

**TRAINING COURSE**

ON

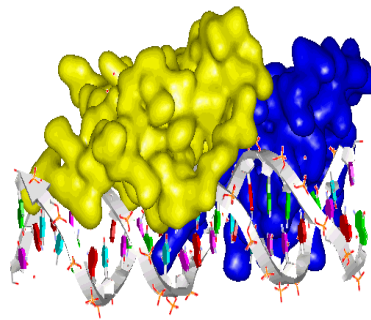
**“Application of Bioinformatics Tools in Biotechnology”**

*25 – 27 March, 2009*



*Organised by*

**BIOINFORMATICS CENTRE  
GAUHATI UNIVERSITY  
GUWAHATI – 781 014, ASSAM, INDIA  
(Member of BTISnet & NEBInet)**



**BIOINFORMATICS INFRASTRUCTURE FACILITY  
(Sponsored by DBT, Govt. of India)**

**VENUE**

**Bioinformatics Centre  
Department of Zoology, Gauhati University, Guwahati-781014**

***Coordinator***

**Dr. D.K. Sharma,  
Prof.& Head, Department of Zoology,  
Gauhati University, Guwahati-781014,  
Assam, India**

**☎ 91-0361-2700470 (O)**

**Fax: 91-0361-2700294**

**E-mail: gauhatiuniv.btisnet@nic.in**

***Co-Coordinator***

**Dr. D.K. Jha, Department of Botany,  
Gauhati University, Guwahati-781014**

**☎ 91-9435047422**

**E-mail: dkjhabot07@gmail.com**

**&**

**Dr. P.J. Handique, Department of Biotechnology,  
Gauhati University, Guwahati-781014**

**☎ 91-9435012920 (M)**

**E-mail: pjhandique@rediffmail.com**

***EDITED BY***

**Chittaranjan Baruah**

**M.Sc. (Zoology), DOEACC Bioinformatics A Level  
Bioinformatics Centre, Zoology Department, G.U.**

***E-mail: chittaranjan\_2004@india.com***

**PARTICIPANTS OF THE TRAINING COURSE ON  
“Application of Bioinformatics Tools in Biotechnology”  
25 – 27 March, 2009  
(Level of Participants: M.Sc. Students)**

1. Ms. Priyanki Das  
Department of Biotechnology  
Gauhati University  
Phone: 9864343702  
E-mail: priyanki85@gmail.com
2. Ms. Pori Deka  
Department of Biotechnology  
Gauhati University  
Phone: 9957864174

- E-mail: *porideka@gmail.com*
3. Ms. Kangkana Katak  
Department of Biotechnology  
Gauhati University  
Phone: 9864802231  
E-mail: *sumki\_1986@gmail.com*
  4. Ms. Karabi Saikia  
Department of Biotechnology  
Gauhati University  
Phone: 9435571399  
E-mail: *karabi.saikia@yahoo.com*
  5. Mr. Ankit Tiwari  
Department of Biotechnology  
Gauhati University  
Phone: 9864448520  
E-mail: *tiwari42@yahoo.com*
  6. Ms. Gayatri Baruah  
Department of Biotechnology  
Gauhati University  
Phone: 9864559125  
E-mail: *gayatrib\_07@rediffmail.com*
  7. Ms. Nitumani Kalita  
Department of Biotechnology  
Gauhati University  
Phone: 0361-2201321  
E-mail: *nitumaina@yahoo.co.in*
  8. Ms. Diptika Tiwari  
Department of Biotechnology  
Gauhati University  
Phone: 9864066508  
E-mail: *goto\_dipti@rediffmail.com*
  9. Ms. Jiumoni Lahkar  
Department of Biotechnology  
Gauhati University  
Phone: 9854329948  
E-mail: *jiumonilahkar@gmail.com*
  10. Ms. Karabi Thakuria  
Department of Biotechnology  
Gauhati University  
Phone: 9859442052  
E-mail: *Karabi.Thakuria09@yahoo.in*
  11. Mr. Jintu Rabha  
Department of Biotechnology  
Gauhati University  
Phone: 99547-07872  
E-mail: n/a
  12. Ms. Moonmoon Das  
Department of Biotechnology  
Gauhati University  
Phone: 9435441389  
E-mail: *suhanee\_87@rediffmail.com*
  13. Ms. Sabira Sultana  
Department of Biotechnology  
Gauhati University  
Phone: 9864200505  
E-mail: *sabbo.87@gmail.com*
  14. Ms. Kumari Dipanjali Saikia  
Department of Biotechnology
  15. Ms. Pallabi Goswami  
Department of Zoology  
Gauhati University  
Phone: 9864912448  
E-mail: *goswami.pallabi@yahoo.com*
  16. Ms. Papari Borah  
Department of Zoology  
Gauhati University  
Phone: 9864952923  
E-mail: n/a
  17. Ms. Sanhita Purkayastha  
Department of Zoology  
Gauhati University  
Phone: 9864089880  
E-mail: *sun\_4u2Day@yahoo.com*
  18. Ms. Jinu Lagachu  
Department of Zoology  
Gauhati University  
Phone: 9957776481  
E-mail: n/a
  19. Ms. Jupitora Deka  
Department of Zoology  
Gauhati University  
Phone: 9854523264  
E-mail: n/a
  20. Ms. Sharmistha Chakravarty  
Department of Zoology  
Gauhati University  
Phone: 9864758010  
E-mail: *senorita1942001@yahoo.com*

**2<sup>nd</sup> Training Course on  
 “Application of Bioinformatics Tools in Biotechnology”  
 25 – 27 March, 2009  
 Bioinformatics Centre, Gauhati University**

**PROGRAMME**

**25-03-09:**

	9:00-9:30 hrs	Registration	
	9:30 – 9:45	Welcome Address	<b>Dr. D.K. Sharma</b> , Coordinator BIF-GU
30	Lecture	Biotechnology & Bioinformatics;	
		<b>Dr. P.J. Handique</b> , Co-coordinator, BIF-GU.	
	10:30-10:45	Tea break	
3:30	Lecture /demo	Emerging concepts of Biotechnology & Bioinformatics; Isolation and cloning of Genes;	<b>Dr. Salvinder Singh</b> , Dept. of Agri Biotech. AAU, Jorhat.
	13:30 – 14:00	Lunch Break	
	14:00-15:00	Lecture /demo	Molecular sampling, Sequencing & Phylogeny; <b>Mr. Udayan Borthakur</b> , Aaranyak
	15:00-16:30	Lecture/demo	<b>Genomics</b> : Genome sequencing techniques, Human Genome Project; <b>Dr. D.K. Sharma</b> , Coordinator BIF-GU

**26-03-09:**

9:45-11:45	Lecture/demo	Basics of Genomics & Proteomics, phylogeny prediction and Cheminformatics;	<b>Mr. S. Mahanta</b> , Assistant Engineer, DOEACC Society Guwahati
11:45-12:00	Tea break		
	12:00-13:30	Practical	<b>Mr. S. Mahanta / Mr. C. Baruah</b> , BIF-GU
	13:30 – 14:00	Lunch Break	
	14:00-16:00	Lecture/demo	Database designing using MS Access/VB/SQL
			<b>Mr. Ripon Biswas</b> , Dy. Manager-IS, <i>IOCL</i>

**27-03-09:**

9:45-13:30	Lecture/demo	Primer designing for PCR	<b>Dr. P Borah</b> , Coordinator, <b>Bioinformatics</b> , College of Veterinary Science, AAU, Khanapara
13:30 – 14:00	Lunch Break		
14:00-16:00	Hands on Training session	<b>Mr. C. Baruah</b> , BIF-GU	
16:00 – 16:30	valedictory Function		

# BIOINFORMATICS AND ITS APPLICATIONS: AN OVERVIEW

*Dr. P. Borah*

Coordinator, Bioinformatics Infrastructure Facility  
and Assoc. Professor, Department of Microbiology  
College of Veterinary Science, Assam Agricultural University  
Khanapara, Guwahati-781022

## Introduction

The term 'Bioinformatics' first introduced in 1987 by Dr. Hwa Lim as: "a new subject of genetic data collection, analysis and dissemination to the research community". Over the years, this emerging field of bioscience has undergone progressive transformation and as its ramifications, new definitions of bioinformatics have emerged. While some people define the term as "An integration of computer, mathematical and statistical methods to manage and analyze biological information", others view it as "The field of science in which Biology, Computer Science, and Information technology merge into a single discipline".

In the present-day context, Bioinformatics involve the use of techniques including applied mathematics, informatics, statistics, computer science, chemistry and biochemistry to solve biological problems usually on the molecular level. Research in computational biology often overlaps with systems biology. Major research efforts in the field include sequence alignment, gene finding, genome assembly, protein structure alignment, protein structure prediction, prediction of gene expression and protein-protein interactions, and the modeling of evolution.

## Bioinformatics and Computational Biology

The terms *bioinformatics* and *computational biology* are often used interchangeably. However *bioinformatics* more properly refers to the creation and advancement of algorithms, computational and statistical techniques, and theory to solve formal and practical problems posed by or inspired from the management and analysis of biological data. *Computational biology*, on the other hand, refers to hypothesis-driven investigation of a specific biological problem using computers, carried out with experimental and simulated data, with the primary goal of discovery and the advancement of biological knowledge.

A common thread in projects in bioinformatics and computational biology is the use of mathematical tools to extract useful information from data produced by high-throughput biological techniques such as genome sequencing. A representative problem in bioinformatics is the assembly of high-quality genome sequences from fragmentary "shotgun" DNA sequencing. Other common problems include the study of gene regulation using data from microarrays or mass spectrometry.

The simplest tasks used in bioinformatics concern the creation and maintenance of databases of biological information. Nucleic acid sequences (and the protein sequences derived from them) comprise the majority of such databases.

The most pressing tasks in bioinformatics involve the analysis of sequence information. **Computational Biology** is the name given to this process, and it involves the following:

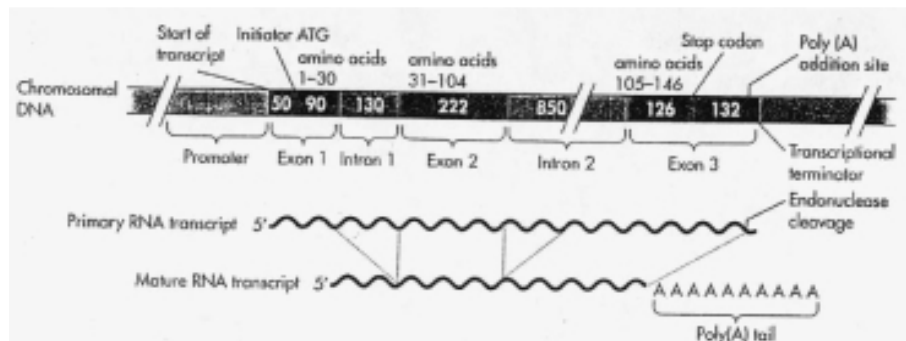
- Finding the genes in the DNA sequences of various organisms
- Developing methods to predict the structure and/or function of newly discovered proteins and structural RNA sequences.

- Clustering protein sequences into families of related sequences and the development of protein models.
- Aligning similar proteins and generating phylogenetic trees to examine evolutionary relationships.

The process of evolution has produced DNA sequences that encode proteins with very specific functions. It is possible to predict the three-dimensional structure of a protein using algorithms that have been derived from our knowledge of physics, chemistry and most importantly, from the analysis of other proteins with similar amino acid sequences.

## Genes and Chromosomes

Each DNA molecule is packaged in a separate *chromosome*, and the total genetic information stored in the chromosomes of an organism is said to constitute its *genome*. With few exceptions, every cell of a Eukaryotic multi-cellular organism contains a complete set of the genome, while the difference in functionality of cells from different tissues is due the variable expression of the corresponding genes. The human genome contains about  $3 \times 10^9$  base pairs (abbreviated *bp*), organized as 46 chromosomes - 22 different autosomal chromosome pairs, and two sex chromosomes: either XX or XY. The 24 different chromosomes range from  $50 \times 10^6$  to  $250 \times 10^6$  bp. The amount of DNA varies between different organisms. The organism *Amoeba dubia* (a single cell organism), for example, has more than 200 times DNA as human.



**Fig.1. The Complexity of the Genome**

Source: J.D. Watson, M. Gilman, J. Witkowski, and M. Zoller. *Recombinant DNA*. W.H. Freeman, New York, 2nd edition, 1992.

The living organisms divide into two major groups: *Prokaryotes*, which are single-celled organisms with no cell nucleus, and *Eukaryotes*, which are higher level organisms, and their cells have nuclei. A gene is a region of DNA that controls a discrete hereditary characteristic, usually corresponding to a single mRNA carrying the information for constructing a protein. In 1977 molecular biologists discovered that most Eukariotic genes have their coding sequences, called *exons*, interrupted by non-coding sequences called *introns*, (See Figure 1). In humans genes constitute approximately 2-3% of the DNA, leaving 97-98% of non-genic *junk DNA*. The role of the latter is as yet unknown, however experiments involving removal of these parts proved to be lethal. Several theories have been suggested, such as physically fixing the DNA in its compressed position, preserving old genetic data, etc.

## The Genetic Code

The rules by which the nucleotide sequence of a gene is translated into the amino acid sequence of the corresponding protein, the so called *genetic code*, were deciphered in the early 1960s. The sequence of nucleotides in the mRNA molecule, that acts as an intermediate was found to be read in serial order in groups of three. Each triplet of nucleotides, called a *codon*, specifies one *amino acid* (the basic unit of a protein, analogous to nucleotides in DNA). Since RNA is a linear polymer of four different nucleotides, there are  $4^3 = 64$  possible codon triplets (See Figure 2). However, only 20 different amino acids are commonly found in proteins, so that most amino acids are specified by several codons. In addition, 3 codons (of the 64) specify the end of translation, and are called *stop codons*. The codon specifying beginning of translation is *AUG*, and is also the codon for the amino acid Methionine. The code has been highly conserved during evolution: with a few minor exceptions, it is the same in organisms as diverse as bacteria, plants, and humans.

		Second base of codon					
		U	C	A	G		
First base of codon	U	UUU } Phe UUC } UUA } Leu UUG }	UCU } UCC } SER UCA } UCG }	UAU } Tyr UAC } UAA } UAG }	UGU } Cys UGC } UGA } UGG } Trp	U C A G	
	C	CUU } CUC } Leu CUA } CUG }	CCU } CCC } Pro CCA } CCG }	CAU } His CAC } CAA } Gln CAG }	CGU } CGC } Arg CGA } CGG }	U C A G	
	A	AUU } AUC } Ile AUA } AUG } Met	ACU } ACC } Thy ACA } ACG }	AAU } Asn AAC } AAA } Lys AAG }	AGU } Ser AGC } AGA } Arg AGG }	U C A G	
	G	GUU } GUC } Val GUA } GUG }	GCU } GCC } Ala GCA } GCG }	GAU } Asp GAC } GAA } Glu GAG }	GGU } GGC } Gly GGA } GGG }	U C A G	
						Third base of codon	

The genetic code, written by convention in the form in which the Codons appear in mRNA. The three terminator codons, UAA, UAG, and UGA, are boxed in red; the AUG initiator codon is shown in green.

Figure 2

## Proteins

A protein is linear polymer of amino acids linked together by peptide bonds. The average protein size is around 200 amino acids long, while large proteins can reach over a thousand amino acids. To a large extent, cells are made of proteins, which constitute more than half of their dry weight. Proteins determine the shape and structure of the cell, and also serve as the main instruments of molecular recognition and catalysis. Proteins have a complex structure, which can be thought of as having four hierarchical structural levels. The amino acid sequence of a protein's chain is called its *primary structure*. Different regions of the sequence form local regular *secondary structures*, such as  $\alpha$ -*helices* which are single stranded helices of amino acids, and  $\beta$ -*sheets* which are planar patches woven from chain segments that are almost linearly arranged. The *tertiary structure* is formed by packing such structures into one or several 3D *domains*. The final, complete, protein may contain several protein domains arranged in a *quaternary structure* (See Figure 3). The whole complex structure (primary to quaternary) is determined by the primary sequence of amino acids and their physico-chemical interaction in the medium. Therefore, its *folding* structure is defined by the genetic material itself, as the three dimensional structure with the minimal free energy. The structure of a

protein determines its functionality. Although the amino acid sequence directly determines the proteins structure, 30% amino acid sequence identity will, in most cases, lead to high similarity in structure.

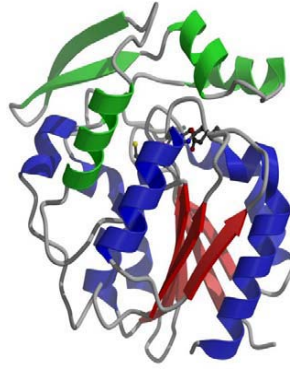


Fig.3. Quaternary structure of protein

### **Major Research Areas in Bioinformatics:**

#### **(a) Sequence analysis:**

Since the bacteriophage  $\Phi$ -X174 was sequenced in 1977, the DNA sequences of hundreds of organisms have been decoded and stored in databases. This data is analyzed to determine genes that code for proteins, as well as regulatory sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematics to construct phylogenetic trees). With the growing amount of data, it became impractical to analyze DNA sequences manually. Today, computer programs are used to search the genome of thousands of organisms, containing billions of nucleotides.

Another aspect of bioinformatics in sequence analysis is the automatic search for genes and regulatory sequences within a genome. Not all of the nucleotides within a genome are genes. Within the genome of higher organisms, large parts of the DNA do not serve any obvious purpose. This so-called junk DNA may, however, contain unrecognized functional elements. Bioinformatics helps to bridge the gap between genome and proteome projects--for example, in the use of DNA sequences for protein identification.

#### **(b) Genome annotation:**

In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence. The first genome annotation software system was designed in 1995 by Owen White, who was part of the team that sequenced and analyzed the first genome of a free-living organism to be decoded, the bacterium *Haemophilus influenzae*. Dr. White built a software system to find the genes (places in the DNA sequence that encode a protein), the transfer RNA, and other features, and to make initial assignments of function to those genes. Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA are constantly changing and improving.

