RECOMBINANT DNA TECHNOLOGY AND BIOTECHNOLOGY

Bioinformatics

Sonika Bhatnagar and A.K.Dubey
Netaji Subhas Institute of Technology
Dwarka, New Delhi - 110075

Revised 08 Nov- 2006

CONTENTS

Keywords

DNA, Protein, Bioinformatics, Database, Sequence, Structure, Gene, National Center for Biotechnology Information (NCBI), Entrez, Expressed Sequence Tag, Sequence Alignment, Global Alignment, Local Alignment, Multiple Sequence Alignment, Phylogenetic Analysis, Motif, Pattern, Profile, Domain, Structural Classification Of Proteins, Homology Modeling, Pubmed, Literature Search, Literature Database, Digital Library, Genome, Sequence Assembly, Genetic Map, Microarray, Gene Expression Analysis, Drug Target, Drug Design

DNA and protein sequences

DNA and proteins are the two main biopolymers. They are linear polymers of simple building blocks constituting the living organisms. The repeating units are the nucleotides in case of DNA or RNA and amino acids in case of proteins. Each nucleotide contains a sugar, a phosphate group and one of four different types of nitrogenous base. While DNA and RNA are made up of four different types of building blocks, proteins contain twenty. The linear arrangement of the building blocks of DNA, RNA or protein is called its sequence. The building blocks and the linear polymer of DNA and protein are shown in Fig. 1 and Fig. 2 respectively. In all higher plants and animals, DNA or deoxyribonucleic acid serves as the genetic blueprint. In other words, DNA constitutes the entire information content required for an organism to exist and function. Proteins, on the other hand, serve as the effector arm of DNA, performing all the cellular functions in addition to being key structural components.

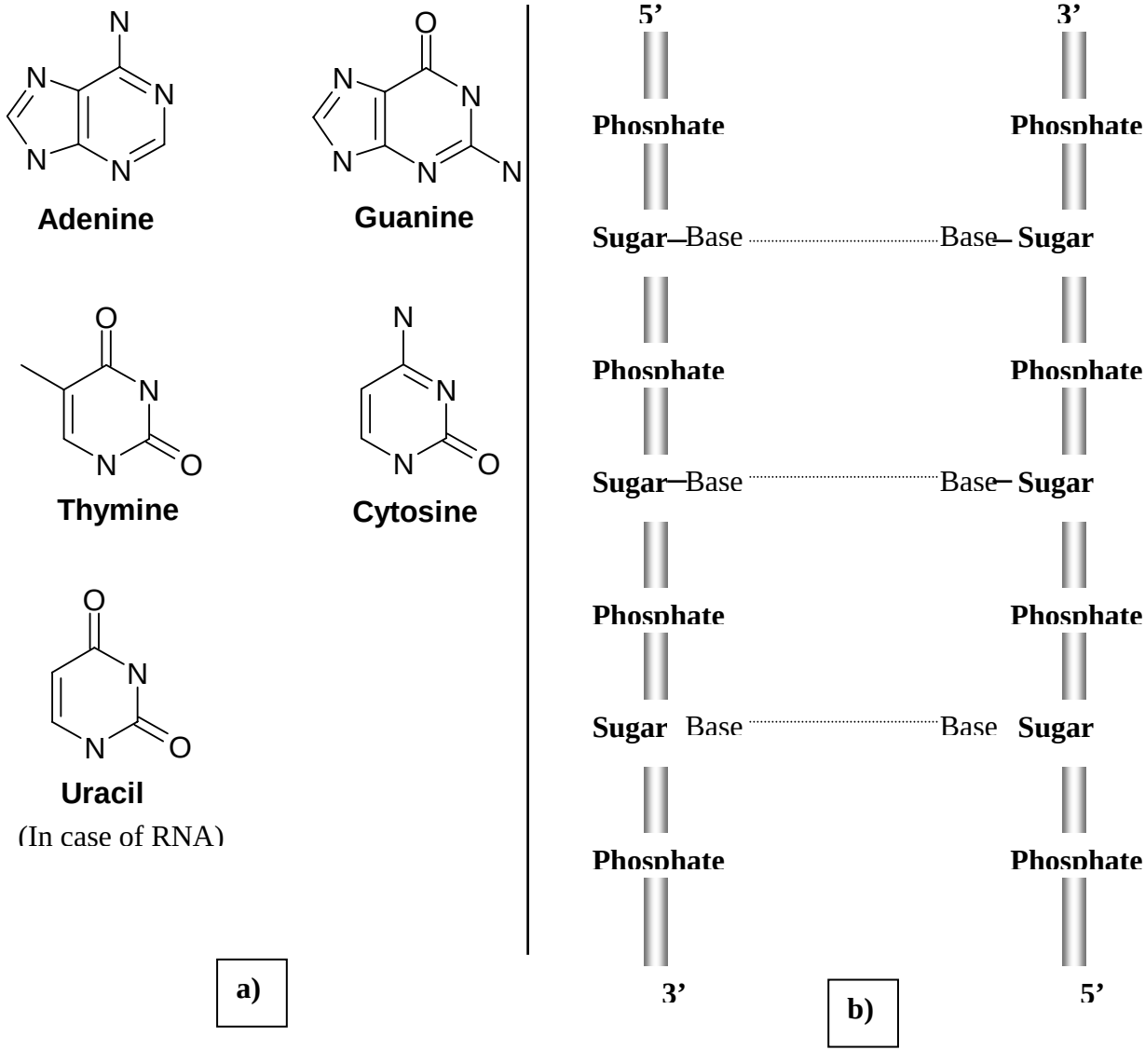

Fig.1: a) The four types of nitrogenous bases present in nucleic acids. b) A linear chain of DNA showing the arrangement of sugar phosphate backbone at the periphery and hydrogen bonded bases at the center. Chemical structures were drawn using MDL ISIS Draw.

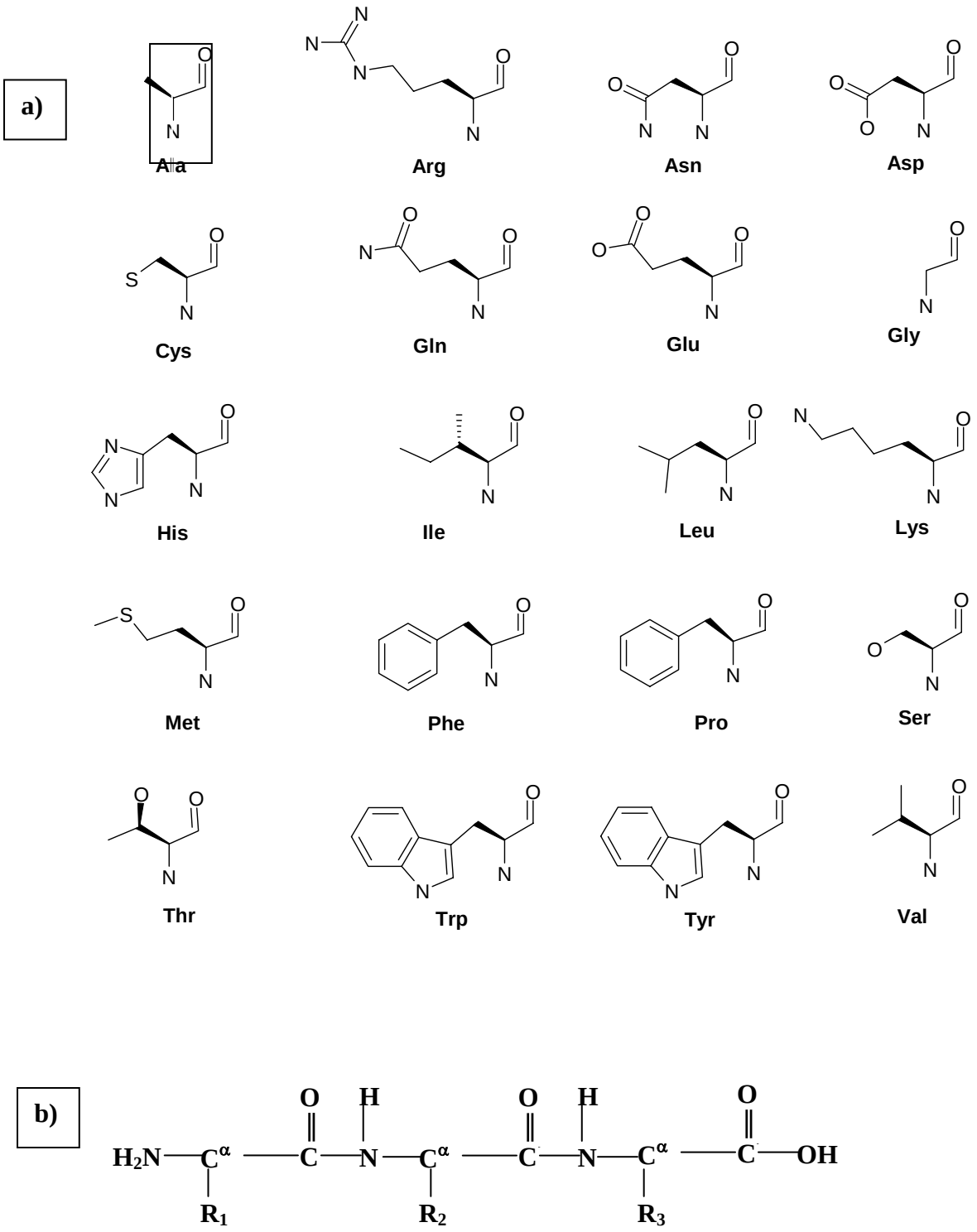

Fig. 2: a) The twenty building blocks of proteins. The main chain atoms are shown in a box in case of Ala. The chemical structures were drawn using MDL ISIS Draw.b) A Poly-Ala chain showing formation of a linear polymer consisting of repeating main chain atoms and variable R Groups

The sequence of a DNA or RNA molecule constitutes its primary information content. Transmission of genetic information by replication depends on the ability of a nucleic acid chain to form specific complementary base pairs with an opposing strand. An adenine can form hydrogen bonds only with thymine and a guanine can make hydrogen bonds only with cytosine. At the time of cell division, the two strands separate and the one of the strands acts as a template for the synthesis of a complementary strand. Thus, identical copies of the DNA are made during replication. In this way the three dimensional structure of DNA is inherently suitable for copying the genetic information. In the cell, this information is used to direct the synthesis of proteins, which in turn carry out cellular processes and determine cellular structure.

In case of a protein, the amino acid sequence directs its spontaneous folding into a three-dimensional structure. Since the function of the protein is directly dependent on its three-dimensional structure, the higher number of building blocks allows the construction of a vast array of molecule with a large number and variety of functions. The twenty amino acids include a large number of functional groups like alcohol, amide, carboxylic acids and others that contribute to enzyme function and specificity.

Conclusion: The sequence of a biopolymer is directly related to its chemical properties, three-dimensional structure and biological function. In turn, these attributes of biomolecules are critical to the flow of information, both from genotype to phenotype as well as from one generation to another.

What is Bioinformatics?

With the improvement and automation of powerful techniques for the sequencing of DNA, a large amount of DNA data from a number of organisms came to be elucidated. The human genome project was one effort and led to the "working draft" of the complete nucleic acid sequence of humans in early 2001. Rapid sequencing of genomes from microorganisms, parasites and higher organisms has led to an explosion of sequence data. Further advances in genomic technologies generated large scale data on protein structure and function, gene expression, protein interactions, etc. This raw data is of immense importance in biological research. However, in order to be useful, this data has to be stored, organized and indexed such that it is easily accessed interpreted and related to other biological data. Therefore, the requirement for computerized databases and analysis tools became apparent. Biological databases were thus created to organize and present a persistent set of information to the user. As an example, a nucleotide sequence record contains information about the molecule, sequence features, source organism, related literature citations, etc.

Bioinformatics is formed from the amalgamation of Biology, computer science and Information Technology. It deals with the storage, retrieval and analysis of biological data *including* nucleic acid and protein sequences, structures, functions and pathways in order to gain new biological insights. To do this, it employs techniques from applied mathematics, computer science and statistics. It is continuously expanding to encompass areas like 3D structure prediction, gene expression analysis, protein interactions, pathway modeling, target identification and drug design. In brief, bioinformatics is a multi-disciplinary field that operates at three levels :-

a) Organization of the biological data to help researchers access, add and modify it. This involves massive efforts towards compilation and maintenance of databases.

b) Development of software tools that help in interpretation of the data {an activity mainly driven by the interests and ideas of the biologists in mining useful information.
c) Use of bioinformatics tools for analysis and interpretation referred to as computational biology.

