BIOINFORMATICS FOR BIOLOGISTS: A CONCEPT OF DIGITAL LABORATORY

Sachin Shukla , Ratnakar Tripathi, Brij Bharti and Rajnikant Mishra*

Abstract

Bioinformatics has been a fast growing interdisciplinary area which has shown great promise in the advancement of research and development in most of the complex areas of biological sciences. Biodiversity of both eukaryotic and prokaryotic organisms makes us focused on functional analysis of their genomes and proteomes. The large-scale analysis of these genomes and proteomes has started to generate huge amounts of data, and it is likely that all genes and proteins would be available as databases electronically. Management and analysis of large volume of accumulated data and to predict their interactions leading to biological activities have become a challenging task. Use of computer based systems have facilitated the process and has resulted into the evolution of the concept of establishing "Digital Library" and "Bioinformatics" as an emerging branch of science.

Nucleic acid and protein sequence databases, digital library of literature have become important resources and storing mechanisms. Their availability to the scientists across the globe has been possible through computer networking. These information resources are highly useful in agriculture, biodiversity, ecology, evolution, biochemistry, genetics, microbiology, molecular biology and many more. Research and development in "Bioinformatics" have improved the techniques of organization of biological information leading to the growth of tools and databases. Therefore, it is necessary for the biologists to have a good understanding of some of the basic bioinformatical tools and techniques which will make their application easier. This review focuses on the concepts of a digital laboratory in the form of bioinformatical tools and their applications.

Introduction

Over the past few decades rapid developments in sophisticated instruments and modern techniques have produced tremendous amount of information (data & literature) related to in general and in biological sciences, genomics and proteomics in particular. The use of computer based technology made easy storage of huge data (databases), quick retrieval of specific data from the library (data retrieval) and their detailed analysis (data analysis) which appeared to be impossible if done manually. It emerged into a new branch of science, "Bioinformatics'. The term bioinformatics was coined by Paulien Hogeweg in 1979 for the study of informatic processes in biotic systems. Since its introduction to the field of science, it has occupied a large number of biological areas where it has emerged as an essential tool for the storage, retrieval and analysis of data. Its primary use began since late 1980s in genomics and genetics, particularly in those areas of genomics involving large-scale DNA sequencing. Bioinformatics now entails the creation and advancement of databases, algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data. It is the name given to these mathematical and computing approaches used to develop

Department of Zoology, Banaras Hindu University, Varanasi-221005, India

understanding of biological processes. Common activities in bioinformatics include mapping and analyzing DNA and protein sequences, alignment of different DNA and protein sequences to compare them and creating and viewing 3-D models of protein structures. The primary goal of bioinformatics is to increase our understanding of biological processes. What sets it apart from other approaches, however, is its focus on developing and applying computationally intensive techniques (e.g., pattern recognition, data mining, machine learning algorithms, and visualization) to achieve this goal. Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, genome-wide association studies and the modeling of evolution.

The scope of bioinformatics is not limited to a particular area or a branch of biological sciences rather it has extended its branches to almost all major areas including genomics, transcriptomics, proteomics, systems biology, functional genomics, metabolomics, structural genomics, nutritional genomics, cheminformatics, phylogenetics and pharmacoinformatics. Hence, for the first time users of bioinformatics, it is important to have a basic knowledge of computers (CPU, I/O units), operating systems (Windows, Linux and UNIX), networks (LAN and WAN) and information technology. It is also beneficial to have basics of home-pages, web-pages and uniform resource locators (URLs).

1. Basics of computer and internet operations: Internet Protocols (TCP/IP)

A protocol is a set of rules that allows the orderly exchange of information. The interconnection of networks is known as internetworking (or an internet). Each part of an internet is a subnetwork (or subnet) and Transmission Control Protocol (**TCP**) and Internet Protocol (**IP**) are a pair of protocols that allow one subnet to communicate with another. The IP part corresponds to the network layer of the Open System Interconnection (OSI) model and the TCP part to the transport layer. Their operation is transparent to the physical and data link layers and can thus be used on Ethernet, Fibre Distributed Data Interface (FDDI) or Token Ring networks. The address of the Data Link layer corresponds to the physical address of the node, such as the Media Access Control (MAC) address (in Ethernet and Token Ring) or the telephone number (for a modem connection). The IP address is assigned to each node on the internet, and is used to identify the location of the network and any subnets. **TCP/IP** was originally developed by the US Defense Advanced Research Projects Agency (DARPA), and its objective was to connect a number of universities and other research establishments to DARPA. Such resultant network is now known as the Internet which uses TCP/IP to transfer data. Any organization can have its own intranets to connect to the Internet and the addresses must conform to the Internet addressing format. Common applications that use TCP/IP communications are remote login and file transfer. File transfer protocol (ftp) for file transfer and telnet allows remote login into another computer using TCP/IP.

As with all other communications protocol, TCP/IP is composed of layers. The **IP** layer is responsible for moving packet of data from node to node. Each packet is based on a four byte destination address called the IP address. The **TCP** layer is responsible for verifying the correct delivery of data from client to server. In addition to these the TCP/IP model comprises following

layers: a) Network access layer specifies the procedures for transmitting data across the network, including how to access the physical medium, such as Ethernet and FDDI; b) Internet layer is responsible for data addressing, transmission, and packet fragmentation and reassembly; c) Transport layer manages all aspects of data routing and delivery including session initiation, error control and sequence checking and d) Application layer is responsible for everything else. Applications must be responsible for all the presentation and part of the session layer (Figure 1).

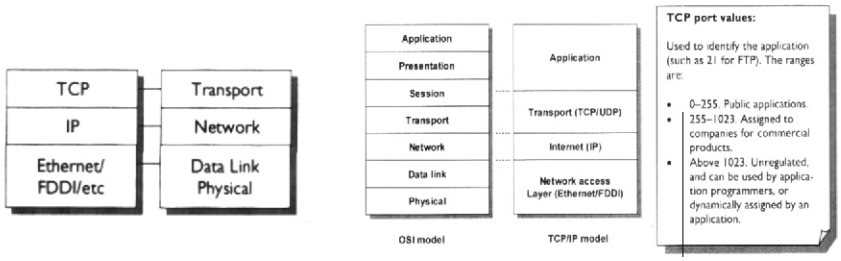

Fig. - 1 TCP/IP and OSI Model

2. Scientific Data archiving systems, Accession and GenInfo Identifier (GI) numbers

Scientific data archiving refers to the long-term storage of scientific data and methods. The various scientific journals have differing policies regarding how much of their data and methods are required to be stored in a public archive. Besides, what is actually archived varies widely between different disciplines. Similarly, the major grant-giving institutions have varying attitudes towards public archival of data. In general, the tradition of science has been for publications to contain sufficient information to allow fellow researchers to replicate and therefore test the results. In recent years this approach has become strained as increasingly research in certain areas depends on large datasets which cannot easily be replicated independently. Data archiving is more important in some fields than others. In a few fields, all of the data necessary to replicate the work is already available in the journal article. In drug development, a great deal of data is generated and must be archived so researchers can verify that the reports that drug companies publish accurately reflect the data.