#### **(c) Computational evolutionary biology:**

Evolutionary biology is the study of the origin and descent of species, as well as their change over time. Informatics has assisted evolutionary biologists in several key ways; it has enabled researchers to:

- trace the evolution of a large number of organisms by measuring changes in their DNA, rather than through physical taxonomy or physiological observations alone,
- more recently, compare entire genomes, which permits the study of more complex evolutionary events, such as gene duplication, lateral gene transfer, and the prediction of bacterial speciation factors,
- build complex computational models of populations to predict the outcome of the system over time
- track and share information on an increasingly large number of species and organisms

**(d) Measuring biodiversity:**

Biodiversity of an ecosystem might be defined as the total genomic complement of a particular environment, from all of the species present, whether it is a biofilm in an abandoned mine, a drop of sea water, a scoop of soil, or the entire biosphere of the planet Earth. Databases are used to collect the species names, descriptions, distributions, genetic information, status and size of populations, habitat needs, and how each organism interacts with other species. Specialized software programs are used to find, visualize, and analyze the information, and most importantly, communicate it to other people. Computer simulations model such things as population dynamics, or calculate the cumulative genetic health of a breeding pool (in agriculture) or endangered population (in conservation). One very exciting potential of this field is that entire DNA sequences, or genomes of endangered species can be preserved, allowing the results of Nature's genetic experiment to be remembered in silico, and possibly reused in the future, even if that species is eventually lost.

**(e) Analysis of gene expression:**

The expression of many genes can be determined by measuring mRNA levels with multiple techniques including microarrays, expressed cDNA sequence tag (EST) sequencing, serial analysis of gene expression (SAGE) tag sequencing, massively parallel signature sequencing (MPSS), or various applications of multiplexed *in-situ* hybridization. All of these techniques are extremely noise-prone and/or subject to bias in the biological measurement, and a major research area in computational biology involves developing statistical tools to separate signal from noise in high-throughput gene expression studies. Such studies are often used to determine the genes implicated in a disorder: one might compare microarray data from cancerous epithelial cells to data from non-cancerous cells to determine the transcripts that are up-regulated and down-regulated in a particular population of cancer cells.

**(f) Analysis of regulation:**

Regulation is the complex orchestration of events starting with an extra-cellular signal and ultimately leading to an increase or decrease in the activity of one or more protein molecules. Bioinformatics techniques have been applied to explore various steps in this process.

**(g) Analysis of protein expression:**

Protein microarrays and high throughput (HT) mass spectrometry (MS) can provide a snapshot of the proteins present in a biological sample. Bioinformatics is very much involved in making sense of protein microarray.

**(h) Analysis of mutations in cancer**

Massive sequencing efforts are currently underway to identify point mutations in a variety of genes in cancer. The sheer volume of data produced requires automated systems to read sequence data, and to compare the sequencing results to the known sequence of the human genome, including

known germline polymorphisms. Further, informatics approaches are being developed to understand the implications of lesions found to be recurrent across many tumors.

Some modern tools (e.g. Quantum 3.1 ) provide tool for changing the protein sequence at specific sites through alterations to its amino acids and predict changes in the bioactivity after mutations.

#### **(i) Prediction of protein structure**

Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called *primary structure*, can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment.

#### **(j) Comparative genomics:**

The core of comparative genome analysis is the establishment of the correspondence between genes or other genomic features in different organisms. It is these intergenomic maps that make it possible to trace the evolutionary processes responsible for the divergence of two genomes.

#### **(k) Modeling biological systems:**

Systems biology involves the use of computer simulations of cellular subsystems (such as the networks of metabolites and enzymes which comprise metabolism, signal transduction pathways and gene regulatory networks) to both analyze and visualize the complex connections of these cellular processes. Artificial life or virtual evolution attempts to understand evolutionary processes via the computer simulation of simple (artificial) life forms.

### **Software tools**

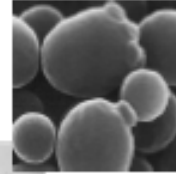
The computational biology tool best-known among biologists is probably BLAST, an algorithm for searching large databases of protein or DNA sequences. National Centre for Biotechnology Information (NCBI) provides a popular implementation that searches their massive sequence databases. Bioinformatic meta search engines (Entrez, Bioinformatics Harvester) help finding relevant information from several databases. There are also free Web-based software designed for structural bioinformatics such as STING.

Computer scripting languages such as Perl and Python are often used to interface with biological databases and parse output from bioinformatics programs. Communities of bioinformatics programmers have set up free/open source projects such as EMBOSS, Bioconductor, BioPerl, BioLinux, BioPython, BioRuby, and BioJava which develop and distribute shared programming tools and objects (as program modules) that make bioinformatics easier.

\*\*\*\*\*

## Eukaryotic model organisms

- *Saccharomyces cerevisiae* (baker's yeast)
- *Caenorhabditis elegans* (C.elegans)
- *Drosophila melanogaster* (fruit fly)
- *Arabidopsis thaliana* (flower)
- *Homo sapiens* (human)



## Basic Concepts and Objectives of Bioinformatics: An Overview

*Dr. Pratap Jyoti Handique*

Dept of Biotechnology,

Gauhati University, Guwahati – 781014.

Email: [pjhandique@rediffmail.com](mailto:pjhandique@rediffmail.com)

Phone: 0361-2700446 (Office –Direct), 9435012920 (mobile).

### What is bioinformatics?

Roughly, bioinformatics describes *any use of computers to handle biological information*. In practice, the definition used by most people is narrower; bioinformatics to them is a synonym for "computational molecular biology"---*the use of computers to characterize the molecular components of living things*.

However *bioinformatics* more properly refers to the creation and advancement of algorithms, computational and statistical techniques, and theory to solve formal and practical problems arising from the management and analysis of biological data.

## Definitions

Bioinformatics is the applications of computational techniques to the management and analysis of biological information. Computers are used to gather, store, analyze and integrate biological and genetic information which can then be applied to gene-based drug discovery and development.

This has implications in diverse areas ranging from artificial intelligence and robotics to genome analysis. In the context of genome initiative, the term was originally applied to the computational manipulation and analysis of biological sequence data of DNA and Protein. In short – bioinformatics is the computational branch of the molecular biology.

## Some related terminologies

- Computational biology: Computational biology is not a "field", but an "approach" involving the use of computers to study biological processes .
- Cheminformatics: The combination of chemical synthesis, biological screening, and data-mining approaches used to guide drug discovery and development"
- Genomics: It is any attempt to analyze or compare the entire genetic complement of a [species](#) or species (plural). It is, of course, possible to compare genomes by comparing more-or-less representative [subsets of genes](#) within genomes.
- Mathematical biology: Mathematical biology also tackles biological problems, but the methods it uses to tackle them need not be numerical and need not be implemented in software or hardware. Indeed, such methods need not "solve" anything; in mathematical biology it would be considered reasonable to publish a result which merely establishes that a biological problem belongs to a particular general class.
- Proteomics: The term proteome was [first coined](#) to describe the set of proteins encoded by the genome<sup>1</sup>. The study of the proteome, called proteomics, now evokes not only all the proteins in any given cell, but also the set of all protein isoforms and modifications, the interactions between them, the structural description of proteins and their higher-order complexes, and for that matter almost everything 'post-genomic'."
- Pharmacogenomics: is the application of genomic approaches and technologies to the identification of drug targets. It is the application of bioinformatics approaches to the cataloguing and processing of information relating to pharmacology and genetics, for example the accumulation of information in databases.
- Pharmacogenetics: It is a subset of pharmacogenomics which uses genomic/bioinformatic methods to identify genomic correlates, for example SNPs (Single Nucleotide Polymorphisms), characteristic of particular patient response profiles and use those markers to inform the administration and development of therapies.

## Importance of Bio-information

The central challenges of bioinformatics is the rationalization of the mass sequence information, to deriving efficient means of data storage and designing more incisive analysis tools. The imperative that drives this analytical process is need to convert sequence information into biochemical and biophysical knowledge, to decipher the structural, functional and evolutionary clues in the language of biological sequences.

To extract biological meaning from sequence information is an exciting science. It is a task of decoding an unknown language. This language may be decomposed into sentences (Protein), words (motifs) and letters (Amino acids) and the code may be tackled at a variety of these levels. The letters have no higher meaning, but their combination into words is important. Sometimes the most subtle of changes a single letter within word can change its meaning and hence the meaning of the entire sentence, so it is vital to decipher the code correctly

**An example:** The single base change in human haemoglobin is an important example. A chain codon for Glutamic acid (GAA) to valine (GUA) is found in homozygous individual and this minute difference results in a change from a normal healthy state to fatal **sickle cell anaemia**.

### Uses

Bioinformatics has a wide range of uses in the management and analysis of biological data. However, it is mainly used for the following:

- ✎ Map genomes and identifies genes of organisms,
- ✎ Determine protein structure & simulate protein interactions,
- ✎ Discover new therapeutic targets,
- ✎ Assess the effects of virtual mutations on gene function,
- ✎ Study of phylogeny of plants and animals including human being.

### More Uses

- Sequence analysis
- Genome annotation
- Computational evolutionary biology
- Measuring biodiversity
- Analysis of gene expression and regulation
- Analysis of protein expression
- Analysis of mutations in cancer
- Prediction of protein structure
- Comparative genomics
- Modeling biological systems
- High-throughput image analysis

### Examples of Uses

Bioinformatics is used to organize and analyze information about cells and biological molecules. Extensive data and information generated through cellular, genetic and molecular experimentation on various cell types found in microbes, plants and animals are now available in various databases. Similarly enormous data on biomolecules of life i.e. nucleic acids (both DNA and RNA) and proteins are managed and analyzed using bioinformatics tools and techniques.

### Use in Nucleic Acid Sequence Databases

Nucleic acids and protein sequence databases have proved to be a valuable resource for the planning and evaluation of results of sequencing experiments. They are the basis for statistical analysis and comparison of large number of sequences. The availability of sequence databases has helped further biological knowledge in a number of areas.

- Example-1: v-sis oncogene found in a simian sarcoma virus is very similar to human gene which give rise to Platelet Derived Growth factor (PGDR). PGDR stimulates epithelial cells to grow which resulted in wound healing. PGRD also plays a role in the proliferation of cells that clog blood vessels creating condition that may lead to heart attack. Thus the computerized comparison established a surprising link between heart disease and cancer. This demonstrates the tremendous importance of information in a database.
- Example-2: Availability of sequence data has also furthered out understanding of evolutionary relationships among many forms of life. Using the aligned RNA or certain enzymes, the phylogenetic relatedness of different organisms can be inferred. These help to develop more stable system of classification, particularly for microbes.

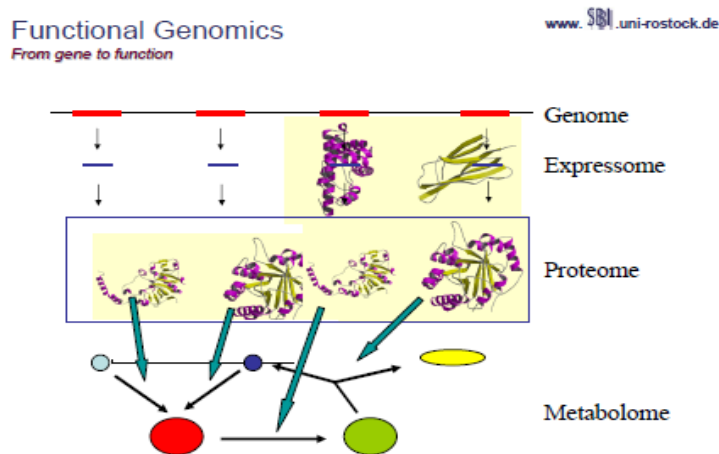
### Bioinformatics in study of phylogenetics

- Studies and comparisons of Nucleic acid and Protein sequences are efficient tool for phylogenetic analysis. Evolutionary relationship is generally represented by a special type of graph called a TREE.

Bioinformatics help in building Phylogenetic trees

- Percentage of matches (similarity table) or differences (distance table) are being prepared which are used for construction of a phylogenetic tree. Distance tables are used for the analysis of macromolecular sequence data.

\*\*\*\*\*



### Historical events of Bioinformatics

*Chittaranjan Baruah*

*Bioinformatics Centre (DBT-BIF)*

*Department of Zoology (UGC-SAP & DST-FIST sponsored Department),*

*Gauhati University, Guwahati – 781 014, Assam, India*

*E-mail: chittaranjan\_2004@india.com*

Historical events	Year
Gregor Mendel ("Father of Genetics") cross-fertilized different colors of the same species of flowers. In a journal, he kept careful records of the colors of flowers that he cross-fertilized and the colors of flowers they produced.	1865
Pauling and Corey propose the structure for the alpha-helix	1951
Watson and Crick propose the double helix model for DNA based on x-ray data obtained by Franklin and Wilkins	1953
The sequence of the first protein to be analyzed, bovine insulin, is announced by F. Sanger	1955
The first integrated circuit is constructed by Jack Kilby at Texas Instruments	1958
Margaret Dayhoff starts the Atlas of Protein Sequence and Structure	1965
The details of the Needleman-Wunsch algorithm for sequence comparison are published	1970
<b>Protein Sequence Database (PSD) by Margaret Dayhoff*</b> *Father of Bioinformatics	<b>1972</b>
The first recombinant DNA molecule is created by Paul Berg and his group	1972

Stanley Cohen invented DNA cloning	1973
Sanger et al. invent cycle sequencing	1977
The first complete gene sequence for an organism (Bacteriophage FX174) is published. The gene consists of 5,386 base pairs which code nine proteins	1980
The Smith-Waterman algorithm for sequence alignment is published	1981
IBM introduces its Personal Computer to the market	1981
The PCR reaction is described by Kary Mullis and co-workers	1983
The FASTP algorithm is published by Lipman & Pearson	1985
The term "Genomics" appeared for the first time. It was coined by Thomas Roderick as a name for the new journal	1986
The SWISS-PROT database is created (University of Geneva and the European Molecular Biology Laboratory)	1986
The Human Genome Initiative is announced by DOE	1986
Perl (Practical Extraction Report Language) is released by Larry Wall.	1987
The National Center for Biotechnology Information (NCBI) is established at the National Cancer Institute in Bethesda	1987
The physical map of <i>E. coli</i> is published	1988
The FASTA algorithm for sequence comparison is published by Pearson & Lupman	1988
The BLAST program (Altschul, et. al.) is implemented	1990
The first reference to the word "bioinformatics" in the scientific literature (source: Bioinformatics.org)	1991 ?
The research institute in Geneva (CERN) announces the creation of the protocols which make-up the World Wide Web	1991
Human Genome Systems, Gaithersburg Maryland, is formed by William Haseltine	1992
The Institute for Genomic Research (TIGR) is established by Craig Venter in Rockville	1992
The PRINTS database of protein motifs is published by Attwood and Beck	1994
Sun releases version 1.0 of Java. Sun and Netscape release version 1.0 of JavaScript	1995
The <i>Haemophilus influenzae</i> genome (1.8 Mb) is sequenced	1995
Affymetrix produces the first commercial DNA chips	1996
Craig Venter forms Celera in Rockville, Maryland	1998
The Swiss Institute of Bioinformatics is established in Geneva	1998
A draft of the human genome (3,000 Mbp) is published	2001

## Genome size comparison

	Species	Chrom.	Genes	Base pairs
	<b>Human</b> (Homo sapiens)	46 (23 pairs)	28-35,000	3.1 billion
	<b>Mouse</b> (Mus musculus)	40	22.5-30,000	2.7 billion
	<b>Puffer fish</b> (Fugu rubripes)	44	31,000	365 million
	<b>Malaria mosquito</b> (Anopheles gambiae)	6	14,000	289 million
	<b>Fruit Fly</b> (Drosophila melanogaster)	8	14,000	137 million
	<b>Roundworm</b> (C. elegans)	12	19,000	97 million
	<b>Bacterium</b> (E. coli)	1	5,000	4.1 million

## PRIMER DESIGNING FOR PCR

*Dr. P. Borah*

Coordinator, BIF and Associate Professor,  
Department of Microbiology, College of Veterinary Science  
AAU, Khanapara, Guwahati-781022

Polymerase chain reaction (PCR) is widely accepted as one of the most important inventions of the 20<sup>th</sup> century in molecular biology. Small amounts of the genetic material can now be amplified to be able to identify and manipulate DNA, detect infectious organisms, detect genetic variations including mutations and numerous other tasks.

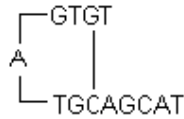
PCR involves three steps: denaturation, annealing and extension. First, the genetic material is denatured, converting double-stranded DNA molecules to single strands. The primers are then annealed to the complementary regions of the single-stranded molecules. In the third step, they are extended by the action of the DNA polymerase. All these steps are temperature-dependent and the common choice of temperatures is 94°C, 60°C and 70°C respectively.

Good primer design is essential for successful PCR. The following is a brief description of the important considerations for designing a primer:

1. **Primer length:** It is generally accepted that the optimal length of PCR primers is 18-22 bp. This length is long enough for adequate specificity, and short enough for primers to bind easily to the template at the annealing temperature.
2. **Melting temperature:** Melting temperature ( $T_m$ ) is defined as the temperature at which one half of the DNA duplex will dissociate to become single-stranded and indicates the duplex stability. Primers with melting temperatures in the range of 65-70°C generally produce the best results. Primers with higher melting temperatures have a tendency for secondary annealing. The GC content of the sequence gives a fair indication of the  $T_m$ .
3. **Primer annealing temperature:** The primer melting temperature is the estimate of the DNA-DNA hybrid stability and critical in determining the annealing temperature. Too high  $T_m$  produces insufficient primer-template hybridization resulting in low PCR product yield. Too low  $T_m$  may possibly lead to non-specific products caused by a high number of base pair mismatches. Mismatch tolerance is found to have the strongest influence on PCR specificity.
4. **GC content:** The GC content (the number of guanine and cytosine bases in the primer as a percentage of the total bases) of primer should be 40-60%.

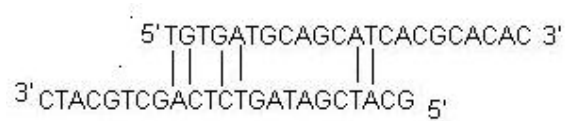
5. **GC Clamp:** The presence of G or C bases within the last five bases from the 3' end of primers (GC clamp) helps to promote specific binding at the 3' end due to the stronger bonding of G and C bases. More than 3 G's or C's should be avoided in the last 5 bases at the 3' end of the primer.
6. **Secondary structures:** Presence of the secondary structures produced by intermolecular or intramolecular interactions can lead to poor or no yield of the product. They adversely affect primer template annealing and thus the amplification. They greatly reduce the availability of primers to the reaction.

i) **Hairpins:** It is formed by intra-molecular interaction within the primer and should be avoided. Presence of hairpins at the 3' end most adversely affects the reaction.



ii) **Self Dimer:** is formed by intermolecular interactions between the two (same sense) primers, where the primer is homologous to itself. Generally a large amount of primers are used in PCR compared to the amount of target gene. When primers form intermolecular dimers much more readily than hybridizing to target DNA, they reduce the product yield.

iii) **Cross Dimer:** Cross dimers are formed by intermolecular interaction between sense and anti-sense primers, where they are homologous.



