BIOINFORMATICS ON POST GENOMIC ERA: FROM GENOMES TO SYSTEMS BIOLOGY

Vihinen, Mauno
Institute of Medical Technology, University of Tampere, Finland and Research Unit, Tampere University Hospital, Tampere, Finland

Keywords: Genome, transcriptome, proteome, data mining, sequence analysis, bioinformatics, functional genomics, systems biology.

Contents

Summary

Genome projects are changing the emphasis of the life sciences. A major part of the human genome containing about 3 billion base pairs in 24 chromosomes was recently published. Genome studies and other laboratory methods are producing mountains of data. Bioinformatics is a relatively new multidisciplinary field that combines methods from biology and biomedicine with computer science, statistics and mathematics. Bioinformatics is discipline to organize, analyze, store, retrieve, share and distribute biological, genomic, proteomic, clinical and biomedical information. In bioinformatics, scientists aim at revealing the mechanisms behind biological phenomena.

This review provides an introduction to the methods and problems tackled by bioinformatics. In addition to identifying all the genes in genomes it is crucial to store and distribute the information in databases. Annotation and identification of genes from genomes are crucial for generating useful genome databases. Ethical, legal, and social implications of genome and gene data are briefly discussed. Currently, data mining is one of the most common bioinformatics tasks. Nucleotide and amino acid sequences contain all the information for genes to function and for proteins to fold and deliver their functions. Sequence analysis methods can derive plenty of information when combined with powerful database search methods.

1. Introduction

The publication of numerous genome sequences, including that for human, has propelled the biosciences into the post genomic era. The human genome contains about 3 billion base pairs organized into 24 chromosomes. The actual number of genes will still remain open for a long time, but current estimates are around 25 000. In addition to identifying all the genes in genomes, it is crucial to store and distribute the information in databases. Genome projects also aim at identifying the genes and proteins critical for life functions, regulation and networks. To make sense of the mountains of biological data, often generated with high-throughput method of –omics approaches, advanced computer based methods are required. Bioinformatics is the field that handles the data. This review provides an introduction to the methods and problems that can be solved with current state-of-the-art bioinformatics, in order to unravel biological information, especially in relation to genome data.

1.1. Implications of Genomes

Genomic data, for the first time, allows systematic analysis and modification of several cellular processes. The availability of microbial genomes facilitates e.g., the identification of new drug targets, the search for new energy sources, monitoring to detect pollutants in environment, and possibly efficient toxic waste cleanup. Medicine will benefit in many ways. Improved and early diagnosis, as well as detection of genetic predispositions to diseases is already available for many disease conditions. Rational drug design based on target information has revolutionized the pharmaceutical industry. In the future gene therapy will be a real therapeutic option for certain disorders.

Bioarchaeology and evolutionary studies of human origin and migration will reveal the history of different population groups. DNA-based forensics will be extensively utilized. Insect, drought and disease resistant crops can be developed, as well as farm animals that are healthier and more productive.

Genome studies also have broad ethical, legal, and social implications (ELSI). These issues are of concern in policy making. Important issues include e.g., (fetal) genetic testing, the effects of genetic counseling and medical practices, the use of genetic information and privacy implications, as well as commercialization and intellectual property protection, including patenting. Safety and environmental issues are of concern for genetically modified foods, feed and microbes.

2. Bioinformatics and Internet

Bioinformatics **(***see also– Bioinformatics***)** is a relatively new multidisciplinary field that combines methods from biology and biomedicine with computer science, statistics and mathematics. Bioinformatics stores, analyses, processes and manipulates diverse information relevant to biological systems. With bioinformatics, scientists aim at revealing the mechanisms lying behind biological phenomena by modeling the dependencies and faint inclinations in massive gene and protein information. These studies generally involve such large data sets that only computer-based methods can be used. Bioinformatics is closely connected to experimental research, usually being an integral part of it.

