# Biological Databases and Protein Sequence Analysis

**M. Madan Babu**, Center for Biotechnology, Anna University, Chennai – 25, India

## 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 (the building blocks of organisms) and nucleic acids (the information carrier). 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.

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 in to:

- Collection of data in a form which can be easily accessed
- Making it available to a multi-user system ( always available for the user)

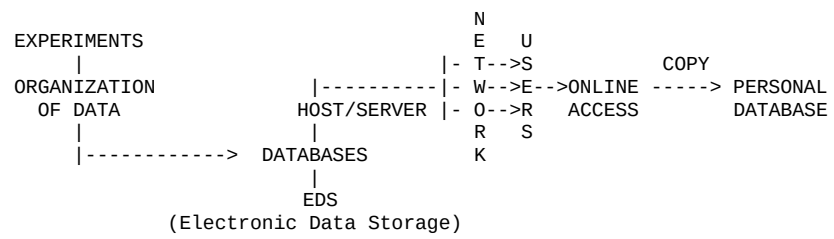



*Figure 1. The network for production, construction and accession of a database*

## Biological Databases

When Sanger first discovered the method to sequence proteins, there was a lot of excitement in the field of Molecular Biology. Initial interest in Bioinformatics was propelled by the necessity to create databases of biological sequences.

Biological databases can be broadly classified in to sequence and structure databases. Sequence databases is applicable to both nucleic acid sequences and protein sequences, whereas structure database is applicable to only Proteins. 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 (analogous to alphabets in a sentence) which characterize the sequence. Around mid nineteen sixties, the first nucleic acid sequence of Yeast tRNA with 77 bases (individual units of nucleic acids) was found out. During this period, three dimensional structure 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. 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 which now has about 70,000 protein sequences from more than 5000 model organisms, a small fraction of all known organisms. These huge varieties of divergent data resources are now available for study and research by both academic institutions and industries. These are made available as public domain

information in the larger interest of research community through Internet (www.ncbi.nlm.nih.gov) and CD-ROMs (on request from www.rcsb.org). These databases are constantly updated with additional entries.

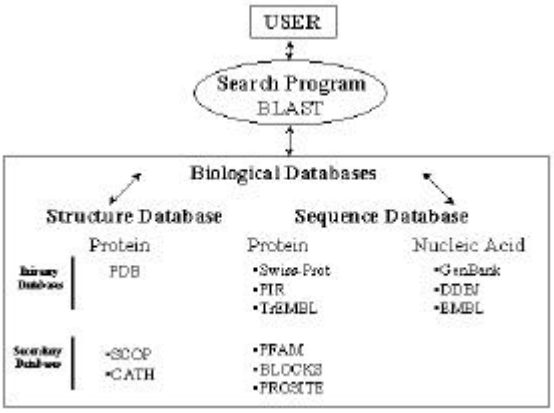



*Figure 2. Biological Databases*

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.


*Figure 3. Examples of Primary databases. Swiss-Prot (left) for the protein sequence database and PDB (right) for the protein structure database*

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 arrived by multiple sequence alignment of a set of related proteins. A secondary structure 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, etc. These also contain information on conserved secondary structure motifs of a particular protein. Some of the secondary database 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.


*Figure 4. Examples of Secondary databases. PROSITE (left) is an example of protein sequence secondary database and SCOP (right) - Structural Classification of Proteins is an example of protein structure secondary database*

Composite database amalgamates a variety of different primary database sources, which obviates the need to search multiple resources. Different composite database use different primary database and different criteria in their search algorithm. Various options for search has also been incorporated in the composite database. The National Center for Biotechnology Information (NCBI) which host these nucleotide and protein databases in their large high available redundant array of computer servers, provides free access to the various persons involved in research. This also has link to OMIM (Online Mendelian Inheritance in Man) which contains information about the proteins involved in genetic diseases.


*Figure 5. Example of composite database. NCBI (National Center for Biotechnology Information)*

The growth of the primary databases gave rise to serious and valid questions on the format of the sequences, reliability and the comprehensiveness of the databases. To address the format issues, in-house software solutions have been developed to convert format of one database to another. A public domain software FORCON can also be used. The newer software tools which are used for analysis accept the data in multiple formats. The problem in the reliability of the data is the possibility of misannotations. The misannotations are some time introduced due to the process of automation of annotation process which are carried out extensively with the help of computers. Misannotations, if introduced, multiplies in subsequent additions and may accumulate to an unbelievable extent and create confusion. A possible solution to prevent this from happening is to flag the protein sequence which has been annotated by sequence comparison but whose function has not been validated by experimental methods.

