Journal of
Computer Science & Systems Biology

Review Article

Open Access

Biological Databases- Integration of Life Science Data

Nishant Toomula[1]*, Arun Kumar[2], Sathish Kumar D[3] and Vijaya Shanti Bheemidi[4]

[1]Department of Biotechnology, GITAM Institute of Technology, GITAM University, Visakhapatnam, India
[2]Department of Biochemistry, GITAM University, Visakhapatnam, India
[3]Department of Biotechnology, University of Hyderabad, Hyderabad, India
[4]Department of Biotechnology, Nottingham Trent University, Nottingham, United Kingdom

Abstract

Over the past few decades, major advances in the field of molecular biology, coupled with advances in genomic technologies, have led to an explosive growth in biological information generated by the scientific community. Biological databases are libraries of life sciences information, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analyses. Information contained in biological databases includes gene function, structure, localization, clinical effects of mutations as well as similarities of biological sequences and structures. This article presents information on some popular bioinformatic databases available online, including sequence, phylogenetic, structure and pathway, and microarray databases.

Keywords: EMBL (European Molecular Biology Laboratory); DDBJ (DNA Data Bank of Japan); NCBI (National Center for Biotechnology Information); Proteomic tools; CATH (Class, architecure, topology, homologous superfamily); Protein domain

Introduction

Bioinformatics is the application of Information technology to store, organize and analyze the vast amount of biological data which is available in the form of sequences and structures of proteins and nucleic acids [1]. The biological information of nucleic acids is available as sequences while the data of proteins is available as sequences and structures. Sequences are represented in single dimension where as the structure contains the three dimensional data of sequences.

The National Center for Biotechnology Information (NCBI 2001) defines bioinformatics as: “Bioinformatics is the field of science in which biology, computer science, and information technology merge into a single discipline. There are three important sub-disciplines within bioinformatics: the development of new algorithms and statistics which 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.”

Relational database concepts of computer science and Information retrieval concepts of digital libraries are important for understanding biological databases. Biological database design, development, and long-term management are a core area of the discipline of bioinformatics [2]. Data contents include gene sequences [3], textual descriptions, attributes and ontology classifications, citations, and tabular data. These are often described as semi-structured data, and can be represented as tables, key delimited records, and XML structures [4,5]. Cross-references among databases are common, using database accession numbers [6,7].

A biological database is a collection of data that is organized so that its contents can easily be accessed, managed, and updated. The activity of preparing a database can be divided into:

- Collection of data in a form which can be easily accessed
- Making it available to a multi-user system

Databases in general can be classified in to primary, secondary and composite databases. A primary database contains information of the sequence or structure alone. Examples of these include Swiss-Prot & PIR for protein sequences, GenBank & DDBJ for Genome sequences and the Protein Databank for protein structures [8].

A secondary database contains derived information from the primary database. A secondary sequence database contains information like the conserved sequence, signature sequence and active site residues of the protein families [9] arrived by multiple sequence alignment of a set of related proteins [10]. A secondary structure [11] database contains entries of the PDB in an organized way. These contain entries that are classified according to their structure like all alpha proteins, all beta proteins, turns, helices [12,13]. These also contain information on conserved secondary structure motifs of a particular protein. Some of the secondary databases created and hosted by various researchers at their individual laboratories include SCOP, developed at Cambridge University; CATH developed at University College of London, PROSITE of Swiss Institute of Bioinformatics, eMOTIF at Stanford.

The first database was created within a short period after the Insulin protein sequence was made available in 1956. Incidentally, Insulin is the first protein to be sequenced. The sequence of Insulin consisted of just 51 residues which characterize the sequence. Around mid nineteen sixties, the first nucleic acid sequence of Yeast tRNA with 77 bases was found out. During this period, three dimensional structures of proteins were studied and the well known Protein Data Bank was developed as the first protein structure database with only 10 entries in 1972. This has now grown in to a large database with over 10,000 entries.

