International Journal of Computing Science and Information Technology, 2014, Vol.2(4), 06-10
ISSN: 2278-9669, October 2014 (http://ijcsit.org)

SOME STUDIES ON MAJOR DATABASES IN BIOINFORMATICS

N.Deepak Kumar [1] Email: nagiletideepakvictor@gmail.com,
Dr.A.Ramamohan Reddy [2] Email: ramamohansvu@yahoo.com,
[12]Dept.Of Computer Science &Engineering, SVUniversity,
Tirupati-517502,India,

Abstract

Making the biological information available for analysis and developing applications is the key and there are a huge number of databases in the public and private domain to do so. Here we discuss some of the major databases that work as the primary sources of informations in bioinformatics. Protein databases are categorized as primary and composite or secondary.

Keywords: DBMS.bioinformatics,Protein,Composite database,SRS

1. Introduction

Bioinformatics represents an interdisciplinary compiling of biosciences and biotechnology with mathematics and informatics. Databases, especially factual databases as for example nucleotide and protein sequence databases, are one of the most important tools in bioinformatics. Bioinformatics includes different subject areas, e.g., development of databases and software. The use of sequence and structure information stored in databases opens together with new concepts, such as neural networks, new approaches in molecular modeling for structure prediction and description, theories and methods of 3D-modeling and of optimizing macromolecules, knowledge-bases sequence analysis and prediction of protein folding, development of knowledge-based systems and artificial intelligence methods for applications in genome research, protein design, drug design, etc.[1].

1.1 European activities in bioinformatics

With the aim to improve the development of bioinformatics in Europe, in November 1990 the European Chemical Industry Foundation (CEFIC) recommended to the Commission of the European Communities the establishment of European Nucleotide Sequence Center. After long discussion, in 1993 the decision has been made to locate the new European Bioinformatics Institute (EBI) at the Genome Campus near Cambridge in the UK. The EBI will work to[2,3]:

- ensure that the Data Library continues, develops in response to advancing biological research, and is capable of exploiting advances in informatics;
- reinforce areas which have been neglected, particularly training and user support;
- make the voice of European bioinformatics heard in the global arena;
- increase the effectiveness of dispersed, high-quality European research and service by entering into collaborations with centers of expertise throughout Europe.

The EBI will provide the EMBL Nucleotide Sequence Database and the SWISSPROT Protein Sequence Database; support and distribute other databases in collaboration with European scientist; help to coordinate the EMBnet nodes and molecular biology network services; carry out research and development in the application of information technology, actively tracking advances to explore their utility to biological problems; provide quality user support; and ensure that Europeans are globally competitive in the profession of bioinformatics[4,5].

1.2 U.S. activities in bioinformatics

In connection with the U.S. Human Genome Program and the U.S. Plant Genome Research Program may large-scale bioinformatics projects were initialized, especially related to the development and improvement of databases for sequence and map data, and software for sequencing and mapping as well as for the creation and use of databases [5,6]

2. Methodology

Protein databases are categorized as primary and composite or secondary. These are discussed here

2.1ProteinPrimarydatabases

The primary sequence databases currently hold over

300,000 non-redundant protein sequences. The most commonly-used are:-

(A).SWISS-PROT:-

http://www.expasy.ch/sprot/sprot-top.html

This is a "curated" database which provides a high level of annotation of each sequence, including: a description of the function of a protein, its domain structure, post-translational modifications, variants, *etc*. It also has extensive links to other databases.

Contains:- **88,757** annotated entries, and **300,152** (TrEMBL) unannotated ones (as at **2 Oct 2000**).

