Bioinformatics –Emerging Area in Medical and Biomedical Informatics

Ashwini Dinodia
Department of Pharmaceutical Education and Research, Bhagat Phool Singh Mahila Vishwavidyalaya, Khanpur Kalan Sonipat-131305 (HR)

Email: ashwinidinodia@gmail.com

ABSTRACT
Bioinformatics is one of the new and emerging scientific fields which deals in the biological data and established itself an important component in the field of science. Bioinformatics is comprised of a large no of software's, tools, statistics which helps us to identify the unknown sequence of the gene, DNA, protein or any biomolecule. Bioinformatics is used to find the 3D structure of protein, DNA or any other unknown substance which may or may not have the biological activity. It reduces the cost as well as time to perform the experiment and used in the drug design of the new molecule and to find out the therapeutic activity of molecule.

Key words: *Bioinformatics, Protein, DNA, Therapeutic activity, Biological activity*

INTRODUCTION

Bioinformatics is a combination of biology and information technology and deals in different computational tools and method for the manage analyze and manipulation of large amount of biological data. It involves three components which are very important and are as follows:

1. Creation of databases for storage and management of big amount of data.
2. Development of algorithms and statics for the determination of relationships among the members of biological data.
3. These tools are used to analyze and interpret the biological data like DNA, RNA, proteins sequences, gene expression, protein structure and the biological pathway(1,2).

The modern molecular biology produces large amount of data and keeps producing data at phenomenal rate(3). Presently the volume of biological data is increasing at a very fast rate and in the case of gene bank this data is growing at an exponential rate nearly doubling in every 10 months. In this era it is quite difficult to understand the interaction between the biological function of the genes and their interactions among them but by the help of bioinformatics and its tools it is quiet relatively easy to understand their nature and their interactions. These databases analyze and help to understand the complex interactions at molecular level in a cell. The latest Nucleic Acid Research Database issue counts about 1000 different molecular biology database(4,5). This field also helps in the finding the sequence of genes in genome and biological information of organism and ecological systems. Bioinformatics organize, analyze and arrange the large amount of biological data in a particular manner acc to the biological database which runs on specific algorithms and latest database technique. The aim of bioinformatics is to organize large amount of data in a particular fashion so that researchers can easily access these data easily and can compare with other entities so that they can arrive at a single conclusion. It should not only search but also find a perfect match for a particular sequence for biological molecule. Its aim is to ensure that there is a proper understanding of biological data in a meaningful manner e.g. FASTA(6) and BLAST (Basic Local Alignment Search Tool)(7). For this purpose Bio DWH is introduced as Jawa based open source toolkit for building life science data ware house by the help of common relation database management system e.g. MySQL, Oracle and Postgre SQL. By the help of the object relation mapping (ORM) technology it had replaced the local database management systems. Bio DWH provides parsers so as to extract the data from public life science data so that it can be stored in data warehouse(8).

BIOLOGICAL DATABASES

There are different type and large no of data bases are used in bioinformatics in order to find the sequence, structure of the gene or amino acid. Some of the biological databases are these can be accessed from the internet at any time.

EMBL(9) http://www.ebi.ac.uk/embl/	The EMBL Nucleotide sequence database for RNA and DNA sequence in which data is collected from scientific literature and patent applications. This is done in collaboration with Gen Bank(USA) and DNA Database of Japan)(DDBJ)
SWISS-PORT(10) http://www.ebi.ac.uk/swissprot/	The SWISS-PORT protein sequence database is a database of protein which is produced collaboratively by Amos Bairoch(University of Geneva) and EMBL Data Library. The data is derived from translations of DNA sequence which is adapted from Protein Identification Resource(PIR) collection and extracted from the literature and are aldo directly submitted by researchers. It contains high quality annotation, non-redundant and cross referenced to several databases notably the EMBL nucleotide sequence database, PROSITE pattern database and PDB.
PROSITE(11) http://www.expasy.org/prosite/	It is dictionary of sites and patters of proteins which is prepared by Amos Bairoch at the university of Geneva.
EC-Enzyme(12) http://www.biochem.ucl.ac.uk/bsm/dbbrowser/protocol/ecenzfrm.html	The ENZYME data bank contains for each type of characterized enzyme for which EC no is there 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(13) http://pir.georgetown.edu/	It is integrated computer system and composed of proteins and amino acid sequence databases and it is designed for the identification and analysis of protein sequence and corresponding coding sequence. It is accessible via online, distributing magnetic tapes and performing off line sequence identification services for researchers.
NCBI/GenBank(14) http://www.ncbi.nlm.nih.gov/Genbank/index.html	Gen Bank is the NIH (National Institute of Health) genetic sequence database, a collection of all known DNA sequences.
OMIM(15) http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM	The Mendelian Inheritance in Man data bank (MIM) which 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 which includes fields of medicine, dentistery, preclinical science, veterinary science journal articles indexed for MEDLINE and their citations by using NLM controlled vocabulary, MeSH (Medical Subject Headings). It contains all citations published in Index Medicus, and corresponds in part to the International Nursing Index and the Index to Dental Literature.
PDP(16) http://www.rcsb.org/pdb/	The X-ray crystallography Protein Data Bank (PDB) is compiled at the Brookhaven National Laboratory.
GDB(17) http://gdbwww.gdb.org/	The GDB Human Genome Data Base supports clinical medicine, biomedical research, and professional and scientific education and provides the storage and dissemination of data for genes and other DNA markers, genetic disease, map location, bibliographic information and locus information.
MGD: The Mouse Genome Database(18) http://www.informatics.jax.org/mgihome/MGD/aboutMGD.shtml	MGD is a comprehensive database of genetic information on the laboratory mouse. It contains information like 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 Data Base(19) http://www.acedb.org/	It contains 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. It can also be obtained via ftp ACeDB and is available for species like C.elegans, human chromosome 21, human chromosome X, mycobacteria, soyabean, rice, maize, grain, forest trees, Gossypium hirsutum, Saccharomyces cerevisiae, Neurospora crassa.

