Bioinformatics - a brief introduction

Lin GAN
Cologne University Bioinformatics Center, Zuelpicher Strasse 47, 50674 Koeln, Germany

ABSTRACT

Biology and computation didn't have so much common language until about nineties of last century. But with the emerging of many new techniques in molecular biology and the enormous volume of data they produces became challenges in computing also. The new term and research area 'Bioinformatics' was introduced with this background. In this article, a brief introduction and current state of bioinformatics is presented. Then a short description of commonly used databases and tools is given. Finally the usage of bioinfomatics in drug design is discussed to illustrate an example of practical application.

Keywords: Bioinformatics, computational molecular biology, drug design.

1. WHAT IS BIOINFIRMATCS

Bioinformatics -- The science of information and information flow in biological systems, esp. of the use of computational methods in genetics and genomics[1].

Modern molecular biology produced and keeps producing data in a phenomenal rate[2]. For example, GenBank grows at an exponential rate, doubling every 10 months[3]. As a result of this surge in data, computers have become indispensable to biological research. Bioinformatics was born on such demand. As a newly emerging interdisciplinary research area, Bioinformatics is not so well-defined, we can just say that it deals with the computational management of all kinds of biological information, whether it may be about genes and their products, whole organisms or even ecological systems. Most of the bioinformatics work that is being done can be described as analyzing biological data, although a growing number of projects deal with the organization of biological information.

The field of bioinformatics seeks to organize, analyse and distribute biological information in specially designed databases powered by smart mathematical algorithms and the latest database techniques. Firstly Bioinformatics organize data in a way that allows researchers worldwide to access existing information and to submit new ones easily. Discrete data are not enough for research work nowadays, so another task of Bioinformatics is to develop tools with which researchers can analyze and compare data. Those programs should not be only text-based search and compare programs, but also should consider biological significant match of the sequence. Examples are FASTA[4] and BLAST[5] . Another aim of Bioinformatics is to use these tools to understand the data in a biologically meaningful manner.

Though being a relative new field, bioinformatics blooms very fast. Volume of already existed biological databases are growing while hundreds, thousands of new databases are coming to life. Varies of sequence analysis tools give us broad choices according to researchers' individual demands. And more specific tools and algorithms are developed for special research purposes.

2. BIOLOGICAL DATABASES

In recent years, many new databases storing biological information have appeared. But this has not only positive effects: nowadays many scientists complain that it gets increasingly difficult to find useful information. This may largely be due to the fact that the information gets more and more scattered over an increasing number of heterogeneous resources. In order to solve this problem, some bioinformatists work hard to organize old databases in a more efficient way, to reduce redundant data and to develop better way in data-mining. Some others, on the other way round retrieve data in other perspect to construct new view of data to meet the need of a group of researchers. As the result of their efforts, you can find all kind of databases (most of them are accessible through internet now), table 1 illustrate some of the commonly used databases.