The Nature and the *Science* follow policies and require the reviewer to determine if all of the supplementary data and methods have been archived. The policy advises reviewers to consider several questions, including: "Should the authors be asked to provide supplementary methods or data to accompany the paper online? (Such data might include source code for modelling studies, detailed experimental protocols or mathematical derivations.)" The *Science* supports the efforts of databases that aggregate published data for the use of the scientific community. Therefore, before publication, large data sets (including microarray data, protein or DNA sequences, and atomic coordinates or electron microscopy maps for macromolecular structures) must be deposited in an approved database and an accession number provided for inclusion in the published paper. Some of the major data archives are: National Archive of Computerized Data on Aging, National

Climatic Data Center, National Geophysical Data Center, National Snow and Ice Data Center, National Oceanographic Data Center, International Tree-Ring Data Bank, ESO/ST-ECF Science Archive Facility, CISL Research Data Archive, World Data Center

Accession and GI numbers

These are the two types of sequence identification numbers which have different formats and are applied at different time points.**GI** (GenInfo Identifier) **number** (sometimes written in lower case, "**gi**") is simply a series of digits that are assigned consecutively to each sequence record processed by NCBI. The GI number bears no resemblance to the Accession number of the sequence record. The nucleotide sequence GI number is shown in the VERSION field of the database record, whereas the protein sequence GI number is shown in the CDS/db_xref field of a nucleotide database record, and the VERSION field of a protein database record. The GI number has been used for many years by NCBI to track sequence histories in GenBank and the other sequence databases it maintains. The **Accession** number system of identifiers was adopted in February 1999 by the International Nucleotide Sequence Database Collaboration (GenBank, European Molecular Biology Laboratory (EMBL), and DNA Data Bank of Japan (DDBJ)). The two systems of identifiers run in parallel to each other. That is, when any change is made to a sequence, it receives a new GI number and an increase to its accession number. The accession number is an unique identifier for a sequence record. An accession number applies to the complete record and is usually a combination of a letter(s) and numbers, such as a single letter followed by five digits (e.g., U12345) or two letters followed by six digits (e.g., AF123456). Some accessions might be longer, depending on the type of sequence record. Accession numbers do not change, even if information in the record is changed at the author's request. Sometimes, however, an original accession number might become secondary to a newer accession number, if the authors make a new submission that combines previous sequences, or if for some reason a new submission supercedes an earlier record. The examples of some of the accession numbers are: NT_123456 constructed genomic contigs, NM_123456 mRNAs, NP_123456 proteins, NC_123456 chromosomes.

The first two letters followed by an underscore now are often used as a curate for combined sequences that correspond to a single molecular type, e.g., **AC_**(Alternate genomic), **AP_**(Alternate protein), **NC_** (chromosome), **NG_** (genomic regions), **NM_** (mRNA), **NP_** (protein), **NR_**(Non-coding RNAs), **NT_** (genomic contigs), **NW_**(Genomic BACS), **NZ_**(Shotgun Genome), **XM_**(modeled mRNA), **XP_**(modeled protein), **XR_**(modeled non-coding transcripts), **YP_** (Proteins without transcripts), and **ZP_**(Proteins annotated from NZ_).

3. Biological Databases and their Applications

Biological databases are developed to perform systemization of results from biological experiments so as to make the data available to the scientific community for the analysis and interpretation. There are a number of biological databases most of which are publically available, whereas some of them are available with copyright or commercially available as follows:

1. **Sequence databases** which can be further be divided into nucleotide sequence databases (NCBI, GenBank, RNAdb, EMBL, Ensembl, MOT, Alternative splicing) and protein sequence

databases (NCBI, Swiss-Prot, TrEMBL, ExPasy),

2. **Structure databases** (PDB, NDB, SCOR, MMDB, FSSP, DALI, M-fold).
3. **Genome databases (**Human-HGMD and HGDB, Mouse-MGI and MGD**,** Yeast-SGD, *C. elegans*-AceDB**,** Drosophila-FlyBase**)**
4. **Protein sequence, structures and interacting proteins database** (PIP, DIP, STRING, BIND, Prosite, Pfam, SCOP, CATH, PDB)
5. **Microarray/ Transcriptome databases** (GEO, ArrayExpress, CGED)
6. **Metabolic and Enzyme databases (**KEGG, BRENDA, MetaCyc, and REBASE**)**
7. **Literature databases (**PubMed, MEDLINE**)**
8. **Disease databases (**OMIM, SNPdb, HGMD**)**
9. **Chemical databases (**TOXNET**)**
10. **Biodiversity and ecosystem based databases (**WBD, EUNIS**)**

3.1 Nucleic acid sequence databases

Some of the important nucleic acid sequence databases include: NCBI, GenBank, RNAdb, EMBL and Ensembl. These are summarized below.

NCBI : Established in 1988 as a national resource for molecular biology information, NCBI **(National Centre for Biotechnology Information)** creates public databases (18), conducts research in computational biology, develops software tools for analyzing genome data, and disseminates biomedical information - all for the better understanding of molecular processes affecting human health and disease (Figure 2A).

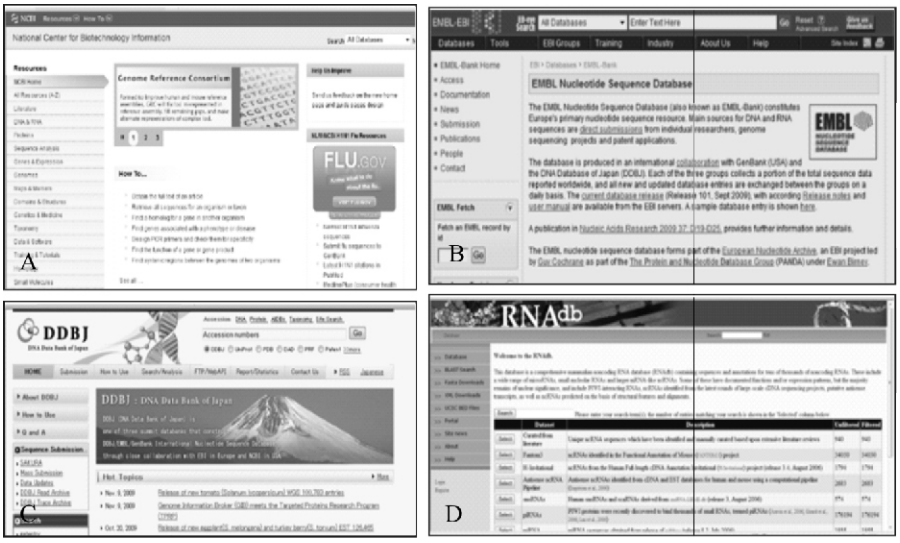

Fig. - 2 Home pages of various nucleotide sequence databases. A. NCBI, B. EMBL, C. DDBJ and D. RNAdb

GenBank : GenBank® is the NIH genetic sequence database, an annotated collection of all publicly available DNA sequences. There are approximately 106,533,156,756 bases in 108,431,692 sequence records in the traditional GenBank divisions and 148,165,117,763 bases in 48,443,067 sequence records in another division as of August 2009. The reported sequences can be explored whereas the newly discovered ones can be submitted authentically to the GenBank.