A factor that has aided the rapid development of the field of bioinformatics is the wide reach of the internet. This has enabled the vast amount of biology related data to be made accessible for analysis through public domain databases and available software tools. Accordingly, this chapter is broadly divided into a discussion of the various publicly available biological databases and a discussion of tools and resources present therein.

Biological Databases

The growing amount of biological data is accompanied by the development of a number of public domain databases. Primary databases contain either raw or curated data that is directly submitted by scientists. It is then filtered and compiled to produce annotated and non-redundant composite databases. A molecular database may pertain to nucleotide sequence, gene structure and location, regulatory elements, protein sequences, structure, motifs, conserved domains, expression data, mutation information, disease linkage, metabolism, evolution, etc. This information is locally divided into many different specialized databases, and then linked in order to facilitate the query and mining of information. The available online Biological databases are compiled yearly by the journal Nucleic acids research in its database issue. Other listings of biological databases can be found in organized and searchable metadatabases (database of database) like MetaDB (Searchable collection of links to Biological databases) , Entrez (Integrated Data access) and harvester (Gene and protein information query).

One of the centers for dissemination of molecular biology information is the NCBI (National Center for Biotechnology Information), a division of the National Library of Medicine at the National Institute of Health, USA. The information available through the NCBI website is especially useful since it follows an integrative data model as discussed before. Therefore, links lead from published literature to the encoded DNA sequences, chromosome maps, proteins and three-dimensional structures of the proteins. This integration of databases makes it easy to navigate complex biological information and forms the basis of the Entrez retrieval system. This data model allows us to access, retrieve and save the data in several different formats by separating it into different domains (like citations, sequences, structures and maps). At the same time, it increases the possibility of making discoveries using this data. Another advantage of dividing data into various integrated databases is that it can be expanded using new links for new data fields as new experiments are carried out.

Collection and storage of sequences

DNA fragments are obtained by fragmentation of plasmid/phage clones or amplified by polymerase chain reaction. They can then be denatured into single strands, hybridized to an oligonucleotide primer and submitted for sequencing using an automated procedure wherein new strands are synthesized from the end of the primer using a heat resistant DNA polymerase enzyme. The DNA polymerase synthesizes DNA complementary to the DNA fragment. Introducing a chain terminating nucleotide of a specific type (e.g. ddATP or dideoxy Adenine TriPhosphate instead of a dATP or deoxyribose Adenine Nucleotide Triphosphate), causes the chain synthesis to be stopped at points of occurrence of A such that a set of nested fragments ending at A can be obtained as shown in Fig. 3. A similar procedure

is adapted for obtaining the corresponding set of nested fragments for the other three bases with a different fluorescent label attached to each of the termination signals. This yields a set of nested fragments for each type of nucleotide (A, T, G and C). When the resulting mixture is subject to electrophoresis, the fragments get separated on the basis of size. A laser beam is used to excite the fluorescent labels that can then be recorded using a detector. This data is fed into the computer and a program is used to determine the probable order of the bands and to predict the sequence. This provides a reliable sequence of up to 500 bases. The resulting sequence can then be used to produce primers downstream in the sequence and the entire procedure outlined above can be repeated to sequence DNA fragments of several kilobases.

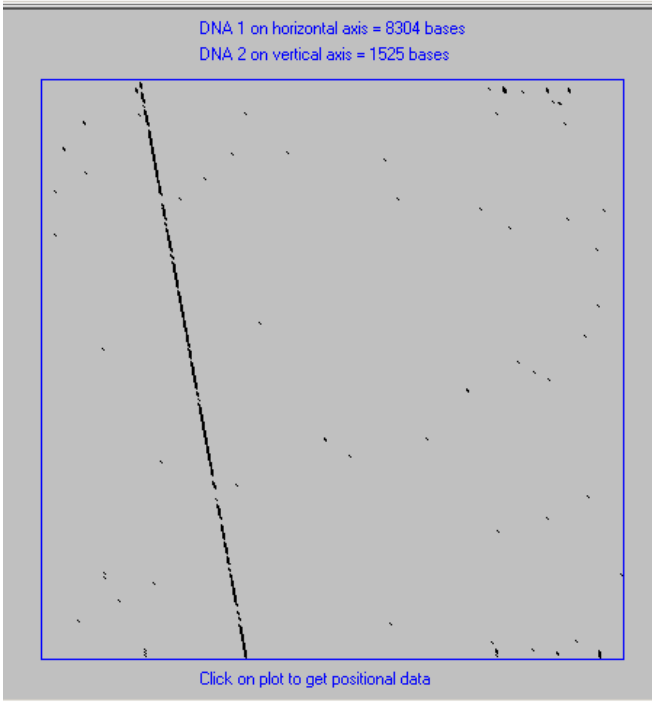

Fig. 3: Dotplot of Mus musculus glycogen synthase kinase 3 beta mRNA with Rattus norvegicus glycogen synthase kinase 3 beta (Gsk3b) mRNA shows that they share a similar stretch of sequence. Clicking on the dotplot indicates that the similar stretch extends from base 1390-2906 in sequence 2 and base 7-1513 in sequence 1. Dotplot drawn using Molecular Toolkit, Colorado State University

Since the genomic sequences are often large, the chromosomes are purified, broken into fragments and cloned. The overlaps in the sequences or contigs are then used for computerized assembly of the sequence.

Nucleotide and protein sequences obtained from experimental procedures are submitted to the databanks like those discussed in the next section via an easy www interface. Here, the sequence records are reviewed, updated and an accession number is allocated, which is required for publishing the sequence. The sequence records and the associated data are converted into a tabular form. The attributes of the sequence are organized into rows and

columns, each with a unique identifier that is carefully indexed and cross-referenced so that it can be located by a search query. A sequence file in text or Ascii format is made available for future analysis. Several interchangeable formats are required by different sequence comparison programs. The most common formats are those of GenBank DNA sequence entry, EMBL data library, Swissprot, FASTA, NBRF/PIR, GCG, plain/ASCII.staden and ASN.1. The NCBI Genbank database allows for easy interconversion among some of the commonly used sequence formats. Alternatively, the READSEQ program can be used to do the interconversions for special type of analyses.

Nucleotide sequence databases

The three main primary nucleotide sequence databases are -

- The EMBL (European Molecular Biology Laboratory) database maintained by EBI (European Bioinformatics Institute). It can be searched using the SRS (Sequence Retrieval and Search) system.
- DDBJ (DNA Data Bank of Japan) maintained by the National Institute of Genetics.
- The Genbank database maintained by the NCBI. It can be searched and accessed through the Entrez search system. Gen Bank is the annotated collection of all publicly available DNA sequences. EMBL, DDBJ, and Gen Bank at NCBI collaborate to exchange data on a daily bases to keep their information current. Genbank also allows for the submission of newly determined sequences to its repository through a www interface Bankit or a standalone software Sequin. It recently announced the important milestone of reaching 100 giga bases, signifying the huge number of DNA sequences now available. The data base records are used to hold raw sequence data and a number of annotations like sequence description, source organism (with taxonomic information) bibliography, known or predicted coding regions and their protein translations. The Entrez Gene database provides a unified resource for information on a gene including -
    - a graphic summary of the gene in the genome with information on intron /exon structure and the flanking genes.
    - View of the transcribed mRNA sequence with locations of Complementary DNA Sites and small sequence variations.
    - Gene structure and phenotypic effects
    - Sequence data of the proteins and their conserved domains
    - Related resources like mutation information etc.

Other important nucleotide databases include –

- Ensembl - This database is a result of the collaboration between the EBI and the WTSI (Wellcome Trust Sanger Institute). It attempts to produce an open access system for automatic analysis and annotation of selected eukaryotic genomes.
- Unigene - The Unigene system attempts to cluster sequences into non-redundant clusters. Each cluster consists of one or more sequences that constitute a unique gene. This is further integrated with its map location and other related information. It can be navigated by organism or searched using keywords. However, since the automated gene clustering methods are still developing, the system is still considered experimental.

Data from Expressed Sequence Tags (EST) in plants and animals having a large number of EST data available has also been included in Unigene in order to aid gene discovery. An EST is a small part of a gene that can be used to identify its location and/or function. They are generated by sequencing either one or both ends of an expressed gene. For an example, 3.6

million ESTs present in the GenBank for *Homo Sapiens* have been reduced to a set of 104,000 sequence clusters that can be searched by gene name, chromosome location, cDNA library, accession number or text search. Presently, the Unigene consists only of the protein coding nuclear genes. The Unigene information can be viewed as a cluster or as a single sequence that incorporates links to related online resources like possible protein products with their title and GenBank accession number, inferred map position & chromosome assignment, tissue source and component sequences of the cluster. Unigene forms the basis of and is linked to three other NCBI resources, namely ProtEST (protein similarity browser), DDD (Digital Differential Display - comparison of EST-based expression profiles) and HomoloGene (information about possible homology relationships). Apart from this, a number of different type of nucleotide sequence databases can be accessed via the www and are summarized in Table 1.

Table 1: Nucleotide databases

S.No.	Database	Description
EBI (European Bioinformatics Institute), a part of EMBL		
1.	EBI genome server	Competed genomes and their translations
2.	ASD	Database of Alternatively spliced exons
3.	ATD	Database of alternate transcripts diversity, which may in turn undergo alternative splicing or polyadenylation
4.	EMBL-Align	Multiple sequence alignments
5.	EMBL-Bank	EMBL nucleotide sequence database
6.	EMBL CDS	EMBL coding sequences with annotations
7.	Ensembl	Automatic annotation of the deciphered eukaryotic genomes
8.	IMGT/HLA and IMGT/LIGM	Immunogenetics databases consisting of the sequences of genes in the human major histocompatibility complex (HLA) and Immunoglobulins and T Cell receptors
9.	IPD	Database of polymorphisms of genes of the immune system
10.	LGIC	Database of Ligand gated ion channels
11.	Mutations	Database of sequence variations and other mutation resources
12.	Parasites	Parasite genome database
NCBI		
1.	Genbank	All publicly available DNA sequences with annotation
2.	DbEST (Database of Expressed Sequence Tags)	A division of GenBank that serves as a separate database for screened and annotated ESTs from a number of organisms.
3.	DbGSS (Database of Genome Survey Sequences)	Similar to dbEST but sequences are genomic in origin rather than cDNA

4.	Umbrella Nucleotide	Composed of three databases, i.e. EST, GSS and core nucleotide (all the remaining nucleotide sequences).
5.	Unigene	Classification of Genbank sequences into non-redundant sets by organism.
6.	DbSTS (Database of Sequence Tagged Sites)	A STS is a 200 to 500 base pair sequence unique to a genome. It defines a specific position on the physical map and thus serve as landmarks.
7.	HomoloGene	A system for automatic comparison of several completed eukaryotic genomes and identification of homologous genes. It is enhanced by homology and phenotype information from several other databases (eg. OMIM, COG, Flybase, Mouse Genome Informatics, Zebrafish Information Network & *Sacchromyces* Genome Database).
8.	MGC (Mammalian Gene Collection)	Listing of all the full length open reading frames for all expressed genes from human, mouse, rat and cow. All the clones available from the cDNA libraries thus generated are available for purchase.
9.	Popset	A set of DNA sequences used to analyze the evolutionary relationships between members of same or different species. These sequences can also be viewed as a multiple sequence alignment.
10.	DbSNP (Database of Single Nucleotide Polymorphism)	SNP is a small variation of the DNA that may not produce a physical change in the organism but may be linked to disease susceptibility and may influence the pharmacological response to certain drugs.
11.	Probe	A database of nucleic acid probes used in biomedical research applications. Useful for analysis of gene expression, gene silencing and genome mapping.
12.	Refseq	Nonredundant set of DNA sequences, tRNA and proteins for more than 3,000 commonly used research organisms.
13.	UniSTS	Unified, nonredundant database of STS that integrates data from maps and markers from several public resources.
14.	TPA (Third party annotation)	A sequence database that provides either experimentally derived or inferred annotation data not directly received from the submitter of the sequence.
15.	Trace archive	A repository for DNA sequence chromatograms for large-scale sequencing projects.
DDBJ (DNA Data Bank of Japan)		
1.	Genome Information Broker	Information on completed genomes
2.	GTOP	Genomes to protein structures and functions – data analyses summary
Genomenet, Japan		
1.	KEGG	Kyoto University's Encyclopedia of Genes and Genomes
2.	KEGG2 Genes	Manual genome annotations

HIV Database		
1.	HIV sequence database	Curated and annotated HIV sequence data
2.	Resistance database	Known muctations associated with drug resistance

***Note**: Other databases pertaining to coding and non-coding DNA sequences, gene structure, transcription regulators, transcription factors, and RNA sequences are also available. A comprehensive list of these can be accessed at http://www.oxfordjournals.org/nar/database/c/*

- **Genome Databases** – Genome is the term used to refer to all the genetic information in the chromosomes of a particular organism. Study of the nucleotide sequence, structure and function in the genome is known as genomics. A number of prokaryotic and eukaryotic genomes have now been sequenced and over a 2000 (either complete or in progress) genomes are currently available for study, analysis and comaprison . A complete listing of publicly funded sequencing efforts is maintained in the NCBI Genomes database. All completed and incomplete genome sequence data can be queried using the Entrez genome resource. The genome databases available through NCBI can easily be searched by keyword or sequence similarity. The GOLD (Genomes online database) currently lists 364 completed genomes. WIT (What is There?) and EBI Completed Genomes can also be referred for genome data. Microbial genomes can additionally be accessed at the websites of TIGR (The Institute of Genome Research) and the Sanger Institute. Databases dedicated to specific organisms include GDB (Genome Database - provides access to annotated human genome data), MGI (Mouse Genome Informatics), RGD (rat Genome Database) and (ZFIN) Zebrafish Information Network. ENCODE (Encyclopedia of DNA Elements) is a project launched by NHGRI (National Human Genome Research Institute) to identify all the functional elements of DNA in selected genomes. Some of the Genome databases are summarized in Table 2.