## Protein Sequence Analysis

Apart from maintaining the large database, mining useful information from these set of primary and secondary databases is very important. Lot of efficient algorithms have been developed for data mining and knowledge discovery. These are computation intensive and need fast and parallel computing facilities for handling multiple queries simultaneously. It is these search tools that integrate the user and the databases. One of the widely used search program is BLAST (Basic Local Alignment Search Tool)

BLAST is a set of similarity search programs designed to explore all of the available sequence databases regardless of whether the query is protein or DNA. The BLAST programs have been designed for speed, with a minimal sacrifice of sensitivity. The scores assigned in a BLAST search have a well-defined statistical interpretation, making real matches easier to distinguish from random background hits. BLAST uses a heuristic algorithm which seeks local as opposed to global alignments and is therefore able to detect relationships among sequences which share only isolated regions of similarity. This is a primary criterion in sequence analysis. Other tool available includes CLUSTALW for multiple sequence alignment.


*Figure 6. BLAST - Basic Local Alignment Search Tool*

**Programs available for the BLAST search include the following**

- **blastp** compares an amino acid query sequence against a protein sequence database
- **blastn** compares a nucleotide query sequence against a nucleotide sequence database
- **blastx** compares a nucleotide query sequence translated in all reading frames against a protein sequence database
- **tblastn** compares a protein query sequence against a nucleotide sequence database dynamically translated in all reading frames
- **tblastx** compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. Note that tblastx program cannot be used with the nr database on the BLAST Web page.

**Databases available for BLAST search include the following**

*Protein Sequence Databases*

- **nr** All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF
- **month** All new or revised GenBank CDS translation+PDB+SwissProt+PIR+PRF released in the last 30 days.
- **swissprot** Last major release of the SWISS-PROT protein sequence database (no updates)
- **Drosophila genome** Drosophila genome proteins provided by Celera and Berkeley Drosophila Genome Project (BDGP). (www.fruitfly.org)
- **yeast** Yeast (Saccharomyces cerevisiae) genomic CDS translations
- **ecoli** Escherichia coli genomic CDS translations
- **pdb** Sequences derived from the 3-dimensional structure from Brookhaven Protein Data Bank (www.pdb.org)
- **kabat** Kabat's database of sequences of immunological interest (http://immuno.bme.nwu.edu)
- **alu** Translations of select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences.

*Nucleotide Sequence Databases*

- **nr** All GenBank+EMBL+DDBJ+PDB sequences (but no EST, STS, GSS, or phase 0, 1 or 2 HTGS sequences). No longer "non-redundant".
- **month** All new or revised GenBank+EMBL+DDBJ+PDB sequences released in the last 30 days.
- **Drosophila genome** Drosophila genome provided by Celera and Berkeley Drosophila Genome Project )
- **dbest** Database of GenBank+EMBL+DDBJ sequences from EST Divisions
- **dbsts** Database of GenBank+EMBL+DDBJ sequences from STS Divisions
- **htgs** Unfinished High Throughput Genomic Sequences: phases 0, 1 and 2
- **gss** Genome Survey Sequence, includes single-pass genomic data, exon-trapped sequences, and Alu PCR sequences.
- **yeast** Yeast (Saccharomyces cerevisiae) genomic nucleotide sequences
- **E. coli** Escherichia coli genomic nucleotide sequences
- **pdb** Sequences derived from the 3-dimensional structure from Brookhaven Protein Data Bank
- **kabat** Kabat's database of sequences of immunological interest
- **vector** Vector subset of GenBank(R), NCBI, in ftp://ncbi.nlm.nih.gov/blast/db/
- **mito** Database of mitochondrial sequences
- **alu** Select Alu repeats from REPBASE, suitable for masking Alu repeats from query sequences. It is available by anonymous FTP from ncbi.nlm.nih.gov (under the /pub/jmc/alu directory).
- **Epd** Eukaryotic Promotor Database found on the web at http://www.genome.ad.jp/dbget-bin/www_bfind?epd

An example of how protein sequence analysis can be useful in diagnosing disease is shown here.