7. **Repeats:** A repeat is a di-nucleotide occurring many times consecutively and should be avoided because they can misprime. For example: ATATATAT. A maximum number of di-nucleotide repeats acceptable is 4 di-nucleotides.
8. **Runs:** Primers with long runs of a single base should generally be avoided as they can misprime. For example, AGCGGGGATGGGG has runs of base 'G' of value 5 and 4. A maximum number of runs accepted is 4 bp.
9. **Avoid Template secondary structure:** A single stranded Nucleic acid sequences is highly unstable and fold into conformations (secondary structures). The stability of these template secondary structures depends largely on their free energy and melting temperatures( $T_m$ ).
10. **Avoid Cross homology:** To improve specificity of the primers it is necessary to avoid regions of homology. Primers designed for a sequence must not amplify other genes in the mixture. Commonly, primers are designed and then BLASTed to test the specificity.

#### Parameters for Primer Pair Design:

1. **Amplicon Length:** The amplicon length is dictated by the experimental goals. For qPCR, the target length is closer to 100 bp and for standard PCR, it is near 500 bp. If you know the positions of each primer with respect to the template, the product is calculated as: Product length = (Position of antisense primer-Position of sense primer) + 1.
2. **Product position:** Primer can be located near the 5' end, the 3' end or any where within specified length. Generally, the sequence close to the 3' end is known with greater confidence and hence preferred most frequently.
3. **T<sub>m</sub> of Product:** Melting Temperature ( $T_m$ ) is the temperature at which one half of the DNA duplex will dissociate and become single stranded. The stability of the primer-template DNA duplex can be measured by the melting temperature ( $T_m$ ).
4. **Optimum Annealing temperature ( $T_a$  Opt):** The formula of Rychlik is most respected. It usually results in good PCR product yield with minimum false product production.

$$T_a \text{ Opt} = 0.3 \times (T_m \text{ of primer}) + 0.7 \times (T_m \text{ of product}) - 25$$

Where,

$T_m$  of primer is the melting temperature of the less stable primer-template pair  
 $T_m$  of product is the melting temperature of the PCR product.

5. **Primer Pair T<sub>m</sub> Mismatch Calculation:** The two primers of a primer pair should have closely matched melting temperatures for maximizing PCR product yield. The difference of 5°C or more can lead no amplification.

#### Summary:

1. Design your PCR primers to be 18-30 oligo nucleotides in length. The longer end of this range allows higher specificity and gives you space to add restriction enzyme sites to the primer end for cloning.
2. Make sure the melting temperatures ( $T_m$ ) of the primers used are not more than 5°C different from each other. You can calculate  $T_m$  with this formula:

$$T_m = 4(G + C) + 2(A + T) \text{ } ^\circ\text{C}$$

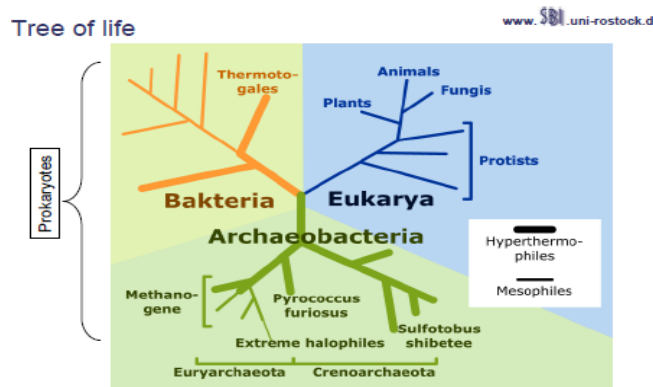
3. Aim for a  $T_m$  between 65 and 70°C for each primer over the region of hybridization
4. Use an annealing temperature ( $T_a$ ) of 10 to 15°C lower than the  $T_m$ .
5. The GC content of each primer should be in the range of 40-60% for optimum PCR efficiency.
6. Try to have uniform distribution of G and C nucleotides, as clusters of G's or C's can cause non-specific priming.

7. Avoid long runs of the same nucleotide.
8. Check that primers are not self-complementary or complementary to the other primer in the reaction mixture, as this will encourage formation of hairpins and primer dimers and will compete with the template for the use of primer and reagent.
9. If you can, make the 3' end terminate in C or A, as the 3' is the end which extends and neither the C nor A nucleotide wobbles. This will increase the specificity.
10. You can avoid mispriming by making the 3' end slightly AT rich.
11. Use the right software. Using the right software is a great way to automate these steps and minimize errors, especially when you have to design primers for many sequences.

References:

- <http://rothlab.ucdavis.edu/protocols/PrimerDesign.html>  
<http://www.biochem.ucl.ac.uk/bsm/nmr/protocols/protocols/oligo.html>  
[http://www.protocol-online.org/prot/Molecular\\_Biology/PCR/PCR\\_Primer/](http://www.protocol-online.org/prot/Molecular_Biology/PCR/PCR_Primer/)  
<http://www.mcb.uct.ac.za/pcroptim.htm>

\*\*\*\*\*



## Biological Databases

*Saurov Mahanta*

Assistant Engineer, Bioinformatics Wing, DOEACC Society, Guwahati Centre, STPI Complex, Near LGBI Airport, Borjhar, Guwahti-781015. Email: saurov@doeaccassam.ac.in.

### *Introduction to Database & Database Management System*

Simply a database is a collection of related data. The meaning of data is known facts that can be recorded and that have implicit meaning. As for example we can have the contact information of people related to us either recorded in a address book by entering their names, telephone numbers, e-mail address, home address etc or can be stored in a computer hard drive using software like Microsoft access, Open Office, Spreadsheets like MS-Excel etc. This is a collection of related data with a implicit meaning and hence a database. Therefore a database is a logically coherent collection of data with some inherent meaning. A random assortment of data cannot correctly be referred to as a database.

Database Management System (DBMS) is a collection of programs that enables users to create and maintain a database. It is therefore a software system that facilitates the process of defining, constructing, manipulating and sharing databases among various users and applications. Defining include specifying data types, structures etc. for the data to be stored in the database. Constructing the

database is the process of storing data on some storage medium that is controlled by DBMS. Manipulation of a database includes such functions as querying the database to retrieve specific data. Sharing a database allows multiple users and programs to access the database concurrently.

There are mainly two types of Database Management Systems: flat file indexing systems and relational Database Management Systems. Now a days a third type that is the object-oriented DBMS is getting popularity. Choosing of the proper database management system is an important decision which have long range implications for capacity and usefulness of the database.

### Biological Databases

Currently, a lot of bioinformatics work is concerned with the technology of databases. These databases include both "public" repositories of gene data like GenBank or the Protein DataBank (the PDB), and private databases like those used by research groups involved in gene mapping projects or those held by biotech companies. Making such databases accessible via open standards like the Web is very important since consumers of bioinformatics data use a range of computer platforms: from the more powerful and forbidding UNIX boxes favoured by the developers and curators to the far friendlier Macs often found populating the labs of computer-wary biologists. RNA and DNA are the proteins that store the hereditary information about an organism. These macromolecules have a fixed structure, which can be analyzed by biologists with the help of bioinformatic tools and databases.

#### Biological databases can be categorized into different categories as follows:

- Nucleotide Sequence Databases
- RNA sequence databases
- Protein sequence databases
- Structure Databases
- Genomics Databases (non-vertebrate)
- Metabolic and Signaling Pathways
- Human and other Vertebrate Genomes
- Human Genes and Diseases
- Microarray Data and other Gene Expression Databases
- Proteomics Resources
- Other Molecular Biology Databases
- Organelle databases
- Plant databases
- Immunological databases

#### Some Molecular Biology Databases and their location in the World wide web.

( Ref. *Nucleic Acid Research, Database issue, Jan, 2009.* )

Sl. no	Database name	Full name and/or description	URL
<b>Nucleotide Sequence Databases 1.1. International Nucleotide Sequence Database Collaboration</b>			
1	DDBJ - DNA Data Bank of Japan	All known nucleotide and protein sequences	<a href="http://www.ddbj.nig.ac.jp">http://www.ddbj.nig.ac.jp</a>
2	EMBL Nucleotide Sequence Database	All known nucleotide and protein sequences	<a href="http://www.ebi.ac.uk/embl.html">http://www.ebi.ac.uk/embl.html</a>
3	GenBank®	All known nucleotide and protein	<a href="http://www.ncbi.nlm.nih.gov/Entre">http://www.ncbi.nlm.nih.gov/Entre</a>

		sequences	<a href="#">z</a>
	<b>DNA sequences: genes, motifs and regulatory sites</b>		
	<b>1.2.1. Coding and coding DNA</b>		
4	Genetic Codes	Genetic codes in various organisms and organelles	<a href="http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi">http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi</a>
5	Entrez Gene	Gene-centered information at NCBI	<a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene</a>
6	TIGR Gene Indices	Organism-specific databases of EST and gene sequences	<a href="http://www.tigr.org/tdb/tgi.shtml">http://www.tigr.org/tdb/tgi.shtml</a>
7	Transterm	Codon usage, start and stop signals	<a href="http://guinevere.otago.ac.nz/transterm.html">http://guinevere.otago.ac.nz/transterm.html</a>
	UniGene	Non-redundant set of eukaryotic gene-oriented clusters	<a href="http://www.ncbi.nlm.nih.gov/UniGene/">http://www.ncbi.nlm.nih.gov/UniGene/</a>
8	UniVec	Vector sequences, adapters, linkers and primers used in DNA cloning, used to check for vector contamination	<a href="http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html">http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html</a>
9	VectorDB	Characterization and classification of nucleic acid vectors	<a href="http://genome-www2.stanford.edu/vectordb/">http://genome-www2.stanford.edu/vectordb/</a>
10	Xpro	Eukaryotic protein-encoding DNA sequences, both intron-containing and intron-less genes	<a href="http://origin.bic.nus.edu.sg/xpro/">http://origin.bic.nus.edu.sg/xpro/</a>
	<b>1.2.2. Gene structure, introns and exons, splice sites</b>		
11	ASAP	<u>A</u> lternative <u>s</u> pliced isoforms	<a href="http://bioinfo.mbi.ucla.edu/ASAP/">http://bioinfo.mbi.ucla.edu/ASAP/</a>
12	ExInt	<u>E</u> xon- <u>i</u> ntron structure of eukaryotic genes	<a href="http://sege.ntu.edu.sg/wester/exint/">http://sege.ntu.edu.sg/wester/exint/</a>
13	Hollywood	Exon annotation database	<a href="http://hollywood.mit.edu">http://hollywood.mit.edu</a>
14	HS3D	<i>Homo sapiens</i> splice sites dataset	<a href="http://www.sci.unisannio.it/docenti/rampone/">http://www.sci.unisannio.it/docenti/rampone/</a>
15	Intronerator	Alternative splicing in <i>C. elegans</i> and <i>C. briggsae</i>	<a href="http://www.cse.ucsc.edu/~kent/intronerator/">http://www.cse.ucsc.edu/~kent/intronerator/</a>
16	SpliceDB	Canonical and non-canonical mammalian splice sites	<a href="http://www.softberry.com/berry.phtml?to pic=splicedb&amp;group=data&amp;subgroup=sp ldb">http://www.softberry.com/berry.phtml?to pic=splicedb&amp;group=data&amp;subgroup=sp ldb</a>
17	SpliceInfo	Modes of alternative splicing in human genome	<a href="http://spliceinfo.mbc.nctu.edu.tw/">http://spliceinfo.mbc.nctu.edu.tw/</a>
18	SpliceNest	A tool for visualizing splicing of genes from EST data	<a href="http://splicenest.molgen.mpg.de/">http://splicenest.molgen.mpg.de/</a>
	<b>1.2.3. Transcriptional regulator sites and transcription factors</b>		
19	DBD	Transcription factor prediction database	<a href="http://stash.mrc-lmb.cam.ac.uk/skk/Cell2/index.cgi?Home">http://stash.mrc-lmb.cam.ac.uk/skk/Cell2/index.cgi?Home</a>
20	DBTBS	<i>Bacillus subtilis</i> promoters and transcription factors	<a href="http://dbtbs.hgc.jp/">http://dbtbs.hgc.jp/</a>
21	DoOP	Database of orthologous promoters: chordates and plants	<a href="http://doop.abc.hu/">http://doop.abc.hu/</a>
22	DPInteract	Binding sites for <i>E. coli</i> DNA-binding proteins	<a href="http://arep.med.harvard.edu/dpinteract">http://arep.med.harvard.edu/dpinteract</a>
23	EPD	<u>E</u> karyotic <u>p</u> romoter <u>d</u> atabase	<a href="http://www.epd.isb-sib.ch">http://www.epd.isb-sib.ch</a>
24	Extra-TRAIN	Extragenic regions and transcriptional regulators in bacteria and archaea	<a href="http://www.era7.com/ExtraTrain/">http://www.era7.com/ExtraTrain/</a>
25	HemoPDB	<u>H</u> ematopoietic <u>p</u> romoter <u>d</u> atabase	<a href="http://bioinformatics.med.ohio-state.edu/HemoPDB">http://bioinformatics.med.ohio-state.edu/HemoPDB</a>
26	HTPSELEX	Transcription factor binding site sequences obtained using high-throughput SELEX method	<a href="http://www.isrec.isb-sib.ch/httpselex/">http://www.isrec.isb-sib.ch/httpselex/</a>
	InsulatorDB	Insulator regulatory elements in vertebrate genomes	<a href="http://insulatordb.utmem.edu">http://insulatordb.utmem.edu</a>

	JASPAR	PSSMs for transcription factor DNA-binding sites	<a href="http://jaspar.cgb.ki.se">http://jaspar.cgb.ki.se</a>
	MAPPER	Putative transcription factor binding sites in various genomes	<a href="http://bio.chip.org/mapper">http://bio.chip.org/mapper</a>
27	MPromDB	Mammalian promoter database	<a href="http://bioinformatics.med.ohio-state.edu/MPromDb">http://bioinformatics.med.ohio-state.edu/MPromDb</a>
28	PLACE	Plant cis-acting regulatory DNA elements	<a href="http://www.dna.affrc.go.jp/htdocs/PLACE">http://www.dna.affrc.go.jp/htdocs/PLACE</a>
29	PlantCARE	Plant promoters and cis-acting regulatory elements	<a href="http://intra.psb.ugent.be:8080/PlantCARE/">http://intra.psb.ugent.be:8080/PlantCARE/</a>
30	PlantProm	Plant promoter sequences for RNA polymerase II	<a href="http://mendel.cs.rhul.ac.uk/">http://mendel.cs.rhul.ac.uk/</a>
31	PRODORIC	Prokaryotic database of gene regulation networks	<a href="http://prodoric.tu-bs.de/">http://prodoric.tu-bs.de/</a>
<b>2. RNA sequence and structure</b>			
32	16S and 23S rRNA Mutation Database	16S and 23S ribosomal RNA mutations	<a href="http://www.fandm.edu/Departments/Biology/Databases/RNA.html">http://www.fandm.edu/Departments/Biology/Databases/RNA.html</a>
33	16S rRNA database	Multiple sequence alignment of prokaryotic 16S rDNA	<a href="http://greengenes.llnl.gov/16S/">http://greengenes.llnl.gov/16S/</a>
34	5S rRNA Database	5S rRNA sequences	<a href="http://biobases.ibch.poznan.pl/5SSData/">http://biobases.ibch.poznan.pl/5SSData/</a>
35	European rRNA database	All complete or nearly complete rRNA sequences	<a href="http://www.psb.ugent.be/rRNA/">http://www.psb.ugent.be/rRNA/</a>
36	miRNAMap	microRNA precursors and their mapping to targets in vertebrate genomes	<a href="http://mirnamap.mbc.nctu.edu.tw">http://mirnamap.mbc.nctu.edu.tw</a>
37	microRNA Registry	Database of microRNAs (small noncoding RNAs)	<a href="http://www.sanger.ac.uk/Software/Rfam/mirna/">http://www.sanger.ac.uk/Software/Rfam/mirna/</a>
38	SARS-CoV RNA SSS	Predicted secondary structures of SARS coronavirus RNA	<a href="http://www.liuweibo.com/sarsdb/">http://www.liuweibo.com/sarsdb/</a>
39	siRNAdb	Functional human siRNA sequences	<a href="http://sirna.cgb.ki.se/">http://sirna.cgb.ki.se/</a>
40	Small RNA Database	Small RNAs from prokaryotes and eukaryotes	<a href="http://condor.bcm.tmc.edu/smallRNA/smallrna.html">http://condor.bcm.tmc.edu/smallRNA/smallrna.html</a>
41	snoRNA-LBME-db	Human C/D box and H/ACA modification guide RNAs	<a href="http://www-snorna.biotoul.fr/">http://www-snorna.biotoul.fr/</a>
42	SRPDB	Signal recognition particle database	<a href="http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html">http://psyche.uthct.edu/dbs/SRPDB/SRPDB.html</a>
43	SSU rRNA Modification Database	Modified nucleosides in small subunit rRNA	<a href="http://medstat.med.utah.edu/SSUmods/">http://medstat.med.utah.edu/SSUmods/</a>
<b>3. Protein sequence databases</b>			
<b>3.1. General sequence databases</b>			
44	NCBI Protein database	All protein sequences: translated from GenBank and imported from other protein databases	<a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein</a>
45	PIR-NREF	PIR's non-redundant reference protein database	<a href="http://pir.georgetown.edu/pirwww/pirnref.shtml">http://pir.georgetown.edu/pirwww/pirnref.shtml</a>
46	PRF	Protein research foundation database of peptides: sequences, literature and unnatural amino acids	<a href="http://www.prf.or.jp/en">http://www.prf.or.jp/en</a>
47	Swiss-Prot	Now UniProt/SwissProt, part of the UniProt knowledgebase	<a href="http://www.expasy.org/sprot">http://www.expasy.org/sprot</a>
48	TCDB	Transporter protein classification database	<a href="http://www.tcdb.org/">http://www.tcdb.org/</a>
49	TrEMBL	Now UniProt/TrEMBL, part of the	<a href="http://www.expasy.org/sprot">http://www.expasy.org/sprot</a>