EMBL : The EMBL (European Molecular Biology Laboratory) Nucleotide Sequence Database (also known as EMBL-Bank) constitutes Europe's primary nucleotide sequence resource (Figure 2B). Main sources for DNA and RNA sequences are direct submissions from individual researchers, genome sequencing projects and patent applications (7).

DDBJ : DNA Data Bank of Japan (DDBJ) is the sole nucleotide sequence data bank in Asia, which is officially certified to collect nucleotide sequences from researchers and to issue the internationally recognized accession number to data submitters. Since it exchanges the collected data with EMBL-Bank/EBI (European Bioinformatics Institute) and GenBank/NCBI (National Center for Biotechnology Information) on a daily basis, the three data banks share virtually the same data at any given time. The virtually unified database is called "the International Nucleotide Sequence Database (INSD)". DDBJ collects sequence data mainly from Japanese researchers, but of course accepts data and issue the accession number to researchers in any other countries (Figure 2C).

RNA Database : This database is a comprehensive mammalian noncoding RNA (ncRNA) database (RNAdb) containing sequences and annotations for tens of thousands of noncoding RNAs (15).These include a wide range of microRNAs, small nucleolar RNAs and larger mRNA-like ncRNAs. Some of these have documented functions and/or expression patterns, but the majority remains of unclear significance, and include interacting RNAs, ncRNAs identified from the latest rounds of large-scale cDNA sequencing projects, putative antisense transcripts, as well as ncRNAs predicted on the basis of structural features and alignments (Figure 2D).

3.2 Genome databases

They include sequence informations derived from the genome projects of different organisms. Human Gene Mutation Database (HGMD), Human Gene Disease database and GeneCards for human; Mouse Genome Informatics (MGI) and Mouse Genome Database (MGD) for mouse, Saccharomyces Genome Database (SGD) for yeast, AceDB for *C. elegans*, and FlyBase for Drosophila (Figure 3 A-D). FlyBase: It is an online bioinformatics database of the biology and genome of the model organism *Drosophila melanogaster* and related Drosophilid dipterans. The FlyBase project is carried out by a consortium of Drosophila researchers and computer scientists at Harvard University and Indiana University in the United States, and University of Cambridge in the United Kingdom. It is one of the organizations contributing to the Generic Model Organism Database (GMOD). It contains a complete annotation of the *Drosophila melanogaster*. It also includes a searchable bibliography of research on Drosophila genetics in the last century. Information on current researchers, and a partial pedigree of relationships between current researchers, is searchable, based on registration of the participating scientist. It also provides

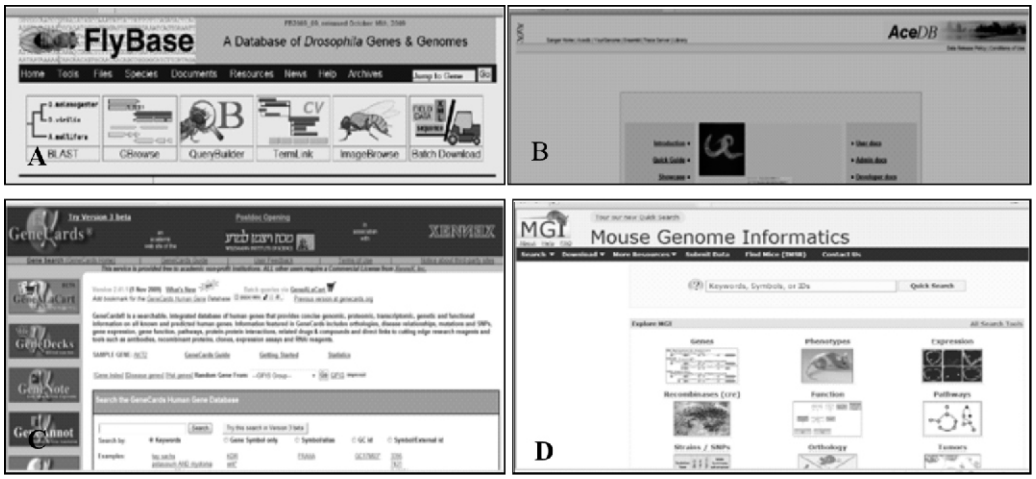

Fig.. - 3 Genome databases **A.** FlyBase (*Drosophila*), **B.** AceDB (*C. elagans*), **C.** GeneCards (Human) and **D.** MGI (Mouse).

a large database of images illustrating the full genome, and several movies detailing embryogenesis (5).

AceDB : AceDB is a genome database system developed since 1989 primarily by Jean Thierry-Mieg (CNRS, Montpellier) and Richard Durbin (Sanger Institute). It was originally developed for the *C.elegans* genome project, from which its name was derived (A C. elegans DataBase). However, the tools in it have been generalized to be much more flexible and the same software is now used for many different genomic databases from bacteria to fungi to plants to man. It is also increasingly used for databases with non-biological content. It provides a custom database kernel, with a non-standard data model designed specifically for handling scientific data flexibly, and a graphical user interface with many specific displays and tools for genomic data. AceDB is used both for managing data within genome projects, and for making genomic data available to the wider scientific community. The AceDB software is primarily developed to run under the Unix operating system, using X-Windows for graphics, with a local copy of the database files. There is also a version of AceDB for Microsoft Windows which, likewise, runs with a local copy of the database.

3.3. Microarray/ Transcriptome databases

GEO: Gene Expression Omnibus is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted (6). Tools are provided to help users query and download experiments and curated gene expression profile (Figure 4A).

ArrayExpress : ArrayExpress is a public archive for functional genomics data compliant with MIAME- and MINSEQE requirements in accordance with compliant data in accordance with MGED recommendations. The Gene Expression Atlas uses curated, re-annotated subset of data from the Archive to provide information about gene expression under various biological conditions (Figure 4B).

Fig. - 4 Web-pages of some transcriptome databases A. Gene Expression Omnibus (GEO) and B. ArrayExpress

3.4. Protein Sequence and interacting protein databases

Protein sequence databases play a vital role as a central resource for storing the data generated by experiments and more conventional efforts, and making them available to the scientific community. A number of important databases are based on the sequence and structure information deposited and archived in the protein data bank (Figure 5).

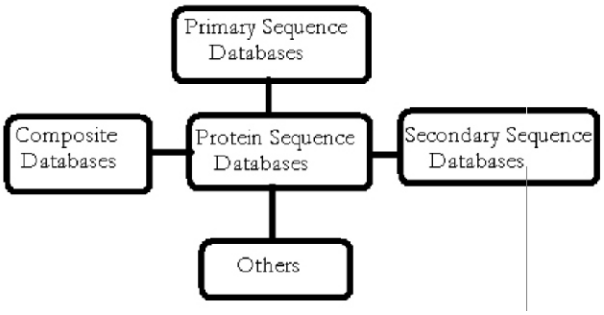

Fig. - 5 Flowchart representing the various types of Protein databases

World wide Protein Data Bank (wwPDB) is the main repository of the protein sequences and associated informations (2). This is a single archive of macromolecular structural data that is freely and publicly available to the global community (Figure 6A). Some of the main protein sequence databases under the wwPDB are RCSB-PDB (Research Collaboratory for Structural Bioinformatics –Protein Data Bank, PDBe (Protein Data Bank in Europe) and PDBj (Protein Data Bank Japan).