***Corresponding author:** Nishant Toomula, Department of Biotechnology, GITAM Institute of Technology, GITAM University, Visakhapatnam, India, E-mail: nishanththumula@gmail.com

Received November 02, 2011; **Accepted** December 05, 2011; **Published** December 09, 2011

Citation: Nishant T, Arun Kumar, Sathish Kumar D, Vijaya Shanti B (2011) Biological Databases- Integration of Life Science Data J Comput Sci Syst Biol 4: 087-092. doi:10.4172/jcsb.1000081

Copyright: © 2011 Nishant T, et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License,which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

While the initial databases of protein sequences were maintained at the individual laboratories, the development of a consolidated formal database known as SWISS-PROT protein sequence database was initiated in 1986.Modern biological databases comprise not only data, but also sophisticated query facilities and bioinformatic data analysis tools [14]; hence, the term "bioinformatic databases" is often used.

Biological databases can be broadly classified in to sequence, structure and pathway databases. Sequence databases are applicable to both nucleic acid sequences and protein sequences, whereas structure databases are applicable to only Proteins.

Sequence databases

Nucleotide and protein sequence databases represent the most widely used and some of the best established biological databases. These databases serve as repositories for wet lab results and the primary source for experimental results. Major public data banks which takes care of the DNA and protein sequences are GenBank [15] in USA, EMBL (European Molecular Biology Laboratory) in Europe and DDBJ (DNADataBank) in Japan [16].

GenBank: The GenBank nucleotide database is maintained by the National Center for Biotechnology Information (NCBI) [17-19], which is part of the National Institute of Health (NIH), a federal agency of the US government.

EMBL: The EMBL nucleotide sequence database [20] is maintained by the European Bioinformatics Institute (EBI) in Hinxton and

DDBJ: DNA Data Bank of Japan Is a biological database that collects DNA sequences submitted by researchers. It is run by the National Institute of Genetics, Japan.

Ensembl: The Ensembl database is a repository of stable, automatically annotated human genome sequences. Ensembl annotates and predicts new genes, with annotation from the InterPro [21] protein family databases and with additional annotations from databases of genetic disease-OMIM [22-24], expression-SAGE [25,26] and gene family [27].

SGD: The Saccharomyces Genome Database (SGD) is a scientific database of the molecular biology and genetics of the yeast Saccharomyces cerevisiae.