Table 2: Selected Genome resources

1.	TIGR	The Institute for Genomic Research Genome Projects for microbes, plants and humans
2.	CoGenT++	Database of Complete Genomes and corresponding protein sequences
3.	EBI Genomes	Complete and unfinished viral, prokaryotic and eukaryotic genomes
4.	EMBL Genome Reviews	Annotated view of complete genomes
5.	GOLD	Genomes online database for completed and ongoing genome projects
6.	PEDANT	Automatic analysis of genome sequences
7.	DOTS	Database of Transcribed sequences in human and mouse
8.	CORG	Comparative Regulatory Genomics elements in multiple species
9.	VEGA	Vertebrate Genome annotation database

10.	WIT	(What is There?) Complete reconstruction of metabolic and signaling pathways
11.	ZFIN	Zebrafish Model Organism Database
12.	RGD and Ratmap	Rat Genome Database and Gene localization
13.	MEGX	Marine Ecological Genomics Portal and Database
14.	MGD	Mouse Genome Database
Viral Genomes		
1.	DPV	Description of Plant viruses and related animal viruses
2.	HCV sequence database	Nucleotide and protein sequences, protein 3D models of Hepatitis C virus
3.	HIV sequence database	Annotated DNA and protein sequences with analysis tools
4.	VBRC	Viral Bioinformatics resource Centre curated viral genomes
5.	Virgen	Comprehensive virus genome resource
6.	Vida	Homologous protein families of sequences from virus genomes
Fungal genomes		
1.	AGD	Ashbya Genome Database
2.	CADRE	Central Aspergillus Data Repository
3.	CGD	Candida Genome Database
4.	SGD	Sachromyces genome Database
Prokaryotic genomes		
1.	Archael genome browser	Features of sequenced archaeal species
2.	BSORF	Bacillus Subtilis Open Reading Frames
3.	CampyDB	Database for analysis of Campylobacter Genome
4.	Ecocyc	E.Coli Pathway/Genome database
5.	GenomeAtlas	Properties of sequenced microbial genomes
Invertebrate Genomes		
1.	C.Elegans project	Accessible through Sanger Institute website or through wormbase
2.	Nematode.net	Nematode Gene sequences
3.	Wormbase	Data repository for information about C.Elegans and related nematodes
Human Genome		
4.	GDB	Annotated Human Genome
5.	GeneCards	Integrated database of human genes
6.	HOWDY	Integrated system for access and analysis of human genome

7.	hmtDB	Human Mitochondrial database
NCBI Website		
1.	Genomes	All publicly available complete and incomplete eukaryotic genome data linked to sequence maps with contigs, genetics and physical maps.
2.	Genome Project	Collection of all large scale genome sequencing, assembly, annotation and mapping projects that is organized by organism.
3.	Gene	Curated and highly integrated database of genes from Refseq genomes. Analysis is supported by available tools and tutorials.
4.	COGs (Clusters of orthologous groups)	Phylogenetic classification of all the proteins encoded by completely sequenced genomes. Each COG lists the evolutionary counterparts of a protein.
5.	Cancer chromosomes	Data from cytogenetic studies on chromosomal aberrations in cancer.

- **Molecular Cytogenetics database -** Karyotyping is the process by which dividing cells are arrested in metaphase of mitosis when chromosomes are condensed and clearly visible. Dyes like giemsa are then added to produce a characteristic banding pattern by which the chromosomes can be identified. Karyotyping finds application in human genetic studies to identify missing or extra chromosomes as well as chromosome extensions and deletions. Thus, it can be used to diagnose chromosomal aberrations like Klinefelter's syndrome, Down's syndrome or Trisomy 13. SKY (Spectral karyotyping) is an improvement on the original technique wherein fluorescent dyes binding to specific areas of the chromosomes are used. A series of specific probes is used with varying amount of dye to lend characteristic spectral properties to the chromosomes. The spectra is measured by an interferometer that can locate even small differences in banding patterns. The spectra is analyzed using specialized software that assigns a distinguishing color to each chromosome, thus producing a colored digital image. M-FISH (Multiplex Fluorescence In Situ Hybridization) also uses spectrally distinguishable fluorescent dyes but employs microscopic filters with narrow bandpass to capture five different images of the chromosomes (corresponding to five different flurochromes), that are then combined by dedicated software. Both techniques allow for improved sensitivity for finding translocations, breakpoints, complex rearrangements, etc. in chromosomes. CGH (Comparative Genomic Hybridization) is complementary to SKY and M-FISH in that it can help in the study of tumors that do not give sufficient metaphase and therefore can not be studied by the previous two techniques. In CGH, chromosomes from normal and tumor DNA are mixed together and hybridized to produce normal metaphase. They are then differentially labeled with separate dyes for normal and tumor DNA. The fluorescence intensities are measured and are used to quantitate the copy numbers of different DNA sequences. Public domain data from all three techniques is housed in the SKY/M-FISH and CGH database at NCBI. The Cytogenetics databases, CytoD and the Mitelman database of chromosomal aberration extract the information on cytogenetic abnormalities from PubMed Abstracts.

Protein databases

Protein databases can broadly be divided into different types depending on the kind of information stored i.e. sequence databases (dealing with the sequence or primary structural information) and structure databases (dealing with the 3-D organization of the proteins). However, these are increasingly becoming integrated with each other and with literature databases.

The main protein sequence databases are SWISS-PROT maintained by SIB (Swiss Institute of Bioinformatics) in collaboration by EBI, PIR (Protein Information Resource) maintained at the Georgetown University Medical Center (GUMC) . Both are curated databases and PRF (Protein Research Foundation, Japan).

- **SWISS-PROT** is a highly annotated, almost non-redundant and integrated with many other databases. It is accompanied by TrEMBL, a computer annotated database of protein sequences derived from the EMBL nucleotide database that is not already present in SWISS-PROT. Together, SWISS-PROT and TrEMBL constitute the UNIPROT database that can be searched by keyword, gene name, organism, etc. Alternatively, sequences can be accessed using SRS.
- **PIR** comprises three separately searchable subsets, namely PIRSF (PIR-Super Family – Classification of proteins by evolutionary family), iProClass (highly integrated comprehensive database for important additional protein information) and iProLINK (integrated Protein Literature, Information and Knowledge – literature search for curating proteins). Additionally, the PIR website allows searching for and identifying peptide sequences up to 30 amino acids in length against the UNIPROT database.
- **PRF** consists of PRFLITDB (PRF Literature Database) and PIRSEQDB (PIR Sequence Database).
- **PDB** (Protein Data Bank) is a worldwide repository for macromolecular structure data founded by RCSB (Research Collaboratory for Structural Bioinformatics), MSD-EBI (Macromolecular Structure Database at EBI) and PDBj (PDB Japan). All three sites offer a number of useful tools for searching and visualizing the structures.
- **Specialized databases** also exist that identify the sites and patterns of biological significance and use this information to classify proteins into different families. These include PROSITE, InterPro (Integrated protein domains and functional sites), BLOCKS (Conserved protein regions), PRINTS (Protein fingerprints or a group of patterns used to identify a protein family), Pfam(Protein families), ProDom (Protein Domains) and PROTOMAP (classification of all SWISS-PROT protein sequences). The entrez Proteins database contains the protein sequences from SWISS-PROT, Protein Information Resource, PDB and translations of annotated sequences from Genbank. Sequence and structure databases available for search and analysis at the NCBI website are summarized in Table 3.

While the genome of an organism is constant, the cellular proteome (or the protein product of the cell's genome at a given time) is constantly changing in different tissues, cellular stages, environmental conditions, etc. In addition to alternative patterns of gene splicing, proteins undergo post-translational modification (e.g. glycosylation, phosphorylation). Therefore, the proteome is larger and much more complex than the genome. The large-scale study of protein structure, function and interactions is called proteomics. One of the techniques that has aided the rapid development of the field of proteomics is peptide mass fingerprinting using a mass spectrometer. For this, a protein is experimentally cleaved using a protease (e.g. trypsin) and the masses of the resulting peptide fragments are identified using a MALDI-TOF

(Matrix-Assisted Laser Desorption/Ionization – Time of Flight) or ESI-TOF (Electron Spray Ionization – Time of Flight) spectrometer. Since trypsin cuts a protein at a specific site, the resulting pattern of peptide masses (or peptide mass fingerprint) can be calculated and is characteristic of a protein. Software programs are available that can calculate the fingerprints from proteins, translated nucleotides or genome databases and compare it with the fingerprint of an unknown protein in order to identify it. The HUPO (Human Proteome Organization) aims to catalog the functions and interactions of all human proteins. One of the resources of the HUPO is the Human Protein Atlas portal that shows the expression and cellular localization of a large variety of proteins. Other proteomics related databases include DIP (Database of Interacting Proteins), AAIndex (Physicochemical properties of peptides), SWISS 2D-PAGE and YPD (Yeast Proteome Database).

Table 3: Selected protein related databases

Protein Sequence databases	
MIPS	Munich Information center for Protein Sequences
ExProt	Database for Protein sequences for which the functions have been Experimentally verified
PIR	Informatics Resource for non-redundant Protein sequences
SwissProt	Protein knowledgebase accessible through the Expasy site
UniProtKB	Central repository for protein sequence and function – contains information from Swiss-Prot, TrEMBL, and PIR with automatically annotation/classification sequences
UniprotKB	Uniprot Knowledgebase
Structure Databases	
PDB	Largest archive of structure data for biological macromolecules
3D Genomics	Structure function annotations of genomes of almost 100 organisms.
MSD	Macromolecular structure Database at EBI
Dali database	3D structure alignment and comparison
SCOP	Structural classification of proteins
Modbase	Comparative protein models
Enzymes and enzyme nomenclature	
BRENDA	Comprehensive enzyme information resource
Enzyme	Swiss prot enzyme nomenclature database
Macie	Mechanism, annotation and activation of enzymes
TECRdb	Themodynamics of Enzyme catalyzed reactions
Proteomics related databases	
AAindex	Physical/Biochemical properties of amino acids
Proteome Database system for microbial research	Interlinked 2D PAGE databases, ICAT-LC/MS, functional classification of proteins and differentially regulated proteins determined by quantitative gel image analysis

Biozon	DNA and protein sequences, structure, conserved domains, family, interactions and pathways.
Open Proteomics Database	Proteomics data obtained by Mass Spectrometry
SWISS 2D PAGE	Protein sequence alignments and structure function predictions.
MIPS	Mammalian protein interaction database
DIP	Database of experimentally determined interactions
Protein sequence motifs and active sites	
Blocks	Ungapped sequence alignments corresponding to most conserved protein region
ASC	Collection of Sequences of amino acids with known Biological Activity
Interpro	Integrated protein families, domains and functional sites
PRINTS	Group of conserved motifs or fingerprints in proteins
e-motif	Database of short conserved sequence stretch or motifs
NCBI protein resources	
Proteins	Sequences of proteins from PIR, PRF, PDB, SWISSPROT and translation of DNA sequences from Gen Bank, EMBL and PDBJ.
PROW (Protein Resources on the Web)	Concise, peer-reviewed information on proteins and protein families.
RefSeq	Comprehension, non-redundant updated set of sequences from genomes, transcripts and proteins for major research organisms.
3D-Domains	Automatically identified structural domains in the Entrez structure database. It is used to identify structural neighbors that can be visualized using Cn3D (See in 3-Dimensions).
MMDB (Molecular Modelling Database)	Entrez's macromolecular database of 3D structures of proteins & nucleotides. It is linked to sequence, bibliographic information, taxonomic information and similar structures in other proteins.
CDD (Conserved Domains Database)	A domain is a structural and functional unit of a protein. The CDD consists of collection of multiple sequence alignments that is linked to 3D structure where possible.