GENE IDENTIFICATION AND SEQUENCE ANALYSIS

Sequence analysis is defined as the study of the different characteristics of biomolecules like nucleic acid or protein which give its unique function. First the molecule to be studied is retrieved from public database. After refinement if required these are subjected to different tools which helps in prediction of their characteristics which is related to their function, structure, history and its homologues. The tool which we use depends upon nature of analysis eg in case of Entrez of PubMed (20) which enables us to search and retrieve data from data domain. Similarly, pattern discovery tools such as Expression Profiler (21), Gene Quiz (22) allows researchers to search out different patterns in the given data other tools like BLAST (Basic Local Alignment Search Tool), Clustal W helps in studying the evolutionary history or origin of the gene or the protein sequence by comparison (7, 23,24). Data visualization tools such as Jal view (25), Gene View (26), Tree View (27), and Genes-Graphs (28) help the researcher to view the data in as graphical representation. These tools generally use advanced mathematical modeling and statistical interference like Hidden Markov Mosel (HMM), regression analysis, Artificial Neural Network (ANN), clustering, dynamic programming, and sequence mining to analyze the sequence of interest. These tools are very simple and can generate a large amount of quantified data which of use that why they are more preferred in biological systems in case of gene or proteins. These analyses are useful in identification of promoter, unrelated translations, terminator regions which is involved in expression regulation, recognition of particular sequence of peptide, intron, exons or Open Reading Frame (ORF) identification of certain coding regions to be used as markers for diagnostic purpose and hence sequential analysis is used in bioinformatics. Stoilov et al. used sequence analysis coupled with homology modeling to investigate the genetic basis of primary congenital glaucoma (PCG) (29). It was also found that a genome-wide sequence analysis (GWSA) of *Mycobacterium tuberculosis* H37Rv found that most of the bacteria protein was the result of exon–shuffling events or repetitive gene duplication (30).

Tools Used in Primary Sequence Analysis

Tool	Description
BLAST(31)	Used for identifying the DNA or protein sequence
HMMER(32)	Identifies the homologous protein sequence from database
Clustal Omega(33)	Multiple sequence alignments can be performed
Sequerome (34)	Used in sequence profiling
Protparam (35)	Predicts the physico-chemical properties of proteins.
JIGSAW(36)	Identification genes and predictation the splicing sites in the selected DNA sequences.
novoSNP(37)	To find the single nucleotide variation in DNA sequence.
Virtual footprint(38)	Prokaryotic genome can be studied with promoter regions with different regulator patterns
WebGeSter (39)	Database contains sequences of transcription terminator sequences and predicts the termination site of genes during transcription
Genscan(40)	Predicts the exon-intron sites in genomic sequences

CONCLUSION AND FUTURE OF BIOINFORMATICS

It is a new field and had progressed at a very fast rate in last decades. It is made possible only due to use of the software's and tools used in bioinformatics and had cut down the cost of the experiment and also saves the time to arrive on a single conclusion. It is now still progressing and everyday new tools and software are added in this field. This field is now used in the pharmacophore, designing of new drug

molecule and studying its physicochemical properties along with its therapeutic effect. It is playing a big role in personalized medicines and helps in flourishing of human kind.

