MBINDSC-1.2 Introduction to Bioinformatics and Computational Biology

BLOCK MB 1.2 A

UNIT 1

INTRODUCTION, HISTORY AND APPLICATIONS OF BIOINFORMATICS; DATABASES – INTRODUCTION, TYPES, APPLICATIONS AND LIMITATIONS; LITERATURE SEARCH DATABASES - PUBMED, MEDLINE.

STRUCTURE

1.1. OBJECTIVES

After reading this unit we will be able to understand:

- Introduction to concepts of bioinformatics
- Aims and scope of bioinformatics
- Applications and limitations of bioinformatics
- Biological data and databases, types of databases
- PubMed and Medline

1.2. INTRODUCTION

The term bioinformatics was coined by Paulien Hogeweg in 1171 for the study of informatics processes in biotic system. Bioinformatics is the application of information technology to mine, visualize, analyze, integrate, and manage biological and genetic information. Such analyzed information has several applications. In simple terms bioinformatics can is application of tools of computation and analysis to the capture and interpretation of biological data.

Bioinformatics has become more important due to various developments in science and technology research. Some of the important contributors to this are exponential expansion of biological information, expansion of multiple types of information, cheaper high throughput technologies, improvement in computation power, lack of standards/quality, need for micro and macro analysis and need for better algorithms.

Bioinformatics data is enormously used in DNA analysis for genome sequencing, sequence assembly, sequence/gene annotations, genefinding/sequence translation tools, sequence similarity searching, comparison between genomes, evolution of sequences (Phylogenetic analysis) and gene expression (Fig. 1). In protein analysis it is used for predicting structure, X-ray crystallography, homology based models, drug designing, sequence similarity, protein family assignments, conserved motifs, proteomics data analysis and protein evolution. Other than these

bioinformatics plays an important role in drug designing, vaccine development, dairy technology, forensics, crop improvement, designing enzymes for detergents, genetic counselling and many others.

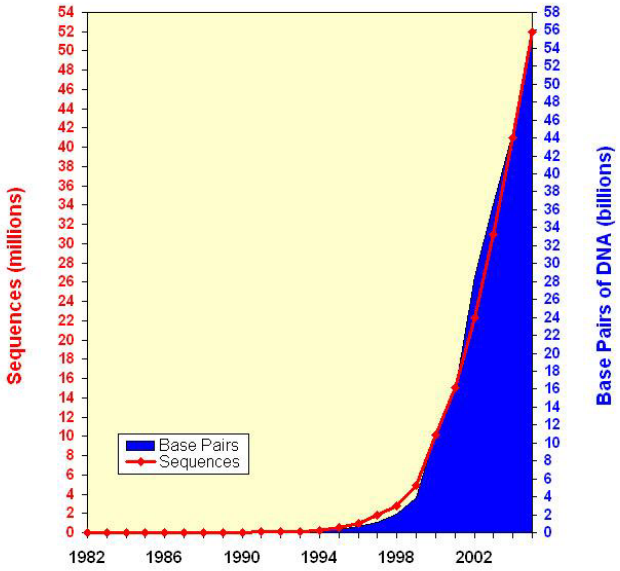


Figure 1: Growth of GenBank due to bioinformatic tools.

1.3. AIMS AND SCOPE OF BIOINFORMATICS

The main aim of bioinformatics is to gain better understanding of the functions of the cell at the molecular level. By analyzing raw molecular sequence and structural data, bioinformatics research gives new insights and provide a overall perspective of the cell. The reason that the functions of a cell can be better understood by analyzing sequence data is ultimately because the flow of genetic information is dictated by the "central dogma" of biology in which DNA is transcribed to RNA, which is translated to proteins. Cellular functions are mainly performed by proteins whose capabilities are ultimately determined by their sequences. Therefore, solving functional problems using sequence and sometimes structural approaches is beneficial.

Bioinformatics consists of two subfields: the development of computational tools and databases and the application of these tools and databases in generating biological knowledge for greater understanding of the living systems. The tool

development includes writing software for sequence, structural, and functional analysis, as well as the construction and maintenance of biological databases. These tools are used in three areas of genomic and molecular biological research: molecular sequence analysis, molecular structural analysis, and molecular functional analysis. The analyses of biological data often generate new problems and challenges that in turn lead to the development of new and better computational tools. The areas of sequence analysis include sequence alignment, sequence database searching, motif and pattern discovery, gene and promoter finding, reconstruction of evolutionary relationships, and genome assembly and comparison. Structural analyses include protein and nucleic acid structure analysis, comparison, classification, and prediction. The functional analyses include gene expression profiling, protein–protein interaction prediction, protein subcellular localization prediction, metabolic pathway reconstruction, and simulation.