Gene expression databases

While genome sequencing provides a static view of the cell, large scale analysis of gene expression is becoming increasingly important to study and analyze the role of different genes at different stages. The two dominant tools for study of genome wide expression studies are DNA microarrays and SAGE (Serial Analysis of Gene Expression). As a large

amount of data is produced from gene expression studies, this is a dominant area in which bioinformatics databases and tools have found use.

DNA microarrays or DNA chips consist of a number of DNA clones or probes tethered to a glass slide. CDNA is prepared from mRNA of sample tissue and labeled with different fluorescent dyes. This cDNA is hybridized to the DNA probes immobilized on the glass slide. The fluorescent dyes are excited using lasers and the resultant image is stored in the digital form for further analysis. In SAGE, a short sequence tag unique to each expressed gene is used. A number of such tags are linked together and sequenced. The number of times a particular tag is observed gives the expression level of the gene. The NCBI GEO (Gene Expression Omnibus) database acts as a repository for microarray, SAGE and mass spectrometric data on gene expression. Two related databases, GEO profiles and GEO datasets allow queries based on gene expression profiles and experimental setup respectively.

The MGED (Microarray Gene Expression Data) society is an international organization that promotes the sharing of microarray data. It has been instrumental in developing MIAME (Minimum Data About a Microarray Experiment), a format that helps authors, reviewers and publishers make Microarray data available to the scientific community in a usable way. Additionally, the MGED supported MAGE-ML (Microarray Gene Expression Markup Language) format aims to provide a standard that facilitates the exchange information between different microarray data systems. This format has been accepted by the public repositories like GEO and ArrayExpress as well a number of scientific journals.

Databases for Drug discovery

One of the final goals of the study of physiological processes and disease mechanisms is to develop drugs against infections, inherited diseases or other errors of metabolism. Bioinformatics tools are involved in both stages of drug development, namely –

a) Target selection and validation
b) Screening or design of drugs using computational and experimental methods

Specific databases like TTD (Therapeutic Targets database) list the details for fully validated and potential therapeutic targets, with information and links to PubMed, known inhibitors, enzyme nomenclature, structure and patent information. Molecular libraries serve as sources of information for pharmaceutical and drug-like compounds, their chemical properties and biological actions. The NLM (National Library of Medicine) maintains a number of chemical databases on drugs, hazardous products, carcinogens and other chemicals that can be searched for a variety of data using ChemIDPlus (Chemical Identification Plus).

The NCBI PubChem system consists of three linked databases i.e. PubChem Substance, Compound and Bioassay. The compound molecular libraries are important cheminformatics resources in screening and design of small molecules that can bind to drug targets. The Bioassay database is an important resource for target validation. Together the PubChem databases can be searched by keyword, structure similarity, compound neighboring properties, etc. The compounds are further linked to the entrez gene, protein, compound structure and literature databases. Other cheminformatics resources include DrugBank, PharmGKB and Superdrug. Table 4 summarizes the Gene expression, small molecules and other molecular databases.

Table 4: Miscellaneous molecular databases

Transcriptional regulatory databases	
ABS	Annotated binding sites for transcription factors
cisRED	Database of predicted regulatory elements
DBD	Database of predicted transcription factors in genomes
TRED	Transcriptional Regulatory element database
Human Genes and Disease	
PMD	Protein Mutation Database
HGMD	Human Gene Mutation database
OMIM	Online Mammalian Inheritance in Man
Cosmic	Catalog of somatic gene mutations in cancer
Cancer chromosomes	Cytogenetic and clinical data resource
Gene Expression Database	
SAGE (Serial Analysis of Gene expression)	Experimental data collected by Identification of short sequence tags in a gene and subsequent quantitation to determine patterns of gene expression.
GEO (Gene Expression Omnibus)	Gene expression data from microarrays, serial analysis and mass spectrometry experiments.
Arrayexpress	Public repository for microarray gene expression data
CGED	Cancer Gene Expression Database
GENSAT	Gene Expression Nervous System Atlas of mouse Central Nervous System using *in situ* hybridization and transgenic methods.
Stanford Microarray database	Microarray data and tools for analysis
Oncomine	Cancer microarray data
Drug target and design related	
Drug bank	Drug and target related information
PharmGKB	Pharmacogenetics Knowledge Base
TTD	Therapeutic Targets Database
Superdrug	Structures of essential marketed drugs
Pubchem Substance, BioAssay and compound	Description of chemical samples from different sources with information about chemical structure, activity, citations, etc., bioassay procedures and chemical content

A SNP (Single Nucleotide Polymorphism) is a small change that can occur within the individual's DNA. It occurs when any one of the A,T,G or C nucleotides is replaced by another. Although this happens very frequently in humans, this variation lies generally in the

non-coding region of DNA as that makes up 95 to 97% of an organism's DNA. These can be associated with the presence of certain diseases and may therefore act as disease markers. Some SNPs lie in the coding DNA regions and can therefore change the protein structure (and therefore function). They have the ability to cause a disease, affect genetic predisposition to certain diseases or change the way a drug is metabolized. The last category of SNPs determine the way an individual responds to a particular drug. Study of the different genes that determine drug effects and behavior is known as pharmacogenomics. It points the way towards personalized drug treatments that are tailored to suit the patient's genetic makeup. The NCBI SNP database maintains an annotated catalog of SNPs and links to data from NCBI and external information sources. Other important SNP resources include The SNP Consortium Database and specialized resources like (IPD) Immunopolymorphism Database and the Database of Genomic Variants in humans.

Literature databases

With the increasing pace of growth in molecular biology, it is critical for a researcher to be familiar with the up to date pre-existing knowledge in the chosen field as derived from published, peer-reviewed literature. Search with popular engines like Google can often provide the desired output. However, specialized search engines like BioNotebook, Biology browser, Infomine, catalog of Biological databases, etc. can be used to search for relevant subject-specific information quickly and efficiently. Some of the biology specific resources are listed online in Search Engine Guide's biology catalog. Scirus and Google scholar are comprehensive search engines specific for science. In addition, The NLM (USA) maintains a number of databases and resources on clinical trials, toxicology (TOXNET), patient information (MedlinePlus), chemical carcinogens (CCRIS – Chemical Carcinogenesis Research Information System), genetic conditions (GHR – Genetics Home Reference) and HIV/AIDS SIS (Specialized Information Service) that can be accessed through its website. Scientific literature from various disciplines can be accessed from Caltech CODA (Collection of Open Digital Archives) and Open archives Initiative.

A growing number of subjects and fields in biology are accompanied by publication of many more journals. The NCBI Entrez PubMed is a database of abstracts from published, peer-reviewed biomedical literature. A keyword searchable interface makes the database convenient and easy to use. The abstracts themselves are further linked with the full-text digital archive of journal articles. Pubmed can be used to locate papers by author, year, journal and citation. Each PubMed entry includes links to nucleotide and protein sequence and structure in addition to books, genetic, mapping information. Clicking on the neighboring link for any article finds other similar articles. PubMed searches can also be saved using 'My NCBI' and links to external providers can be incorporated using 'Linkout'. Specific clinical research areas can be searched using the PubMed clinical queries database. Additionally the special queries option may be used to limit the search to specialized database subsets like cancer topics, AIDS, Bioethics, History of Medicine, Toxicology, etc. Alternatively, all molecular and literature databases can be queried simultaneously using the 'All databases' option. The Entrez 'E-utilities' options allows for programmable specific queries that may not be covered by the regular web interfaces. The IEB (Information Engineering Branch) is responsible for developing new tools and databases and is primarily meant for those interested in software development as well as for announcements of new resources. PubMed also offers links to related gateways and databases on consumer health, clinical trials, toxicology, etc. The detailed information on the Entrez life science journals and can be retrieved via FTP, journal search or journal browser options. The NCBI bookshelf is

fast growing to include popular textbooks providing background information definitions and insights into many molecular biology related areas.

The Biological Abstracts database is a complete collection of bibliographic references to life sciences journal literature covering all areas of Biology including ecology, plant sciences, zoology or literature. The web of science provides seamless access to 8,700 journals along with search and navigation tools. The Agricola literature database is specific to agriculture while CAB Abstracts additionally cover veterinary and animal science. However, as the open access movement has gained momentum, content from many important and high impact journals are now in the public domain. The DOAJ (Directory of Open Access Journals) is a list of free full text peer-reviewed journals while the Stanford University's Highwire Press is the largest archive of free full-text scientific journal articles. Biomed Central is an open access publisher for peer-reviewed biomedical research that publishes more than 150 freely and permanently accessible journals. The PloS (Public Library of Science) also publishes open access, peer-reviewed journals, the contents of which are deposited in PubMed Central's free public archive. Cogprints is a collection of self-archived postprints in the areas of biology, computer science, neuroscience and other subjects related to the study of cognition. Other open source resources include CURATOR, DIVA, HKUST, MIT Open courseware, Scopus and NASA through the Open Archives Initiative.

Several general and specialized portals are also available for Biotechnology related news and information. Portals like Sciweb, Bio.com and biospace.com. Specialized Agribiotech based portals include CropBiotech.net, Pew initiative on Food and Biotechnology and Council for Biotechnology Information while Bioplanet.com and 2Can are specialized Bioinformatics portals. The Bioinformatics Links Directory is an online resource for useful tools, databases and resources for the molecular biologist organized by functional classification and accompanied by a brief synopsis as well as relevant citations. The life science literature databases and resources with brief descriptions are summarized in Table 5.

Bioinformatics Tools

A number of Bioinformatics programs and packages are now publicly available for search, comparison, prediction, modeling and analysis. Starting with the sequence, rapidly developing databases are accompanied by programs for quickly locating the similar sequences, generating optimal views required for facilitating the analysis and identifying the taxonomic relationships. Tools accompanying Gene maps are necessary for localization and display of genes. At the level of the structure, newly developed tools allow us to retrieve the structures of interest, locate similar structures and classify them by overall topology. Detailed tutorials for some of these can be found listed at serial no. 16 of the suggested reading for this chapter.

The tools are rapidly developing along with the newly evolving technologies for genomics and proteomics, their development being driven by the scientist's requirement and gaining momentum from the Open Source movement. A number of tools are now available for sequence & structure comparisons, pattern finding and prediction, structure prediction, gene expression analysis, functional characterization of proteins, drug design and pathway modeling. Some of these are summarized under this topic. Most of the bioinformatics search and analysis tools are now available as web servers, a complete list of which is compiled annually by Nucleic Acids Research, an open access journal. This list can be accessed at http://bioinformatics.ubc.ca/resources/links_directory/narweb2006/categorized.php .