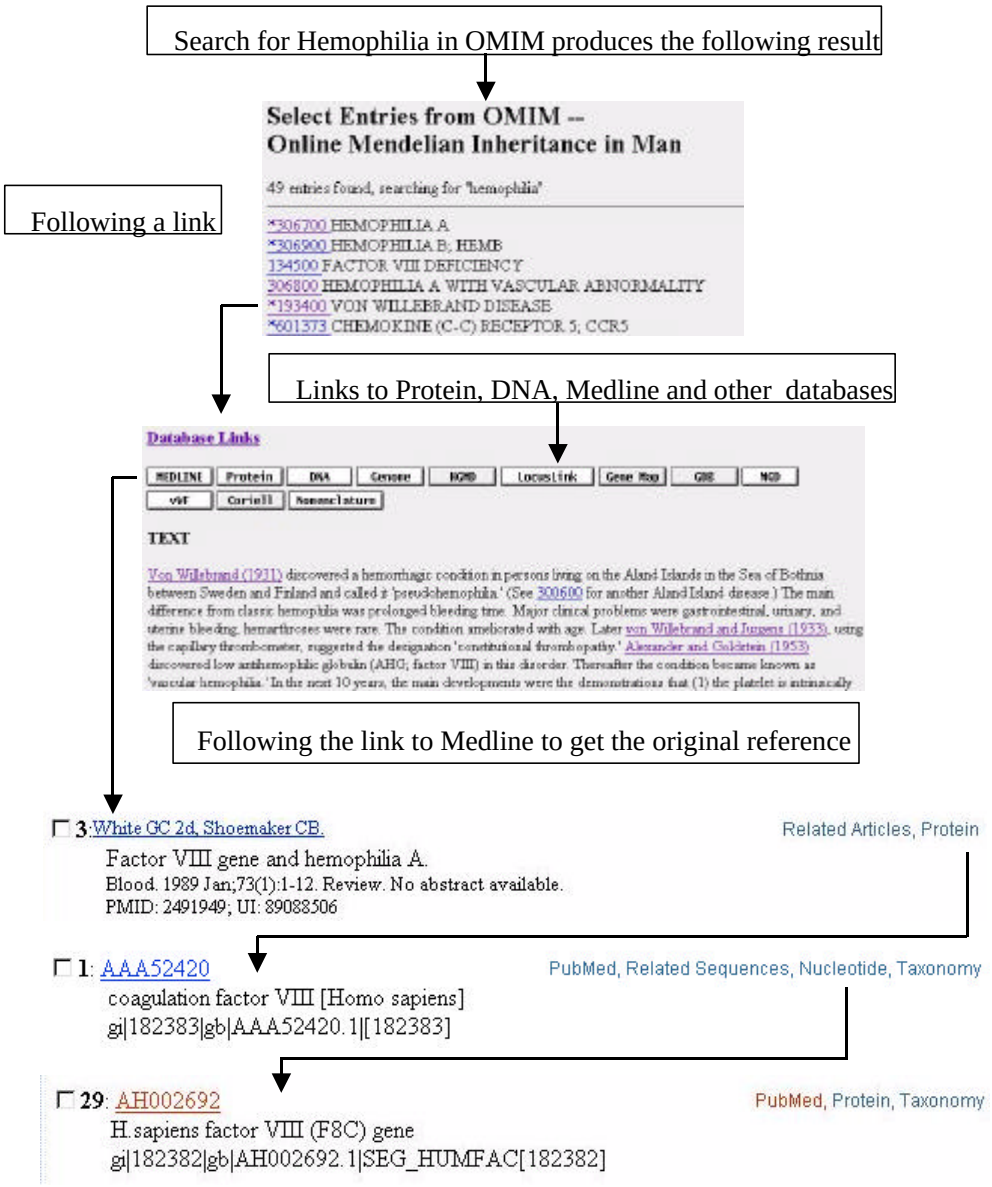



*Figure 7. Application of Bioinformatics in Medical Science*

## Conclusion

The present challenge is to handle a huge volume of data, such as the ones generated by the human genome project, to improve database design, develop software for database access and manipulation, and device data-entry procedures to compensate for the varied computer procedures and systems used in different laboratories. New academic programs which train students in Bioinformatics are providing them with background in molecular biology and in computer science, including database design and analytical approaches. Realizing the importance of Bioinformatics, the Govt. of India through Dept. of Biotechnology has set up distributed information center & research facility across the country to provide access to the databases and computing facilities for research. There is no doubt that Bioinformatics tools for efficient research will have significant impact in biological sciences and betterment of human lives.