Fig. - 6 Home Pages of various protein data banks including: **A**.World wide Protein Data Bank, **B**. RCSB-PDB, **C**. PDBe and **D.** PDBj.

Three main contributors of the world-wide protein data bank are as follows:

1. **RCSB-PDB:** The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the world wide Protein Data Bank, the RCSB PDB curates and annotates PDB data according to agreed upon standards. The RCSB PDB also provides a variety of tools and resources (Figure 6B). Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

2. **PDBe:** EBI Protein Structure Database in Europe, is a project for the collection, management and distribution of data about macromolecular structures, derived from the Protein Data Bank (PDB) (17). The PDBe is one of the founding members of Worldwide Protein Data Bank (wwPDB) which consists of organizations that act as deposition, data processing and distribution centers for PDB data (Figure 6C).

3. **PDBj:** PDBj (Protein Data Bank Japan) maintains a centralized archive of macromolecular structures and provides integrated tools, in collaboration with the RCSB in USA and the PDBe in EU. PDBj is supported by JST-BIRD (Figure 6D) (10).

NCBI protein sequence databse

It can be accessed by typing the URL **http://www.ncbi.nlm.nih.gov/** in the address bar of the internet explorer (Figure 7A).

Click on the caption **All Databases.** A new window containing the list of all the available databases in the NCBI will be opened (Figure 7B).Click on the **Protein: Sequence database.**Type the name of protein of interest in the search bar provided adjacent to the Protein, and press "Enter" or click on "Go". This action will provide a list of all the protein sequences for a protein of interest in various categories of animals submitted in the NCBI database till date for a protein of interest. For example:Pax6 (Figure 7C).

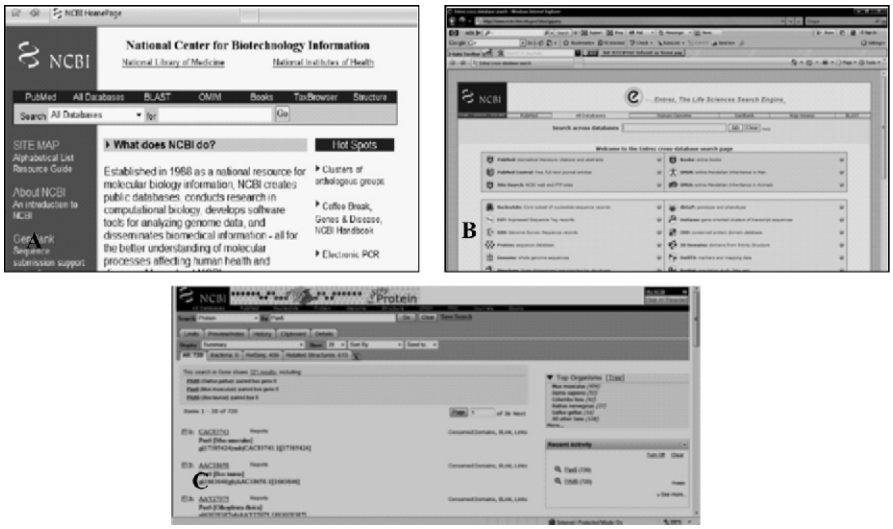

Fig. - 7 Accessing NCBI protein sequence database.

Alignment of Sequences

In bioinformatics, a **sequence alignment** is a way of arranging the sequences of DNA, RNA, or protein to identify regions of similarity that may be a consequence of functional, structural, or evolutionary relationships between the sequences. Aligned sequences of nucleotide or amino acid residues are typically represented as rows within a matrix. Gaps are inserted between the residues so that identical or similar characters are aligned in successive columns. Sequence alignments can be stored in a wide variety of text-based file formats, many of which were originally developed in conjunction with a specific alignment program or implementation. Most web-based tools allow a limited number of input and output formats, such as FASTA format and GenBank format and the output is not easily editable. Several conversion programs are available for conversion of results.

A DNA, RNA or protein sequence can be aligned in following ways:

1. **Global and Local Alignment**: **G**lobal alignments, which attempt to align every residue in every sequence, are most useful when the sequences in the query set are similar and of roughly equal size. A general global alignment technique is the Needleman-Wunsch algorithm, which is based on dynamic programming. **L**ocal alignments are more useful for dissimilar sequences that are suspected to contain regions of similarity or similar sequence motifs within their larger sequence context. The Smith-Waterman algorithm is a general local alignment method also based on dynamic programming. With sufficiently similar sequences, there is no difference between local and global alignments.

2. **Pairwise Alignment**: Pairwise sequence alignment methods are used to find the best-matching piecewise (local) or global alignments of two query sequences. Pairwise alignments can only be used between two sequences at a time, but they are efficient to calculate and are often used for methods that do not require extreme precision (such as searching a database for sequences with high homology to a query). The three primary methods of producing pairwise alignments

are dot-matrix methods, dynamic programming, and word methods. e.g., Basic Local Alignment Search Tool (BLAST) (**http://blast.ncbi.nlm.nih.gov/Blast**) (Figure 8A-C) (1).

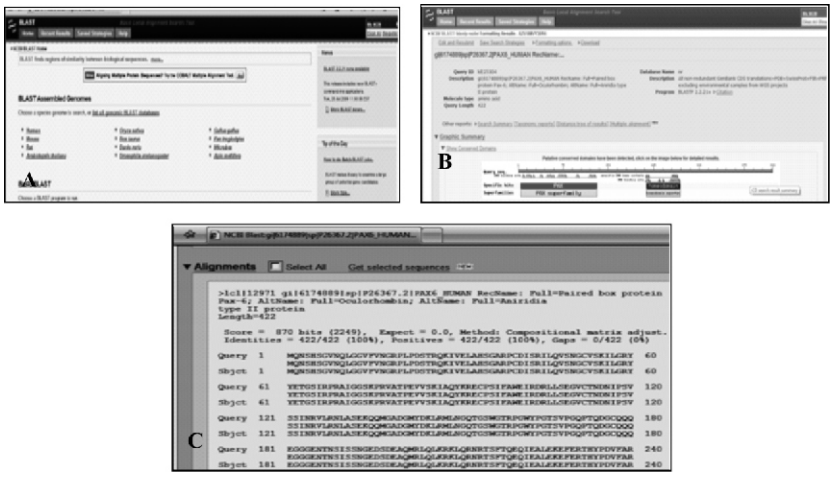

Fig. - 8 Steps of protein sequence alignment using NCBI-BLAST.