Table 5: Selected Literature Databases

Open Access resources	
Public Library of Science	Peer-reviewed scientific and medical literature
Biomed Central	Peer-reviewed open access journals
DOAJ	Directory of Open Access Journals
Open Archives Initiative	Access to archived eprints
Dspace	Digital archive system
Bioline International	Not for profit electronic journals publisher
SPARC	Scholarly Publishing and Resources Coalition
INASP	International network for the availability of scientific publications
CODA	Caltech Collection of Open Digital Archives
EBI resources	
Medline, EBIMed	SRS interface to search more than 11 million life science citations and abstracts updated weekly
OMIM	Online Mendelian inheritance in Man database of genes and genetic disease
Patent abstracts	Abstracts of European patent applications
Taxonomy	Taxonomy database of international sequence database collaboration
NLM resources	
Medline Plus	Medical encyclopedia, health related topics and drug information
Cilinicaltrials.gov	Ongoing evaluation of new treatments
GHR	Genetics Home Reference
AIDSinfo	AIDS prevention, treatment and clinical trials
CCRIS	Chemical carcinogenesis Research Information System
NCBI resources	
PubMed	Abstracts of published journal articles with links to full text articles and information about library holdings. It can be searched by keyword, journal name, author name, PubMed ID, etc.
OMIM	Online Mendelian Inheritance in Man
OMIA	Online Mendelian Inheritance in Animals
PMC	(PubMed Central) Full text articles from Life Science Journals.
Journals	Detailed information on Biomedical journals.
Educational resources	
Biology Project	Biology lessons and learning resources

ActionBioscience	Educational resources
Biointeractive	Biology teaching materials
Biology Online	Dictionary, tutorials and articles in Biology
Biovisa	E-books, Free Journals, protocols and forum
VSNS	Biocomputing course, text-book and other resources
Kimball's Biology pages	Online Biology Textbook
Molecular Biology Web book	Online Textbook
Medconnect	Online resource for professionals in the field of Medicine
Science Gems	Links to science resources
World lecture Hall	Online course materials
Bionotebook	Directory of Biology web pages
Biosciences	Virtual Library of Biotechnology and life science
Bioinformatics Links Directory	Tools and databases for Molecular Biology
Cytogenetic resources	Images, links and software
Biotechnology related Portals	
Bioexchange	Industry related e-business service, tools, software and protocols
HUPO	Human Protein Atlas
Sciweb Bioweb Biospace	Generalized Biotechnology news and information portals
CropBiotechnet Pew Initiative Agricola	Agribiotech related portals
Bioplanet 2can	Bioinformatics portals
Search engines	
Google scholar Scirus Infomine	Science specific search engines
Bionotebook Biology Browser	Biology search Engines

Sequence alignment

Once a nucleotide or protein has been sequenced, the most common next step is to compare it with the known sequences present in the database. This involves alignment of the query sequence with those present in the database and represents a way of inferring structural, functional and evolutionary relationships between them. Comparison of DNA sequence between members of different species is based on the hypothesis of a common ancestor from which different organisms have been derived by mutation during evolution.

Similarity between two sequences can be expressed as an observable quantity. A threshold value of similarity can be used to infer a common evolutionary origin or homology between those sequences. Thus, a high level of similarity between two sequences is indicative of homology. This, in turn, indicates a shared ancestry and similarity in protein 3-D structure.

Of special interest to the biologist is the search and retrieval of previously characterized gene or genome sequences that are homologous to a new or unknown sequence -

a) For a nucleotide sequence, this can help in identification of a single gene, derivation of evolutionary relationships, identification of functional elements, etc.
b) Additionally, in case the sequence is expressed, this can help in predicting secondary or tertiary structure and binding sites of a protein.
c) In case of a genome, sequence similarity searches can be used for annotation. This includes prediction of gene structure and function, finding potential splice sites, introns, exons, promoter locations, etc. Sequence comparison of entire genomes from different organisms can be carried out for identification of similar or different sequences.

Global vs local alignments

If two similar sequences are used as the x and y coordinates of graph and identical residues are represented by a point on the graph, stretches of similarity in the sequence will appear as a diagonal as shown in Fig. 4. This is known as a dot matrix representation. In some versions, dots above different cut-off similarities are coded in different colors. Dot plots are used as visual representations of sequence alignments. They are powerful tools for finding multiple regions of local sequence similarity.

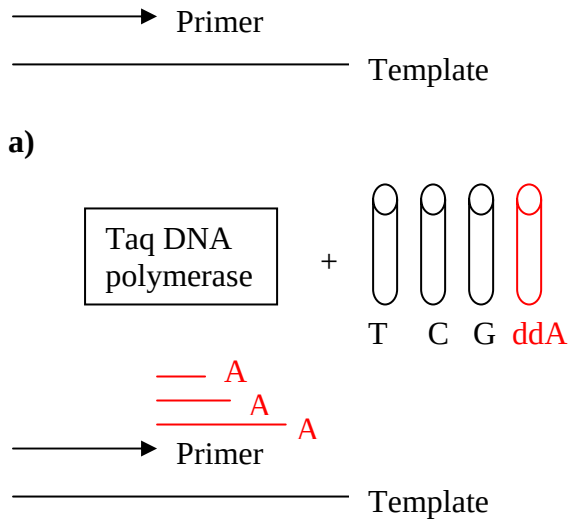

Fig 4: a) Hybridization of Template DNA strand with an oligonucleotide primer b) Synthesis of complementary strand in the presence of a pool of NTPs and a fluorescently labeled chain terminating ddATP, yielding nested fragments (red), which can then be separated by electrophoresis and identified by laser excitation

The likelihood of regions of local similarity being present is higher in case of protein sequences, since proteins from different families often share the same structural or functional sub units. Therefore, it is more helpful to do a local similarity search like BLAST when searching DNA and protein databases. Alignment of sequences along their entire length is achieved by global sequence alignment programs like FASTA. Global alignments are more useful once homology has been established. These are then used to generate a multiple sequence alignment as when building an evolutionary tree.

As there are a number of possible alignments for a sequence, optimal alignment programs determine the best possible alignment between two sequences using a scoring function that assigns positive values for identical residues and negative values for gaps or substitutions. The negative score for a gap in the alignment is known as gap penalty. It is found that some mutations do not affect the structure of a protein, yielding a functionally unchanged product. These are known as conservative mutations as against non-conservative mutations that yield proteins with altered structural and functional properties. A Substitution matrix assigns higher positive values for conservative or accepted mutations as compared to unconserved changes. It can enhance the sensitivity of an alignment, which Examples of such matrices include the PAM (Point Accepted Mutation) matrices and the BLOSUM matrices (Blocks Substitution Matrix). It is usually a fixed value and an addition deduction proportional to the length of the gap. Considering the large number of sequences in a database, it is also useful to find the statistical significance of an alignment, i.e. the chances of the similarity being coincidental. This is represented by the E-value of the alignment. The lower the E-value, the higher the statistical significance of the alignment. Lastly, amino acid sequence alignments are more sensitive and less error prone due to their higher information content. Therefore, protein sequence searches are preferred unless non-coding regions are being analyzed.

Database Searching

As discussed before, one of the easiest ways to identify a newly sequenced gene or protein is to compare it to previously sequenced genes. Due to the large amount of sequence information now available, it is now stored in databases. A number of gene, protein and genome databases are available for comparison as summarized in Tables 1, 2 & 3. Most of these databases can be queried with a sequence using database search programs that use heuristic methods to comb through the vast amount of sequence data. Thus, certain approximations are introduced into the program to increase the search speed at a small cost to the reliability of the results. Commonly used sequence alignment programs FASTA (Fast-All) and BLAST (Basic Local Alignment Search Tool) options at the NCBI website are based on heuristic algorithms. They break both the query and the database sequences into fragments (words) and initially seek matches between these fragments and then attempt to extend the word hits in either direction. Each Entrez protein entry shows a Blink that shows the pre-computed Blastp output for that entry against the nr (non redundant) database. The Blink display allows easy navigation of 200 sequence hits to find out the best results, organisms from which best hits have been reported, evolutionary relationships of those organisms, similar sequences with known structure, conserved domains, etc. Blast and Fasta services now form an integral part of database searches. They can also be used to search and analyze different types of nucleotide or protein sequence and structure data through the EBI toolbox or Expasy servers. Apart from Blast, NCBI offers public domain tools for aligning mRNA and cDNA sequences to a genomic sequence (Spidey and Splign respectively). Spidey attemts to determine the intron/exon structure of the genomic DNA and reports the mRNA alignments with the exons. Splign uses a heuristic algorithm to efficiently report the optimal local alignment of cDNA to genomic DNA.

The Blast output

Similar sequences in the database are returned as hits in order of their scores and statistical significance. Each such hit obtained is linked to its annotation, literature citation and structure information through a clickable link. Thus, the blast search allows us to search several specific nucleotide and protein databases. The search can also be refined to a specific organism, database field, molecule type, gene location and date. The limits can also be set to exclude certain kinds of sequences or to combine two or more of the aforementioned options. One of the preliminary information output by NCBI Blast is the putative conserved domains detected in the sequence. This tells us about important protein features and functions. Subsequent blast output contains a RID (Request ID) that can be used to retrieve the Blast search for future reference. Query sequence information and databases searched are then summarized. The link to the Taxonomy reports shows an Organism report which groups the results by organism, a Lineage report that shows a simplified view of the relationships between the organisms according to their taxonomic classification and the Taxonomy report gives detailed taxonomic information about the organisms. The Graphical overview shows a visual alignment of the top 50 hits with the query sequence. It is color coded to reflect the degree of similarity detected. Mousing over any of the bars representing the hits displays sequence and score information. The results are then displayed in a hit table ranked by statistical significance. The hit table contains four distinct columns i.e. a hyperlink to the sequence record with a brief description, a bit score calculated from the observed gaps and substitutions hyperlinked to the respective sequence alignment, the E-value and an icon that links each sequence to records in outside databases like [L] Locuslink or [S] 3D structure. A number of formats are also available for viewing the output results. The default parameter is pairwise alignment format. It returns the query sequence lined up with each of the hits found. Sequences can also be anchored to the query sequence with or without display of information on identical residues. Each pairwise alignment of the query and the hit sequence shows a letter between the two sequences for identity and a blank in the positions that do not match. The N (nucleotide) and X (proteins) string denotes the low complexity regions and dashes denote gaps. The number identical residues, conserved substitutions and gaps are reported for each alignment.

Multiple Sequence Alignment

Once one or more hits have been found in the database for a query protein sequence, simultaneous alignment of multiple sequences is done. Additionally a multiple sequence alignment may also be attempted for proteins with converging function or structure that have evolved independently. Typically, redundant or identical sequences should be excluded from an alignment and those with high but comparable similarity should be used. Multiple sequence alignments can be either global (with gaps) or local (aligning only the region between gaps).

Including multiple sequences with high level of similarity improves the accuracy and sensitivity of an alignment. An alignment that corresponds to the structure is the most chemically and biologically relevant one. It can be used to predict secondary structure, accessibility and function for a novel protein. Another application of a multiple sequence alignment is construction of PCR primer design using known DNA sequences. Multiple sequence alignments can also be used to identify new members of a protein family by identifying the conserved pattern or Blocks. A similar application is to search the sequence database by constructing a profile (possible sequence variation at each position of the sequence) from the multiple sequence alignment and using it to search the database for new

members of the same family. The sequence alignment program PSI-BLAST (Position Specific Iterated BLAST) at NCBI uses a position specific scoring matrix constructed from a gapped MSA of the hits found during the search. This can increase the sensitivity of the search, allowing for distantly related sequences to be located. The PROSITE method finds characteristic patterns (sequence motifs) for some of the protein facilities and can identity uncharacterized proteins.

To do a multiple sequence alignment, the database hits containing the region of interest are first edited to similar sequence length. Automatic alignment can then be done using programs like CLUSTALW, a hierarchical program that works by first generating pair-wise alignment of pairs of sequences. An ungapped alignment can be made using BlockMaker. The alignment is manually inspected, especially in the regions with the gaps. The quality of the alignment can be determined by a test of statistical significance. However, the most accurate alignments are constructed by taking experimental data like catalytic site and conserved secondary structure elements into consideration. A useful option offered with database searches at the EBI website is Mview, a program that allows the search output to be directly converted to a color coded multiple sequence alignment. Mview can also be obtained and run in standalone mode. Various visualization tools are available to identify the residues with similar physico-chemical properties using different coloring / shading schemes. A guide tree is then constructed by cluster analysis such that similar sequences are closer together than dissimilar ones, thus allowing for deduction of evolutionary relationships using molecular phylogenetic analysis.

Phylogenetic Ananlysis

Using the different DNA patterns in this technique, it is possible to study the evolution of an organism. This is based on the premise of a common ancestral DNA and evolution of genomes by slow accumulation of mutations. Therefore, genomes with fewer differences will have a recently shared common ancestor. Additionally, tracking the evolution pattern of individual genes can tell us about which genes have been conserved in the genomes and which have been horizontally transferred. Phylogenetic analysis can be used to map the genes in two organisms that may have similar functions. It can also be used to map the changes in a rapidly changing genome, like that of a virus. Commonly used phylogenetic analysis programs include PHYLIP (Phylogenetic Inference Package) and PAUP (Phylogenic Analysis using Parsimony). Three main methods used for phylogenetic analysis are parsimony, distance and maximum likelihood. The prediction can be made using either of these methods for either DNA or protein sequences. The reliability of the predictions can then be evaluated.