		UniProt knowledgebase	
50	UniParc	UniProt archive, a repository of all protein sequences	<a href="http://www.uniprot.org/database/archive.shtml">http://www.uniprot.org/database/archive.shtml</a>
51	UniProt	Universal protein knowledgebase	<a href="http://www.uniprot.org/">http://www.uniprot.org/</a>
52	UniRef	Clustered sets of related sequences from UniProt	<a href="http://www.uniprot.org/database/nref.shtml">http://www.uniprot.org/database/nref.shtml</a>
<b>3.2. Protein properties</b>			
53	PINT	Protein-protein interactions thermodynamic database	<a href="http://pintdb.dyndns.org/index.html">http://pintdb.dyndns.org/index.html</a>
54	PPD	Protein pKa database	<a href="http://www.jenner.ac.uk/ppd/">http://www.jenner.ac.uk/ppd/</a>
55	ProNIT	Thermodynamic data on protein-nucleic acid interactions	<a href="http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html">http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html</a>
56	ProTherm	Thermodynamic data for wild-type and mutant proteins	<a href="http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html">http://gibk26.bse.kyutech.ac.jp/jouhou/protherm/protherm.html</a>
57	REFOLD	Experimental data on protein refolding and purification	<a href="http://refold.med.monash.edu.au">http://refold.med.monash.edu.au</a>
<b>3.3. Protein localization and targeting</b>			
58	DBSubLoc	Database of protein subcellular localization	<a href="http://www.bioinfo.tsinghua.edu.cn/dbsubloc.html">http://www.bioinfo.tsinghua.edu.cn/dbsubloc.html</a>
59	LOCATE	Membrane organization and subcellular localization of mouse proteins	<a href="http://mpdb.imb.uq.edu.au">http://mpdb.imb.uq.edu.au</a>
60	NOPdb	Nucleolar proteome database	<a href="http://www.lamondlab.com/NOPdb/">http://www.lamondlab.com/NOPdb/</a>
61	NURSA	Nuclear receptor signaling atlas	<a href="http://www.nursa.org">http://www.nursa.org</a>
62	PSORTdb	Protein subcellular localization in bacteria	<a href="http://db.psорт.org/">http://db.psорт.org/</a>
<b>3.4. Protein sequence motifs and active sites</b>			
	eMOTIF	Protein sequence motif determination and searches	<a href="http://motif.stanford.edu/emotif">http://motif.stanford.edu/emotif</a>
	Metalloprotein Site	Metal-binding sites in metalloproteins	<a href="http://metallo.scripps.edu/">http://metallo.scripps.edu/</a>
	O-GlycBase	O- and C-linked glycosylation sites in proteins	<a href="http://www.cbs.dtu.dk/databases/OGLYCBASE/">http://www.cbs.dtu.dk/databases/OGLYCBASE/</a>
	PDBSite	3D structure of protein functional sites	<a href="http://srs6.bionet.nsc.ru/srs6/">http://srs6.bionet.nsc.ru/srs6/</a>
	Phospho.ELM	S/T/Y protein phosphorylation sites (former PhosphoBase)	<a href="http://phospho.elm.eu.org/">http://phospho.elm.eu.org/</a>
	PROMISE	Prosthetic centers and metal ions in protein active sites	<a href="http://metallo.scripps.edu/PROMISE">http://metallo.scripps.edu/PROMISE</a>
	ProRule	Functional and structural information on PROSITE profiles	<a href="http://www.expasy.org/tools/scanprosite">http://www.expasy.org/tools/scanprosite</a>
	PROSITE	Biologically-significant protein patterns and profiles	<a href="http://www.expasy.org/prosite">http://www.expasy.org/prosite</a>
<b>3.5. Protein domain databases; protein classification</b>			
	InterPro	Integrated resource of protein families, domains and functional sites	<a href="http://www.ebi.ac.uk/interpro">http://www.ebi.ac.uk/interpro</a>
	Pfam	Protein families: Multiple sequence alignments and profile hidden Markov models of protein domains	<a href="http://www.sanger.ac.uk/Software/Pfam/">http://www.sanger.ac.uk/Software/Pfam/</a>
	PIR-ALN	Curated database of protein sequence alignments	<a href="http://pir.georgetown.edu/pirwww/dbinfo/piraln.html">http://pir.georgetown.edu/pirwww/dbinfo/piraln.html</a>
	SIMAP	Similarity matrix of proteins: precomputed similarity data	<a href="http://mips.gsf.de/services/analysis/simap/">http://mips.gsf.de/services/analysis/simap/</a>
	SMART	Simple modular architecture research tool: signalling, extracellular and chromatin-associated protein domains	<a href="http://smart.embl-heidelberg.de/">http://smart.embl-heidelberg.de/</a>

	SUPFAM	Grouping of sequence families into superfamilies	<a href="http://pauling.mbu.iisc.ernet.in/~supfam">http://pauling.mbu.iisc.ernet.in/~supfam</a>
<b>3.6. Databases of individual protein families</b>			
	Histone Database	Histone fold sequences and structures	<a href="http://research.nhgri.nih.gov/histones/">http://research.nhgri.nih.gov/histones/</a>
	Homeobox Page	Homeobox proteins, classification and evolution	<a href="http://www.biosci.ki.se/groups/tbu/homeo.html">http://www.biosci.ki.se/groups/tbu/homeo.html</a>
	Hox-Pro	Homeobox genes database	<a href="http://www.iephb.nw.ru/labs/lab38/spirov/hox_pro/hox-pro00.html">http://www.iephb.nw.ru/labs/lab38/spirov/hox_pro/hox-pro00.html</a>
	Homeodomain Resource	Homeodomain sequences, structures and related genetic and genomic information	<a href="http://research.nhgri.nih.gov/homeodomain/">http://research.nhgri.nih.gov/homeodomain/</a>
	NPD	Nuclear protein database	<a href="http://npd.hgu.mrc.ac.uk/">http://npd.hgu.mrc.ac.uk/</a>
	NucleaRDB	Nuclear receptor superfamily	<a href="http://www.receptors.org/NR/">http://www.receptors.org/NR/</a>
	Nuclear Receptor Resource	Nuclear receptor superfamily	<a href="http://nrr.georgetown.edu/NRR/nrrhome.htm">http://nrr.georgetown.edu/NRR/nrrhome.htm</a>
	Ribonuclease P Database	RNase P sequences, alignments and structures	<a href="http://www.mbio.ncsu.edu/RNaseP/home.html">http://www.mbio.ncsu.edu/RNaseP/home.html</a>
	RPG	Ribosomal protein gene database	<a href="http://ribosome.miyazaki-med.ac.jp/">http://ribosome.miyazaki-med.ac.jp/</a>
	TrSDB	Transcription factor database	<a href="http://bioinf.uab.es/cgi-bin/trsdb/trsdb.pl">http://bioinf.uab.es/cgi-bin/trsdb/trsdb.pl</a>
	V K C D B	Voltage-gated potassium channel database	<a href="http://vkcdb.biology.ualberta.ca/">http://vkcdb.biology.ualberta.ca/</a>
<b>4. Structure Databases</b>			
<b>4.1. Small molecules</b>			
	LIGAND	Chemical compounds and reactions in biological pathways	<a href="http://www.genome.ad.jp/ligand/">http://www.genome.ad.jp/ligand/</a>
	PDB-Ligand	3D structures of small molecules bound to proteins and nucleic acids	<a href="http://www.idrtech.com/PDB-Ligand/">http://www.idrtech.com/PDB-Ligand/</a>
	PubChem	Structures and biological activities of small organic molecules	<a href="http://pubchem.ncbi.nlm.nih.gov/">http://pubchem.ncbi.nlm.nih.gov/</a>
<b>4.2. Carbohydrates</b>			
	Monosaccharide Browser	Space-filling Fischer projections of monosaccharides	<a href="http://www.beechtreecommon.org/biochemistry/monosaccharide/">http://www.beechtreecommon.org/biochemistry/monosaccharide/</a>
	SWEET-DB	Annotated carbohydrate structure and substance information	<a href="http://www.dkfz-heidelberg.de/spec2/sweetdb/">http://www.dkfz-heidelberg.de/spec2/sweetdb/</a>
<b>4.3. Nucleic acid structure</b>			
	NDB	Nucleic acid-containing structures	<a href="http://ndbserver.rutgers.edu/">http://ndbserver.rutgers.edu/</a>
	NTDB	Thermodynamic data for nucleic acids	<a href="http://ntdb.chem.cuhk.edu.hk/">http://ntdb.chem.cuhk.edu.hk/</a>
<b>4.4. Protein structure</b>			
	BALiBASE	A database for comparison of multiple sequence alignments	<a href="http://www-igbmc.u-strasbg.fr/BioInfo/BALiBASE2/index.html">http://www-igbmc.u-strasbg.fr/BioInfo/BALiBASE2/index.html</a>
	CATH	Protein domain structures database	<a href="http://www.biochem.ucl.ac.uk/bsm/cath_new">http://www.biochem.ucl.ac.uk/bsm/cath_new</a>
	CE	3D protein structure alignments	<a href="http://cl.sdsc.edu/ce.html">http://cl.sdsc.edu/ce.html</a>
	Dali	Protein fold classification using the Dali search engine	<a href="http://www.bioinfo.biocenter.helsinki.fi:8080/dali/">http://www.bioinfo.biocenter.helsinki.fi:8080/dali/</a>
	HOMSTRAD	Homologous structure alignment database: curated structure-based alignments for protein families	<a href="http://www-cryst.bioc.cam.ac.uk/homstrad">http://www-cryst.bioc.cam.ac.uk/homstrad</a>
	IMB Jena Image Library	Visualization and analysis of 3D biopolymer structures	<a href="http://www.imb-jena.de/IMAGE.html">http://www.imb-jena.de/IMAGE.html</a>
	RCSB PDB	Protein structure databank: all	<a href="http://www.rcsb.org/pdb">http://www.rcsb.org/pdb</a>

		publicly available 3D structures of proteins and nucleic acids	
	PDB-REPRDB	Representative protein chains, based on PDB entries	<a href="http://mbs.cbrc.jp/pdbreprdb/cgi/reprdb_menu.pl">http://mbs.cbrc.jp/pdbreprdb/cgi/reprdb_menu.pl</a>
	PDBsum	Summaries and analyses of PDB structures	<a href="http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/">http://www.ebi.ac.uk/thornton-srv/databases/pdbsum/</a>
	PDB_TM	Transmembrane proteins with known 3D structure	<a href="http://pdbtm.enzim.hu/">http://pdbtm.enzim.hu/</a>
	PMDB	3D protein models obtained from structure predictions	<a href="http://a.caspar.it/PMDB/">http://a.caspar.it/PMDB/</a>
	Protein Folding Database	Experimental data on protein folding	<a href="http://pfd.med.monash.edu.au">http://pfd.med.monash.edu.au</a>
	SCOP	Structural classification of proteins	<a href="http://scop.mrc-lmb.cam.ac.uk/scop">http://scop.mrc-lmb.cam.ac.uk/scop</a>
	SUPERFAMILY	Assignments of proteins to structural superfamilies	<a href="http://supfam.org/">http://supfam.org/</a>
	SURFACE	Surface residues and functions annotated, compared and evaluated: a database of protein surface patches	<a href="http://cbm.bio.uniroma2.it/surface">http://cbm.bio.uniroma2.it/surface</a>
<b>5. Genomics Databases (non-human)</b>			
<b>5.1. Genome annotation terms, ontologies and nomenclature</b>			
	GO	Gene ontology consortium database	<a href="http://www.geneontology.org/">http://www.geneontology.org/</a>
	GOA	EBI's gene ontology annotation project	<a href="http://www.ebi.ac.uk/GOA">http://www.ebi.ac.uk/GOA</a>
<b>5.1.1. Taxonomy and Identification</b>			
	ICB	<i>gyrB</i> database for identification of bacteria	<a href="http://seasquirt.mbio.co.jp/icb/index.php">http://seasquirt.mbio.co.jp/icb/index.php</a>
	NCBI Taxonomy	Names of all organisms represented in GenBank	<a href="http://www.ncbi.nlm.nih.gov/Taxonomy/">http://www.ncbi.nlm.nih.gov/Taxonomy/</a>
	PANDIT	Protein and associated nucleotide domains with inferred trees	<a href="http://www.ebi.ac.uk/goldman-srv/pandit/">http://www.ebi.ac.uk/goldman-srv/pandit/</a>
<b>5.2. General genomics databases</b>			
	Comparative Genomics	Nucleotide frequencies and the GC and TA skews in complete genome sequences	<a href="http://www.unil.ch/comparativegenomics/">http://www.unil.ch/comparativegenomics/</a>
	DEG	Database of essential genes from bacteria and yeast	<a href="http://tubic.tju.edu.cn/deg">http://tubic.tju.edu.cn/deg</a>
	EBI Genomes	EBI's collection of databases for the analysis of complete and unfinished viral, pro- and eukaryotic genomes	<a href="http://www.ebi.ac.uk/genomes">http://www.ebi.ac.uk/genomes</a>
	Entrez Genomes	NCBI's collection of databases for the analysis of complete and unfinished viral, pro- and eukaryotic genomes	<a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Genome</a>
	Genome Information Broker	DDBJ's collection of genome databases	<a href="http://gib.genes.nig.ac.jp">http://gib.genes.nig.ac.jp</a>
	KEGG	Kyoto encyclopedia of genes and genomes: databases on genes, proteins, and metabolic pathways	<a href="http://www.genome.jp/kegg">http://www.genome.jp/kegg</a>
	TIGR Microbial Database	Lists of completed and ongoing genome projects with links to complete genome sequences	<a href="http://www.tigr.org/tdb/mdb/mdbcomplete.html">http://www.tigr.org/tdb/mdb/mdbcomplete.html</a>
	TIGR Comprehensive Microbial Resource	Various data on complete microbial genomes: Uniform annotation, properties of DNA and predicted proteins	<a href="http://www.tigr.org/CMR">http://www.tigr.org/CMR</a>
<b>6. Metabolic Enzymes and Pathways; Signaling Pathways</b>			

	Pathguide	A listing of pathway, signal transduction and protein-protein interaction databases	<a href="http://pathguide.org">http://pathguide.org</a>
<b>6.1. Enzymes and Enzyme Nomenclature</b>			
	Enzyme Nomenclature	IUBMB Nomenclature Committee recommendations	<a href="http://www.chem.qmw.ac.uk/iubmb/enzyme">http://www.chem.qmw.ac.uk/iubmb/enzyme</a>
<b>6.2. Metabolic Pathways</b>			
	BioSilico	Integrated access to various metabolic databases	<a href="http://biosilico.kaist.ac.kr/">http://biosilico.kaist.ac.kr/</a>
	KEGG Pathway	Metabolic and regulatory pathways in complete genomes	<a href="http://www.genome.jp/kegg/pathway.html">http://www.genome.jp/kegg/pathway.html</a>
<b>7. Human and other Vertebrate Genomes</b>			
<b>7.1. Human genome databases, maps and viewers</b>			
	Ensembl	Annotated information on eukaryotic genomes	<a href="http://www.ensembl.org/">http://www.ensembl.org/</a>
	GenAtlas	Human genes, markers and phenotypes	<a href="http://www.genatlas.org/">http://www.genatlas.org/</a>
	GeneCards	Integrated database of human genes, maps, proteins and diseases	<a href="http://bioinfo.weizmann.ac.il/cards/">http://bioinfo.weizmann.ac.il/cards/</a>
	Human Genome Segmental Duplication Database	Segmental duplications in the human genome	<a href="http://projects.tcag.ca/humandup">http://projects.tcag.ca/humandup</a>
	Map Viewer	Display of genomic information by chromosomal position	<a href="http://www.ncbi.nlm.nih.gov/mapview/">http://www.ncbi.nlm.nih.gov/mapview/</a>
	MGC	Mammalian genome collection: Full-length ORFs for human, mouse, and rat genes	<a href="http://mgc.nci.nih.gov/">http://mgc.nci.nih.gov/</a>
	NCBI RefSeq	Non-redundant collection of naturally-occurring biological molecules	<a href="http://www.ncbi.nlm.nih.gov/RefSeq/">http://www.ncbi.nlm.nih.gov/RefSeq/</a>
	UCSC Genome Browser	Genome assemblies and annotation	<a href="http://genome.ucsc.edu/">http://genome.ucsc.edu/</a>
<b>8. Human Genes and Diseases</b>			
<b>8.1. General Databases</b>			
	OMIA	Online Mendelian inheritance in animals: A catalog of animal genetic and genomic disorders	<a href="http://www.angis.org.au/omia">http://www.angis.org.au/omia</a>
	OMIM	Online Mendelian inheritance in man: A catalog of human genetic and genomic disorders	<a href="http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM">http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM</a>
<b>9. Microarray Data and other Gene Expression Databases</b>			
	ArrayExpress	Public collection of microarray gene expression data	<a href="http://www.ebi.ac.uk/arrayexpress">http://www.ebi.ac.uk/arrayexpress</a>
	GEO	Gene expression omnibus: Gene expression profiles	<a href="http://www.ncbi.nlm.nih.gov/geo/">http://www.ncbi.nlm.nih.gov/geo/</a>
	GermOnline	Gene expression in mitotic and meiotic cell cycle	<a href="http://www.germonline.org/">http://www.germonline.org/</a>
	GXD	Mouse gene expression database	<a href="http://www.informatics.jax.org/menu/expression_menu.shtml">http://www.informatics.jax.org/menu/expression_menu.shtml</a>
	H-ANGEL	Human anatomic gene expression library	<a href="http://www.jbirc.aist.go.jp/hinv/index.jsp">http://www.jbirc.aist.go.jp/hinv/index.jsp</a>
	HemBase	Genes expressed in differentiating human erythroid cells	<a href="http://hembase.niddk.nih.gov/">http://hembase.niddk.nih.gov/</a>
	HugeIndex	Expression levels of human genes in normal tissues	<a href="http://zlab.bu.edu/HugeSearch">http://zlab.bu.edu/HugeSearch</a>
	IGTC	International mouse Gene Trap Consortium data	<a href="http://wwwtest.genetrap.org">http://wwwtest.genetrap.org</a>