The three aspects of bioinformatics analysis are not isolated but often interact to produce integrated results. For example, protein structure prediction depends on sequence alignment data; clustering of gene expression profiles requires the use of phylogenetic tree construction methods derived in sequence analysis. Sequence-based promoter prediction is related to functional analysis of foundation of all bioinformatics analysis. Sequence-based promoter prediction is related to functional analysis of coexpressed genes. Gene annotation involves a number of activities, which include distinction between coding and noncoding sequences, identification of translated protein sequences, and determination of the gene's evolutionary relationship with other known genes; prediction of its cellular functions employs tools from all three groups of the analyses.

1.4. KEY REQUIREMENTS FOR BIOINFORMATICS ANALYSIS

Key requirements for bioinformatics analysis include

•Knowledge of data/input sources

•Knowledge of methods and their assumptions

•Plan to get from point a to point b

•Knowledge of equipment

•Knowledge of potential sources of error

•Interpretation of results

•Reproducibility and reliability of results

1.5. APPLICATIONS OF BIOINFORMATICS

Bioinformatics has not only become essential for basic genomic and molecular biology research, but is having a major impact on many areas of biotechnology and biomedical sciences. It has applications, for example, in knowledge-based drug design, forensic DNA analysis, and agricultural biotechnology. Computational studies of protein–ligand interactions provide a rational basis for the rapid identification of novel leads for synthetic drugs. Knowledge of the three-dimensional structures of proteins allows molecules to be designed that are capable of binding to the receptor site of a target protein with great affinity and specificity. This informatics-based approach significantly reduces the time and cost necessary to develop drugs with higher potency, fewer side effects, and less toxicity than using the traditional trial-and-error approach. In forensics, results from molecular phylogenetic analysis have been accepted as evidence in criminal courts. Some sophisticated statistics and likelihood-based methods for analysis of DNA have been applied in the analysis of forensic identity. It is worth mentioning that genomics and bioinformtics are now poised to revolutionize our healthcare system by developing personalized and customized medicine. The high speed genomic sequencing coupled with sophisticated informatics technology will allow a doctor in a clinic to quickly sequence a patient's genome and easily detect potential harmful mutations and to engage in early diagnosis and effective treatment of diseases. Bioinformatics tools are being used in agriculture as well. Plant genome databases and gene expression profile analyses have played an important role in the development of new crop varieties that have higher productivity and more resistance to disease.

1.6. LIMITATIONS OF BIOINFORMATICS

Bioinformatics has a number of inherent limitations. Bioinformatics predictions are not formal proofs of any concepts. They do not replace the traditional experimental research methods of actually testing hypotheses. In addition, the quality of bioinformatics predictions depends on the quality of data and the sophistication of the algorithms being used. Sequence data from high throughput analysis often contain errors. If the sequences are wrong or annotations incorrect, the results from the downstream analysis are misleading as well. That is why it is so important to maintain a realistic perspective of the role of bioinformatics.

Most algorithms make incorrect predictions that make no sense when placed in a biological context. Errors in sequence alignment, for example, can affect the outcome of structural or phylogenetic analysis. The outcome of computation also depends on the computing power available. Many accurate but exhaustive algorithms cannot be used because of the slow rate of computation. Instead, less accurate but faster algorithms have to be used. This leads to a compromise between accuracy and computational feasibility. Therefore, it is important to keep in mind the potential for errors produced by bioinformatics programs. Caution should always be exercised when interpreting prediction results. It is necessary to use multiple programs and perform multiple evaluations. A more accurate prediction can often be obtained if one draws a consensus by comparing results from different algorithms.

1.7. DATABASES IN BIOLOGY

A database is a collection of information that is organized so that it can easily be accessed, managed, and updated. A database is basically a collection of information organized in such a way that a computer program can quickly select desired pieces of data. A database as an electronic filing system.

What is biological data?

Biological data comes in many different forms and formats like texts, sequences, structures, links, numbers, images and biological matter.