The tree of life web project is an online collaboration of the world's biologists that provides information on the diversity, evolutionary relationships and characteristics of the earth's organisms. Treebase is a relational database of phylogenetic knowledge hosted by the University at Buffalo. Starting with the root of all life on earth, each species (branch) is linked in a hierarchical structure, constituting the tree of life. The tree thinking group provides various educational resources on phylogenetic perspectives. The NCBI Taxonomy project attempts to provide phylogenetic taxonomy classification for all the organisms (including extinct ones) that are represented by sequence data in the nucleotide or protein database. Each blast hit is accompanied by a detailed taxonomy report. The taxonomy browser may be used to find the taxonomic position and to retrieve the associated sequence and structure data on an organism. Taxplot is a tool for comparison of translated proteins of one reference genome with two others using Blast and plotting the output. The NCBI COG

database is an attempt for the phylogenetic classification of proteins as it identifies the members of an orthologous (common ancestral origin) group.

Protein structure related tools

As discussed before, every protein has a three-dimensional spatial organization or 3-D structure. There are four broad levels of protein structure, namely primary (sequence), secondary (arrangement of polypeptide backbone), tertiary (overall three-dimensional structure) and quaternary (spatial arrangement of two or more polypeptide chains). In addition, secondary structure elements often exist in small combinations named as structural motifs e.g. helix turn-helix. A profile generated from a multiple sequence alignment highlights the conserved secondary structural elements and thus helps to identify the structural motifs. The polypeptide chain of a protein is folded into structurally compact and functionally independent domains.

Analysis of the known protein structures shows that in spite of the huge number of protein sequences, there are fewer number of structural folds (core 3D structure) in the protein database. This allows us to predict the three-dimensional structure of a protein from its sequence if the structure of a homologous protein is known. In this context, it is useful to classify proteins by structure. The SCOP (Structural Classification of Proteins) database and CATH (Class Architecture, Topology and Homologous super family) are an effort in this direction. VAST (Vector Alignment Search Tool) and DALI (Distance Matrix Alignment) are alignment tools used to compare and identify protein structures.

A number of public domain molecular visualization tools are available e.g. Rasmol, Chime, Protein Explorer, Pymol etc. Cn3d is the NCBI visualization software. Molecular visualization tools help us view, measure and render proteins in many different ways in order to identify the secondary structure elements, charge, hydrophobicity, binding site, accessibility etc. A collection of the molecular visualization resources and tutorials (MolviZ.org) is maintained by E. Martz.

Some useful predictive tools for proteins are:-

1. Identification of a protein with a given amino acid composition (e.g. AACompident), isoelectric point (pI) and molecular weight (TagIdent), Short Sequence tags, mass or a combination of all these (Multident). Identification of a protein family is also possible from its amino acid composition (Propsearch). This may be useful if querying a database with a sequence does not return any significant hits.
2. Identification of a protein from its peptide mass finger print (Mascot) or raw MS/MS data (PepMapper, Mascot and Peptidesearch). MS/MS spectra can also be used to identify peptides and proteins by searching libraries of known sequences using OMSSA (Open Mass Spectrometry Search Algorithm) at NCBI.
3. Prediction of PI or Molecular weight of a protein from its amino acid sequence (pI Tool), Trans membrane regions (Tmpredict, a database of transmembrane domains) and various other physical and chemical parameters (ProtParam, SAPS-Statistical Analysis of Protein Sequences)
4. A number of tools for protein study and analysis are listed in the Expasy Proteomics Tools list. Prediction of some of the structural features of a protein can be done in the following ways –

- Prediction of solvent accessibility (Predictprotein, Protscale)
- Prediction of Coiled coil regions from the sequence (Paircoil)
- Prediction of secondary structure (GOR- Garnier, Osguthorpe and Robson, SIMPA96, Pfam, nnpredict).

5. Identification of single motifs (PROSITE), Multiple motifs (Prints, Blocks) and Profiles (Profilescan) in a given sequence.
6. Tertiary structure prediction of a given sequence –The protein database (PDB) can be searched with a query sequence to find a homologous protein for which the structure has been solved. Subsequently, Homology or Comparative modeling can be done (SwissModel). In those cases where no significant identity with the known protein structure is found, remote homology modeling or threading can be used (Threader). Threading involves querying the structure fold database with a protein sequence for which the structure has not been solved. The sequence is aligned to a set of 3D environment descriptors (like accessibility, local secondary structure, structural interactions, etc.) for each residue of the solved structure according to a pre-computed scoring table. The best fit is then determined. Alternatively, *ab initio* prediction methods may also be used to predict the structure. Structure Prediction with Online Resources (SPORes) is an online service for protein structure prediction.
7. Prediction of protein function from its sequence and structure – This includes prediction of biochemical function and cellular location. For this, the functional domains are identified and their function is assigned (HNB-Helmholtz Network for Bioinformatics). The CDART tool from NCBI allows fast annotation of protein domains by using a RPS (Reverse Position Specific) Blast against the Conserved domain database. The RPS blast searches a database of pre-calculated PSSMs with the query sequence.

Genomics related tools

Sequencing of a genome produces a large amount of data since each stretch has to be sequenced many times to minimize the errors. Apart from large size, genomic data is also very complex because of presence of various type of repetitive elements like VNTRs (Variable number tandem repeats), satellite DNA and multiple copies. Therefore, reconstruction of the entire genomic sequence from overlapping strings of subsequences is a complex computational task. The sequencing of the human genome necessitated the development of novel assembly tools to efficiently and accurately interpret the sequence information. Phrap (Phil's Revised Assembly Program) is a commonly used assembling programs. It is useful for assembling small sequences into a contig (Contiguous – adjacent, connecting without a break) while calculating an error probability for each position in the sequence. It also identifies the potential SNPs and possible misassembly sites. Other assembly tools include Assembler 2.0, PCAP (Parallel Contig Assembly program) and CAP3 (Contig Assembly Program). The NCBI ModelMaker option allows a view of the contig sequence records, mRNA and EST records, putative exons and other evidence on which the gene sequence has been assembled or modeled.

With the completion of the human genome project, the genetic map of the entire human genome has become available. A genetic map is constructed on the basis of meiotic crossover frequencies and shows the relative positions of the various genes with respect to each other on the chromsome. This is useful for identification of a gene causing an inherited disease. A genetic map is constructed with the help of various DNA markers. Some of the identified DNA markers can be linked to disease in affected individuals.

Once a genome has been sequenced and assembled, the protein coding regions and genes need to be identified and annotated. Many of the annotation programs work on the principle of sequence similarity to known genes or ESTs. They are able to identify sequence patterns characteristics of regulatory and splice sites in expressed genes. One of the most popular ORF identification tool is GRAIL (Gene Recognition and Analysis Internet Link). It can be accessed as part of GrailEXP (Grail Experimental gene discovery suite). Several other approaches can be used to predict and analyze genes. Some are based on Hidden Markov Models e.g. GLIMMER (Gene Locator and Interpolated Marker Modeler) while others like Genscan use probabilistic models that incorporate information on the basic transcriptional, translational and splicing signals, gene length and compositional features. Apart from prediction of genes and ORFs, several tools are available for predicting splice sites, promoter signals, protein binding sites, repetitive DNA and tRNA genes. The NCBI e-PCR tool is useful for identifying the STS sites in a given nucleotide sequence and is thus useful for constructing genetic and physical maps of the genome. The NCBI ORF Finder locates all the possible exons in a given sequence by searching for standard and alternative start and stop codons. Vecscreen is a utility for screening nucleic acid sequences before analysis or submission to detect regions of vector, linker or adapter origin.

Whereas a genetic map is based on meiotic crossover frequencies and is measured in centimorgan, a physical map describes the characteristics of the DNA sequence and is measured in base pairs (bp). A physical map therefore helps in locating a specific gene for cloning. The DNA map consists of several molecular markers e.g. RFLPs (restriction Fragment Length Polymorphisms), VNTRs, STSs and ESTs. Based on these, a number of different maps are available. DbSTS and dbEST are two such map resources at the NCBI website.

Genome analysis and structure-function annotation involves the following steps:-

1. Identifying sequence homologs with database searching programs.
2. Identifying relevant motifs and domains in the Pfam, Prosite, Blocks, Prints or Smart database.
3. Predicting structural features like transmembrane regions
4. Predicting Secondary and tertiary structure of expressed proteins

A public domain software package for performing the steps listed above is SEALS (A system for Easy analysis of Lots of Sequences).

The availability of complete genomes allows for –

a. **Identification of various previously unknown genes –** Tools for whole genome comparison and functional annotation include KEGG (Kyoto Encyclopedia of Genes and Genomes) and WIT. The MBGD (Microbial Genome database) focuses on the completed microbial genomes and allows for comparison among them.

   Prediction of protein function involves not only sequence similarity but also the cross-genome similarities in terms of gene number, organization and divergence. The COG database at NCBI is especially useful for functional assignment. It groups proteins from conserved families into orthologous groups. Orthologs are descended from the same gene and are thus expected to have the same function. Thus, searching for sequence similarity in the COG database can predict protein function. Additionally, each genome in the organism specific genome databases at the NCBI website is linked to its detailed online information resources, the detailed genetic map view information, its entry in the genomes project database and allows

organism/species/class specific blast searches. Of special interest to virologists is the retrovirus resources at NCBI that aid in genotyping of a strain, global alignment of multiple sequences, automatic sequence annotation for HIV-1 and detailed annotated maps for sixteen other retroviruses.

b. Understanding Evolutionary relationships between different species – The COG database can be searched using a tool that allows us to identify a phylogentic profile, a set of genomes in which that cluster is present. COGs with a particular pattern can be selected. Since members of the same pathway show the same pattern, they can be grouped together. It is also possible to improve functional prediction for proteins using this method.
c. Comparison of the similarities and differences between genomes of different organisms – The pattern search tool of the COG database allows for the use of logical operators like AND, OR and NOT. It can be used to delineate subsets of a genome that are specific to a particular trait of the organism e.g. bacterial pathogenesis.
d. Reconstruction of metabolic pathways in different organisms – Using the steps described above protein functional prediction, grouping and analysis of phylogenetic profiles can help in deciphering the metabolic pathways of an organism.

Metabolic Pathways

Within the cell, the basic unit of life, a series of chemical processes are responsible for synthesis and breakdown of various essential molecules like glucose, fatty acids, amino acids and many others. These reactions are collectively known as the cellular metabolic pathways. Various other signal transduction pathways are responsible for relay of signals within the cell. Each pathway consists of one or more ligands and molecular effectors. It usually has several layers of regulatory control and often integrates with other pathways for fine control and coordination of biological processes, forming a metabolic network. The knowledge and understanding of the processes controlling the pathways of metabolism and signal transduction is crucial to our understanding of health and their changes leading to disease.

Several online pathway resources are available. Main among these is the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathway resource for metabolism, information processing, cellular processes and human diseases. Digitalized version of Roche Applied Science's biochemical pathways wall charts are also maintained by Expasy server. The Metacyc database maintains a nonredundant curated collection of experimentally elucidated pathways from more than six hundred different organisms. It has a query and visualization interface for prediction of metabolic pathways for newly sequenced genomes. It also finds application in metabolic engineering and comparison of biochemical pathway networks. The open source Biocarta pathways resource provides a graphic rich visual format for both classical and newly suggested pathways as well as information for over 120,000 genes from different organisms. It also provides tools and templates for pathway modeling.

Microarray data analysis

Gene expression analysis techniques have been used to study changes in the cell's internal environment during a host of physiological and pathological processes. Examples include fruit ripening, stages of cell cycle, environmental shock and development of cancer. These are then compared with standards to study the different genes that are involved during these metabolic changes. Genes that are expressed or repressed together are considered to be "guilty by association". Analysis of a gene's expression pattern under different conditions

provides clues to its interactions with other cellular proteins and thus helps in making predictions about the gene's function. It can also help in disease prognosis, diagnosis, studying the effect of a drug or developing new drugs.