3. **Multiple-sequence alignment**: Multiple sequence alignment is an extension of pairwise alignment to incorporate more than two sequences at a time. Multiple alignment methods try to align all of the sequences in a given query set. Multiple alignments are often used in identifying conserved sequence regions across a group of sequences hypothesized to be evolutionarily related. Such conserved sequence motifs can be used in conjunction with structural and mechanistic information to locate the catalytic active sites of enzymes. Alignments are also used to aid in establishing evolutionary relationships by constructing phylogenetic trees. Nevertheless, the utility of these alignments in bioinformatics has led to the development of a variety of methods suitable for aligning three or more sequences. Some of the commonly used softwares are ClustalW (**http://www.ebi.ac.uk/Tools/clustalw2/**) and T-Coffee (**http://www.ebi.ac.uk/t-coffee/**)

4. **Structural alignment** : Structural alignments, which are usually specific to protein and sometimes RNA sequences, use information about the secondary and tertiary structure of the protein or RNA molecule to aid in aligning the sequences. These methods can be used for two or more sequences and typically produce local alignments; however, because they depend on the availability of structural information, they can only be used for sequences whose corresponding structures are known (usually through X-ray crystallography or NMR spectroscopy). Structural alignments are used as the "gold standard" in evaluating alignments for homology-based protein structure prediction because they explicitly align regions of the protein sequence that are structurally similar rather than relying exclusively on sequence information. e.g., DALI (Distance matrix alignment; **http://ekhidna.biocenter.helsinki.fi/dali**) which is used for constructing the FSSP (Families of structurally similar proteins) structural alignment database and SSAP (Sequential Structure Alignment Program) which is used in construction of CATH (Class, Architechture, Topology, Homology) database **http://www.cathdb.info/.**

5. **Phylogenetic Analysis:** Phylogenetics and sequence alignment are closely related fields due to the shared necessity of evaluating sequence relatedness. The field of phylogenetics makes extensive use of sequence alignments in the construction and interpretation of phylogenetic trees, which are used to classify the evolutionary relationships between homologous genes represented in the genomes of divergent species. The degree to which sequences in a query set differ is qualitatively related to the sequences' evolutionary distance from one another. Roughly speaking, high sequence identity suggests that the sequences in question have a comparatively young most recent common ancestor, while low identity suggests that the divergence is more ancient. PHYLIP (http://www.phylip.com/) is commonly used software for phylogenetic analysis. Pairwise alignment can be performed using BLAST (Figure 8A-C), EMBOSS (a European Molecular Biology Laboratory tool) (Figure 9A-B). Similarly multiple sequence alignment is explored by EMBL-EBI (European Bioinformatics Institute)'s T-Coffee program (http://www.ebi.ac.uk/t-coffee/) (Figure 9B) and ClustalW (http://www.ebi.ac.uk/Tools/clustalw2).

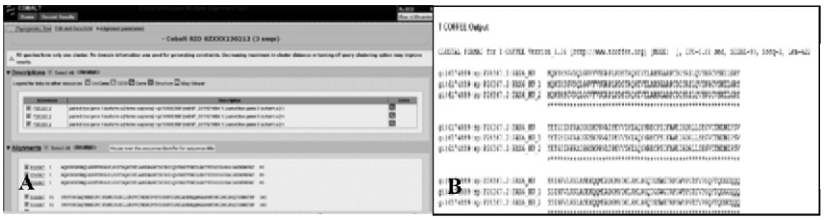

Fig. - 9 Results of multiple sequence alignment using **A.** COBALT (NCBI) and **B**. T-Coffee (EMBL)

3.5 Structure Databases

The structure databases provide information and analysis of structural features of DNA, RNA and proteins. Some of the major ones are PDB, Nucleic acid database (NDB), SCOR, Molecular Modelling Database (MMDB), FSSP, DALI, M-fold. If PDB-Id is known (e.g., PDB-Id for "PAX6" is 6pax) informations in the database will be displayed by using **(http://www.rcsb.org/pdb/home/home.do)** and visualized by **Jmol** molecular visualization software (Figure 10 A-C).

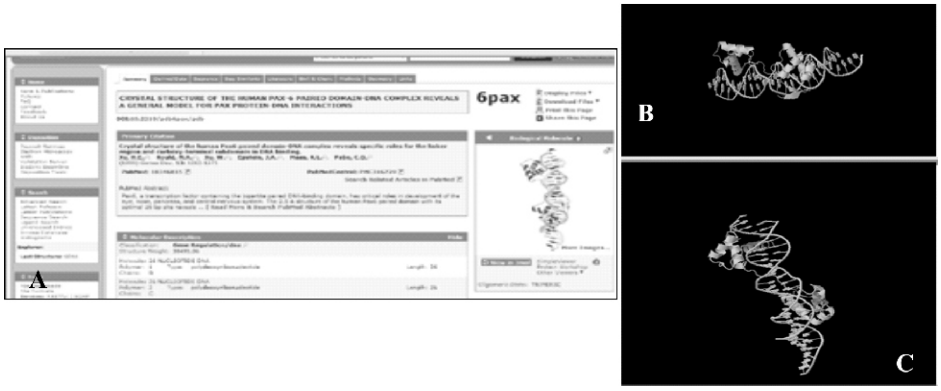

Fig. - 10 Accessing the PDB for exploring the structure of PAX6, a transcription factor.
A. Results for the PDB-Id-6pax, **B-C**. Images of reported crystal structure of PAX6 in Jmol, molecular visualization software.

A. Molecular Modelling and Visualization of Proteins

Molecular modeling is a collective term that refers to theoretical methods and computational techniques to model or mimic the behavior of molecules. The techniques are used in the fields of computational chemistry, computational biology and materials science for studying molecular systems ranging from small chemical systems to large biological molecules and material assemblies. The Molecular modeling methods are now routinely used to investigate the structure, dynamics and thermodynamics of inorganic, biological, and polymeric systems. The types of biological activity that have been investigated using molecular modeling include protein folding, enzyme catalysis, protein stability, conformational changes associated with biomolecular function, and molecular recognition of proteins, DNA, and membrane complexes. Some of the commonly used molecular modeling softwares are: MMDB (Molecular Modelling Database http://www.ncbi.nlm.nih.gov/Structure/MMDB/mmdb.shtml), InsightII (Now part of Discovery studio software) and WhatIF (URL:_**http://swift.cmbi.kun.nl/whatif/**).

The Molecular Modeling Database (MMDB) contains 3D macromolecular structures, including proteins and polynucleotides. MMDB contains over 40,000 structures and is linked to the rest of the NCBI databases, including sequences, bibliographic citations, taxonomic classifications, and sequence and structure neighbors (14). Following actions can be performed using MMDB. There are two conditions for retrieving 3D structures for a gene or protein product of interest:**A.**If the 3D structure of a protein of interest is already resolved it can be visualized by Cn3D 4.1 molecular visualization software or in RasMol (e.g., Structure view in Cn3D or structure view in RasMol) (Figure 11A-B). **B.** If the appropriate information about the 3-D structure of a