1. Text: Biological data available in the form of textual information and related references. E.g., PubMed and OMIM.
2. Sequence: Biological data available in the form of databases containing DNA and protein sequences. E.g., GenBank and UniProt.
3. Protein structure: Biological data available in the form of structures (protein). E.g., PDB, SCOP and CATH.
4. Links: These kinds of databases consist of a collection of links from protein domains and families to other databases providing related resources. E.g., InterPro database.
5. Images: Biological data available in the form of 2D gel and microscopic images E.g. Reference gel images.
6. Numerical data: Gene expression data as well as other microarray data are also accessible from a number of databases. E.g., ArrayExpress of the European Bioinformatics Institute, EBI.
7. Biological matter: Data available in the form of frozen bacterial strains, vectors etc. Along with collecting information on each of these specific biological matters, e.g. the UniVec database (of NCBI).

1.7.1. Biological databases

Due to the latest technological development large scale research projects are being undertaken worldwide and therefore both technical and scientific knowledge is received and accumulated rapidly. As the amount of data gained from research projects all over the world is heavily increasing there is a growing need for databases storing and handling the biological data. Databases and the ability to organize data are needed in order to keep research efficient and to get optimal output and information from data obtained in the lab.

Databases related to different data types and subjects as e.g. nucleotides, proteins, genomes and taxonomy are available (Fig. 2). Within each field several databases present their content in different ways, different file formats and with different purposes. Some databases are large and contain global data collections maintained and kept up to date by the responsible organization, while others are

small and local and maybe only maintained for a limited period of time while a specific project is going on.

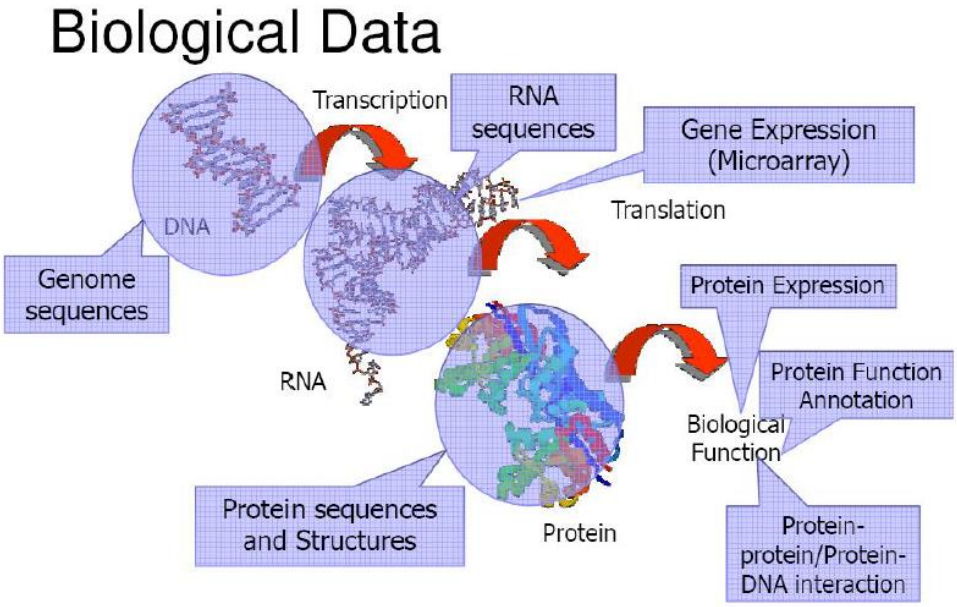


Figure 2: Different types of biological databases.

1.7.2. Necessity of databases

The biological data bases are required for a number of reasons like:

•For storing and communicating large datasets.
•For dissemination biological information.
•For providing organized data for analysis and retrieval.
•For making biological data available in computer-readable form.

1.7.3. Types of databases

Databases are of different types and they are classified based on their types, derivation, content, availability and technical design.

Databases based on their contents:

Based on their contents, biological databases can be roughly divided into three categories: primary databases, secondary databases, and specialized databases.

Primary databases contain original biological data. They are archives of rawsequence or structural data submitted by the scientific community. GenBank and Protein Data Bank (PDB) are examples of primary databases.

Secondary databases contain computationally processed or manually curated information, based on original information from primary databases. Translated protein sequence databases containing functional annotation belong to this category. Examples are SWISS-Prot and Protein Information Resources (PIR) (successor of Margaret Dayhoff's Atlas of Protein
Sequence and Structure.

Specialized databases are those that cater to a particular research interest. For example, Flybase, HIV sequence database, and Ribosomal Database Project are databases that specialize in a particular organism or a particular type of data.