Many bioinformatics services, programs and databases are available in the Internet. The majority of basic bioinformatics analyses can generally be performed over the Internet, as fast as or even faster than in local computers. For the distribution of the biological information, the Internet is the primary means. All the most important registries are freely available. There are already hundreds of biological and biomedical databases available on the Web. Internet databases provide up to date information, providing a great many benefits over many locally maintained services. Recently commercial databases and services with restricted access and/or high price have also emerged. In addition to the most important databases, the majority of the most frequently used software tools can either be run on the Internet or downloaded for free.

3. Genomes

Genomes contain all the information necessary for an organism to live. The draft sequence is largely useless without further bioinformatic analyses. Functional genomics aims at the identification of the locations of genes and their functions. Some 326 complete genomes have been fully sequenced, and approximately 1000 prokaryotic and about 610 eukaryotic genome projects are ongoing as of July 2006. A list of the central genome databases is given in Table 1. Genome sequences facilitate for the first time the gaining of a complete picture of the processes occurring within an organism. In genome research, computers are needed at all stages. Genomes are sequenced by using the so-called shotgun sequencing approach, which generates a large number of randomly selected sequence stretches, which are computationally combined to contigs, and eventually to complete genomes. In fact, the computations last much longer than the actual data collection in the big sequencing centers. By analyzing the genomes, it is possible to address several questions. Instead of comparing organisms based on single sequences, it can now be made by comparing complete reaction pathways and metabolic networks.

GOLD Genomes OnLine Database	http://www.genomesonline.org/
KEGG	http://www.genome.ad.jp/kegg/

Sanger Center	http://www.sanger.ac.uk/
MIPS	http://mips.gsf.de/
GOBASE	http://www.bch.umontreal.ca/gobase/
Ensembl	http://www.ensembl.org/
Entrez Gene	http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene

Table 1. Gene and Genome Information in the Internet.

The genome sequences have to be further analyzed, in the first place to identify the locations of genes. For example, human genes comprise less than 5 percent of the genome. Further, eukaryotic genes are so-called mosaic genes composed of exons and intervening introns. Some genes can be fragmented to hundreds of exons and introns, which makes the identification of coding genes far from trivial, especially because the raw sequencing data can contain errors.

In the analysis and annotation of the genomes, functional reconstruction is used for the conceptual assembly of metabolic pathways, transport units and signal transduction pathways. Special tools are available for the characterization of gene structures and functions. These methods rely on extensive database analysis and comparison to other genes and genomes. Naturally, the functions have to be verified by experimentalists. However, functional annotation of individual genes is still largely incomplete.

-
-
-

Bibliography

Blundell, T. L., Sibanda, B. L., Montalvao, R. W., Brewerton, S., Chelliah, V., Worth, C. L., Harmer N. J., Davies, O., Burke, D. (2006) *Philos Trans R Soc Lond B Biol Sci* **29**, 413:423.

Fox J. A.,. Butland, S. L., McMillan, S., Campbell, G, Ouellette, B. F. F.(2005) The Bioinformatics Links Directory: a Compilation of Molecular Biology Web Servers. Nucl. Acids Res. **33**: W3-W24; doi:10.1093/nar/gki594

Galperin M. Y. (2006). The Molecular Biology Database Collection: 2006 update. *Nucleic Acids Research* **34**: D3-D5; doi:10.1093/nar/gkj162

Biographical Sketch

Mauno Vihinen received his PhD in Biochemistry from the University of Turku, Finland (1990), followed by appointments at Turku Centre for Biotechnology (1990-1993), and Karolinska Institute, Stockholm (1993-1995). From 1995-1997, Vihinen was at University of Helsinki, Finland (Associate Professor). Since 1998, he has been Professor of Bioinformatics at Institute of Medical Technology, University of Tampere, Finland. The scientific interests of his research group are studying protein structure-function relationships, especially in relation to human disease, and the analysis and development of software tools for both gene and protein expression studies, microarrays and proteomics, respectively. The group has applied bioinformatics methods, especially in the research of primary immunodeficiencies and certain cancers. In addition to applied bioinformatics, he has also developed programs on diverse areas of bioinformatics, and has contributed to the development and maintenance of numerous databases (mutation databases and knowledge bases for diseases). Gene and expression data analysis and method development for the data analysis in these fields are among major interests of his research group as well as systems biology, especially in analysis of immunological processes and immunodeficiencies.