Biological Databases- Integration of Life Science Data

Nishant Toomula1, Arun Kumar2, Sathish Kumar D3 and Vijaya Shanti Bheemidi4

1Department of Biotechnology, GITAM Institute of Technology, GITAM University, Visakhapatnam, India
2Department of Biochemistry, GITAM University, Visakhapatnam, India
3Department of Biotechnology, National Center for Biotechnology Information, Bethesda, Bethesda, USA
4Department 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, architecture, 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 into 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 just 51 residues which characterize the sequence. Around mid nineteen sixties, the first nucleic acid sequence of Yeast tRNA with 77 bases was just 51 residues which characterize the sequence. Around mid nineteen sixties, the first nucleic acid sequence of Yeast tRNA with 77 bases was just 51 residues which characterize the sequence. Around mid nineteen sixties, the first nucleic acid sequence of Yeast tRNA with 77 bases was just 51 residues which characterize the sequence.
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 into 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 take care of the DNA and protein sequences are GenBank [15] in USA, EMBL (European Molecular Biology Laboratory) in Europe and DDBJ (DNA DataBank) 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) provides database 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 complex 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, architecture, 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/embreln/                                         | 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.
efficient manner to further the growth of systems biotechnology [56,57]. Some examples of the pathway databases 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 Curriculm 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) Classification of protein 3D structures in a hierarchical scheme of structural classes. 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 IT51-5.8S-IT52 Regions of the Three Species of Fructose Eudivoices 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. Shanithi 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 Haemocytos 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.

**Table 3:** Summary of Structure 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 4:** Summary of Pathway 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 |
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 5: 1–17.