Based on type of data they are classified as:

–Nucleotide sequences
–Protein sequences
–Proteins sequence patterns or motifs
–Macromolecular 3D structure
–Gene expression data
–Metabolic pathways
–Proteomics data

Based on Technical design they are classified as:

–Flat-files

–Relational database (SQL)

–Exchange/publication technologies (HTML, CORBA, XML)
•Each one of the above are inter convertible

Based on Availability of data they are classified as:

–Publicly available, no restrictions

–Available, but with copyright

–Accessible, but not downloadable

–Academic, but not freely available

–Proprietary, commercial; possibly free for academics

1.7.4. Access to databases

There are more than 1500 databases and 50-100 databases are added every year. These databases can be accessed by several methods like

•Search Engines

•Journals related to Bioinformatics

•Websites like:

–http://www.biophys.uni-duesseldorf.de/BioNet/Pedro/rt_all.html

–www.expasy.ch

–Several others websites

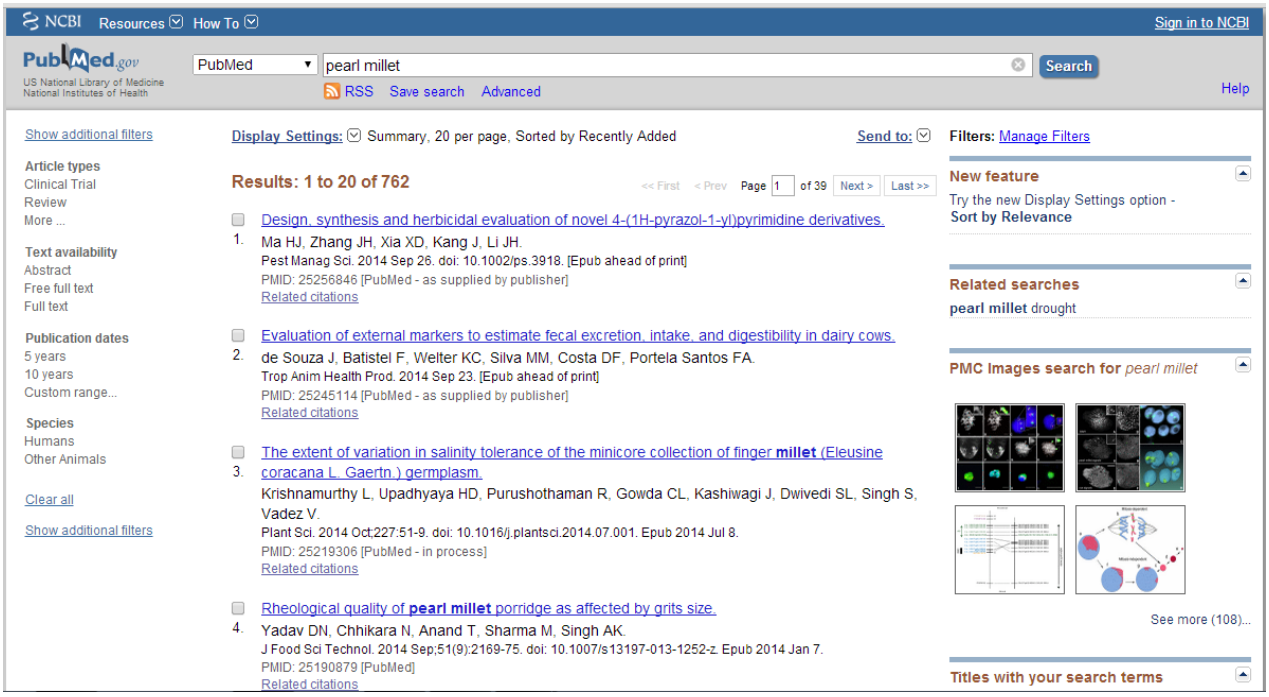
1.8. LITERATURE SEARCH DATABASES

Literature search refers to the process in which people use tools to search for literature relevant to their individual needs. typical user information needs include, but are not limited to, finding the bibliographic information about a specific article, or searching for publications pertinent to a specific topic (e.g. a disease). With the ease of Internet access, the amount of biomedical literature in electronic format is on the rise. the size of the bibliome has grown exponentially over the past few years.

1.8.1. PUBMED

PubMed is a free Web literature search service developed and maintained by the National Center for Biotechnology Information (NCBI).