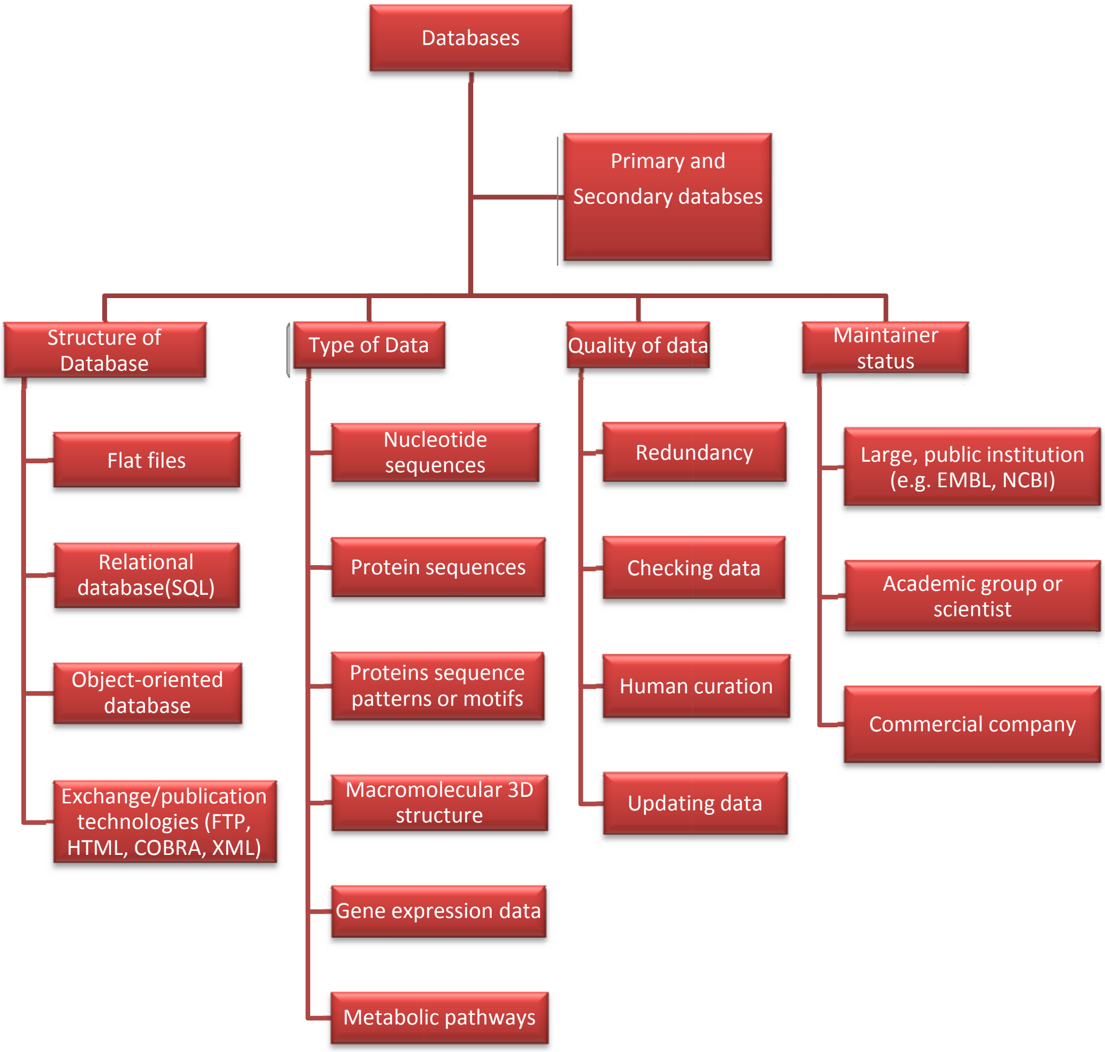

Figure 1: Schematic representation of Database.

dbEST: dbEST [28] is a division of GenBank that contains sequence data and other information on short, "single-pass" cDNA sequences, or Expressed Sequence Tags (ESTs) [29], generated from randomly selected library clones. Expressed Sequence Tags (ESTs) are currently the most widely sequenced nucleotide commodity in the terms of number of sequences and total nucleotide count [30].

PIR: The Protein Information Resource (PIR) is an integrated public bioinformatics resource that supports genomic and proteomic research and scientific studies [34]. PIR has provided many protein databases and analysis tools to the scientific community, including the PIR-International Protein Sequence Database (PSD) of functionally annotated protein sequences. The PIR-PSD, originally created as the Atlas of Protein Sequence and Structure edited by Margaret Dayhoff, contained protein sequences that were highly annotated with functional, structural, bibliographic, and sequence data [35,36].

Swiss-Prot: Swiss-Prot [37,38] is a protein sequence and knowledge database. It is well known for its minimal redundancy, high quality of annotation, use of standardized nomenclature, and links to specialized databases. As Swiss-Prot is a protein sequence database, its repository contains the amino acid sequence, the protein name and description, taxonomic data, and citation information [39].

TrEMBL: The European Bioinformatics Institute, collaborating with Swiss-Prot, introduced another database, TrEMBL (translation of EMBL nucleotide sequence database) [40]. This database consists of computer annotated entries derived from the translation of all coding sequences in the nucleotide databases.

UniProt: UniProt database is organized into three layers. The UniProt Archive (Uni-Parc) stores the stable, nonredundant, corpus of publicly available protein sequence data. The UniProt Knowledge base (UniProtKB) consists of accurate protein sequences with functional annotation [41,42]. Finally, the UniProt Reference Cluster (UniRef) datasets provide nonredundant reference clusters based primarily on UniProtKB. UniProt also offers users multiple tools, including searches against the individual contributing databases, BLAST [43,44] and multiple sequence alignment, proteomic tools, and bibliographic searches [45].