Table 1. Some of the commonly used biological databases

Database Name	**Description**
EMBL[6] http://www.ebi.ac.uk/embl/	The EMBL Nucleotide Sequence Database is a comprehensive database of DNA and RNA sequences collected from the scientific literature and patent applications and directly submitted from researchers and sequencing groups. Data collection is done in collaboration with GenBank (USA) and the DNA Database of Japan (DDBJ).
SWISS-PROT[7] http://www.ebi.ac.uk/swissprot/	The SWISS-PROT Protein Sequence Database is a database of protein sequences produced collaboratively by Amos Bairoch (University of Geneva) and the EMBL Data Library. The data in Swiss-Prot are derived from translations of DNA sequences from the EMBL Nucleotide Sequence Database, adapted from the Protein Identification Resource (PIR) collection, extracted from the literature and directly submitted by researchers. It contains high-quality annotation, is non-redundant, and cross-referenced to several other databases, notably the EMBL nucleotide sequence database, PROSITE pattern database and PDB.
PROSITE[8] http://www.expasy.org/prosite/	The PROSITE dictionary of sites and patterns in proteins prepared by Amos Bairoch at the University of Geneva.
EC-Enzyme[9] http://www.biochem.ucl.ac.uk/bsm/dbbrowser/protocol/ecenzfrm.html	The 'ENZYME' data bank contains the following data for each type of characterized enzyme for which an EC number has been provided: EC number, Recommended name, Alternative names, Catalytic activity, Cofactors, Pointers to the SWISS-PROT entrie(s) that correspond to the enzyme, Pointers to disease(s) associated with a deficiency of the enzyme.
PIR[10] http://pir.georgetown.edu/	The Protein Identification Resource consists of an integrated computer system composed of a number of protein and nucleic acid sequence databases and software designed for the identification and analysis of protein sequences and their corresponding coding sequences. The PIR serves the scientific community through on-line access, distributing magnetic tapes, and performing off-line sequence identification services for researchers.
NCBI/GenBank[11] http://www.ncbi.nlm.nih.gov/Genbank/index.html	GenBank is the NIH (National Institute of Health) genetic sequence database, a collection of all known DNA sequences.
OMIM [12] http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM	The Mendelian Inheritance in Man data bank (MIM) is prepared by Victor Mc Kusick with the assistance of Claire A. Francomano and Stylianos E. Antonarakis at John Hopkins University.
MEDLINE http://www.ncbi.nlm.nih.gov/PubMed/	MEDLINE is NLM's (National Library of Medicine) premier bibliographic database covering the fields of medicine, nursing, dentistry, veterinary medicine, and the preclinical sciences. Journal articles are indexed for MEDLINE, and their citations are searchable, using NLM's controlled vocabulary, MeSH (Medical Subject Headings). MEDLINE contains all citations published in Index Medicus, and corresponds in part to the International Nursing Index and the Index to Dental Literature.
PDB[13] http://www.rcsb.org/pdb/	The X-ray crystallography Protein Data Bank (PDB) is compiled at the Brookhaven National Laboratory.
GDB[14] http://gdbwww.gdb.org/	The GDB Human Genome Data Base supports biomedical research, clinical medicine, and professional and scientific education by providing for the storage and dissemination of data about genes and other DNA markers, map location, genetic disease and locus information, and bibliographic information.
MGD: The Mouse Genome Database[15] http://www.informatics.jax.org/mgihome/MGD/aboutMGD.shtml	MGD is a comprehensive database of genetic information on the laboratory mouse. This initial release contains the following kinds of information: Loci (over 15,000 current and withdrawn symbols), Homologies (1300 mouse loci, 3500 loci from 40 mammalian species), Probes and Clones (about 10,000), PCR primers (currently 500 primer pairs), Bibliography (over 18,000 references), Experimental data (from 2400 published articles).
ACeDB (A Caenorhabditis elegans DataBase)[16] http://www.acedb.org/	Containing data from the Caenorhabditis Genetics Center (funded by the NIH National Center for Research Resources), the C. elegans genome project (funded by the MRC and NIH), and the worm community. ACeDB is also the name of the generic genome database software in use by an increasing number of genome projects. The software, as well as the C. elegans data, can be obtained via ftp. ACeDB databases are available for the following species: C. elegans, Human Chromosome 21, Human Chromosome X, Drosophila melanogaster, mycobacteria, Arabidopsis, soybeans, rice, maize, grains, forest trees, Solanaceae, Aspergillus nidulans, Bos taurus, Gossypium hirsutum, Neurospora crassa, Saccharomyces cerevisiae, Schizosaccharomyces pombe, and Sorghum bicolor.

3. COMPUTATIONAL TOOLS

Computational tools are needed to collect and analyze data in the most efficient manner. Tools like BLAST[5], FASTA[4], BLAT[17] are developed to compare sequences. You can either compare two sequences to see the similarity of them, or compare your entry sequence with the sequences in database to retrieve the sequences with high homology to the entry sequence. Tools like Phred[18,19], Phrap[20] and Consed[21] are grown up with the genome projects. These tools work with shotgun sequencing data. They read out the data automatically, evaluate the data and try to construct contigs with the help of overlapping. With program like Genscan[22], we can predict genes out of a genome sequence. Many bioinformaticists are also working on the prediction of the biological functions of genes and proteins (or parts of them) based on structural data.

4. BIOINFORMATICS AND DRUG DESIGN

Bioinformatics technology is used to solve complex biological questions related to metabolic pathways, genes, protein function and pharmacological/developmental aspects of drugs and medicines. We all know that drug design is a very time and money consuming procedure. Companies invest millions of money and decades of time to develop a new drug. Bioinformatics helps to accelerate this process and make the drug more efficient and specific at the same time. It has significant advantages over traditionally expensive and time consuming "wet lab" research methods, because computational tools give the most predictive and accurate information about genes and proteins with regards to mediating aspects of drug action.

To enhance the design and development of new drugs, the field has diversified into different sections as follows:

- Proteomics: A field involved in studying different proteomes and their protein expression in cells, which assists in identifying disease mechanisms for therapeutic drug targets.
- Pharmacogenomics: A study of individual’s genetic inheritance which is responsible for his/her reaction to different drugs. So as to prescribe the effective and least toxic drugs and decrease overall medical costs.
- Computer based drug design: Design of drug molecules is done on the basis of the structure of drug receptors, structure activity relationship, toxicity assessment, physical/chemical properties of drugs. The formulation aspects of polymorphism and compatibility with different formulation additives are computed on specially designed software.

Together these fields aid in the development of new drug molecules, which reduces cost of conducting clinical trials, which often fail due to variations in animal or human models.

5. CONCLUSION

Bioinformatics is a young and prosperous research area. It encompasses a wide range of subject areas including structural biology, genomics and gene expression studies etc. Here I can only give a very brief introduction, and describe some popular databases and tools used in current study. A short discussion of bioinformatics used in drug design is also give though many important and useful areas are still left uncovered.