REFERENCES

1. H. C. van Kampen and A. J. G. Horrevoets. The Role of Bioinformatics in Genomic Medicine. Cardiovascular Research: New Technologies, Methods, and Applications, edited by Gerard Pasterkamp and Dominique P. V. de Kleijn. Springer, New York, 2005.
2. N. M. Luscombe, D. Greenbaum, M. Gerstein. Method of Archieve 2001:40:(4): 346-358.
3. T. Reichhardt, It's sink or swim as a tidal wave of data approaches, *Nature* (1999) 399:517-520.
4. M. Y. Galperin. The molecular biology database collection: 2008 update.2008 Nucleic Acids Research, 36(Database issue): D2–D4.
5. NCBI handbook, Oct 2002.
6. W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison, PNAS:1988 85:2444-2448.
7. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, *J. Mol. Biol.* (1990):215,403-410.
8. T. Etzold, A. Ulyanov, P. Argos. SRS: Information retrieval system for molecular biology data banks. Methods in Enzymology, 1996:266:114–128.
9. 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.* 2003:31,:17-22.
10. A. Bairoch, B. Boeckmann B, The SWISS-PROT protein sequence data bank, recent developments, *Nucleic Acids Res.*,(1993):21:3093-3096.
11. A. Bairoch, The PROSITE dictionary of sites and patterns in proteins, its current status, *Nucleic Acids Res.* (1993):21:3097-3103.
12. A. Bairoch, The ENZYME data bank, *Nucleic Acids Res*. (1993):21:3155-3156.
13. K.E. Sidman, D.G. George, W.C Barker, L.T. Hunt, The protein identification resource (PIR),*Nucleic Acids Res.* (1988):16:1869-1871.
14. D. Benson, D.J. Lipman, J. Ostell, Genbank, *Nucleic Acids Res.* 21,(1993):2963-2965.
15. V.A. McKusick, *Catalogs of autosomal dominant, autosomal recessive, and X-linked phenotypes*, Tenth Edition, Johns Hopkins University Press, Baltimore (1991).
16. 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.
17. 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.*(1993):21:3003-3006.
18. 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.* (2003):31:193-195.
19. S. Kelley, Getting started with Acedb, *Brief Bioinform.*(2000):1:131-137.
20. Geer RC, Sayers EW Entrez: making use of its power. Brief Bioinform (2003):4:179-184.
21. Parmigiani G, Garrett ES, Irizarry RA, Zeger SL The analysis of gene expression data: an overview of methods and software, Springer, New York. (2003):1-45.
22. Hoersch S, Leroy C, Brown NP, Andrade MA, Sander C The Gene Quiz web server: protein functional analysis through the Web Trends Biochem Sci (2000):25:33-35.
23. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res:(1997):25:3389-3402.
24. Thompson JD, Higgins DG, Gibson TJ CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res: (1994):22:4673-4680.
25. Clamp M, Cuff J, Searle SM, Barton GJ, The Jalview Java alignment editor. Bioinformatics (2004):20:426-427.
26. Thomas P, Starlinger J, Vowinkel A, Arzt S, Leser U: Gene View: a comprehensive semantic search engine for PubMed. Nucleic Acids Res:(2012):40:W585-591.

27. Page RD (2001) Tree View. Glasgow University, Glasgow, UK.
28. Zhang Y, Phillips CA, Rogers GL, Baker EJ, Chesler EJ, et al.On finding bicliques in bipartite graphs: a novel algorithm and its application to the integration of diverse biological data types. BMC Bioinformatics (2014):15:110.
29. Stoilov I, Akarsu AN, Alozie I, Child A, Barsoum-Homsy M, et al. Sequence analysis and homology modeling suggest that primary congenital glaucoma on 2p21 results from mutations disrupting either the hinge region or the conserved core structures of cytochrome P4501B1. Am J Hum Genet:(1998):62:573-584.
30. Tekaia F, Gordon SV, Garnier T, Brosch R, Barrell BG, et al. Analysis of the proteome of Mycobacterium tuberculosis in silico. Tuber Lung Dis (1999):79:329-342.
31. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: Basic local alignment search tool. J Mol Biol (1990):215: 403-410.
32. Finn RD, Clements J, Eddy SR HMMER web server: interactive sequence similarity searching. Nucleic Acids Res (2011):39:W29-37.
33. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, et al. (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol 7:539.
34. Ganesan N, Bennett NF, Velauthapillai M, Pattabiraman N, Squier R, et al. Web-based interface facilitating sequence-to-structure analysis of BLAST alignment reports. Bio techniques (2005):39:186,188.
35. Gasteiger E, Hoogland C, Gattiker A, Ron D Appel, Amos Bairoch, et al. In: The proteomics protocols handbook; Protein identification and analysis tools on the ExPASy server. Springer (2005):571-607.
36. Allen JE, Salzberg SL JIGSAW: integration of multiple sources of evidence for gene prediction. Bioinformatics (2005):21:3596-3603.
37. Weckx S, Del-Favero J, Rademakers R, Claes L, Cruts M, et al. novo SNP, a novel computational tool for sequence variation discovery. Genome Res (2005):15:436-442.
38. Münch R, Hiller K, Grote A, Scheer M, Klein J, et al. Virtual Footprint and PRODORIC: an integrative framework for regulon prediction in prokaryotes. Bioinformatics (2005):21:4187-4189.
39. Unniraman S, Prakash R, Nagaraja V. Conserved economics of transcription termination in eubacteria. Nucleic Acids Res (2002):30: 675-684.
40. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol (1997):268:78-94.