Structure databases

Knowledge of protein structures and of molecular interactions is key to understanding protein functions and com-plex regulatory mechanisms underlying many biological processes [50].

Protein Data Bank: The PDB (Protein Data Bank) is the single worldwide archive of Structural data of Biological macromolecules, established in Brookhaven National Laboratories in 1971. It contains Structural information of the macromolecules determined by X-ray crystallographic, NMR methods [51-53]. PDB is maintained by the Research Collaboratory for Structural Bioinformatics (RCSB). It allows the user to view data both in plain text and through a molecular viewer using Jmol.

SCOP: The SCOP (Structural Classification of Proteins) [54] database was started by Alexey Murzin in 1994. Its purpose is to classify protein 3D structures in a hierarchical scheme of structural classes.

CATH: The CATH database (Class, architecure, topology, homologous superfamily) is a hierarchical classification of protein domain structures, which clusters proteins at four major structural levels.

NDB: Nucleic Acid Database, also curated by RCSB and similar to the PDB and the Cambridge Structural Database [55], is a repository for nucleic acid structures. It gives users access to tools for extracting information from nucleic acid structures and distributes data and software.

Pathway databases

Development of metabolic databases derived from the comparative study of metabolic pathways will cater the industrial needs in more

Database	URL	Feature
GenBank [31]	http://www.ncbi.nlm.nih.gov/	NIH's archival genetic sequence database
EMBL	http://www.ebi.ac.uk/embl/	EBI's archival genetic sequence database
DDBJ	http://www.ddbj.nig.ac.jp/	NIG's archival genetic sequence database
SGD	http://www.yeastgenome.org/	A repository for baker's yeast genome and biological data
EBI genomes	http://www.ebi.ac.uk/genomes/	It provides access and statistics for the completed genomes [32]
Ensembl	http://www.ensembl.org/	Database that maintains automatic annotation on selected eukaryotic genomes [33]
UniGene	http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene	Each UniGene cluster contains sequences that represent a unique gene, as well as related information.
dbEST	http://www.ncbi.nlm.nih.gov/dbEST/	Division of GenBank that contains expression tag sequence data

Table 1: Summary of Nucleotide sequence databases.

Database	URL	Feature
Swiss-Prot/TrEMBL [46,47]	http://www.expasy.org/sprot/	Description of the function of a protein, its domains structure, post-translational modifications etc,
UniProt [48]	http://www.pir.uniprot.org/	Central repository for PIR, Swiss-Prot, and TrEMBL
PIR	http://pir.georgetown.edu/	It strives to be comprehensive, well-organized, accurate, and consistently annotated.
Pfam	pfam.sanger.ac.uk/	Database of protein families defined as domains [49]
PROSITE	www.expasy.ch/prosite/	Database of protein families and domains

Table 2: Summary of Protein sequence databases.

Database	URL	Feature
PDB	www.rcsb.org/pdb/	Protein structure repository that provides tools for analyzing these structures
SCOP	scop.mrc-lmb.cam.ac.uk/scop/	Classification of protein 3D structures in a hierarchical scheme of structural classes
CATH	www.cathdb.info	Hierarchical classification of protein domain structure
NDB	http://ndbserver.rutgers.edu/	Database housing nucleic acid structural information

Table 3: Summary of Structure databases.

Database	URL	Feature
KEGG	http://www.genome.jp/kegg/	Protein structure repository that provides tools for analyzing these structures
BioCyc	http://www.biocyc.org/	Classification of protein 3D structures in a hierarchical scheme of structural classes
BRENDA	http://www.brenda-enzymes.org/	Hierarchical classification of protein domain structure
EMP	http://emp.mcs.anl.gov/	Database of Enzymes and Metabolic pathways public server
BRITE	http://www.genome.jp/kegg/brite.html	Biomolecular Relations in Information, Transmission and Expression

Table 4: Summary of Pathway databases.

efficient manner to further the growth of systems biotechnology [56,57]. Some examples of the pathway databses are KEGG [58], BRENDA, Biocyc.

