**
   Swiss-Prot is a manually curated biological database of protein sequences. Swiss-Prot was created in 1986. It is maintained by the Swiss Institute of Bioinformatics and the European Bioinformatics Institute. Swiss-Prot strives to provide reliable protein sequences associated with a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases.
5. **TrEMBL**
   It is a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot. UniProtKB/TrEMBL contains the translations of all coding sequences (CDS) present in the EMBL/GenBank/DDBJ Nucleotide Sequence

Databases and also protein sequences extracted from the literature or submitted to UniProtKB /Swiss-Prot. The database is enriched with automated classification and annotation.

6. PIR (http://pir.georgetown.edu/)

The Protein Information Resource (PIR), located at Georgetown University Medical Center (GUMC), is an integrated public bioinformatics resource to support genomic and proteomic research, and scientific studies. PIR was established in 1984 by the National Biomedical Research Foundation (NBRF) as a resource to assist researchers in the identification and interpretation of protein sequence information. Prior to that, the NBRF compiled the first comprehensive collection of macromolecular sequences in the Atlas of Protein Sequence and Structure, published from 1965-1978 under the editorship of Margaret Dayhoff.

7. PROSITE (http://www.expasy.org/prosite/)

PROSITE is a database of protein families and domains. It consists of entries describing the domains, families and functional sites as well as amino acid patterns, signatures, and profiles in them. These are manually curated by a team of the Swiss Institute of Bioinformatics and tightly integrated into Swiss-Prot protein annotation. PROSITE was created in 1988 by Amos Bairoch. It is part of the ExPASy (Expert Protein Analysis System) proteomics analysis servers.

Conclusion:

In last few decades, bioinformatics has become an integral part of research and development in the biomedical sciences. With the current deluge of data, computational methods have become indispensable to biological investigations. Originally developed for the analysis of biological sequences, bioin-formatics now encompasses a wide range of subject areas including struc-tural biology, genomics and gene ex-pression studies. Bioinformatics has not only provided greater depth to biological investigations, but added the dimension of breadth as well. In this way, we are able to examine individual systems in detail and also compare them with those that are related in order to uncover common principles that apply across many systems and highlight unusual features that are unique to some.

References:

1. http://www.nepjol.info/index.php/AV/article/viewFile/8294/6762. Last retrieved on September 12, 2015
2. http://www.ncbi.nlm.nih.gov/Class/MLACourse/Modules/MolBioReview/bioinformatics.html Last retrieved on September 12, 2015
3. N.M. Luscombe, D. Greenbaum, M. Gerstein. What is bioinformatics? An introduction and overview. Yearbook of Medical Informatics 2001; 83-100.
4. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF, Jr., Brice MD, Rodgers JR, et al. The Protein Data Bank. A computer-based archival file for macromolecular structures. Eur J Biochem 1977;80(2):319-24.
5. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28(1):235-42.
6. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 1988;85(8):2444-2448.
7. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation ofproteindatabasesearchprograms.Nucleic Acids Res. 1997;25(17):3389-3402.
8. https://en.wikipedia.org/wiki/Bioinformatics Last retrieved on September 12, 2015
9. https://en.wikipedia.org/wiki/Proteomics Last retrieved on September 12, 2015
10. https://en.wikipedia.org/wiki/Human_Genome_Project Last retrieved on September 13, 2015

❖❖❖❖