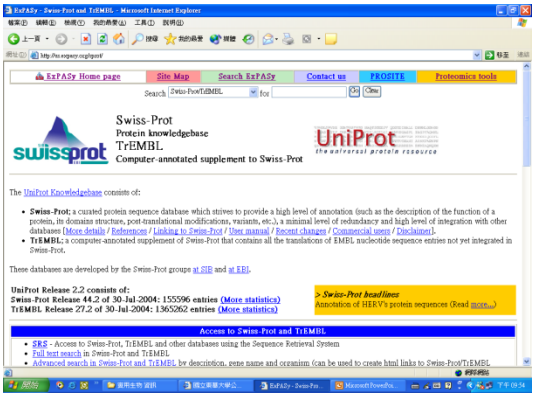

Fig 1.Swiss prot HOME PAGE

(B).Protein Information Resource (PIR) :-

- **at the *National Biomedical Research Foundation* (NBRF)**

http://www-nbrf.georgetown.edu/pir/

- *Contains:-* **20,397** annotated and fully classified entries, and **195,726** verified and classified ones (as at **6 Oct 2000**).
- **PIR-International (PIR) - at the *National Biomedical Research Foundation* (NBRF)**

http://www.mips.biochem.mpg.de/proj/protseqdb/

An annotated, curated, and largely non-redundant protein sequence database created by a common effort of PIR (USA), MIPS (Munich Information Centre for Protein Sequences, Germany), and JIPID (Japan). In addition to sequences, it provides a variety of biological information, *e.g.* protein function, homology information, or sequence-related information (features). *Contains:-* **178,320** entries (as at **20 April 2000**).

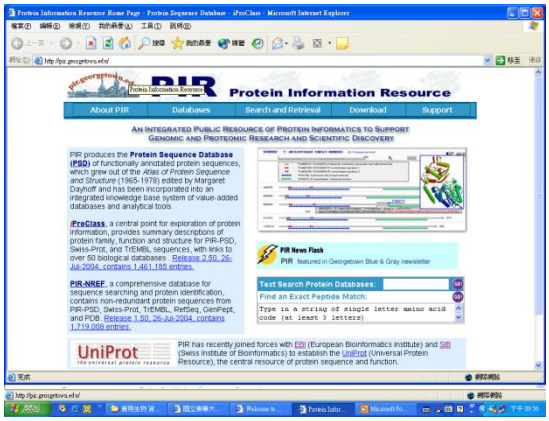

Fig 2.PIR HOME PAGE

2.2. CompositeDatabases

There are a number of "composite" databases of protein sequences. These compile their sequence data from the primary sequence databases and filter them to retain only the non-redundant sequences. The best-known are:-

(A).owl

http://www.bioinf.man.ac.uk/dbbrowser/OWL/

Contains:- 279,796 entries, from **SWISS-PROT**, **PIR** (1-3), **GenBank** (translation) and **NRL-3D** (as at **30 Nov 1998**).

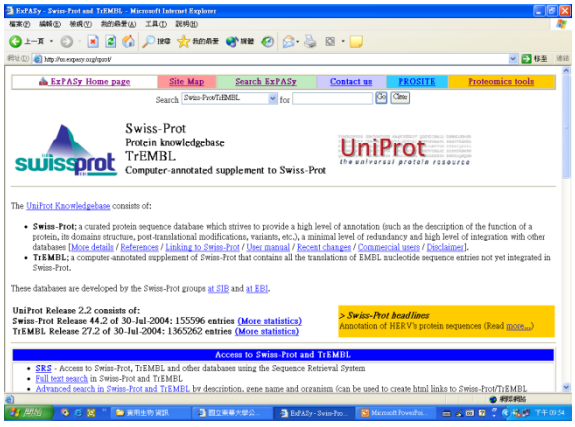

Fig 3.Swiss prot HOME PAGE

(B).Non-Redundant DataBases (NRDB) - at the National Center for Biotechnology Information (NCBI)

http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein

The Protein sequences database contains sequence data from the translated coding regions from DNA sequences in GenBank, EMBL and DDBJ as well as protein sequences submitted to PIR, SWISSPROT, PRF, Protein Data Bank (PDB) (sequences from solved structures). Searchable by Entrez.