KEGG: The Kyoto Encyclopedia of Genes and Genomes (KEGG) is the primary resource for the Japanese GenomeNet service that attempts to define the relationships between the functional meanings and utilities of the cell or the organism and its genome information [59-61]. KEGG contains three databases: PATHWAY, GENES, and LIGAND. The PATHWAY database stores computerized knowledge on molecular interaction networks. The GENES database contains data concerning sequences of genes and proteins generated by the genome projects. The LIGAND database holds information about the chemical compounds and chemical reactions that are relevant to cellular processes.

BRENDA: It is the main collection of enzyme functional data [62] available to the scientific community. It is maintained and developed at the Institute of Biochemistry and Bioinformatics at the Technical University of Braunschweig, Germany.

BioCyc: The BioCyc Database Collection is a compilation of pathway and genome information for different organisms [63]. It includes two other databases, EcoCyc [64], which describes Escherichia coli K-12; MetaCyc [65], which describes pathways for more than 300 organisms.

Conclusion

As biology has increasingly turned into a data rich science, the need for storing and communicating large datasets has grown tremendously. The obvious examples are the nucleotide sequences, the protein sequences, and the 3D structural data produced by X-ray crystallography and NMR. Biological databases are an important tool in assisting scientists to understand and explain a host of biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. This knowledge helps facilitate the fight against diseases, assists in the development of medications and in discovering basic relationships amongst species in the history of life.

References

1. Ragunath PK, Venkatesan P, Ravimohan R (2009) New Cur-riculum Design Model for Bioinformatics Postgraduate program using Systems Biology Approach. J Comput Sci Syst Biol 2: 300-305.
2. Bourne P (2005). "Will a biological database be different from a biological journal?" *PLoS Comput Biol* 1: 179–81.
3. Neha S, Vrat BS, Kumud J, Thakur PD, Rajinder K, et al. (2011) Comparative In silico Analysis of Partial Coat Protein Gene Sequence of Zucchini Yellow Mosaic Virus Infecting Summer Squash (Cucurbita pepo L.) Isolated From India. J Proteomics Bioinform 4: 068-073.
4. Riad AM, Hassan AE, Hassan QF (2009) Investigating Perfor-mance of XML Web Services in Real-Time Business Systems. J Comput Sci Syst Biol 2: 266-271.
5. Henneges C, Hinselmann G, Jung S, Madlung J, Schütz W, et al. (2009) Ranking Methods for the Prediction of Frequent Top Scoring Peptides from Proteomics Data. J Proteomics Bioinform 2: 226-235.
6. Karthick RNS, Muthukumaran J (2008) Prediction of Three Dimensional Model and Active Site Analysis of Inducible Serine Protease Inhibitor -2 (ISPI -2) in Galleria Mellonella. J Comput Sci Syst Biol 1: 119-125.
7. Liu Z, Liu Y, Liu S, Ding X, Yang Y, et al. (2009) Analysis of the Sequence of ITS1-5.8S-ITS2 Regions of the Three Species of Fructus Evodiae in Guizhou Province of China and Identification of Main Ingredients of Their Medicinal Chem-istry. J Comput Sci Syst Biol 2: 200-207.
8. Singh S, Gupta SK, Nischal A, Khattri S, Nath R, et al. (2010) Comparative Modeling Study of the 3-D Structure of Small Delta Anti-gen Protein of Hepatitis Delta Virus. J Comput Sci Syst Biol 3: 001-004.
9. Varsale AR, Wadnerkar AS, Mandage RH, Jadhavrao PK (2010) Cheminformatics. J Proteomics Bioinform 3: 253-259.
10. Sahoo GC, Rani M, Dikhit MR, Ansari WA, Das P (2009) Structural Modeling, Evolution and Ligand Interaction of KMP11 Protein of Different Leishmania Strains. J Comput Sci Syst Biol 2: 147-158.
11. Shanthi V, Ramanathan K, Sethumadhavan R (2009) Role of the Cation-π Interaction in Therapeutic Proteins: A Comparative Study with Conventional Stabilizing Forces. J Comput Sci Syst Biol 2: 051-068.
12. Dawson W, Kawai G (2009) Modeling the Chain Entropy of Biopolymers: Unifying Two Different Random Walk Models under One Framework. J Comput Sci Syst Biol 2: 001-023.
13. Vaseeharan B, Valli SJ (2011) In silico Homology Modeling of Prophenoloxidase activating factor Serine Proteinase Gene from the Haemocytes of Fenneropenaeus indicus. J Proteomics Bioinform 4: 053-057.
14. Hoskeri JH, Krishna V, Amruthavalli C (2010) Functional Annotation of Conserved Hypothetical Proteins in Rickettsia Massiliae MTU5. J Comput Sci Syst Biol 3: 050-052.