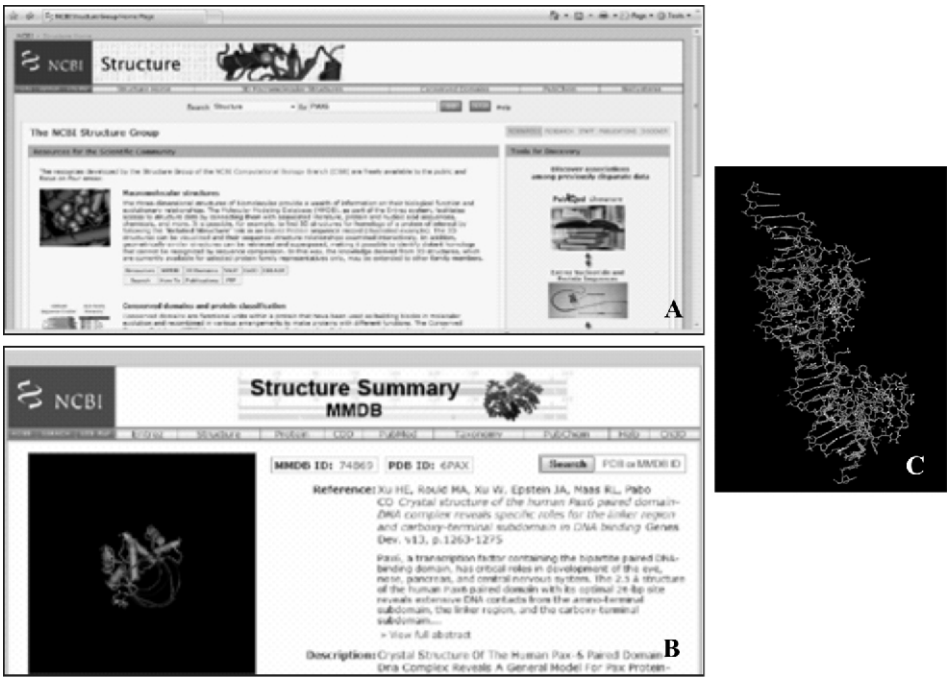

Fig. - 11 Retrieving 3D structures using MMDB. **A.** Home page of MMDB, **B.** Results of search for the structure of PAX6 and **C** The 3D structure of PAX6 in RasMol visualization software.

protein of interest is not clearly known Related Structures can be visualized accordingly by RasMol visualization software (Figure 11C).

B. Secondary Structure Prediction

Secondary structure prediction is a set of techniques in bioinformatics that aim to predict the local secondary structures of proteins and RNA sequences based only on knowledge of their primary structure - amino acid or nucleotide sequence, respectively. For proteins, a prediction consists of assigning regions of the amino acid sequence as likely alpha helices, beta strands (often noted as "extended" conformations), or turns. The success of a prediction is determined by comparing it to the results of the DSSP algorithm applied to the crystal structure of the protein. The best modern methods of secondary structure prediction in proteins reach about 80% accuracy(3-4); this high accuracy allows the use of the predictions in fold recognition and *ab initio* protein structure prediction, classification of structural motifs, and refinement of sequence alignments (11). The predict protein is also a useful software are secondary structure prediction (16).

Some of the commonly used protein secondary structure prediction servers are summarized in the figure 12 and they are as follows:

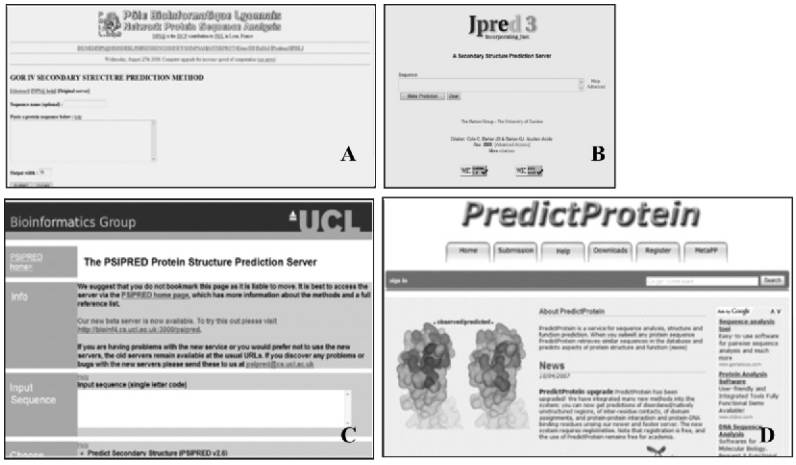

Fig.12. Web-pages of some major secondary structure prediction servers for proteins.
A. GOR IV, B. Jpred, C. PSIPRED and D. PredictProtein

GOR (**http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html**)
Jpred, (**http://www.compbio.dundee.ac.uk/www-jpred/**) (4)
PredictProtein (**http://www.predictprotein.org**/) (16)
PSIPRED (**http://bioinf.cs.ucl.ac.uk/psipred/**). (9)

M-fold

Michael Zuker, professor of mathematical sciences, develops tools for predicting the secondary structure of RNA and DNA, mainly by using thermodynamic methods. Much of his work has been on RNA structure, which is important in understanding many biological processes, including translation regulation in messenger RNA, replication of single-stranded RNA viruses,

and the function of structural RNAs and RNA/protein complexes (Figure 13). His algorithms have been widely used for drug design, and work on DNA folding has been very popular with the biotechnology community. Recent work in his laboratory includes the development of methods to predict folding hybridization and melting curves for two strands of RNA or DNA, and he is developing statistically based rules for RNA folding. His algorithms are available on this website, which is so popular that the server registers as many as 800,000 hits a month. His papers outlining his algorithms get cited almost every day of the year.

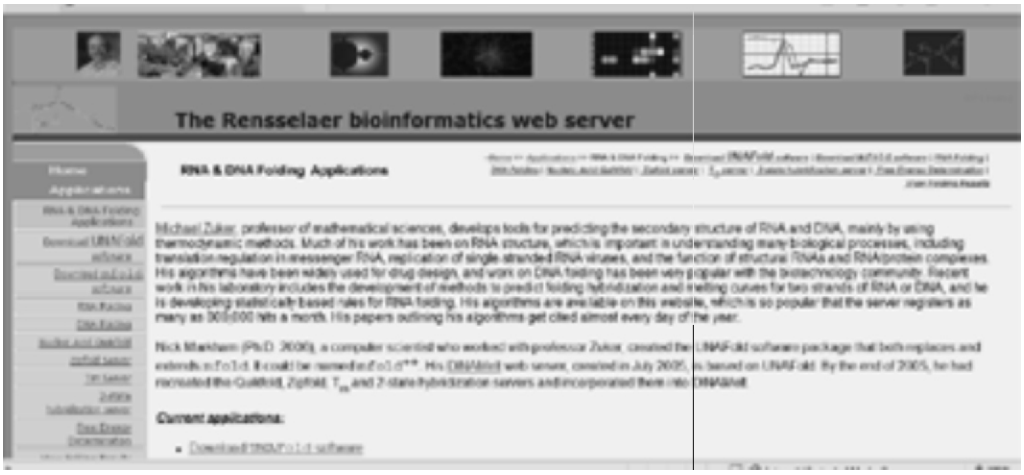

Fig. - 13 M-fold, a software for predicting the secondary structures of nucleic acids (DNA and RNA).

3.6 Metabolic and Enzyme databases

Metabolomics is the "systematic study of the unique chemical fingerprints that specific cellular processes leave behind" - specifically, the study of their small-molecule metabolite profiles. The metabolome represents the collection of all metabolites in a biological organism, which are the end products of its gene expression. Thus, while mRNA gene expression data and proteomic analyses do not tell the whole story of what might be happening in a cell, metabolic profiling can give an instantaneous snapshot of the physiology of that cell. One of the challenges of systems biology and functional genomics is to integrate proteomic, transcriptomic, and metabolomic information to give a more complete picture of living organisms.