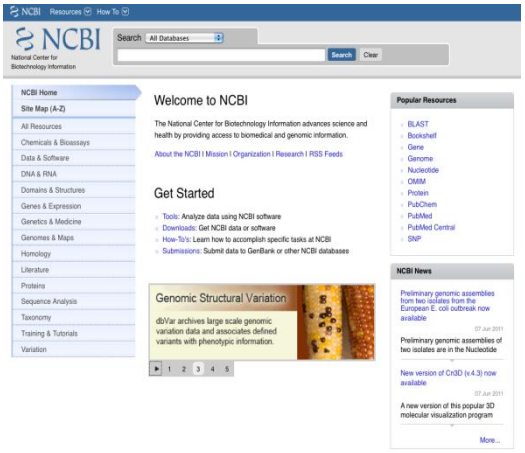

Fig 4.NCBI HOME PAGE

FIG 5.The NCBI in the USA has its own search engine called Entrez.

2.3. Secondary databases

Secondary databases are those that contain information **derived** from the primary sequence databases.

*(A)*PROSITE

http://www.expasy.ch/prosite

PROSITE is a database of residue patterns and profiles that characterise biologically significant sites in proteins and can help reliably identify to which known protein family (if any) a new sequence belongs.

The database allows you to scan a given sequence against all the **PROSITE** patterns and profiles at:-

- **ScanProsite - Scan protein against PROSITE**

http://www.expasy.ch/tools/scnpsit1.html

Contains:- 1,072 documentation entries that describe **1,445** different patterns, rules and profiles/matrices (as at 4 Oct 2000).

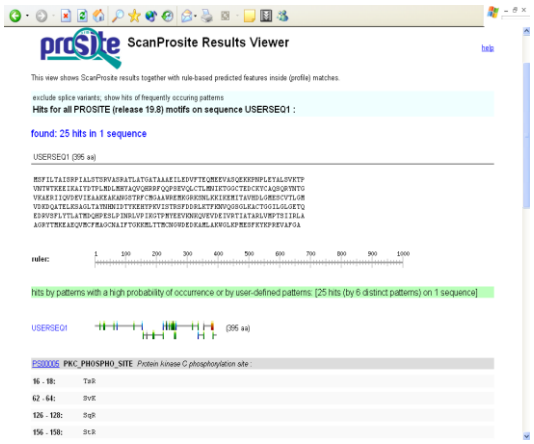

Fig 6.PROSITE HOME PAGE

(B).PRINTS - Protein Motif Fingerprint Database:-

http://bioinf.mcc.ac.uk/dbbrowser/PRINTS/PRINTS.html

PRINTS is a compendium of protein fingerprints. A fingerprint is a group of conserved motifs used to characterise a protein family; its diagnostic power is refined by iterative scanning of OWL. Usually the motifs do not overlap, but are separated along a sequence, though they may be contiguous in 3D-space. Fingerprints can encode protein folds and functionalities more flexibly and powerfully than can single motifs: the database thus provides a useful adjunct to PROSITE. The database allows you to scan a given sequence against all the PRINTS patterns and profiles at:-

- FingerPRINTScan

http://bioinf.man.ac.uk/fingerPRINTScan

The InterPro database uses a collection of profiles from PRINTS, Prosite, ProDom, Pfam and SWISS-PROT for whole genome analysis:-

- InterPro

http://www.ebi.ac.uk/interpro/

Contains:- 1,410 entries (as at 25 Sept 2000).

(C).Pfam :- Protein families database of alignments and HMMs

http://www.sanger.ac.uk/Pfam/

Pfam is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains. There are two part: Pfam-A are accurate human crafted multiple alignments whereas Pfam-B is an automatic clustering of the rest of SWISS-PROT and TrEMBL using the program Domainer. It differs from the other databases in aiming to get as complete as possible a set of *accurate* protein multiple alignments, which require human skills in making the "seed" alignments.

To scan a given sequence against the Pfam library, go to:-

- Search Pfam

http://www.sanger.ac.uk/Pfam/search.shtml

Pfam is a large collection of multiple sequence alignments and profile hidden Markov models (HMMs) covering many common protein domains. It is semi-automatically generated and aims to be comprehensive as well as accurate.

Contains:- 2,478 families (as at Sept 2000).