15. Shi Huang (2008) The Genetic Equidistance Result of Molecular Evolution is Independent of Mutation Rates. J Comput Sci Syst Biol 1: 092-102.

16. Chandra Sekhara Rao A, Somayajulu DVLN (2011) Influenza Classification from Nucleotide Sequence Database. J Comput Sci Syst Biol 4: 077-080.

17. Mudunuri SB, Rao AA, Pallams etty S, Mishra P, Nagarajaram HA (2009) VMD: Viral Microsatellite Database-A Compre-hensive Resource for all Viral Microsatellites. J Comput Sci Syst Biol 2:283-286.

18. Ingale A (2010) In Silico Homology Modeling and Epitope Prediction of Nucleocapsid Protein region from Japanese Encephalitis Virus. J Comput Sci Syst Biol 3: 053-058.

19. Nath M, Goel A, Taj G, Kumar A (2010) Molecular Cloning and Comparative In silico Analysis of Calmodulin Genes from Cereals and Millets for Understanding the Mechanism of Differential Calcium Accumulation. J Proteomics Bioinform 3: 294-301.

20. Garg N, Pundhir S, Prakash A, Kumar A (2008) PCR Primer Design: DREB Genes. J Comput Sci Syst Biol 1: 021-040.

21. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bate-man A, D. et al. (2006) InterPro, progress and status in 2005. Nuc Acids Res 33: 201–205.

22. Antonarakis SE, McKusick VA (2000) OMIM passes the 1,000-disease-gene mark. Nat Genet 25: 11.

23. Gao S, Wang X (2009) Predicting Type 1 Diabetes Candidate Genes using Human Protein-Protein Interaction Networks. J Comput Sci Syst Biol 2: 133-146.

24. Uría GTJ, Mora DM, Saladié RJM (2009) BIOUROLOGY.COM. J Proteomics Bioinform 2: 064-066.

25. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270: 484–487.

26. Wheeler DL, Church DM, Lash AE, Leipe DD, Madden TL, et al. (2002) Database resources of the National Center for Biotechnology Information. Nuc Acids Res 29 : 11–16.

27. Enright AJ, Iliopoulos I, Kyrpides NC, Ouzounis CA (1999) Protein interaction maps for complete genomes based on gene fusion events. Nature 402: 86–90.

28. Boguski MS, Lowe TM, Tolstoshev CM (1993) dbEST —database for expressed sequence tags. Nature Genet 4: 332–333.

29. Miller NA, Kingsmore SF, Farmer AD, Langley RJ, Mudge J, et al. (2008) Management of High-Throughput DNA Sequencing Projects: Alpheus. J Comput Sci Syst Biol 1: 132-148.