Kidney Development Database	Kidney development and gene expression	<a href="http://golgi.ana.ed.ac.uk/kidhome.html">http://golgi.ana.ed.ac.uk/kidhome.html</a>
LOLA	List of lists annotated: a comparison of gene sets identified in different microarray experiments	<a href="http://www.lola.gwu.edu/">http://www.lola.gwu.edu/</a>
MAGEST	Ascidian ( <i>Halocynthia roretzi</i> ) gene expression patterns	<a href="http://www.genome.ad.jp/magest">http://www.genome.ad.jp/magest</a>
MAMEP	Molecular anatomy of the mouse embryo project: Gene expression data on mouse embryos	<a href="http://mamep.molgen.mpg.de/">http://mamep.molgen.mpg.de/</a>
MEPD	Medaka (freshwater fish <i>Oryzias latipes</i> ) gene expression pattern database	<a href="http://www.embl.de/mepd/">http://www.embl.de/mepd/</a>
MethDB	DNA methylation data, patterns and profiles	<a href="http://www.methdb.de/">http://www.methdb.de/</a>
Mouse SAGE	SAGE libraries from various mouse tissues and cell lines	<a href="http://mouse.biomed.cas.cz/sage">http://mouse.biomed.cas.cz/sage</a>
NASCarrays	Nottingham <i>Arabidopsis</i> Stock Centre microarray database	<a href="http://affymetrix.arabidopsis.info">http://affymetrix.arabidopsis.info</a>
NetAffx	Public Affymetrix probesets and annotations	<a href="http://www.affymetrix.com/">http://www.affymetrix.com/</a>
Osteo-Promoter Database	Genes in osteogenic proliferation and differentiation	<a href="http://www.opd.tau.ac.il">http://www.opd.tau.ac.il</a>
PEDB	Prostate expression database: ESTs from prostate tissue and cell type-specific cDNA libraries	<a href="http://www.pedb.org/">http://www.pedb.org/</a>
PEPR	Public expression profiling resource: Expression profiles in a variety of diseases and conditions	<a href="http://pepr.cnmcresearch.org">http://pepr.cnmcresearch.org</a>
RECODE	Genes using programmed translational recoding in their expression	<a href="http://recode.genetics.utah.edu/">http://recode.genetics.utah.edu/</a>
RefExA	Reference database for human gene expression analysis	<a href="http://www.lsbm.org/db/index_e.html">http://www.lsbm.org/db/index_e.html</a>
rOGED	Rat ovarian gene expression database	<a href="http://app.mc.uky.edu/kolab/rogedendo.asp">http://app.mc.uky.edu/kolab/rogedendo.asp</a>
SAGEmap	NCBI's resource for SAGE data from various organisms	<a href="http://www.ncbi.nlm.nih.gov/SAGE">http://www.ncbi.nlm.nih.gov/SAGE</a>
SIEGE	Smoking Induced Epithelial Gene Expression	<a href="http://pulm.bumc.bu.edu/siegeDB">http://pulm.bumc.bu.edu/siegeDB</a>
Stanford Microarray Database	Raw and normalized data from microarray experiments	<a href="http://genome-www.stanford.edu/microarray">http://genome-www.stanford.edu/microarray</a>
TmaDB	Tissue microarray database	<a href="http://www.bioinformatics.leeds.ac.uk/tmadb/">http://www.bioinformatics.leeds.ac.uk/tmadb/</a>
Tooth Development Database	Gene expression in dental tissue	<a href="http://bite-it.helsinki.fi/">http://bite-it.helsinki.fi/</a>
<b>10. Proteomics Resources</b>		
2D-PAGE	Proteome database system for microbial research	<a href="http://www.mpiib-berlin.mpg.de/2D-PAGE">http://www.mpiib-berlin.mpg.de/2D-PAGE</a>
dbPTM	Information on post-translational modification of proteins	<a href="http://dbptm.mbc.nctu.edu.tw/">http://dbptm.mbc.nctu.edu.tw/</a>
DynaProt 2D	Proteome database of <i>Lactococcus lactis</i>	<a href="http://www.wzw.tum.de/proteomik/lactis/">http://www.wzw.tum.de/proteomik/lactis/</a>
GelBank	2D gel electrophoresis patterns of	<a href="http://gelbank.anl.gov/">http://gelbank.anl.gov/</a>

		proteins from complete microbial genomes	
	Open Proteomics Database	Mass-spectrometry-based proteomics data for human, yeast, <i>E.coli</i> and <i>Mycobacterium</i>	<a href="http://bioinformatics.icmb.utexas.edu/OPD/">http://bioinformatics.icmb.utexas.edu/OPD/</a>
	PEP	Predictions for entire proteomes: Summarized analyses of protein sequences	<a href="http://cubic.bioc.columbia.edu/pep/">http://cubic.bioc.columbia.edu/pep/</a>
	PepSeeker	Peptide identification and ion information from proteome experiments	<a href="http://nwsr.bms.umist.ac.uk/cgi-bin/pepseeker/pepseek.pl">http://nwsr.bms.umist.ac.uk/cgi-bin/pepseeker/pepseek.pl</a>
	PeptideAtlas	Peptides identified in LC-MS/MS proteomics experiments	<a href="http://www.peptideatlas.org">http://www.peptideatlas.org</a>
	PRIDE	Proteomics identification database	<a href="http://www.ebi.ac.uk/pride/">http://www.ebi.ac.uk/pride/</a>
	RESID	Pre-, co- and post-translational protein modifications	<a href="http://www-nbrf.georgetown.edu/pirwww/dbinfo/resid.html">http://www-nbrf.georgetown.edu/pirwww/dbinfo/resid.html</a>
	SWISS-2DPAGE	Annotated 2D gel electrophoresis database	<a href="http://www.expasy.org/ch2d/">http://www.expasy.org/ch2d/</a>
<b>11. Other Molecular Biology Databases</b>			
<b>11.1. Drugs and drug design</b>			
	ANTIMIC	Database of natural antimicrobial peptides	<a href="http://research.i2r.a-star.edu.sg/Templar/DB/ANTIMIC/">http://research.i2r.a-star.edu.sg/Templar/DB/ANTIMIC/</a>
	AOBase	Antisense oligonucleotide selection and design	<a href="http://www.bioit.org.cn/ao/aobase">http://www.bioit.org.cn/ao/aobase</a>
	APD	Antimicrobial peptide database	<a href="http://aps.unmc.edu/AP/main.php">http://aps.unmc.edu/AP/main.php</a>
	BSD	Biodegradative strain database: Microorganisms that can degrade aromatic and other organic compounds	<a href="http://bsd.cme.msu.edu/">http://bsd.cme.msu.edu/</a>
	DART	Drug adverse reaction target database	<a href="http://xin.cz3.nus.edu.sg/group/drt/dart.asp">http://xin.cz3.nus.edu.sg/group/drt/dart.asp</a>
	DrugBank	Combined information on drugs and drug targets	<a href="http://redpoll.pharmacy.ualberta.ca/drugbank/">http://redpoll.pharmacy.ualberta.ca/drugbank/</a>
	GLIDA	G-protein coupled receptors ligand database	<a href="http://gdds.pharm.kyoto-u.ac.jp:8081/glida/">http://gdds.pharm.kyoto-u.ac.jp:8081/glida/</a>
	MetaRouter	Compounds and pathways related to bioremediation	<a href="http://pdg.cnb.uam.es/MetaRouter">http://pdg.cnb.uam.es/MetaRouter</a>
	Peptaibol	Peptaibol (antibiotic peptide) sequences	<a href="http://www.cryst.bbk.ac.uk/peptaibol/welcome.html">http://www.cryst.bbk.ac.uk/peptaibol/welcome.html</a>
	PharmGKB	Pharmacogenomics knowledge base: effect of genetic variation on drug responses	<a href="http://www.pharmgkb.org/">http://www.pharmgkb.org/</a>
	SuperDrug	2D and 3D chemical structures of various drugs	<a href="http://bioinformatics.charite.de/superdrug">http://bioinformatics.charite.de/superdrug</a>
	SuperNatural	Natural compounds and their suppliers	<a href="http://bioinformatics.charite.de/supernatural">http://bioinformatics.charite.de/supernatural</a>
	TTD	Therapeutic target database	<a href="http://xin.cz3.nus.edu.sg/group/cjttd/ttd.asp">http://xin.cz3.nus.edu.sg/group/cjttd/ttd.asp</a>
<b>11.2. Probes</b>			
	IMGT/PRIMER-DB	Immunogenetics oligonucleotide primer database	<a href="http://imgt3d.igh.cnrs.fr/PrimerDB/Query_PrDB.pl">http://imgt3d.igh.cnrs.fr/PrimerDB/Query_PrDB.pl</a>
	MPDB	Synthetic oligonucleotides useful as primers or probes	<a href="http://www.biotech.ist.unige.it/interlab/mpdb.html">http://www.biotech.ist.unige.it/interlab/mpdb.html</a>
	PrimerPCR	PCR primers for eukaryotic and prokaryotic genes	<a href="http://bioinfo.ebc.ee/PrimerStudio/">http://bioinfo.ebc.ee/PrimerStudio/</a>
	probeBase	rRNA-targeted oligonucleotide probe sequences, DNA microarray layouts, and associated information	<a href="http://www.microbial-ecology.net/probebase">http://www.microbial-ecology.net/probebase</a>

	QPPD	Quantitative PCR Primer Database for human and mouse	<a href="http://web.ncifcrf.gov/rtp/GEL/primerdb/default.asp">http://web.ncifcrf.gov/rtp/GEL/primerdb/default.asp</a>
	RTPrimerDB	Real-time PCR primer and probe sequences	<a href="http://medgen.ugent.be/rtpprimerdb/">http://medgen.ugent.be/rtpprimerdb/</a>
<b>11.3. Unclassified databases</b>			
	PubMed	Citations and abstracts of biomedical literature	<a href="http://pubmed.gov/">http://pubmed.gov/</a>
	BioImage	Database of multidimensional biological images	<a href="http://www.bioimage.org/">http://www.bioimage.org/</a>
	BioModels	Published mathematical models of biological interest	<a href="http://www.ebi.ac.uk/biomodels/">http://www.ebi.ac.uk/biomodels/</a>
<b>12. Organelle Databases</b>			
	ChloroplastDB	Chloroplast genome database	<a href="http://chloroplast.cbio.psu.edu/">http://chloroplast.cbio.psu.edu/</a>
	FUGOID	Functional genomics of organelle introns database	<a href="http://web.austin.utexas.edu/fugoid/introndata/main.htm">http://web.austin.utexas.edu/fugoid/introndata/main.htm</a>
	GOBASE	Organelle genome database	<a href="http://megasun.bch.umontreal.ca/gobase/">http://megasun.bch.umontreal.ca/gobase/</a>
	OGRe	Organelle genome retrieval system	<a href="http://ogre.mcmaster.ca">http://ogre.mcmaster.ca</a>
	Organelle genomes	NCBI's organelle genome resource	<a href="http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/organelles.html">http://www.ncbi.nlm.nih.gov/genomes/ORGANELLES/organelles.html</a>
	Organelle DB	Organelle proteins and subcellular structures	<a href="http://organelledb.lsi.umich.edu/">http://organelledb.lsi.umich.edu/</a>
	PLprot	<i>Arabidopsis thaliana</i> chloroplast protein database	<a href="http://www.pb.ipw.biol.ethz.ch/proteomics">http://www.pb.ipw.biol.ethz.ch/proteomics</a>
<b>12.1. Mitochondrial Genes and Proteins</b>			
	AMPDB	<i>Arabidopsis</i> mitochondrial protein database	<a href="http://www.mitoz.bcs.uwa.edu.au/AMPDB/">http://www.mitoz.bcs.uwa.edu.au/AMPDB/</a>
	HMPD	Human mitochondrial protein database	<a href="http://bioinfo.nist.gov:8080/examples/servlets/index.html">http://bioinfo.nist.gov:8080/examples/servlets/index.html</a>
	Human MtDB	Human mitochondrial genome database	<a href="http://www.genpat.uu.se/mtDB">http://www.genpat.uu.se/mtDB</a>
	HvrBase	Primate mitochondrial DNA control region sequences	<a href="http://www.hvrbase.org/">http://www.hvrbase.org/</a>
	MamMiBase	Mammalian mitochondrial genome database	<a href="http://xavante.fmrp.usp.br/mammibase/">http://xavante.fmrp.usp.br/mammibase/</a>
	Mitochondriome	Metazoan mitochondrial genes	<a href="http://www.ba.itb.cnr.it/mitochondriome/index.html">http://www.ba.itb.cnr.it/mitochondriome/index.html</a>
	MitoDat	Mitochondrial proteins (predominantly human)	<a href="http://www-lecb.ncifcrf.gov/mitoDat/">http://www-lecb.ncifcrf.gov/mitoDat/</a>
	MitoDrome	Nuclear-encoded mitochondrial proteins of <i>Drosophila</i>	<a href="http://www2.ba.itb.cnr.it/MitoDrome/">http://www2.ba.itb.cnr.it/MitoDrome/</a>
	MitoMap	Human mitochondrial genome	<a href="http://www.mitomap.org/">http://www.mitomap.org/</a>
	MitoNuc	Nuclear genes coding for mitochondrial proteins	<a href="http://www2.ba.itb.cnr.it/MitoNuc/">http://www2.ba.itb.cnr.it/MitoNuc/</a>
	MITOP2	Mitochondrial proteins, genes and diseases	<a href="http://ihg.gsf.de/mitop2/start.jsp">http://ihg.gsf.de/mitop2/start.jsp</a>
	MitoPD	Yeast mitochondrial protein database	<a href="http://bmerc-www.bu.edu/projects/mito/">http://bmerc-www.bu.edu/projects/mito/</a>
	MitoProteome	Experimentally described human mitochondrial proteins	<a href="http://www.mitoproteome.org">http://www.mitoproteome.org</a>
	MPIMP	Mitochondrial protein import machinery of plants	<a href="http://millar3.biochem.uwa.edu.au/~lister/index.html">http://millar3.biochem.uwa.edu.au/~lister/index.html</a>
	PLMitRNA	Plant mitochondrial tRNA	<a href="http://bighost.area.ba.cnr.it/PLMitRNA/">http://bighost.area.ba.cnr.it/PLMitRNA/</a>
<b>13. Plant Databases</b>			
<b>13.1. General plant databases</b>			
	BarleyBase	Expression profiling of plant genomes	<a href="http://www.barleybase.org/">http://www.barleybase.org/</a>
	CR-EST	Crop ESTs: barley, pea, wheat and potato	<a href="http://pgrc.ipk-gatersleben.de/cr-est/">http://pgrc.ipk-gatersleben.de/cr-est/</a>
	CropNet	Genome mapping in crop plants	<a href="http://ukcrop.net/">http://ukcrop.net/</a>
	DRASTIC	Database resource for analysis of	<a href="http://www.drastic.org.uk">http://www.drastic.org.uk</a>

	signal transduction in plant cells	
FLAGdb++	<a href="http://genoplante-info.infobiogen.fr/FLAGdb/">Integrative database about plant genomes</a>	<a href="http://genoplante-info.infobiogen.fr/FLAGdb/">http://genoplante-info.infobiogen.fr/FLAGdb/</a>
GénoPlante-Info	Plant genomic data from the Génoplante consortium	<a href="http://genoplante-info.infobiogen.fr/">http://genoplante-info.infobiogen.fr/</a>
GrainGenes	Genes and phenotypes of wheat, barley, rye, triticale, oats	<a href="http://wheat.pw.usda.gov">http://wheat.pw.usda.gov</a> or <a href="http://www.graingenes.org">http://www.graingenes.org</a>
Gramene	A resource for comparative grass genomics	<a href="http://www.gramene.org/">http://www.gramene.org/</a>
openSputnik	Plant EST clustering and functional annotation	<a href="http://www.opensputnik.org/">http://www.opensputnik.org/</a>
Phytome	Comparative genomics of plant species	<a href="http://www.phytome.org">http://www.phytome.org</a>
PhytoProt	Clusters of (predicted) plant proteins	<a href="http://genoplante-info.infobiogen.fr/phytoprot">http://genoplante-info.infobiogen.fr/phytoprot</a>
PlantMarkers	A database of predicted molecular markers from plants	<a href="http://markers.btk.fi/">http://markers.btk.fi/</a>
Plant MPSS	Massively parallel signature sequencing of plant genes	<a href="http://mpss.udel.edu">http://mpss.udel.edu</a>
PlantGDB	Plant genome database: Actively-transcribed plant genes	<a href="http://www.plantgdb.org/">http://www.plantgdb.org/</a>
PLANT-PIs	Plant protease inhibitors	<a href="http://bighost.area.ba.cnr.it/PLANT-PIs">http://bighost.area.ba.cnr.it/PLANT-PIs</a>
PlantsP/PlantsT	Plant proteins involved in phosphorylation and transport	<a href="http://plantsp.sdsc.edu/">http://plantsp.sdsc.edu/</a>
TIGR plant repeat database	Classification of repetitive sequences in plant genomes	<a href="http://www.tigr.org/tdb/e2k1/plant.repeats">http://www.tigr.org/tdb/e2k1/plant.repeats</a>
TropGENE DB	Genes and genomes of sugarcane, banana, cocoa	<a href="http://tropgenedb.cirad.fr/">http://tropgenedb.cirad.fr/</a>
<i>13.2. Arabidopsis thaliana</i>		
AGNS	<i>Arabidopsis</i> GeneNet supplementary: Gene expression and phenotypes of mutants and transgens	<a href="http://emj-pc.ics.uci.edu/mgs/dbases/agns">http://emj-pc.ics.uci.edu/mgs/dbases/agns</a>
AGRIS	<i>Arabidopsis</i> gene regulatory information server: promoters, transcription factors and their target genes	<a href="http://arabidopsis.med.ohio-state.edu/">http://arabidopsis.med.ohio-state.edu/</a>
<i>Arabidopsis</i> MPSS	<i>Arabidopsis</i> gene expression detected by massively parallel signature sequencing	<a href="http://mpss.udel.edu/at/">http://mpss.udel.edu/at/</a>
<i>Arabidopsis</i> Nucleolar Protein Database	Comparative analysis of human and <i>Arabidopsis</i> nucleolar proteomes	<a href="http://bioinf.scri.sari.ac.uk/cgi-bin/atnopdb/proteome_comparison">http://bioinf.scri.sari.ac.uk/cgi-bin/atnopdb/proteome_comparison</a>
ASRP	<i>Arabidopsis thaliana</i> small RNA project	<a href="http://asrp.cgrb.oregonstate.edu/">http://asrp.cgrb.oregonstate.edu/</a>
ARAMEMNON	<i>Arabidopsis thaliana</i> membrane proteins and transporters	<a href="http://aramemnon.botanik.uni-koeln.de/">http://aramemnon.botanik.uni-koeln.de/</a>
ARTADEdb	<i>Arabidopsis</i> tiling-array-based detection of exons	<a href="http://omicspace.riken.jp/ARTADE/">http://omicspace.riken.jp/ARTADE/</a>
AthaMap	Genome-wide map of putative transcription factor binding sites in <i>Arabidopsis thaliana</i>	<a href="http://www.athamap.de/">http://www.athamap.de/</a>
CATMA	Complete <i>Arabidopsis</i> transcriptome microarray	<a href="http://www.catma.org">http://www.catma.org</a>
DATF	Database of <i>Arabidopsis</i> transcription factors	<a href="http://datf.cbi.pku.edu.cn/">http://datf.cbi.pku.edu.cn/</a>
GabiPD	Central database of the German Plant Genome Project	<a href="http://gabi.rzpd.de/">http://gabi.rzpd.de/</a>
GeneFarm	Expert annotation of <i>Arabidopsis</i> gene and protein families	<a href="http://genoplante-info.infobiogen.fr/Genefarm/">http://genoplante-info.infobiogen.fr/Genefarm/</a>