KEGG Metabolic Pathway Database

KEGG (Kyoto Encyclopedia of Genes and Genomes) PATHWAY is a collection of manually drawn pathway maps (10) representing our knowledge on the molecular interaction and reaction networks for following (Figure 14A-B):

1. Metabolism
2. Genetic Information Processing
3. Environmental Information Processing
4. Cellular processes
5. Human Diseases

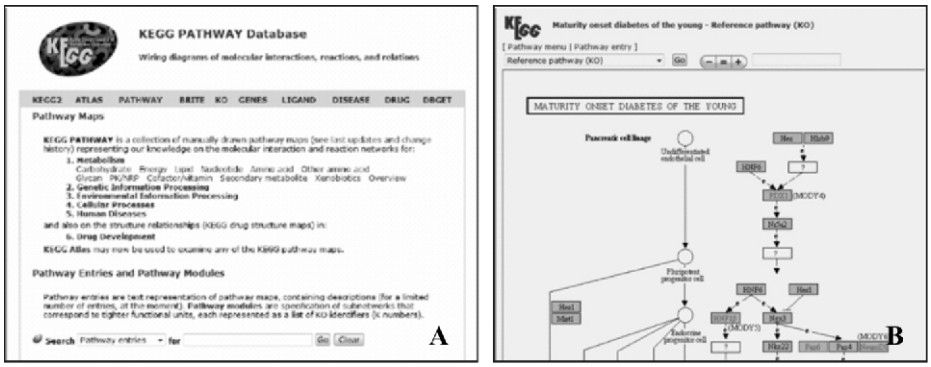

Fig. - 14 Application of KEGG to search for a metabolic pathway**. A.** Home page and **B.** A pathway for diabetes explored with the help of KEGG.

3.7 Literature Databases

PubMed and MEDLINE: PubMed is a service of the US National Library of Medicine that includes over 18 million citations from MEDLINE and other life science journals. MEDLINE® is the world's most comprehensive source of life sciences and biomedical bibliographic information. It contains nearly eleven million records from over 7,300 different publications from 1965 to November 16, 2005. In addition, there are electronic libraries which provide free online access to large number of books and journals.

3.8 Disease Databases and Clinical Informatics

Clinical informatics is the study of information systems (computers and programs) used in the clinical practice of medicine. The following examples demonstrate a few aspects of the field Data entry, Telemedicine, Imaging, Data display and online informations about the diseases (Figure 15A). Some of the commonly used on-line databses related to various diseases and genetic disorders

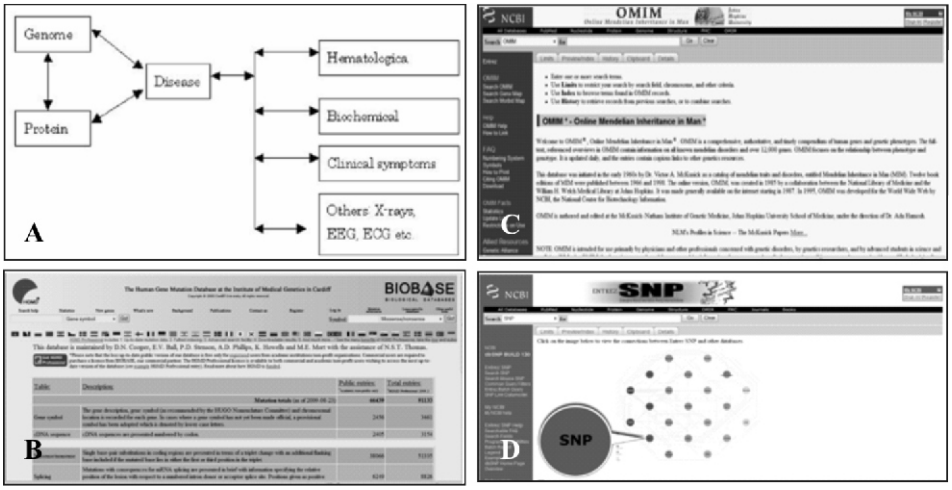

Fig. - 15 Clinical Informatics and Disease databases. **A.** Model for integrated clinical informatics database, B-D: Home pages of some major disease databases, **B.** HGMD, **C.** OMIM and **D.** dbSNP.

include: **OMIM** (Online Mendelian Inheritance in Man), **HGMD** (Human Gene Mutation Database) and **dbSNP** (Database for Single Nuclotide Polymorphism).

OMIM : This database is a catalog of human genes and genetic disorders authored and edited by Dr. Victor A. McKusick and his colleagues at Johns Hopkins and elsewhere, and developed for the World Wide Web by NCBI, the National Center for Biotechnology Information. The database contains textual information, pictures, and reference information (Figure 15C). It also contains copious links to NCBI's Entrez database of MEDLINE articles and sequence information.

HGMD: This constitutes a comprehensive core collection of data on germ-line mutations in nuclear genes underlying or associated with human inherited disease (http://www.hgmd.org). Data cataloged include single base-pair substitutions in coding, regulatory, and splicing-relevant regions, microdeletions and microinsertions, indels, and triplet repeat expansions, as well as gross gene deletions, insertions, duplications, and complex rearrangements. Each mutation is entered into HGMD only once, in order to avoid confusion between recurrent and identical-by-descent lesions.. HGMD includes cDNA reference sequences, now provided for more than 90% of the listed genes, splice junction data, disease-associated and functional polymorphisms, and links to data present in publicly available online locus-specific mutation databases. Accessing HGMD (URL: **http://www.hgmd.cf.ac.uk/**) requires registration through a non-profit domain like @bhu.ac.in, @jnu.ac.in. The home page appears like figure 15B.

dbSNP: This is a NCBI archive of reported single nucleotide polymorphisms in various genes. The home-page can be accessed through the following URL:www.ncbi.nlm.nih.gov/projects/SNP/. The informations about the reported single nucleotide polymorphisms (SNPs) in a particular gene of interest can be accessed by typing the name of gene of interest in the search bar of the database followed by a click on **Go** (Figure 15D).

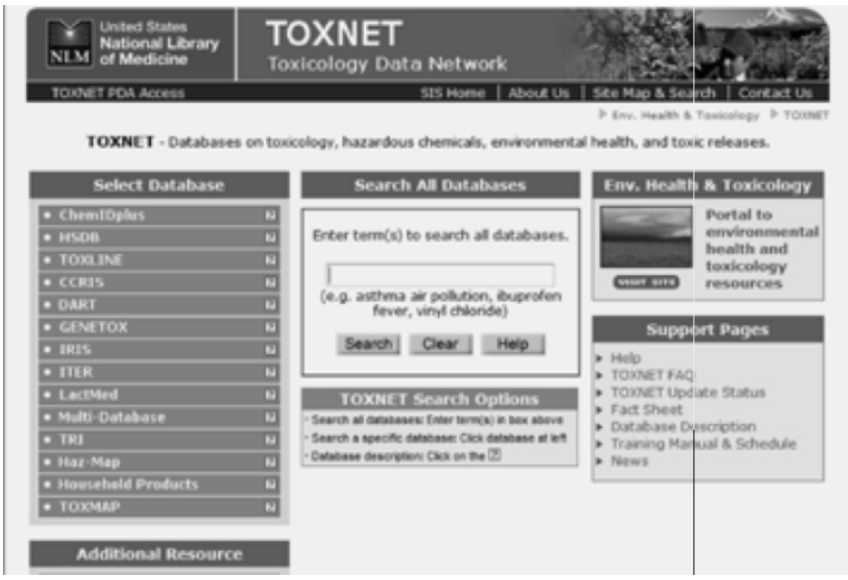

Fig. - 16 TOXNET, a database for toxic and hazardous chemicals.