REFERENCES

1. Definition of Oxford English dictionary.
2. T. Reichhardt, ″It’s sink or swim as a tidal wave of data approaches,″ *Nature* **399**, 517-520 (1999).
3. NCBI handbook, 2002 Oct.
4. W.R. Pearson, D.J. Lipman, ″Improved tools for biological sequence comparison,″ *PNAS* **85**, 2444-2448 (1988).
5. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, ″Basic local alignment search tool,″ *J. Mol. Biol.* **215**, 403-410 (1990).
6. G. Stoesser, W. Baker, A. van den Broek, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, F. Nardone, P. Stoehr, M. A. Tuli, K. Tzouvara, R. Vaughan, ″The EMBL Nucleotide Sequence Database: major new developments,″ *Nucleic Acids Res.* **31**, 17-22 (2003).
7. A. Bairoch, B. Boeckmann B, ″The SWISS-PROT protein sequence data bank, recent developments,″ *Nucleic Acids Res.* **21**, 3093-3096 (1993).
8. A. Bairoch, ″The PROSITE dictionary of sites and patterns in proteins, its current status,″ *Nucleic Acids Res.* **21**, 3097-3103 (1993).
9. A. Bairoch, ″The ENZYME data bank,″ *Nucleic Acids Res.* **21**, 3155-3156 (1993).
10. K.E. Sidman, D.G. George, W.C Barker, L.T. Hunt, ″The protein identification resource (PIR),″ *Nucleic Acids Res.* **16**, 1869-1871 (1988).

11. D. Benson, D.J.. Lipman, J. Ostell, ″Genbank,″ *Nucleic Acids Res.* **21**, 2963-2965 (1993).
12. V.A. McKusick, *Catalogs of autosomal dominant, autosomal recessive, and X-linked phenotypes*, Tenth Edition, Johns Hopkins University Press, Baltimore (1991).
13. F.E. Abola, F.C. Bernstein, T.F. Koetzle, In: A.M. Lesk Ed. *Computational molecular biology. Sources and methods for sequence analysis*, Oxford University Press, Oxford, (1988), pp. 69-81.
14. A.J. Cuticchia, K.H. Fasman, D.T. Kingsbury, R.J. Robbins, P.L. Pearson, ″The GDB(TM) Human Genome Data Base Anno,″ *Nucleic Acids Res.* **21**, 3003-3006 (1993).
15. J.A. Blake, J.E. Richardson, C.J. Bult, J.A. Kadin, J.T. Eppig, ″Mouse Genome Database Group MGD: the Mouse Genome Database,″ *Nucleic Acids Res.* **31**, 193-195 (2003).
16. S. Kelley, ″Getting started with Acedb,″ *Brief Bioinform.* **1**, 131-137 (2000).
17. W.J. Kent, ″BLAT--the BLAST-like alignment tool,″ *Genome Res.* **12**, 656-664 (2002).
18. B. Ewing, L. Hillier, M.C. Wendl, P. Green, ″Base-calling of automated sequencer traces using phred. I. Accuracy assessment,″ *Genome Res.* **8**, 175-185 (1998).
19. B. Ewing, P. Green, ″Base-calling of automated sequencer traces using phred. II. Error probabilities,″ *Genome Res.* **8**, 186-194 (1998).
20. D. Gordon, C. Desmarais, P. Green, ″Automated finishing with autofinish,″ *Genome Res.* **11**, 614-625 (2001).
21. D. Gordon, C. Abajian, P. Green, ″Consed: a graphical tool for sequence finishing,″ *Genome Res.* **8**, 195-202 (1998).
22. C. Burge, S. Karlin, ″Prediction of complete gene structures in human genomic DNA,″ *J. Mol. Biol.* **268**, 78-94 (1997).

NEW AUSTRALIAN RESEARCH CENTRE PURSUES PHOTONICS

Sydney, Australia, July 21, 2003. A new collaboration between Australia's leading optical communications experts will develop photonic chips. CUDOS - the ARC Centre of Excellence for Ultrahigh bandwidth Devices for Optical Systems -- has been established with A$11.5 million of funding from the Australian Research Council (ARC) to undertake fundamental research aimed at underpinning the development of photonic chips.

The Centre is a collaboration between the University of Sydney, the Australian National University (ANU), the University of Technology Sydney (UTS), Macquarie University, Swinburne University of Technology and CSIRO.

Over A$40 million of cash and in kind support will be provided by the ARC, the State Government of New South Wales and the collaborators. Around 70 researchers will be involved in the exciting research programs of the Centre.

ARC Federation Fellow Professor Ben Eggleton has returned from a directorial position with Lucent Technologies, Bell Laboratories to head the Centre at the University of Sydney.

"This is a tremendously exciting opportunity," Professor Eggleton said. "We have a strong and exciting vision, and to accomplish it we are combining six strong research groups with tremendous experimental infrastructure and well-established theoretical programs, and building brand new capabilities in micro-photonic device development."

"We also have established formal collaborative links with researchers in the US, Japan and Europe."

The Centre will demonstrate all-optical processing applications and devices for ultra-high bandwidth optical systems. These will derive from fundamental research in the most exciting and vibrant areas of photonics science-non-linear optical materials leading to nonlinear optical devices, photonic crystals, micro-structured optical fibres and micro-photonics.

For more information, contact Professor Ben Eggleton +612 9351 3604, mailto:egg@physics.usyd.edu.au, or visit www.cudos.org.au .

Laser Focus World