MAtdB	MIPS <i>Arabidopsis thaliana</i> database	<a href="http://mips.gsf.de/proj/thal/db">http://mips.gsf.de/proj/thal/db</a>
RARGE	RIKEN <i>Arabidopsis</i> genome encyclopedia: cDNAs, mutants and microarray data	<a href="http://rarge.gsc.riken.jp/">http://rarge.gsc.riken.jp/</a>
SeedGenes	Genes essential for <i>Arabidopsis</i> development	<a href="http://www.seedgenes.org/">http://www.seedgenes.org/</a>
TAIR	The <i>Arabidopsis</i> information resource	<a href="http://www.arabidopsis.org/">http://www.arabidopsis.org/</a>
WAtDB	Wageningen <i>Arabidopsis thaliana</i> database: mutants, transgenic lines and natural variants	<a href="http://www.watdb.nl/">http://www.watdb.nl/</a>
<b>13.3. Rice</b>		
BGI-RISe	Beijing genomics institute rice information system	<a href="http://rise.genomics.org.cn/">http://rise.genomics.org.cn/</a>
INE	Integrated rice genome explorer	<a href="http://rgp.dna.affrc.go.jp/giot/INE.html">http://rgp.dna.affrc.go.jp/giot/INE.html</a>
IRIS	International rice information system	<a href="http://www.iris.irri.org/">http://www.iris.irri.org/</a>
MOsDB	MIPS <i>Oryza sativa</i> database	<a href="http://mips.gsf.de/proj/plant/jsf/rice/index.jsp">http://mips.gsf.de/proj/plant/jsf/rice/index.jsp</a>
OryGenesDB	Rice genes, T-DNA and Ds flanking sequence tags	<a href="http://orygenesdb.cirad.fr/">http://orygenesdb.cirad.fr/</a>
Oryzabase	Rice genetics and genomics	<a href="http://www.shigen.nig.ac.jp/rice/oryzabase/">http://www.shigen.nig.ac.jp/rice/oryzabase/</a>
Oryza Tag Line database	T-DNA insertion mutants of rice	<a href="http://genoplante-info.infobiogen.fr/OryzaTagLine/">http://genoplante-info.infobiogen.fr/OryzaTagLine/</a>
RAD	Rice annotation database	<a href="http://golgi.gs.dna.affrc.go.jp/SY-1102/rad/index.html">http://golgi.gs.dna.affrc.go.jp/SY-1102/rad/index.html</a>
RAP-DB	Rice annotation project database	<a href="http://rapdev.lab.nig.ac.jp/">http://rapdev.lab.nig.ac.jp/</a>
RiceGAAS	Rice genome automated annotation system	<a href="http://ricegaas.dna.affrc.go.jp/">http://ricegaas.dna.affrc.go.jp/</a>
Rice PIPELINE	Unification tool for rice databases	<a href="http://cdna01.dna.affrc.go.jp/PIPE">http://cdna01.dna.affrc.go.jp/PIPE</a>
Rice proteome database	Rice proteome database	<a href="http://gene64.dna.affrc.go.jp/RPD/main_en.html">http://gene64.dna.affrc.go.jp/RPD/main_en.html</a>
RMD	Rice mutant database	<a href="http://rmd.ncpgr.cn/">http://rmd.ncpgr.cn/</a>
<b>13.4. Other plants</b>		
Brassica ASTRA	A database for <i>Brassica</i> genomic research	<a href="http://hornbill.cspp.latrobe.edu.au/cgi-bin/pub/index.pl">http://hornbill.cspp.latrobe.edu.au/cgi-bin/pub/index.pl</a>
MaizeGDB	Maize genetics and genomics database	<a href="http://www.maizegdb.org/">http://www.maizegdb.org/</a>
LIS (formerly MGI)	Legume information server (formerly Medicago genome initiative): ESTs, gene expression and proteomic data	<a href="http://www.comparative-legumes.org/">http://www.comparative-legumes.org/</a>
MtDB	<i>Medicago trunculata</i> genome database	<a href="http://www.medicago.org/MtDB">http://www.medicago.org/MtDB</a>
Panzea	Maize genome project data	<a href="http://www.panzea.org">http://www.panzea.org</a>
PoMaMo	Potato Maps and More: Potato genome data	<a href="https://gabi.rzpd.de/PoMaMo.html">https://gabi.rzpd.de/PoMaMo.html</a>
SGMD	Soybean genomics and microarray database	<a href="http://psi081.ba.ars.usda.gov/SGMD/default.htm">http://psi081.ba.ars.usda.gov/SGMD/default.htm</a>
SoyGD	Soybean genome database	<a href="http://soybeanome.siu.edu">http://soybeanome.siu.edu</a>
TED	Tomato expression database	<a href="http://ted.bti.cornell.edu">http://ted.bti.cornell.edu</a>
TIGR Maize database	Maize genome sequencing consortium site	<a href="http://maize.tigr.org">http://maize.tigr.org</a>
<b>14. Immunological Databases</b>		
BCIpep	A database of B-cell epitopes	<a href="http://bioinformatics.uams.edu/mirror/bcip/ep/">http://bioinformatics.uams.edu/mirror/bcip/ep/</a>
dbMHC	Genetic and clinical database of the human MHC	<a href="http://www.ncbi.nlm.nih.gov/mhc/">http://www.ncbi.nlm.nih.gov/mhc/</a>
Epitome	Antigenic epitopes in proteins and antibodies that bind them	<a href="http://predictprotein.org/AifoEpi/">http://predictprotein.org/AifoEpi/</a>

FIMM	Functional molecular immunology data	<a href="http://research.i2r.a-star.edu.sg/fimm">http://research.i2r.a-star.edu.sg/fimm</a>
GPX	Macrophage expression atlas	<a href="http://darwin.gti.ed.ac.uk/GPX/cgi-bin/Scripts/selectexperiment.cgi">http://darwin.gti.ed.ac.uk/GPX/cgi-bin/Scripts/selectexperiment.cgi</a>
HLA Ligand/Motif	A database and search tool for HLA sequences	<a href="http://hlaligand.ouhsc.edu/">http://hlaligand.ouhsc.edu/</a>
IL2Rgbase	X-linked severe combined immunodeficiency mutations	<a href="http://research.nhgri.nih.gov/scid/">http://research.nhgri.nih.gov/scid/</a>

\*\*\*\*\*

## Basics of Cheminformatics

Chittaranjan Baruah

Bioinformatics Centre (DBT-BIF)

Department of Zoology (UGC-SAP & DST-FIST sponsored Department),

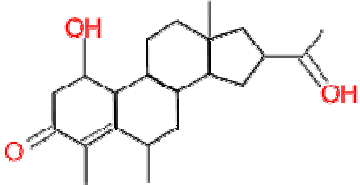
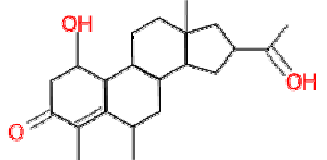
Gauhati University, Guwahati – 781 014, Assam, India

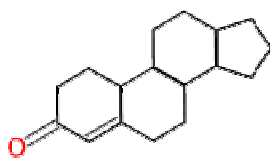
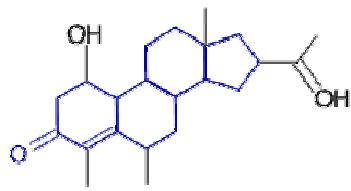
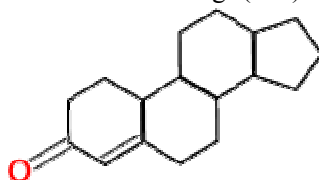
E-mail: chittaranjan\_2004@india.com

### What is Cheminformatics?

**Cheminformatics** is a cross between Computer Science and Chemistry: The process of storing and retrieving information about chemical compounds. The term Chemoinformatics was defined by F.K. Brown in 1998. *In silico* Cheminformatics techniques are used in [pharmaceutical](#) companies in the process of [drug discovery](#). These methods can also be used in chemical and allied industries in various other forms.

*Information Systems* in Cheminformatics are concerned with storing, retrieving, and searching information, and with storing *relationships* between bits of data. For example:

Operation	Classical System	Information	Chemical Information System	
<b>Store</b>	Name = 'Jimmy Carter'	Stores text, numbers, dates, ...		Stores chemical compounds and information about them.
<b>Retrieve</b>	Find record #13282	Retrieves 'Jimmy Carter'	Find: CC(=O)C4CC3C2CC(C)C1=C(C) C(=O)CC(O)C1C2CCC3(C)C4	Retrieves: 

<b>Search</b>	Find Presidents named 'Bush'	George Bush and George W. Bush	Find molecules containing: 	Retrieves: 
<b>Relationship</b>	Year Carter was elected	Answer: Elected in 1976	What's the logP(o/w) of: 	logP(o/w) = 2.62

## How is Cheminformatics Different?

There are four key problems a cheminformatics system solves:

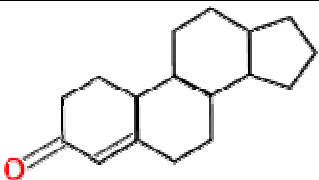
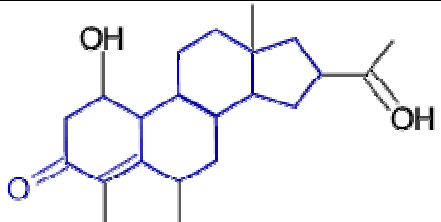
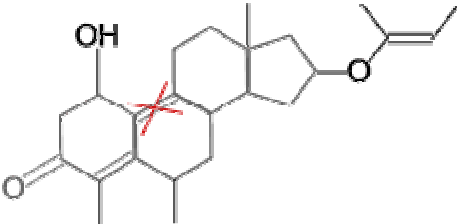
- 1. Store a Molecule** Computer scientists usually use the *valence model* of chemistry to represent compounds. Section 2, Representing Molecules, discusses this at length.
- 2. Find exact molecule** If you ask, "Is Abraham Lincoln in the database?" it's not hard to find the answer. But, given a specific molecule, is it in the database? What do we know about it? This may seem simple at first glance, but it's not, as we'll see when we discuss tautomers, stereochemistry, metals, and other "flaws" in the valence model of chemistry.
- 3. Substructure search** If you ask, "Is anyone named Lincoln in the database?" you usually expect to find the former President and a number of others - this is called a *search* rather than a *lookup*. For a chemical informatics system, we have a *substructure search*: Find all molecules containing a partial molecule (the "substructure") drawn by the user. The substructure is usually a functional group, "scaffold", or core structure representing a class of molecules. This too is a hard problem, *much* harder than most text searches, for reasons that go to the very root of mathematics and the theory of computability.
- 4. Similarity search** Some databases can find similar-sounding or misspelled words, such as "Find Lincon" or "find Cincinnati", which respectively might find Abraham Lincoln and Cincinnati. Many chemical information systems can find molecules similar to a given molecule, ranked by similarity.

## Query Languages in Cheminformatics : SMARTS

In addition to a typographical way to represent molecules, we also need a way to enter *queries* about molecules, such as, "Find all molecules that contain a phenol."

With text, we're familiar with the concept of typing a partial word, such as "ford" to find "Henry Ford" as well as "John Hartford". For chemistry, we can also specify partial structures, and find anything that contains them.

For example:

Query	Database	Matches?
		<b>YES</b> (matched portion highlighted in blue)
		<b>NO</b> (double bond indicated doesn't match)

The simplest query language for chemistry is SMILES itself: Just specify a structure, such as "Oc1cccc1", and search. This is how eMolecules' basic searching works. It's simple and, because of the high-performance indexes in eMolecules, it is very fast.

However, for general-purpose cheminformatics, one needs more power. What if the substructure you're looking for isn't a valid molecule? For example ClccBr (1,2- substitution on an aromatic ring) isn't a whole molecule, since the concept of aromaticity is only sensible in the context of a whole ring system.

Or what if the thing we're looking for isn't a simple atom such as Br, but rather a concept like "Halogen"? Or, "A terminal methyl"?

To address this, cheminformatics systems have special *query languages*, such as SMARTS (SMiles ARbitrary Target Specification). SMARTS is a close cousin to SMILES, but it has *expressions* instead of simple atoms and bonds. For example, [C,N] will find an atom that is either carbon or nitrogen.

### IUPAC Names, Trade Names, Common Names

Chemistry also has three other important name systems:

- **IUPAC Names** (from IUPAC, the International Union of Pure and Applied Chemistry) established a naming convention that is widely used throughout chemistry. Any chemical can be named, and all IUPAC names are unambiguous. This textual representation is aimed at humans, not computers: Chemists versed in IUPAC



When a query is entered, the cheminformatics system breaks apart the query using the same technique, to find all of the fragments in the query. It then checks its index for each fragment, and combines the lists it finds to get only those molecules that have *all* of those fragments.

This doesn't mean that all molecules returned by the index actually are matches. In the language of databases, we say the index will return *false positives*, candidate molecules that don't actually match the substructure search.

Consider our example of searching for "John Hartford" - the index might return many pages that have both "John" and "Hartford", yet have nothing to do with bluegrass music or steamboats. For example, it might return a page containing, "President John F. Kennedy visited Hartford, Connecticut today...". To confirm that the search system has found something relevant, it must check the pages return from the index to ensure that the specific phrase "John Hartford" is present. However, notice that this is *much* faster than searching every page, since the overwhelming majority of web pages were instantly rejected because they have neither "John" nor "Hartford" on them.

Similarly, a chemical fragment index serves to find only the most *likely* molecules for our substructure match - anything that the index didn't find is definitely not a match. But we still have to examine each of the molecules returned by the indexing system and verify that the complete substructure for which we are searching is present.

### **NP-Complete - A Little about Computability**

Searching through a page of text for the words, "John Hartford" is pretty easy for a modern computer. Although false positives returned by the index are a nuisance and impair performance, they are not a catastrophe. Not so for substructure matching. Unfortunately, substructure matching falls into a category of "hard" mathematical problems, which means false positives from the index are a big problem.

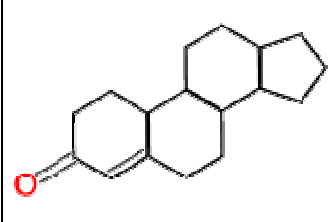
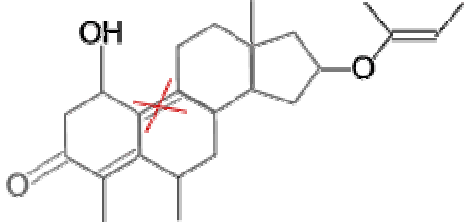
Substructure matching (finding a certain functional group within a molecule) is an example of what mathematicians call *graph isomorphism*, and is in a class of problems called *NP Complete*. Roughly speaking, this means the time it takes to do a substructure search is non-polynomial, i.e. exponential in the number of atoms and bonds. To see why this is a computational disaster, compare two tasks, one that takes polynomial time,  $k_1 * N^2$ , versus one that takes exponential time  $k_2 * 2^N$ . Our polynomial task is bad enough: If we double N, it takes *four times* as long to solve. But the exponential task is worse: *Every time we add an atom it doubles*. So going from one atom to two doubles the time, and going from 100 atoms to 101 atoms doubles the time. Even if we can get  $k_2$  down to a millionth of  $k_1$ , we're still in trouble - a million is just  $2^{20}$  or twenty atoms away.

It has been mathematically proven that substructure searching is in the set of NP Complete problems, so there's no point wasting our time searching for a polynomial algorithm. The good news is that most molecules have "low connectivity", meaning most atoms have fewer than four bonds, unlike the weird and twisted graphs that mathematicians consider. In practice, most substructure matching can be done in polynomial time around  $N^2$  or  $N^3$ . But even with this improvement, substructure matching is an "expensive" time-consuming task for a computer.

The key point is that indexing is particularly important for cheminformatics systems. The typical modern computer can only examine a few thousand molecules per second, so examining millions of molecules one-by-one is out of the question. The indexing done by a modern cheminformatics system is the key to its performance.

### **Molecular Similarity**

Substructure searching is a very powerful technique, but sometimes it misses answers for seemingly trivial differences. We saw this earlier with the following:

Query	Target
	
We're looking for steroids...	But we don't find this one because of the double bond.

It is somewhat like searching for "221b Baker Street" and finding nothing because the database contains "221B Baker Street" and the system doesn't consider "b" and "B" a match.

A good similarity search would find the target structure shown above, because even though it is not a substructure match, it is highly similar to our query.

There are many ways to measure similarity.

**2D topology** The best-known and most widely used similarity metrics compare the two-dimensional topology, that is, they only use the molecule's atoms and bonds without considering its shape.

Tanimoto similarity is perhaps the best known as it is easy to implement and fast to compute. An excellent summary of 2D similarity metrics can be found in section 5.3 of the Daylight Theory Manual.

**3D configuration** One of the most important uses of similarity is in the discovery of new drugs, and a molecule's shape is critical to its medicinal value (see QSAR).