3.9 Chemical databases

Toxnet: It is a network of databases designed and developed by SIS's Toxicology and Environmental Health Information Program (TEHIP, http://tox.nlm.nih.gov). It is a collection of toxicology and environmental health databases. TOXNET includes the Hazardous Substances Data Bank (HSDB®), a database of potentially hazardous chemicals, TOXLINE® (8) (containing references to the world's toxicology literature), and ChemIDplus® (a chemical dictionary and structure database) (Figure 16).

3.10 Biodiversity and Ecosystem based databases

WBD (World Biodiversity database): The World Biodiversity Database (WBD) is a continuously growing taxonomic database and information system that allows one to search and browse a number of online species banks covering a wide variety of organisms (Figure 17). The 21 species banks accessible through the WBD offer taxonomic information, species names, synonyms, descriptions, illustrations and literature references, as well as online identification keys and interactive geographical information systems. The WBD currently includes 25493 unique taxa, plus 4149 synonyms.

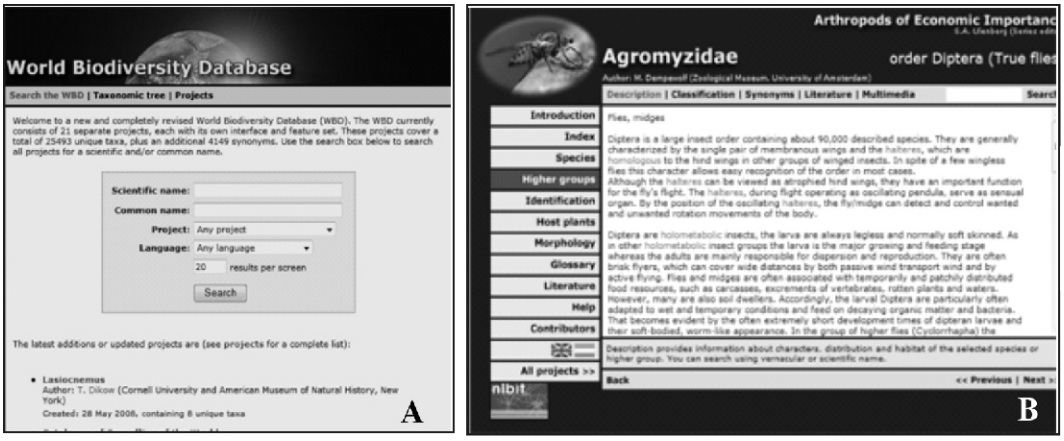

Fig. - 17. WBD, world biodiversity database. **A.** Home-page of WBD and **B.** An example of family Agromyzidae explored through WBD.

The twentieth century science is believed to be dominated by contributions from biology and information science. The combination of these two has proven to be one of the most essential and fascinating scientific tools called the bioinformatics. With the fast growing technologies in artificial intelligence and with the generation of huge amount of data from various genome projects it is becoming more and more essential to have a knowledge of such important databases. World wide interest is gaining momentum in understanding and applying the algorithms and basics of bioinformatics for the advancement in analysis of data. Many bioinformatical tools from the areas of genomics, transcriptomics, proteomics and metabolomics have accelerated the exploration power of biologists in various areas (Appendix-I). This review provides a brief overview of biological databases and their applications particularly in the field of genomics and proteomics.

References:

[1] Altschul SF., Gish W., Miller, W., Myers, EW., and Lipman, DJ. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

[2] Berman H, Henrick K, Nakamura H, Markley JL (2007) The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data..*Nucleic Acids Res.* (Database issue):D301-3.

[3] Chou, PY, Fasman, GD, (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* 47, 45-148.

[4] Cuff, JA., Clamp, ME., Siddiqui, AS., Finlay, M. and Barton, GJ. (1998). Jpred: A Consensus Secondary Structure Prediction Server, *Bioinformatics* 14, 892-893.

[5] Drysdale R, FlyBase Consortium (2008) FlyBase : a database for the Drosophila research community.*Methods Mol Biol.* 420:45-59.

[6] Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* 30(1):207-10.

[7] Emmert D.B., Stoehr P.J., Stoesser G., Cameron G.N. (1994) The European Bioinformatics Institute (EBI) databases *Nucleic Acids Research* 26(1): 3445-3449.

[8] Hochstein C, Szczur M. (2006)TOXMAP: a GIS-based gateway to environmental health resources. *Med Ref Serv Q.* 25(3):13-31.

[9] Jones, DT., McGuffin, LJ, Bryson, K., (1999). The PSIPRED Protein Structure Prediction Server.

[10] Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M (2009) KEGG for representation and analysis of molecular networks involving diseases and drugs..*Nucleic Acids Res.*

[11] Kneller, DG., Cohen, FE., and Langridge, R., (1990). Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network. *J. Mol. Biol.* 214, 171-182.

[12] Nakamura H, Ito N, Kusunoki M. Tanpakushitsu Kakusan Koso (2002)[Development of PDBj: Advanced database for protein structures]; *Japanese Review,*47(8):1097-101.

[13] Notredame C., Higgins D., Heringa J. (2000) T-Coffee: A novel method for multiple sequence alignments. *Journal of Molecular Biology* 302: 205-217.

[14] Ohkawa H, Ostell J, Bryant S (1995) MMDB: an ASN.1 specification for macromolecular structure..*Proc Int Conf Intell Syst Mol Biol.* 3:259-67.

[15] Pang KC, Stephen S, Engström PG, Tajul-Arifin K, Chen W, Wahlestedt C, Lenhard B, Hayashizaki Y, Mattick JS. (2005) RNAdb—a comprehensive mammalian noncoding RNA database.*Nucleic Acids Res.* 33 (Database issue):D125-30.

[16] Rost, B., Yachdav, G., and Liu, J., (2004). The PredictProtein Server. *Nucleic Acids Research* 32,(Web Server issue), W321-W326.

[17] Velankar S, Best C, Beuth B, Boutselakis CH, Cobley N, Sousa Da Silva AW, Dimitropoulos D, Golovin A, Hirshberg M, John M, Krissinel EB, Newman R, Oldfield T, Pajon A,

Penkett CJ, Pineda-Castillo J, Sahni G, Sen S, Slowley R, Suarez-Uruena A, Swaminathan J, van Ginkel G, Vranken WF, Henrick K, Kleywegt GJ.(2009) PDBe: Protein Data Bank in Europe. *Nucleic Acids Res.*

[18] Woodsmall RM, Benson DA. (1993) Information resources at the National Center for Biotechnology Information. *Bull Med Libr Assoc* 81(3):282-284.