30. Sablok G, Shekhawat NS (2008) Bioinformatics Analysis of Distribution of Microsatellite Markers (SSRs) / Single Nucleotide Polymorphism (SNPs) in Expressed Transcripts of Prosopis Juliflora: Frequency and Distribution. J Comput Sci Syst Biol 1: 087-091.

31. Vijai S, Indramani, Dharmendra KC, Pallavi S (2009) HLA Class I and II Binding Promiscuity of the T-cell Epitopes in Putative Proteins of Hepatitis B Virus. J Comput Sci Syst Biol 2: 069-073.

32. Manikandakumar K, Kumaran MS, Srikumar R (2009) Matrix Frequency Analysis of Oryza Sativa (japonica cultivar-group) Complete Genomes. J Comput Sci Syst Biol 2: 159-166.

33. Pandarinath P, Shashi M, Appa Rao A (2010) Bioinformatic Approach for the Identification of Hepatitis B Viral Insert in The Exon Region of Human Genome. J Comput Sci Syst Biol 3: 089-090.

34. Anandakumar S, Shanmughavel P (2008) Computational Annotation for Hypothetical Proteins of Mycobacterium Tuberculosis. J Comput Sci Syst Biol 1: 050-062.

35. Galperin MY (2006) The molecular biology database collection: 2006 update. Nuc Acids Res 34 : 3–5.

36. Wu CH, Yeh LS, Huang H, Arminski L, Castro-Alvear J, et al. (2003) The protein information resource, Nuc Acids Res 31: 345–347.

37. Donovan CO, Martin MJ, Gattiker A, Gasteiger E, Bairoch A, et al. (2002) High-quality protein knowledge resource: SWISSPROT and TrEMBL. Brief Bioinform 3: 275–284.

38. Murty USN, Amit KB, Neelima A (2009) An In Silico Approach to Cluster CAM Kinase Protein Sequences. J Proteomics Bioinform 2: 097-107.

39. Satpathy R, Guru RK, Behera R, Priyadarshini A (2010) Homology Modelling of Lycopene Cleavage Oxygenase: The Key Enzyme of Bixin Production. J Comput Sci Syst Biol 3: 059-061.

40. Ramón A, Javier GG, Baldo O (2008) Integration and Prediction of PPI Using Multiple Resources from Public Databases. J Proteomics Bioinform 1: 166-187.

41. Zhang ZH , Hongzhan H, Amrita C, Mira J, Anatoly D, et al. (2008) Integrated Bioinformatics for Radiation-Induced Pathway Analysis from Proteomics and Microarray Data. J Proteomics Bioinform 1: 047-060.

42. Nikhil S, Rekha K, Sodhi JS, Bhalla TC. (2009) In Silico Analysis of Amino Acid Sequences in Relation to Specificity and Physiochemical Properties of Some Microbial Nitrilases. J Proteomics Bioinform 2: 185-192.

43. Lakshmi PTV, Uma MS, Karthikeyan PP, Annamalai A (2008) Sequence and Structure Comparison Studies of Phycocyanin in Spirulina Platensis. J Comput Sci Syst Biol 1: 063-072.

44. Gupta SK, Akhoon BA, Srivastava M, Gupta SK (2010) A Novel Algorithm to Design an Efficient siRNA by Combining the Pre Proposed Rules of siRNA Designing. J Comput Sci Syst Biol 3: 005-009.

45. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, et al. (2006) The Universal Protein Resource (Uni-Prot): AN expanding universe of protein information. Nuc. Acids Res 34: 187–191.

46. Oxana VG, Eugeniya ID, Igor NS (2008) Phylogenetic Analyses of the Loops in Elongation Factors EF1A: Stronger Support for the Grouping of Animal and Fungi. J Comput Sci Syst Biol 1: 073-080.

47. Brum IJB, Martins-de-Souza D, Smolka MB, Novello JC, Galembeck E (2009) Web Based Theoretical Protein pI, MW and 2DE Map. J Comput Sci Syst Biol 2: 093-096.