3D similarity searches compare the configuration (also called the "conformation") of a molecule to other molecules. The "electronic surface" of the molecule is the important bit - the part that can interact with other molecules. 3D searches compare the surfaces of two molecules, and how polarized or polarizable each bit of the surface is.

3D similarity searches are uncommon, for two reasons: It's difficult and it's slow. The difficulty comes from the complexity of molecular interactions - a molecule is not a fixed shape, but rather a dynamic object that changes according to its environment. And the slowness comes from the difficulty: To get better results, scientists employ

more and more complex programs.

**Physical Properties** The above 2D and 3D similarity are based on the molecule's structure. Another technique compares the properties - either computed or measured or both - and declares that molecules with many properties in common are likely to have similar structure. It is the idea of QSAR taken to the database.

**Clustering** "Clustering" is the process of differentiating a set of things into groups where each group has common features. Molecules can be clustered using a variety of techniques, such as common 2D and/or 3D features.

Note that clustering is not a similarity metric *per se* (the topic of this section), but it may use various similarity metrics when computing clusters. It is included here because it can be used as a "cheap substitute". That is, when someone wants to find compounds similar to a known compound, you can show them the group (the cluster) to which the compound belongs. It allows you to pre-compute the clusters, spending lots of computational time up front, and then give answers very quickly.

Many cheminformatics databases have one or more similarity searches available.

## The Chemical Registration Systems

Chemical Registration is the "big brother" of cheminformatics.

A cheminformatics system is primarily devoted to recording chemical structure. Chemical Registration systems are additionally concerned with:

- Structural novelty - ensure that each compound is only registered once
- Structural normalization - ensure that structures with alternative representations (such as nitro groups, ferrocenes, and tautomers) are entered in a uniform way.
- Structure drawing - ensure that compounds are drawn in a uniform fashion, so that they can be quickly recognized "by eye".
- Maintaining relationships among related compounds. For example, all salt forms of a compound should be recognized as being related to one another, and compounds in different solvates are also related.
- Registering mixtures, formulations and alternative structures.
- Registering compounds the structure of which is unknown.
- Roles, responsibilities, security, and company workflow.
- Updates, amendments and corrections, and controlling propagation of changes (e.g. does changing a compound change a mixture containing that compound?)

The scope of Chemical Registration Systems is far beyond the goals of this brief introduction to cheminformatics. However, to illustrate just one of the points above, let's consider structural novelty. In real life, chemical structure can be very ambiguous. Imagine you have five bottles of a particular compound that has a stereo center:

1. The contents of the first bottle were carefully analyzed, and found to be a single stereoisomer.
2. The contents of the second bottle were carefully analyzed and found to contain a racemic mixture of the stereoisomers.
3. The stereoisomers of the third bottle are unknown. It may be pure, or have one predominant form, or be a racemic mixture.
4. The fourth bottle was obtained by running part of the contents of bottle #2 through a chromatographic separation. It is isotopically pure, but you don't know which stereoisomer.
5. The fifth bottle is the other fraction from the same separation of #4. It is also isotopically pure, but you don't know which stereoisomer, *but you know it's the opposite of #4.*

Which of these five bottles contain the same compound, and which are different? That is the essential task of a chemical registry system, which would consider all five to be different. After all, you probably have data about each bottle (that's why you have them), and you must be able to record it and not confuse it with the other bottles.

In this example above, consider what is known and not known:

Bottle	Known	Not Known
1	Everything	Nothing
2	Everything	Nothing
3	Compound is known	Stereochemistry
4	Compound and purity known, stereochemistry is opposite of #5	Specific stereochemistry
5	Compound and purity known, stereochemistry is opposite of #4	Specific stereochemistry

A cheminformatics system has no way to record the contents of the five bottles; it is only concerned with structure. By contrast, a chemical registration system can record both *what is known* as well as *what is not known*. This is the critical difference between the two.

## CERTAIN APPLICATIONS OF CHEMINFORMATICS

### Storage and retrieval in [Chemical databases](#)

The primary application of cheminformatics is in the storage of information relating to compounds. The efficient search of such stored information includes topics that are dealt with in computer science as [data mining](#) and [machine learning](#). Related research topics include:

- [Unstructured data](#)
- [Structured Data Mining](#) and mining of [Structured data](#)
  - Database mining
  - Graph mining
  - [Molecule mining](#)
  - [Sequence mining](#)
  - Tree mining

### [Chemical file format](#)

The *in silico* representation of chemical structures uses specialized formats such as the [XML](#)-based [Chemical Markup Language](#), or [SMILES](#). These representations are often used for storage in large [chemical databases](#). While some formats are suited for visual representations in 2 or 3 dimensions, others are more suited for studying physical interactions, modeling and docking studies.

### **Virtual screening**

In contrast to [high-throughput screening](#), virtual screening involves the creation of large *in silico* virtual libraries of compounds, which are then submitted to a [docking](#) program in order to identify the most active members. In some cases, [combinatorial chemistry](#) is used in the development of the library to increase the efficiency in mining the chemical space. More commonly, a diverse library of small molecules or [natural products](#) is screened.

### **Quantitative structure-activity relationship (QSAR)**

This is the calculation of [quantitative structure-activity relationship](#) and quantitative structure property relationship values, used to predict the activity of compounds from their structures. In this context there is also a strong relationship to [Chemometrics](#). Chemical [expert systems](#) are also relevant, since they represent parts of chemical knowledge as an *in silico* representation.

\*\*\*\*\*

## **CREATION AND MANAGEMENT OF RELATIONAL DATABASES**

*Sri Arunava Gupta*

Research Associate

BIF, CVSc, AAU, Khanapara, Guwahati-22

### **Introduction**

Today's computing world is driven by databases. Whether it be a transaction processing system of a bank or commercial institution or a simple guest book on a personal website, a search-engine serving links to web-pages over the internet (e.g. Google) or online information repositories (e.g. Wikipedia, ScienceDirect) - ranging from 'passive' text to sound and video (e.g. YouTube) - databases have come to stay. The field of Bioinformatics (computational biology) too has not remained untouched by the database phenomenon.

Biological databases (Knowledgebases) are libraries of life-sciences information, collected from scientific experiments, published literature, high throughput experiment technology, and computational analyses. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetics.

Databanks like GenBank (National Center for Biotechnology Information) store a colossal amount of information pertaining to the sequences of organisms, in many special machines called database servers.

Biological database design, development, and long-term management is, in fact, a core area of the discipline of Bioinformatics.

Since *Relational Database* concepts of computer science are important for understanding the creation and management of biological databases, it is imperative for researchers, scientists and others in the biological field to have an insight into these concepts.

### 1. Database - What is it?

A Database is simply a collection of related data. By data, we mean known facts that can be recorded and that have some implicit meaning. Again, this 'collection' may be of a varying size and complexity.

For example, an organization might store information about their organization, their employees and financial information etc. on computers using special software like

MS-ACCESS. This is a collection of related data and hence is a database.

A Database generally has:

- a source from which the data are derived.
- a degree of interaction with the events in the real world and
- an audience that is actively interested in the contents of the database, e.g. in the example above, the company management might be interested in knowing the details of the company employees.

The storage of the data is a key question in the making of a database. Merely recording facts and figures on computer does not, strictly speaking, make a database. For instance, storing certain statistics in a MS-WORD table does not make a database. In this light, a database is essentially a way of structuring and organizing the data, taking into account the relationships within it. To accomplish this task, special database software known as DBMS are used.

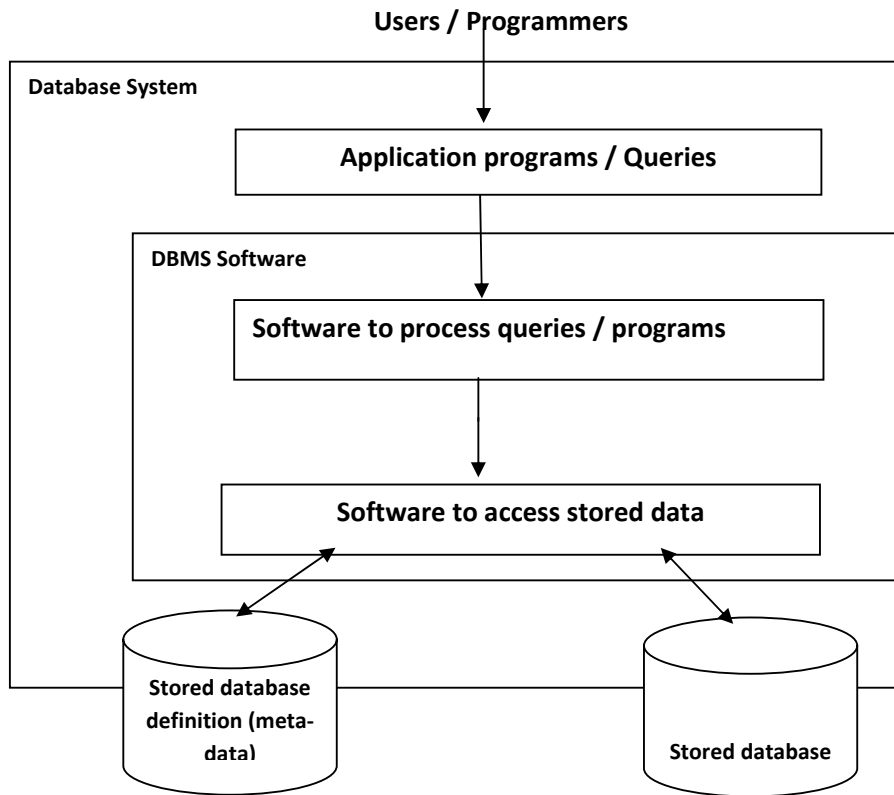
### 2. Database Management System (DBMS)

A DBMS is a collection of programs that enables the creation and management of *computerized* databases. It facilitates the following processes:

- **Database definition** – involves specifying the data types, structures and constraints for the data i.e. what values they are allowed to take
- **Database construction** – the process of storing the data itself on some storage medium that is controlled by the DBMS.
- **Database manipulation** – includes such functions as updating and querying the database.

### 3. The Database Environment

The database and the DBMS software together are referred to as a Database System. The Database Environment consists of the Database System *plus* the end users and/or programmers. The operation of a simplified database environment is depicted in the following self-explanatory diagram:



The database administrator writes and enforces the procedures and standards that are then used by designers, analysts, programmers, and end users. The end users use the application programs such as software developed in any platform- MS Access, Visual Basic, Java et al - by analysts and programmers to retrieve the data stored in the database. In turn, the application programs make use of the DBMS, which manages the data.

The interface is the gateway to the database, which resides within the hardware. Finally, the System Administrator manages the entire system. The main purpose of the creation of a database environment is to help an organization to perform its mission and to achieve its goals.

#### 4. Why are Databases so popular?

The main reason is elimination of redundancy.

In traditional file processing, each user defines and implements the files needed for a specific application. To take the example of a student information system, one user, say, the academic cell, may keep a file (e.g. in MS-WORD format) on students and their grades. A second user, the accounting office, may keep track of students' fees and their payments (e.g. in MS-EXCEL format). Although both users are interested in data about students, each user maintains separate files. This redundancy in defining and storing data results in wasted storage space and in redundant efforts to maintain common data up-to-date.

A major characteristic of the database approach is that it maintains a single repository of data that is defined once and then is accessed by various users. The other advantages of the database approach are:

- **Program-Data insulation** – the data and the programs that use the data are kept separate.
- **Homogeneity** in the way the data are structured.
- Support of **multiple views** of the data.
- **Sharing of data** and **multi-user** transaction processing.

All these advantages are not available in either manual or traditional (computerized) file-processing systems.

#### 5. Data Models

Irrespective of the way in which the data is actually stored in the database, there must be some way to visualize or describe the structure of a database. The concept of data model enables us to achieve this abstraction. There are several data models like the hierarchical data model and the network data model, but the most famous data model that enjoys universal popularity in the database world is the *relational* data model.

**6. The Relational Model:** The 'relation' in 'relational model' comes from the mathematical notion of relations from the field of set theory. The basic data structure of the relational model is a '**relation**'. This is a *table* where information about a particular entity (say, 'employee') is represented in columns and rows. The columns enumerate the various attributes of an entity (e.g. employee\_name, designation, salary, etc). Rows (also called *records*) represent instances of an entity (e.g. specific employees). Each record again is a collection of related *fields* or attribute-values (e.g. "John Smith", "Manger", \$1,500.00, etc). A field is the smallest (atomic) unit of information in database theory.

A database which is designed in accordance with the principles of the relational model is referred to as a **relational database**. The corresponding DBMS is known as Relational Database Management System (RDBMS).

## 7. Relational Database Management System

A RDBMS implements the features of the relational model outlined above. More precisely, a **Relational Database Management System (RDBMS)** is a program that lets us create, update, and administer a relational database.

### RDBMS Examples:

- **Microsoft Access**
- **MySQL**
- **PostgreSql**
- **SQL Server**
- **Oracle**

These RDBMSs are designed to help in organizing large amounts of information in a way where the data can be easily searched, sorted, and updated by the end user. In addition, they have features like indexing, joins, aggregates, and an overall table structure that makes the job of the application programmer easier.

Certain **Relational Operations** of the RDBMS enable the users (or programs) to interact with the data contained in a relational database through *queries* written in a special language, usually **SQL (Structured Query Language)**.

## 8. How an RDBMS organizes data

RDBMS data is invariably structured in database tables, fields and records. Each RDBMS table consists of database table rows. Each database table row consists of one or more database table fields. Again, the different database tables are related by common fields (database table columns).

EmployeeID	EmployeeName	Department	Grade	DOJ	Salary
2213	Harish Khare	Sales	A	10/22/1979	\$5,000.00
2214	Ashoka de Silva	Marketing	B	1/25/1973	\$1,900.00
2215	Rohit Mathur	R&D	C	2/28/1983	\$500.00
2216	Vijayendra Rao	Management	A	10/10/1970	\$500,000.00
3310	Gurmeet Singh	Sales	B	8/19/2000	\$1,000.00
4321	Prithviraj Chauhan	Sales	B	9/25/2008	\$1,750.00

Fig: A Table in MS-ACCESS

## 9. What is the difference between DBMS and RDBMS?

- RDBMS is a Relational Data Base Management System - Relational DBMS. This adds the additional condition that the system supports a tabular structure for the data, with enforced relationships between the tables. This excludes the databases that don't support a tabular structure or don't enforce relationships between tables. The main advantage of an RDBMS is that it checks for referential integrity (relationship between related records using Foreign Keys). One can set the constraints in an RDBMS such that when a particular record is changed, related records are updated/deleted automatically.
- DBMS are for smaller organizations with small amount of data, where security of the data is not of major concern and RDBMS are designed to take care of large amounts of data and also the security of this data.

## 10. Microsoft ACCESS:

Microsoft ACCESS™ is a development environment used to create computer databases for the Microsoft Windows family of operating systems. This is a Relational Database Management System which combines the relational Microsoft Jet Database engine with a Graphical user Interface (GUI) and software development tools. The main reason behind the popularity of ACCESS is undoubtedly its user-friendly GUI.

## 11. Frequently Encountered Terms in ACCESS:

### a. Data

Raw facts from which the required information is derived. Data have little meaning unless they are grouped in a logical manner.

### b. Field

A character or a group of characters (numeric or alphanumeric) that describes a specific characteristic. A field may define a telephone number, a date, or other specific characteristics that the end user wants to keep track of.

### c. Record

A logically connected set of one or more fields that describes a person, place, event, or thing. For example, an EMPLOYEE record may be composed of the fields EMPLOYEE\_ID, EMPLOYEE\_NAME, DATE\_OF\_JOIN, BASIC\_SALRY, etc.

A '**primary key**' is a key that has a unique value for each record in the table, e.g. EMPLOYEE\_ID.

### d. File

All the tables, forms, reports queries are stored in a single computer file properly tagged and kept in a **.mdb** (MS-ACCESS) file

### e. Table

A table is a container that holds information about like items. For example, an **Employee** table would contain the same basic details on each employee: name, title, department and so on. Each detail, each chunk of information you need to store lives in a field.

#### f. Forms

Forms are sometimes referred to as 'data entry screens.' They are the interfaces you use to work with your data, and they often contain command buttons that perform various commands. You can create a database without using forms by simply editing your data in the table datasheets. However, most database users prefer to use forms for viewing, entering and editing data in tables.

#### g. Reports

Reports are what you use to summarize and present data in the tables. A report usually answers a specific question, such as "How much money did we receive from each customer this year?" or "What cities are our customers located in?" Each report can be formatted to present the information in the most readable way possible and according to requirement of the organization.

#### h. Queries

Queries are the real workhorses in a database, and can perform many different functions. Their most common function is to retrieve specific data from the tables. The data you want to see is usually spread across several tables, and queries allow you to view it in a single datasheet. Also, since you usually don't want to see all the records at once, queries let you add criteria to "filter" the data down to just the records you want.

#### i. Macros

Macros in Access can be thought of as a simplified programming language which you can use to add functionality to your database. For example, you can attach a macro to a command button on a form so that the macro runs whenever the button is clicked. Thus, they can be great time-saving devices.

#### j. Modules

A module is a collection of declarations, statements, and procedures that are stored together as a unit. A module can be either a class module or a standard module. Class modules are attached to forms or reports, and usually contain procedures that are specific to the form or report they're attached to. Standard modules contain general procedures that aren't associated with any other object.