Each biological sample in a microarray experiment is capable of yielding 4,000 to 50,000 gene profiles. A complete experiment may involve hundreds of microarrays. Therefore, large expression data sets are often the norm. After data collection, results from multiple microarray experiments are first normalized to make them comparable in order to integrate them into a single analysis. Additionally, the data is scanned for noise and artifacts. Subsequent expression level analysis can be divided into two main categories, namely supervised or unsupervised. Supervised approaches use knowledge (e.g. gene function or regulation pattern) from outside the microarray experiment to drive the analysis. They are generally used to identify genes with expression levels that are significantly different between subgroups and to find genes that accurately predict a characteristic of the sample. Unsupervised methods, however, are geared to determine patterns within the dataset and do not require any outside knowledge. They determine those genes that have an interesting pattern and groups with similar patterns of gene expression. After the analysis, a considerable amount of work still remains in order to interpret the biological significance of the analysis. Some of the freely available software for microarray analysis include Cluster and Treeview, Genepattern and Multi Expression Viewer. Powerful data analysis packages for statistical analysis are BRB tools and ICBR AnalyzeIt tools.

Apart from this, SageMap is an NCBI tool for differential analysis of SAGE data in the GEO database. The Sage data in the GEO database is collected from individual labs as well as the NCI's CGAP (Cancer Genome Anatomy Project), an effort to map the expression profiles of normal, pre-cancerous and cancerous cells in order to aid their diagnosis and treatment. Unigene DDD (Digital Differential Display) from NCBI allows online comparison of gene expression profiles of different cDNA libraries. DDD uses a statistical test to identify those genes that vary in expression from one tissue to another. The SageGenie available at the CGAP website is a tool for visual display and analysis of human and mouse gene expression profiles. The CGAP website also hosts a number of gene, chromosome, tissue, pathway and RNAi (interference RNA) resources. Gene Ontology based tools for expression data analysis are also available that cluster the genes and to identify the enriched pathways. These include downloadable programs like Genmapp and integrative online tools like DAVID (The Database for Annotation, Visualization and Integrated Discovery) provided by NIAID (National Institute for Allergic and Infectious Diseases).

Tools for Drug discovery

Drug discovery is an expensive and time-consuming process that requires input from many diverse experimental fields. HTS (High Throughput Screening) is an experimental method that allows the screening of large compound libraries against selected drug targets to allow selection of lead compounds. However, this is not only cost intensive but also does not yield compounds with favorable ADME (Absorption, Distribution, Metabolism and Excretion) properties. In recent times, *in silico* methods have been used to aid the process of drug discovery at many stages, thus optimizing cost and time of development. Finding suitable drug candidates with virtual screening and design procedures has become an attractive option. Energy minimization programs (e.g. GAMESS – General Atomic and Molecular Electronic Structure System, Ghemical) are available for optimizing the structure. Docking programs (e.g. DOCK, AutoDock, ArgusDock) are used to predict binding of virtual libraries of drug-like compounds to selected targets. Visualization tools (e.g. Jmol, Molvis) are

available for visualization of drug binding and interaction. QSAR (Quantitative Structure Activity Relationship) studies are used to relate activities of compounds to their chemical and structural properties. This requires generation of various types of molecular descriptors related to polar surface area, hydrophobicity, solubility, hydrogen bond donating and accepting properties, etc. and can be carried out with online resources like VCCL (Virtual Computational Chemistry Laboratory), MarvinBeans and SOMFA (Self Analyzing Molecular Field Analysis). Chemistry drawing software ACD/Labs ChemSketch and ISIS Draw from ACD (Advanced Chemical Laboratory) and MDL (Molecular Design Limited) respectively is available for structure illustrations of small molecules and can also be used to calculate molecular properties. Open source software libraries like MMTK (Molecular Modelling Tool Kit) provide easy to use integrated interfaces for molecular simulation and modeling.

Conclusion

Bioinformatics is a rapidly evolving field that lies at the interface of chemistry, mathematics, biology and computer science. It requires input from many diverse areas and is aiding the rapid development of areas in biotechnology, genetics, genomics, proteomics and drug design by providing a more global perspective in design of experiments. It allows us to transfer information to new systems from other, well- characterized organisms and experiments using an integrative approach. As an example, classification of proteins into structurally and functionally related groups allows us to make intuitive inferences about the possible evolutionary relationships between them as well as to predict the structure and functions of newly discovered proteins. Similarly genomic maps can help us visualize the exact positions of genes on a given chromosome and to examine the details of the upstream and downstream coding and non-coding regions to determine regulation and linkage in health and disease. Small changes in DNA sequences can not only be related to disease propensities but can potentially be used to tailor pharmacological treatment and dosage as per individual needs. Finally, as an end result of all the biological studies conducted, biologists hope to create a composite picture of the dynamic cell and the organism as a whole. Thus, by understanding the basic biological processes, we are better positioned to exploit them for industrial use. Additionally, the information from biological pathogens leads to better tools for preventing infectious diseases while the understanding of the human system and processes positions us for diagnosis, treatment and prevention of complex disorders like diabetes and cancer. The future of Bioinformatics is the integration of genomics, proteomics and bioinformatics generated information to produce a complete view of the entire biological system in a discipline known as Systems Biology.

The biological databases and bioinformatics tools and applications are growing at an enormous pace. They will continue to grow as newer and emerging technologies provide a plethora of information about the biological systemsof different organisms. The topics and resources covered in this chapter only reflect some of the dominant areas for the use of computational tools and are intended as introductory reading. The future of Bioinformatics is aptly reflected in the following finding of the Science 2020 Group (Venice, July 2005) as reported in “Towards 2020 Science”.

Indeed we believe computer science is poised to become as fundamental to biology as mathematics has become to physics. We postulate this because there is a growing awareness among biologists that to understand cells and cellular systems requires viewing them as information processing systems……We believe this is a potential starting point for fundamental new developments in biology, biotechnology and medicine.

List of Internet resources in alphabetical order

2Can	http://www.ebi.ac.uk/2can/home.html
3D Domains	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Domains
3D Genomics	http://www.sbg.bio.ic.ac.uk/3dgenomics
AACompIdent	www.expasy.org/tools/aacomp/
AAIndex	www.genome.ad.jp/dbget/
ABS	http://genome.imim.es/datasets/abs2005/abs.html
ActionBioscience	http://www.actionbioscience.org/
Archaeal genome browser	http://archaea.ucsc.edu/
ArgusDock	www.planaria-software.com/index.htm
Arrayexpress	http://www.ebi.ac.uk/arrayexpress/?
ASC	http://bioinformatica.isa.cnr.it/ASC/
ASDB	http://hazelton.lbl.gov/~teplitski/alt/
Assembler2.0	www.tigr.org/software/assembler/
AutoDock	www.scripps.edu/mb/olson/doc/autodock/
Bankit	www.ncbi.nln.nih.gov/BankIt/
Bio.com	http://www.bio.com/
Biocarta	http://www.biocarta.com/genes/index.asp
Bioexchange	http://www.bioexchange.com/
Bioinformatics Links Directory	http://bioinformatics.ubc.ca/resources/links_directory/
Biointeractive	http://www.hhmi.org/biointeractive/index.html
Bioline	http://www.bioline.org.br/
Biology Online	http://www.biology-online.org/
Biology Project	http://www.biology.arizona.edu/
Bionotebook	http://www.pasteur.fr/recherche/BNB/bnb-en.html
Biosciences	http://vlib.org/Biosciences
Biovisa	http://biovisa.net/index.php3
Bioweb	http://cellbiol.com/
Biozon	http://biozon.org/
BLAST	www.ncbi.nlm.nih.gov/BLAST
Blink	www.ncbi.nlm.nih.gov/sutils/static/blinkhelp.html
BLOCKS	blocks.fhcrc.org/
BMC	http://www.biomedcentral.com/
Bookshelf	ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed
BRB Tools	http://linus.nci.nih.gov/BRB-ArrayTools.html
BRENDA	http://www.brenda.uni-koeln.de/
BSORF	http://bacillus.genome.jp/
CampyDB	http://campy.bham.ac.uk/
Cancer chromosomes	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=cancerchromosomes
CATH	cathwww.biochem.ucl.ac.uk/latest/index.html
CCRIS	toxnet.nlm.nih.gov/cgi-bin/sis/htmlgen?CCRIS
CDART	www.ncbi.nlm.nih.gov/Structure/lexington/lexington.cgi?cmd=rps
CDD	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=cdd
CGAP	cgap.nci.nih.gov/

CGED	http://cged.genes.nig.ac.jp/cged2/cgi-bin/input.cgi
ChemIDPlus	chem2.sis.nlm.nih.gov/chemidplus/chemidlite.jsp
ChemSketch	www.acdlabs.com/download/chemsk.html
Chime	www.umass.edu/microbio/chime/getchime.htm
ChimerDB	http://genome.ewha.ac.kr/ChimerDB/
cisRED	http://genome.imim.es/datasets/abs2005/abs.html
Clinical trials	clinicaltrials.gov/
CLUSTALW	ftp://ftp.ebi.ac.uk/pubsoftware
Cluster and Treeview	http://rana.lbl.gov/EisenSoftware.htm
Cn3D	www.ncbi.nlm.nih.gov/Structure/CN3D/cn3d.shtml
CODA	http://library.caltech.edu/digital/index.html
Cogent	http://cgg.ebi.ac.uk/services/cogent/
Cogprints	http://cogprints.org/
COGs	www.ncbi.nlm.nih.gov/COG/
CORG	http://corg.molgen.mpg.de/
Cosmic	http://www.sanger.ac.uk/genetics/CGP/cosmic/
Council for Biotechnology Information	http://www.whybiotech.com/
CropBiotech	http://www.isaaa.org/kc/
Curator	http://mitizane.ll.chiba-u.jp/curator/index_e.html
CytoD	http://www.changbioscience.com/cytogenetics/cyto.htm
Cytogenetics resources	http://www.kumc.edu/gec/prof/cytogene.html
DALI	www.ebi.ac.uk/dali/
Database of genomic variants	http://projects.tcag.ca/variation/
DAVID	http://niaid.abcc.ncifcrf.gov/
DBD	http://dbd.mrc-lmb.cam.ac.uk/DBD/index.cgi?Home
DbEST	www.ncbi.nlm.nih.gov/dbEST/
DbGSS	www.ncbi.nlm.nih.gov/dbGSS/index.html
DbSNP	www.ncbi.nlm.nih.gov/SNP/
DbSTS	www.ncbi.nlm.nih.gov/dbSTS/
DDBJ	www.ddbj.nig.ac.jp
DDD	www.ncbi.nlm.nih.gov/UniGene/ddd.cgi?
DIP	dip.doe-mbi.ucla.edu/
DIVA	http://www.diva-portal.org/
DOAJ	http://www.doaj.org/
DOCK	dock.compbio.ucsf.edu/
Drugbank	http://redpoll.pharmacy.ualberta.ca/drugbank/
Dspace	http://dspace.org/
EBI Completed Genomes	www.ebi.ac.uk/genomes
EBI databases & Tools	http://www.ebi.ac.uk/services/
Ecocyc	http://ecocyc.org/
EID	http://hsc.utoledo.edu/bioinfo/eid/
EMBL	www.ebi.ac.uk/embl/
Ensembl	www.ensembl.org
Entrez Proteins	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein
Enzyme	http://www.expasy.org/enzyme/

e-PCR	www.ncbi.nlm.nih.gov/sutils/e-pcr/
Expasy Biochemical Pathways	http://www.expasy.org/tools/pathways/
Exprot	http://www.cmbi.kun.nl/EXProt/
FASTA	www.ebi.ac.uk/fasta33
Flybase	http://flybase.bio.indiana.edu/
FUGOID	http://www.oxfordjournals.org/nar/database/summary/782
GAMESS	www.msg.ameslab.gov/GAMESS/GAMESS.html
GDB	www.gdb.org/
Genbank	www.ncbi.nlm.nih.gov/Genbank/
Genecards	http://www.genecards.org/index.shtml
Genepattern	http://www.broad.mit.edu/cancer/software/genepattern/
Genmapp	www.genmapp.org
GenomeAtlas	http://www.cbs.dtu.dk/services/GenomeAtlas/
Genomenet	http://www.genome.jp/
GENSAT	www.ncbi.nlm.nih.gov/projects/gensat/
Genscan	genes.mit.edu/GENSCANinfo.html
GEO	ncbi.nlm.nih.gov/geo/
Ghemical	www.uku.fi/~thassine/ghemical/
GHR	ghr.nlm.nih.gov/
GOLD	www.genomesonline.org/
Google	www.google.com
GOR	npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_gor4.html
GrailEXP	compbio.ornl.gov/grailexp/
GUMC	gumc.georgetown.edu/
Harvester	http://harvester.embl.de/
HGMD	http://www.hgmd.org/
Highwire Press	http://highwire.stanford.edu/lists/freeart.dtl
HIV/AIDS SIS	http://sis.nlm.nih.gov/hiv.html
HKUST	http://repository.ust.hk/dspace/
hmtDB	http://www.genpat.uu.se/mtDB/
HNB	www.hnbioinfo.de/
HomoloGene	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=homologene
HOWDY	http://www-alis.tokyo.jst.go.jp/HOWDY/
HPID	http://wilab.inha.ac.kr/hpid/
HS3D	http://www.sci.unisannio.it/docenti/rampone/
Human Protein Atlas	www.proteinatlas.org/
HUMHOT	http://www.jncasr.ac.in/humhot/
HUPO	www.hupo.org/
ICBR AnalyzeIt	http://genomics3.biotech.ufl.edu/AnalyzeIt/AnalyzeIt.html
INASP	http://www.inasp.info/peri/free.html
Infomine	infomine.ucr.edu/
Interpro	www.ebi.ac.uk/interpro/
IPD	http://www.ebi.ac.uk/ipd/
Isfinder	http://www-is.biotoul.fr/
ISIS Draw	www.mdli.com/
Islander	http://129.79.232.60/cgi-bin/islander/islander.cgi
Jmol	jmol.sourceforge.net/