## Assignments

---

- **1 Sequence Retrieval from Swiss-Prot (fasta format)**

1. Get the Protein sequence of the Human Hemoglobin - á Chain
2. Get the Nucleic Acid sequence of Human Hemoglobin - á Chain

- **2 Structure Retrieval from PDB**

1. Get the Crystal structure of the Human - Hemoglobin á Chain
2. Get the structural Classification of Human - Hemoglobin á Chain from SCOP database

- **3 Sequence Analysis (BLAST)**

1. Get the sequences of the Hemoglobin á Chain from 5 different sources
2. Do a Multiple Sequence Alignment
3. Analyze the conserved regions

- **4 Structure Analysis (using WEBLAB)**

1. Display the Protein in the 'Solid Ribbon' mode
2. Mark the conserved regions on the structure
3. Display them in the Ball and Stick Model

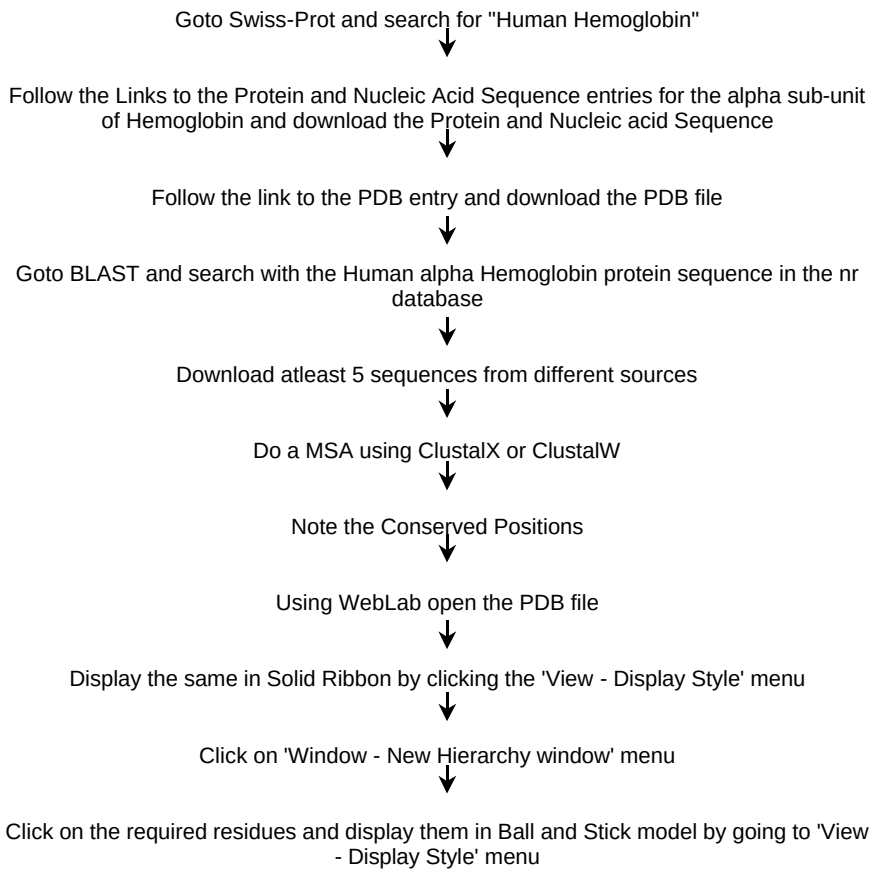
**Site address of the Databases:**

| | | |
|---|---|---|
| Swiss-Prot | - | http://www.expasy.ch |
| PDB | - | http://www.pdb.org |
| BLAST | - | http://www.ncbi.nlm.nih.gov/blast |
| SCOP | - | http://scop.mrc-lmb.cam.ac.uk/scop/ |

---

**Answer key to the Assignments**

---




---