48. Rossetti RAM, Lorenzi JCC, Giuliatti S, Silva CL, Coelho-Castelo AAM (2008) In Silico Prediction of the Tertiary Structure of M. leprae Hsp65 Protein Shows an Unusual Structure in Carboxy-terminal Region. J Comput Sci Syst Biol 1: 126-131.

49. Razia M, Raja KR, Padmanaban K, Sivaramakrishnan S, Chellapandi P (2010) A Phylogenetic Approach for Assigning Function of Hypothetical Proteins in Photorhabdus luminescens Subsp. Laumondii TT01 Genome. J Comput Sci Syst Biol 3: 021-029.

50. Aghdam MH, Tanha J, Naghsh-Nilchi AR, Basiri ME (2009) Combination of Ant Colony Optimization and Baye-sian Classification for Feature Selection in a Bioinformatics Dataset. J Comput Sci Syst Biol 2: 186-199.

51. Prakash N, Patel S, Faldu NJ, Ranjan R, Sudheer DVN (2010) Molecular Docking Studies of Antimalarial Drugs for Malaria. J Comput Sci Syst Biol 3: 070-073.

52. Liang K, Wang X (2011) Protein Secondary Structure Prediction using Deterministic Sequential Sampling. J Data Mining in Genom Proteomics 2:107.

53. Kumar S, Sahu BB, Tripathy NK, Shaw BP (2009) In Silico Identification of Putative Proton Binding Sites of a Plasma Membrane H+ -ATPase Isoform of Arabidopsis Thaliana, AHA1. J Proteomics Bioinform 2: 349-359.

54. Subramanian R, Muthurajan R, Ayyanar M (2008) Comparative Modeling and Analysis of 3-D Structure of EMV2, a Late Embryogenesis Abundant Protein of Vigna Radiata (Wilczek). J Proteomics Bioinform 1: 401-407.

55. Allen FH, Bellard S, Brice MD, Cartwright BA, Doubleday A, et al. (1979) The Cambridge crystallographic data centre: Computer-based search, retrieval, analysis and display of information. Acta Cryst., 35: 2331–2339.

56. Chellapandi P, Sivaramakrishnan S, Viswanathan MB (2010) Systems Biotechnology: an Emerging Trend in Metabolic Engineering of Industrial Microorganisms. J Comput Sci Syst Biol 3: 043-049.

57. Lakshminarasimhan N, Tagore S (2009) Damages in Metabolic Pathways: Computational Approaches. J Comput Sci Syst Biol 2: 208-215.

58. Sarangi AN, Aggarwal R, Rahman Q, Trivedi N (2009) Sub-tractive Genomics Approach for in Silico Identification and Character-ization of Novel Drug

Targets in Neisseria Meningitidis Serogroup B. J Comput Sci Syst Biol 2: 255-258.

59. Nicolas T (2009) Data Mining, a Tool for Systems Biology or a Systems Biology Tool. J Comput Sci Syst Biol 2: 216-218.
60. Rao VS, Das SK, Umari EK (2009) Glycomics Data Mining. J Comput Sci Syst Biol 2: 262-265.
61. Morya VK, Dewaker V, Mecarty SD, Singh R (2010) In silico Analysis Metabolic Pathways for Identification of Putative Drug Targets for Staphylococcus aureus. J Comput Sci Syst Biol 3: 062-069.
62. Parveen S, Asad UK (2008) Proteolytic Enzymes Database. J Proteomics Bioinform 1: 109-111.
63. Hua D, Yanghua X, Wei W, Li J, Momiao X (2008) Symmetry of Metabolic Network. J Comput Sci Syst Biol 1: 001-020.
64. Caspi R, Foerster H, Fulcher CA, Hopkinson R, Ingraham J, et al. (2006) MetaCyc: A multi organism database of metabolic pathways and enzymes. Nuc Acids Res 34: 511–516.
65. Romero P, Wagg J, Green ML, Kaiser D, Krummenacker M, et al. (2004) Computational prediction of human metabolic pathways from the complete human genome. Genome Biology 6: 1–17.