PubMed is a literature search engine providing access to all published medically related articles, abstracts and journals. This database is a biomedical literature database which contains abstracts and in some cases the full text articles fromnearly 5,000 journals.Animportant feature ofPubMed is the retrieval of information based on medical subject headings (MeSH) terms. TheMeSH system consists of a collection of more than 20,000 controlled and standardized vocabulary termsused for indexing articles. In other words, it is a thesaurus that helps convert search keywords into standardized terms to describe a concept. By doing so, it allows "smart" searches in which a group of accepted synonyms are employed so that the user not only gets exact matches, but also related matches on the same topic that otherwise might have been missed. Another way to broaden the retrieval is by using the "Related Articles"option. PubMed uses a word weight algorithm to identify related articles with similar words in the titles, abstracts, and MeSH. By using this feature, articles on the same topic that were missed in the original search can be retrieved.



PubMed gives you biological data in text format and this service provided by the U.S. National Library of Medicine links to more than 23 million resources from different journals within the field of life science. A relatively new functionality at the NCBI website is the possibility to sign up for an account at My NCBI which is a service offering a customized and automated PubMed update. After registration at My NCBI you can save your searches and set up automated searches alerting you by e-mail. You can also customize e.g. filtering options on the searches.
PubMed can be accessed at http://www.ncbi.nlm.nih.gov/pubmed/.

1.8.2. MEDLINE

MEDLINE is the U.S. National Library of Medicine® (NLM) premier bibliographic database that contains over 21 million references to journal articles in life sciences with a concentration on biomedicine. A distinctive feature of MEDLINE is that the records are indexed with NLM Medical Subject Headings (MeSH®). MEDLINE is the online counterpart to MEDLARS® (MEDical Literature Analysis and Retrieval System) that originated in 1164.

MEDLINE is the primary component of PubMed®, part of the Entrez series of databases provided by the NLM National Center for Biotechnology Information (NCBI).

Source: Currently, citations from over 5,600 worldwide journals in about 40 languages; about 60 languages for older journals. Citations for MEDLINE are created by the NLM, international partners, and collaborating organizations.

Broad subject coverage: The subject scope of MEDLINE is biomedicine and health, broadly defined to encompass those areas of the life sciences, behavioral sciences, chemical sciences, and bioengineering needed by health professionals and others engaged in basic research and clinical care, public health, health policy development, or related educational activities. MEDLINE also covers life sciences vital to biomedical practitioners, researchers, and educators, including aspects of biology, environmental science, marine biology, plant and animal science as well as biophysics and chemistry. Increased coverage of life sciences began in 2000.

The majority of the publications covered in MEDLINE are scholarly journals; a small number of newspapers, magazines, and newsletters considered useful to particular segments of the NLM broad user community are also included. For citations published in 2010 or later, over 40% are for cited articles published in the U.S., about 13% are published in English, and about 84% have English abstracts written by authors of the articles.

Availability: MEDLINE is the primary component of PubMed (http://pubmed.gov); a link to PubMed is found on the NLM homepage (http://www.nlm.nih.gov). The result of a MEDLINE/PubMed search is a list of citations (including authors, title, source, and often an abstract) to journal articles and an indication of free electronic full-text availability. Searching is free of charge and does not require registration.

A growing number of MEDLINE citations contain a link to the free full text of the article archived in PubMed Central® or to other sites.

1.1. SUMMARY

Bioinformatics is the application of information technology to mine, visualize, analyze, integrate, and manage biological and genetic information. Such analyzed information has several applications. In simple terms bioinformatics can is application of tools of computation and analysis to the capture and interpretation of biological data.

Bioinformatics has not only become essential for basic genomic and molecular biology research, but is having a major impact on many areas of biotechnology and biomedical sciences. It has applications, for example, in knowledge-based drug design, forensic DNA analysis, and agricultural biotechnology. In forensics, results from molecular phylogenetic analysis have been accepted as evidence in criminal courts. Plant genome databases and gene expression profile analyses have played an important role in the development of new crop varieties that have higher productivity and more resistance to disease.

Databases in Biology: A database is a collection of information that is organized so that it can easily be accessed, managed, and updated. A database is basically a collection of information organized in such a way that a computer program can quickly select desired pieces of data. A database as an electronic filing system.

Types of databases: Databases are of different types and they are classified based on their types, derivation, content, availability and technical design.

Databases based on their contents: Based on their contents, biological databases can be roughly divided into three categories: primary databases, secondary databases, and specialized databases.

Primary databases contain original biological data. They are archives of rawsequence or structural data submitted by the scientific community. GenBank and Protein Data Bank (PDB) are examples of primary databases.