KEGG	http://www.genome.jp/kegg/
Kimball's Biology pages	http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/
Macie	http://www.expasy.org/enzyme/
Marvin Beans	www.chemaxon.com/marvin/do-download.html
Mascot	www.matrixscience.com/
Medconnect	http://www.medconnect.com/
Medline	http://srs.ebi.ac.uk/srsbin/cgi-bin/wgetz?-page+LibInfo+-lib+MEDLINE
MedlinePlus	medlineplus.gov/
MEGX	http://www.megx.net
Metacyc	http://metacyc.org/
MetaDB	http://www.neurotransmitter.net/metadb/metadb.php
MethDB	http://www.methdb.de/
MGC	mgc.nci.nih.gov/
MGED society	http://www.mged.org/
MGI	www.informatics.jax.org/
MIPS	http://mips.gsf.de/
MIPS-PPI	http://mips.gsf.de/proj/ppi/
MIT open courseware	http://ocw.mit.edu
Mitelman database	http://cgap.nci.nih.gov/Chromosomes/Mitelman
MMTK	starship.python.net/~hinsen/MMTK/
Modbase	http://modbase.compbio.ucsf.edu/modbase-cgi-new/index.cgi
ModelMaker	www.ncbi.nlm.nih.gov/mapview/static/ModelMakerHelp.html
Molecular Biology Web book	http://www.web-books.com/MoBio/
Molecular Toolkit	www.vivo.colostate.edu/molkit/index.html
MolVis	molvis.sdsc.edu/visres/index.html
MolviZ	www.umass.edu/microbio/chime/
MSD-EBI	www.ebi.ac.uk/msd/
Multi Expression Viewer	http://www.tm4.org/
MultiIdent	www.expasy.org/tools/multiident/
NCBI	www.ncbi.nlm.nih.gov/
NCBI Gene	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
NCBI Genome Project	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=genomeprj
NCBI Genomes	www.ncbi.nlm.nih.gov/Genomes
NCBI retrovirus resources	www.ncbi.nlm.nih.gov/retroviruses/
NCBI Tools	ncbi.nlm.nih.gov/Tools/
NHGRI	www.genome.gov/
NLM	www.nlm.nih.gov/databases/
nnpredict	www.cmpharm.ucsf.edu/~nomi/nnpredict.html
Nucleotide	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide
OMIA	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=omia
OMIM	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM
OMSSA	pubchem.ncbi.nlm.nih.gov/omssa/

Oncomine	http://www.oncomine.org/main/index.jsp
Open Archives Initiative	http://www.openarchives.org/
ORF Finder	www.ncbi.nlm.nih.gov/gorf/gorf.html
Paircoil	theory.lcs.mit.edu/paircoil
Paup	http://paup.csit.fsu.edu/
PCAP & CAP3	seq.cs.iastate.edu
PDB	www.wwpdb.org/
PDBj	www.pdbj.org/
PEDANT	http://pedant.gsf.de/
PepMapper	wolf.bms.umist.ac.uk/mapper/
Peptidesearch	www.narrador.embl-heidelberg.de/GroupPages/PageLink/peptidesearchpage.html
Pew Initiative	http://pewagbiotech.org/
Pfam	www.sanger.ac.uk/Pfam/ www.sanger.ac.uk/Software/Pfam/
Pharmgkb	http://www.pharmgkb.org
Phred	www.phrap.org/
Phylip	http://evolution.genetics.washington.edu/phylip.html
PI Tool	www.expasy.org/tools/pi_tool.html
PIR	pir.georgetown.edu/
PloS	http://www.plos.org/
PMC	www.pubmedcentral.nih.gov/
PMD	http://pmd.ddbj.nig.ac.jp/
PopSet	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Popset
Predictprotein	www.predictprotein.org/
PRINTS	bioinf.man.ac.uk/dbbrowser/PRINTS/
Prints	umber.sbs.man.ac.uk/dbbrowser/PRINTS/
Probe	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=probe
ProDom	protein.toulouse.inra.fr/prodom.html
Profilescan	myhits.isb-sib.ch/cgi-bin/motif_scan
Propsearch	abcis.cbs.cnrs.fr/propsearch/
Prosite	www.expasy.org/prosite/ us.expasy.org/prosite
Protein Explorer	www.umass.edu/microbio/chime/pe/protexpl/frntdoor.htm
Proteome Database system	http://www.mpiib-berlin.mpg.de/2D-PAGE
ProtEST	www.ncbi.nlm.nih.gov/UniGene/ProtEST/
PROTOMAP	protomap.stanford.edu/
Protscale	www.expasy.org/tools/protscale.html
PROW	www.ncbi.nlm.nih.gov/prow/
PubChem BioAssay	www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pcassay
PubChem Compund	www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pccompound
PubChem Substance	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pcsubstance
PubMed	www.ncbi.nlm.nih.gov/pubmed
PubMed Journals	ncbi.nlm.nih.gov/entrez/query.fcgi?db=Journals
PyMol	pymol.sourceforge.net/
Rasmol	www.umass.edu/microbio/rasmol/index2.htm
Ratmap	http://ratmap.org/

RCSB	www.rcsb.org/pdb/Welcome.do
ReadSeq	http://thr.cit.nih.gov/molbio/readseq/
RefSeq	www.ncbi.nlm.nih.gov/RefSeq/
RGD	rgd.mcw.edu/
SAGE	www.ncbi.nlm.nih.gov/SAGE/
SageGenie	cgap.nci.nih.gov/SAGE
Sanger Institute	www.sanger.ac.uk/
Science gems	http://www.sciencegems.com/
Scirus	www.scirus.com
Sciweb	http://www.sciweb.com/
SCOP	scop.mrc-lmb.cam.ac.uk/scop/
Scopus	www.scopus.com
Seals	www.ncbi.nlm.nih.gov/CBBresearch/SEALS/
Search Engine Guide	www.searchengineguide.com/pages/Science/Biology/
Sequin	www.ncbi.nih.gov/projects/sequin/
SIB	www.isb-sib.ch/
SIMPA96	hnb.mpi-sb.mpg.de/hnb-cgi/HNBMenu?BP/tmp/109398302600896000000000+BP+Menu.rc+SecStruc
SKY/M-FISH & CGH Db	www.ncbi.nlm.nih.gov/sky/
SmarDB	http://smartdb.bioinf.med.uni-goettingen.de/
SNP Consortium Database	http://snp.cshl.org/
SOMFA	bellatrix.pcl.ox.ac.uk/Downloads/
SPARC	http://www.arl.org/sparc/
Spidey	www.ncbi.nlm.nih.gov/IEB/Research/Ostell/Spidey/
Splign	www.ncbi.nlm.nih.gov/sutils/splign
SPORes	cgat.ukm.my/spores/
Stanford microarray database	http://genome-www5.stanford.edu/
STR	http://www.cstl.nist.gov/div831/strbase/
Structure(MMDB)	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Domains
SWISS-2DPAGE	www.expasy.ch/ch2d/
Swissmodel	www.expasy.org/swissmod/SWISS-MODEL.html
SWISSPROT/TrEMBL /UniProt	www.expasy.org/sprot/
TagIdent	www.expasy.org/tools/tagident.html
Taxonomy	ncbi.nlm.nih.gov/entrez/query.fcgi?db=Taxonomy
TECRdb	http://xpdb.nist.gov/enzyme_thermodynamics
Threader	bioinf.cs.ucl.ac.uk/threader/
TIGR	www.tigr.org/db.shtml
Tmpredict	www.isrec.isb-sib.ch/tmbase/TMPRED_doc.html
TOXNET	toxnet.nlm.nih.gov/
TPA	www.ncbi.nih.gov/Genbank/TPA.html
Trace Archive	www.ncbi.nlm.nih.gov/Traces/trace.cgi?
TRED	http://rulai.cshl.edu/cgi-bin/TRED/tred.cgi?process=home
Tree of life	http://www.tolweb.org/tree/
Tree thinking group	http://tree-thinking.org/index.html

Treebase	http://www.treebase.org/treebase/
UniGene	www.ncbi.nlm.nih.gov/UniGene/
Uniprot	http://www.ebi.uniprot.org/index.shtml
UniSTS	www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=unists
VAST	www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html
VBRC	http://www.brc-central.org/cgi-bin/brc-central/brc_central.cgi
VEGA	http://vega.sanger.ac.uk/index.html
VIDA	http://www.biochem.ucl.ac.uk/bsm/virus_database/VIDA.html
Virgen	http://bioinfo.ernet.in/virgen/virgen.html
VSNS	http://www.techfak.uni-bielefeld.de/bcd/welcome.html
Wellcome Trust	www.wellcome.ac.uk/
WIT	wit.mcs.anl.gov/
World lecture Hall	http://web.austin.utexas.edu/wlh/
YPD	www.proteome.com/YPDhome.html
ZFIN	zfin.org/cgi-bin/webdriver?MIval=aa-ZDB_home.apg

Suggested Readings

1. Introduction to DNA Structure. Hallick, R.B. URL: www.blc.arizona.edu/Molecular_Graphics/DNA_Structure/DNA_Tutorial.HTML
2. DNA Structure Tutorial. Martz, E., URL: molvis.sdsc.edu/dna/index.htm
3. Introduction To Protein Structure. URL: webhost.bridgew.edu/fgorga/proteins/default.htm
4. Bioinformatics-A Practical Guide to the Analysis of Genes and Proteins. Baxevanis, A.D. and Ouebette, B.F.F. Pub: John Wiley & Sons Inc.
5. Bioinformatics Methods and Applications - Geneomics, Proteomics and Drug Discovery. Rastogi, S.C., Mendiratta, N. and Rastogi, P. Pub: Prentice Hall of India Private Ltd.
6. NCBI Guide – URL: www.ncbi.nlm.nih.gov/Sitemap/ResourceGuide.html
7. Coffee break – URL: www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowSection&rid=coffeebrk.TOC
8. Genomics : The Science and Technology behind the Human Genome Project. Cantor, C.R. and Smith, C.L. Pub: Wiley Interscience.
9. Sequence-Evolution-Function. Koonin, E.V. and Galpain, M.- Computational Approaches in Comparative Genomics. Pub: Kluver Academic.
10. Bioinformatics: A Practical Approach. Higgins, D. and Taylor, W. Pub: Oxford University Press.
11. Introduction to Computational Biology : Maps Sequences and Genomes. Waterman, M.S. Pub: CRC press.
12. Entrez : Making Use of its Power. Geer, R.C. and Sangers E.W. Briefings in Bioinformatics. 2003 June; 4 (2) 1779-84.
13. Structural Bioinformatics: Bourne, P.E., Weissig, H. Pub: Wiley-Leiss.
14. Bioinformatics: Sequence and Genome Analysis. Mount, D., Pub: Cold Spring Harbor Laboratory Press.
15. NCBI Hand book. URL: www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowTOC&rid=handbook.TOC&depth=2
16. http://www.ornl.gov/sci/techresources/Human_Genome/posters/chromosome/tools.shtml