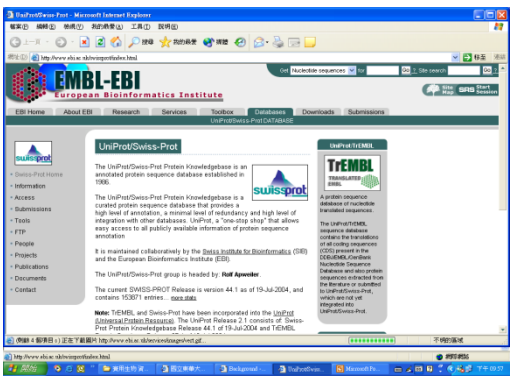

Fig 7.EMBL HOME PAGE

2.4. PROTEIN STRUCTURE DATABASE

In biology, a **protein structure database** is a database that is modeled around the various experimentally determined protein structures. The aim of most protein structure databases is to organize and annotate the protein structures, providing the biological community access to the experimental data in a useful way. Data included in protein structure databases often includes three-dimensional coordinates as well as experimental information, such as unit cell dimensions and angles for x-ray crystallography determined structures. Though most instances, in this case either proteins or a specific structure determinations of a protein, also contain sequence information and some databases even provide means for performing sequence based queries, the primary attribute of a structure database is structural information, whereas sequence databases focus on sequence information, and contain no structural information for the majority of entries. Protein structure databases are critical for many efforts in computational biology such as structure based drug design, both in developing the computational methods used and in providing a large experimental dataset used by some methods to provide insights about the function of a protein.

A. (A).The Protein Data Bank

The Protein Data Bank (PDB) was established in 1971 as the central archive of all experimentally determined protein structure data. Today the PDB is maintained by an international consortia collectively known as the Worldwide Protein Data Bank (wwPDB). The mission of the wwPDB is to maintain a single archive of macromolecular structural data that is freely and publicly available to the global community.

Fig 8.PDB HOME PAGE

3. Result

The use of databases and other information sources as an aid to increase the public perception of biotechnology: The lack of public acceptance of biotechnology is also a result of the lack of information, since there is a direct link between information and attitude. Researchers in the public perception of biotechnology agree that attempts to improve access to scientific information are highly desirable (GRINDLEY and BENNETT, 1993). Specific measures should be taken to enhance public perception mostly through the availability of objective information, especially in connection with biotechnology's impact on human health through the development of new pharmaceutical, diagnostic, and other medical products.

4. Conclusion

The complexity of data and databases connected with the necessity to crosslink different (types of) information which are part of different (types of) databases, to combine different forms of information representation, to extend the numeric data by means of supplementary, descriptive information, to use standardized or easily translatable formats which must be interconnected in order to integrate individual databases in a global concept. The integration of databases from various producers and structures in systems which have a single, unified administration and allow a homogeneous access to the various heterogeneous data present. For bibliographic databases, the Commission of the European Communities recommends the creation of a "Common core Database" by the bunching of central biotechnology databases for the prevention of duplicates, overlapping, etc.

References

[1].ALSTON, y., COOMBS, j. (1992), Biosciences, Information Sources and Services, New York: Stockton Press.

[2].CRAFTS-LIGHTLY, A. (1986), Information Sources in Biotechnology, Weinheim: VCH Verlagsgesell-schaft.

[3].GRINDLEY, J.N., BENNETT, D.J. (1993), Public perception and the socio-economic integration of biotechnology, in: Biotechnologia 20, 89-102.

[4].LŰCKE, E.-M., POETZSCH, E. (1993), Biotechnology Directory Eastern Europe, Berlin-New York: de Gruyter.

[5].MARCACCIO, K. Y. (1993), Gale Directory of Databases, Vol. 1: Online Database, Detroit: Gale Research Inc.

[6].MEWES, H.-W. (1990), Workshop Computer Applications in Biosciences, Book of Abstracts, p. 11, Martinsried.

[7].POETZSCH, E. (1986), Faktographische informationsfonds zur Biotechnologie, Berlin: WIZ.

[8]. POETZSCH. E. (1993), Bio Technologie Das Jahrund Adreβbuch 93/94, Berlin: polycom Verlags-gesellschaft

[9].Wren JD, Bateman A (2008). "Databases, data tombs and dust in the wind.". Bioinformatics 24 (19): 2127–8. doi:10.1093/bioinformatics/btn464. PMID 18819940.

[10].http://ezgenome.ezbiocloud.net/