Secondary databases contain computationally processed or manually curated information, based on original information from primary databases. Translated protein sequence databases containing functional annotation belong to this category. Examples are SWISS-Prot and Protein Information Resources (PIR) (successor of Margaret Dayhoff's Atlas of Protein
Sequence and Structure.

Specialized databases are those that cater to a particular research interest. For example, Flybase, HIV sequence database, and Ribosomal Database Project are databases that specialize in a particular organism or a particular type of data.

PUBMED: PubMed is a free Web literature search service developed and maintained by the National Center for Biotechnology Information (NCBI). PubMed is a literature search engine providing access to all published medically related articles, abstracts and journals. This database is a biomedical literature database which contains abstracts and in some cases the full text articles fromnearly 5,000 journals.

MEDLINE: MEDLINE is the U.S. National Library of Medicine® (NLM) premier bibliographic database that contains over 21 million references to journal articles in life sciences with a concentration on biomedicine. A distinctive feature of

MEDLINE is that the records are indexed with NLM Medical Subject Headings (MeSH®). MEDLINE is the primary component of PubMed®, part of the Entrez series of databases provided by the NLM National Center for Biotechnology Information (NCBI).

1.10. CHECK YOUR PROGRESS

1. Define bioinformatics and list its applications.
2. What are the aims and scope of bioinformatics?
3. What are the limitations of bioinformatics?
4. Give an account of different forms of databases in biology.
5. How are databases classified based on their contents?
6. How databases are classified based on type of data?
7. Briefly describe PubMed.
8. Add a note on Medline.

1.11. KEY WORDS

Bioinformatics, applications and limitations, biological databases, PubMed, Medline.

1.12. FURTHER SUGGESTED READING

1. Apweiler, R. 2000. Protein sequence databases. Adv. Protein Chem. 54:31–71.

2. Attwood, T. K., and Miller, C. J. 2002. Progress in bioinformatics and the importance of being earnest. Biotechnol. Annu. Rev. 8:1–54.

3. Goodman, N. 2002. Biological data becomes computer literature: New advances in bioinformatics. Curr. Opin. Biotechnol. 13:68–71.

4. Geer, R. C., and Sayers, E.W. 2003. Entrez:Making use of its power. Brief. Bioinform. 4:171–84.
5. Hagen. J. B. 2000. The origin of bioinformatics. Nat. Rev. Genetics 1:231–6.

6. Hughes, A. E. 2001. Sequence databases and the Internet.MethodsMol. Biol. 167:215–23.

7. Luscombe, N. M., Greenbaum, D., and Gerstein, M. 2001. What is bioinformatics? A proposed definition and overview of the field.Methods Inf.Med. 40:346–58.

8. Ouzounis,C. A., andValencia, A. 2003. Early bioinformatics: The birth of a discipline –Apersonal view. Bioinformatics 11:2176–10.

9. Stein, L. D. 2003. Integrating biological databases. Nat. Rev. Genet. 4:337–45.

1.13. SOURCES

1. Altschul, S. F., Gish, W., Miller, W.E. , Myers, W., and Lipman, D. J. 1110. Basic local alignment search tool. J. Mol. Biol, 215:403–410.
2. Arthur M. Lesk. 2002. Introduction to Bioinformatics. Oxford University Press, Great Clarendon Street, Oxford OX2 6DP.
3. Changmin Liao. 2013. Introduction on Several Popular Nucleic Acids Databases. Journal of Computer Engineering and Informatics 1:82-87.
4. Jin Xiong. 2006. Essential Bioinformatics Cambridge University Press, The Edinburgh Building, Cambridge cb2 2ru, UK. www.cambridge.org.
5. Heitor Silvério Lopes and Leonardo Magalhães Cruz (Eds). 2011. Computational Biology and Applied Bioinformatics, Published by InTech Janeza Trdine 1, 51000 Rijeka, Croatia.
6. Lewitter, F. 1118. Text-based database searching. Trends Guide to Bioinformatics, pages 3–5.
7. Pagni M. and Jongeneel CV. 2001. Making sense of score statistics for sequence alignments. Briefings in Bioinformatics, 2:51–67.
8. Pearson, R. W. and Lipman, D. J. 1118. Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. USA, 85:2444–2448.
9. Smith, T. F. and Waterman, M. S. 1181. Identification of common molecular subsequences. J. Mol. Biol., 147:115–117.