### 12. MS-ACCESS – A Typical Session

To launch ACCESS (2003), *Start->Programs->Microsoft Office-> Microsoft Access 2003*

Below is an example of a table in a relational database for a particular company:

#### Employee Table

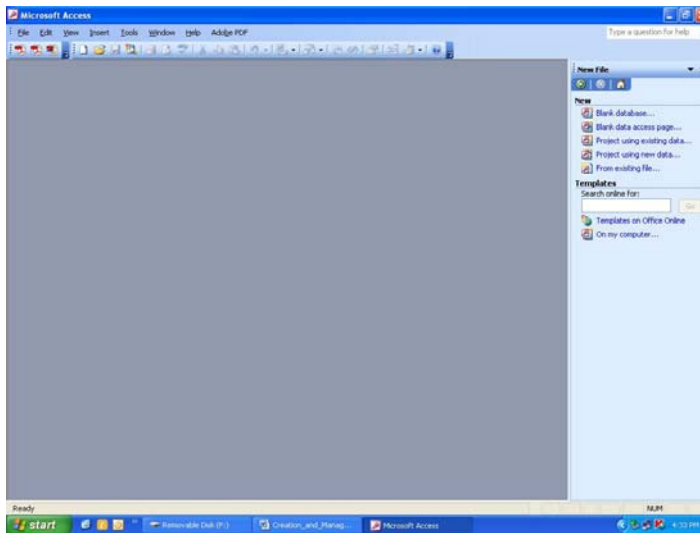
<b>EmployeeID</b>	<b>EmployeeName</b>	<b>Department</b>	<b>Grade</b>	<b>DOJ</b>	<b>Salary</b>
<i>Number</i>	<i>Number</i>	<i>Character</i>	<i>Character</i>	<i>Date</i>	<i>Currency</i>

2213	Harish Khare	Sales	A	10/22/1979	\$5,000.00
2214	Ashoka de Silva	Marketing	B	1/25/1973	\$1,900.00
2215	Rohit Mathur	R&D	C	02/28/1983	\$500.00
2216	Vijayendra Rao	Management	A	10/10/1970	\$5,00,000.00
3310	Gurmeet Singh	Sales	B	08/19/2000	\$1,000.00
4321	Prithviraj Chauhan	Sales	B	09/25/2008	\$1,750.00

The **Employee** table has 6 columns (*EmployeeID*, *EmployeeName*, *Department*, *Grade*, *DOJ* (Date of Joining in mm-dd-yyyy format) and *Salary*) and 6 rows (or records) of data. Each of the columns conforms to the **data types** ‘Number’, ‘Character’, ‘Date’ and ‘Currency’ (there are other data-types too). The data type for a column indicates the type of data values that may be stored in that column:

- Number - may only store numbers, possibly with a decimal point.
- Character - may store numbers, letters and punctuation. Access calls this data type **Text**.
- Date - may only store date and time data.
- Currency – ideal for storing salary and similar information.

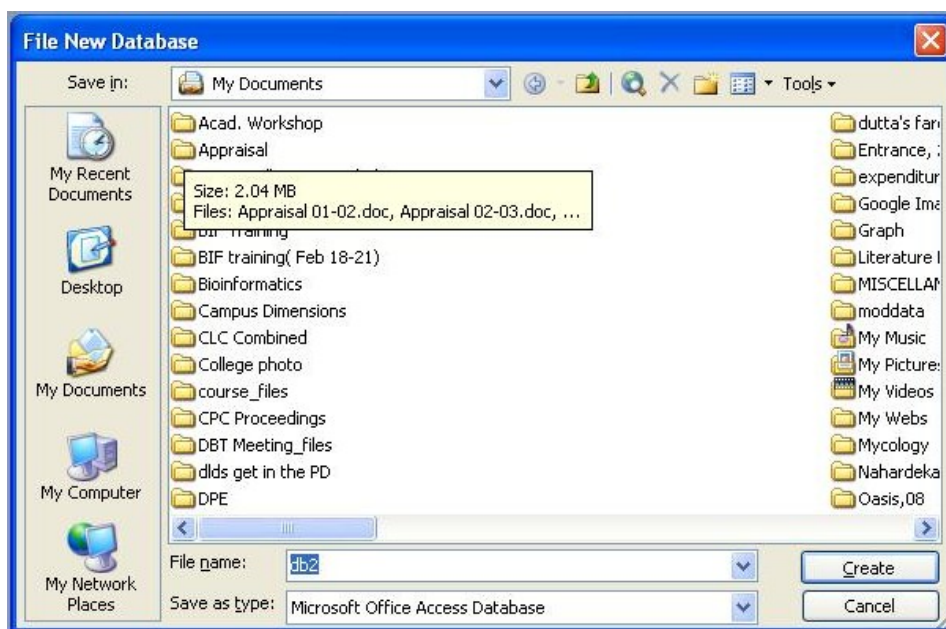
Once Access is running, an initial screen will be displayed as follows:



From this initial screen, the user can create a new database (either blank or with some tables created with the database wizard), or open an existing database.

In general, the first time one begins a project, a new, blank database should be created. In subsequent sessions, use the *Open existing database* option to re-open the database created previously.

By selecting **Blank Database**, the 'File New Database' dialog will appear' and prompt for a file-name. Assign any name you want and then click on the Create button to create the database like so:



In the above example, db2 is the name chosen for this particular database and the 'Save as type' is Microsoft Office Access Database. (By default, **.mdb** is the three letter extension given for Microsoft Database files.)

It is advisable to keep the name of the database (db2 in the above example) relatively short. The tabs in the database window, which comes up next, include:

- Tables - Displays any tables in the database.
- Queries - Displays any queries saved in the database.
- Forms - Displays any forms saved in the database.
- Reports - Displays any reports saved in the database.
- Macros - Displays any macros (short programs) stored in the database.
- Modules - Displays any modules (Visual Basic for Applications procedures) stored in the database.

### Creating and Viewing Tables

Tables are the main units of data storage in Access. A table is made up of one or more *columns* (or *fields*) and a given column may appear in more than one table in order to indicate a relationship between the tables. We will give the step-by-step instructions for creating a sample table in Microsoft Access.

There are a number of ways to create a table in Access. Access provides *wizards* that guide the user through creating a table by suggesting names for tables and columns. The other main way to create a table is by using the *Design View* to manually define the columns (fields) and their data types. While using the wizards is a fast way to create tables, the user has less control over the column names (fields) and data types. In this tutorial, we will describe the steps to create a table using the *Design View*.

### Creating a Table Using the Design View

To create a table in Access using the Design View, make sure the Table tab is displayed (that is, Access should be set to work with tables rather than with queries, forms, reports, etc.) and perform the following steps:

1. Double-click on Create table in Design view



2. A new spreadsheet-like window, the table 'Design View' will now appear. Fill in the **Field Name**, **Data Type** and **Description** for each column/field in the table. A few fields are shown filled below:

Field Name	Data Type	Description
EmployeeID	Number	unique identifier for an employee; no 2 employees will have same ID
EmployeeName	Text	Name of employee
Department	Text	Name of department


Fill in the information for the fields as follows:

Field Name	Data Type	Description
EmployeeID	Number	unique Identifier for an employee; no 2 employees will have same ID
EmployeeName	Text	Name of employee
Department	Text	Name of department
Grade	Text	Designation Grade
DOJ	Date/Time	Date of Joining
Salary	Currency	'take-home pay' of the employee

A figure showing the design view with the complete table definition is given below:

Field Name	Data Type	Description
EmployeeID	Number	unique identifier for an employee; no 2 employees will have same ID
EmployeeName	Text	Name of employee
Department	Text	Name of department
Grade	Text	Designation Grade
DOJ	Date/Time	Date of Joining
Salary	Currency	'take-home pay' of the employee

- Now that all of the fields have been defined for the table, a Primary Key should be defined.

Position the cursor on the EmployeeID field and click on the Primary Key  icon on the Table Design toolbar.

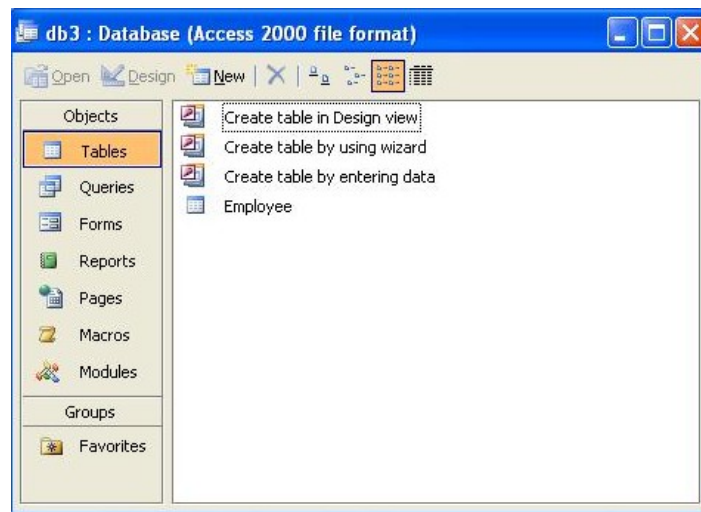
**Notice that a small key appears next to the field name on the left side.**

- As a final step, the table must be saved. Pull down the File menu and choose the Save menu item. A dialog box will appear where the name of the new table should be specified. Note that Access gives a default name such as **Table1** or **Table2**. Simply type over this default name with the name of the table.

For this example, name the table: **Employee**. Then click on the OK button.



At this point, the new Customer table has been created and saved. Switch back to the Access main screen by pulling down the File menu and choosing the Close menu item. This will *close* the Design View for the table and display the Access main screen. Notice that the new **Employee** table appears alongside the 'Tables' tab:



### Viewing and Adding Data to a Table

Data can be added, deleted or modified in tables using a simple spreadsheet-like display. To bring up this view of a single table's data, highlight the name of the table and then click on the Open button.

In this view of the table, shown in the figure below, the fields (columns) appear across the top of the window and the rows or records appear below. This view is similar to how a spreadsheet would be designed:

EmployeeID	EmployeeName	Department	Grade	DOJ	Salary
0					\$0.00

Note at the bottom of the window the number of records is displayed. In this case, since the table was just created, only one blank record appears. To add data to the table, simply type in values for each of the fields (columns). Press the Tab key to move between fields within a record. Use the up and down arrow keys to move between records. Enter the data as given below:

EmployeeID	EmployeeName	Department	Grade	DOJ	Salary
<i>Number</i>	<i>Number</i>	<i>Character</i>	<i>Character</i>	<i>Date</i>	<i>Currency</i>
2213	Harish Khare	Sales	A	10/22/1979	\$5,000.00
2214	Ashoka de Silva	Marketing	B	1/25/1973	\$1,900.00
2215	Rohit Mathur	R&D	C	02/28/1983	\$500.00
2216	Vijayendra Rao	Management	A	10/10/1970	\$5,00,000.00
3310	Gurmeet Singh	Sales	B	08/19/2000	\$1,000.00
4321	Prithviraj Chauhan	Sales	B	09/25/2008	\$1,750.00

EmployeeID	EmployeeName	Department	Grade	DOJ	Salary
2213	Harish Khare	Sales	A	10/22/1979	\$5,000.00
2214	Ashoka de Silva	Marketing	B	1/25/1973	\$1,900.00
2215	Rohit Mathur	R&D	C	2/28/1983	\$500.00
2216	Vijayendra Rao	Management	A	10/10/1970	\$500,000.00
3310	Gurmeet Singh	Sales	B	8/19/2000	\$1,000.00
4321	Prithviraj Chauhan	Sales	B	9/25/2008	\$1,750.00
0					\$0.00

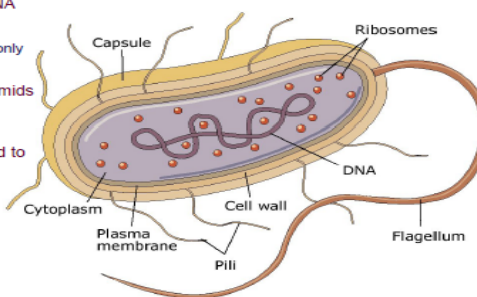
To save the new data, pull down the File menu and choose Save. To navigate to other records in the table, use the Navigation bar at the bottom of the screen:

Record:  2 of 4

To modify (i.e. edit, delete or replace) existing data, simply navigate to the record of interest and tab to the appropriate field. Then make the required modification. You can also use the arrow keys to navigate and the delete or backspace keys to change the existing data.

----OOO----

- the genome of a *prokaryote* comes as a single double-stranded DNA molecule in ring-form
  - in average 2mm long
  - whereas the cells diameter is only 0.001mm
  - < 5 Mb
- *prokaryotic* cells can have plasmids as well (see next slide)
- protein coding regions have no *introns*
- little non-coding DNA compared to eukaryotes
  - in *E.coli* only 11%



## Hands on Training

### Basic Bioinformatics practical skills

*Chittaranjan Baruah*

Bioinformatics Centre (DBT-BIF)

Department of Zoology (UGC-SAP & DST-FIST sponsored Department),

Gauhati University, Guwahati – 781 014, Assam, India

E-mail: Chittaranjan\_2004@india.com

1. Download Cn3D and Rasmol and install in your computer. 2
2. Download the nucleotide sequence of cytochrome b (cytb) gene of Gangetic dolphin (*Platanista gangetica*) and save the sequence in GenBank, FASTA, ASN.1 and XML Format. Perform a BLAT search and identify two most similar sequences 2+8+5= 15
3. Download the sequence and structure of Myoglobin from SWISSPROT and RCSB-PDB respectively. Visualize the structure with Rasmol and determine the domain property. Display the Ramachandran plot for glycine using SWISS PDB Viewer. 5+3+2= 10
4. Download the structure of Nucleosome from NCBI-MMDB. Download and install Cn3D from NCBI and visualize the structures. 2+3 = 5
5. Following is the amino acid sequence of a protein (fragment): “ **KGDIMVFPR** ”
  - a. Calculate the total number of atoms.
  - b. Calculate hydrophobicity.
  - c. Determine molecular weight.
  - d. Determine the Theoretical Isoelectric point (pI)
 (Hints: Use ExPasy tools like ProtParam, ProtScal) 4x 2 = 8
6. Download a compound of biomedical importance from PubChem and Draw out the structure using ISIS Draw. 2+3 = 5
7. Find out 2 research papers of “Biotechnology” from NCBI **PubMed** database. (1marks).
8. Scientists have isolated a virus and sequenced. The sequence is found as given below. As a Bioinformatics professional the sequence is send to you. Identify the virus. (Hints : Use similarity and Homology searching tools like BLAST and FASTA). 5

ATCATGCTCCTCAAACGTGTCGGTCGCGCACGATGCATCTGGCAAGCGGGTGTATTA  
CCTCACCCGCGACCCACCACCCCTTGCACGAGCTGCATGGGAGACAGCCAGACA  
CACTCCAGTCAACTCCTGGCTAGGCAACATCATCATGTATGCGCCCACCTTGTGGGCG  
AGGATGATCCTGATGACCCACTTCTTCTCCATCCTCCTAGCCCAGGAGCAACTTGAAA  
AAACC

## Molecular Biology Freeware for Windows

*Chittaranjan Baruah*

*Bioinformatics Centre (DBT-BIF)*

*Department of Zoology (UGC-SAP & DST-FIST sponsored Department),*

*Gauhati University, Guwahati – 781 014, Assam, India*

*E-mail: chittaranjan\_2004@india.com*

### 🔴 *DNA, RNA and genomic analysis:*

● [DNA Club](#) - DNA analysis software, features include remove vector sequence, find, find ORF, sequence editing, translate to protein sequence, protein sequence editing, RE Map, RE Map with translation, PCR primer selection, primer or probe evaluation etc.

● [DNA for Windows](#) is a compact, easy to use DNA analysis program, ideal for small-scale sequencing projects.

● [Gene Designer](#) - a brilliant software tools that allows one to combine building blocks such as regulatory DNA elements (promoters, ribosome-binding sites) with amino acid sequences, affinity & protease cleave tags and cloning features and codon optimize for any expression host.

[CLC Free Workbench](#) - allows basic sequence analysis such as open reading frame determination, restriction site analysis, translation from DNA/ RNA to proteins, alignments, and tree reconstruction in a single window format.

● - [Geneious](#) (Alexei Drummond Biomatters Ltd. Auckland, New Zealand) provides an automatically-updating library of genomic and genetic data; for organizing and visualizing data. It provides a fully integrated, visually-advanced toolset for: [sequence alignment](#) and phylogenetics; sequence analysis including BLAST; protein structure viewing, NCBI, EMBL, Pubmed auto-find & much more ...

● [Genome2D](#) - is for the visualization of transcriptome and other customized data sets (visualizes a bacterial genome with all relevant information such as gene orientation, operon structure, transcriptional terminators or regulator binding sites) on linear chromosome maps constructed from annotated bacterial genome sequences. For a complete list of what this incredible program will do see [here](#). Images and data tables from Genome2D can easily be exported for further use in other presentation programs. (Reference: Baerends, R.J.S., et al. (2004) [Genome Biology, 5: R37](#)).

● EMBOSS (European Molecular Biology Open Source Software Suite) can be downloaded from [here](#).

● [SEQtools](#) is a program package for routine handling and analysis of DNA and protein sequences. The package includes general facilities for sequence and contig editing, restriction enzyme mapping, translation, and repeat identification.

● [RNA draw](#): is an integrated program for RNA secondary structure calculation and analysis by Ole Matzura & Anders Wennborg (1996) Computer Applications in the Biosciences (CABIOS) **12**: 247-249

● [RNAstructure](#) - RNA Secondary Structure Prediction and Analysis for Microsoft Windows. This program includes a secondary structure prediction algorithm, a sequence editor, an integrated drawing tool, the OligoWalk program, OligoScreen, Dynalign, and a partition function calculator. (Reference: D.H. Mathews (2005) Bioinformatics **21**: 2246 - 2253.)

● [loopDloop](#) is a tool for drawing RNA secondary structures in molecular biology.

● [FinchTV](#) - Another useful tool for viewing and editing electropherograms.

● [WinGene](#) - Used to give the reverse complement a given sequence, translate the sequence in all reading frames and identify ORFs (Export polypeptides encoded by a given ORF into WinPep), perform an oligomer analysis (base composition and melting temperature), and perform a restriction analysis.

● [GeneDoc](#) is a full featured multiple sequence alignment editor, analyzer and shading utility for Windows.

● [Ridom TraceEdit](#) - displays chromatogram files from Applied Biosystems automated sequencers and files in the Staden SCF format. Incorrect base calls can be edited and saved. (Reference: J. Rothgänger et al. 2006. Bioinformatics **22**: 493-494).

● [DNA Master](#) - is "perhaps the world's greatest sequence editor" and analysis package. Find under "computer."

● [GeSTer](#) (V. Nagaraja, Indian Institute of Science, Bangalore, India) - is extremely useful in locating stem-loop structures, including rho-independent terminators in annotated genomes. Since it does not run conveniently on Windows XP see how you can [modify](#) the \*.gbk file so that it works.

● [Staden Package](#) - consists of a series of tools for DNA sequence preparation (pregap4), assembly (gap4), editing (gap4) and DNA/protein sequence analysis (spin). The package was originally developed at the MRC-LMB in Cambridge. It is now open source (BSD licence) and is hosted on sourceforge.net.

### ● *Plasmid graphics and drawing:*

● [pDRAW32](#) DNA analysis software by AcaClone software (Kjeld Olesen). pDRAW lets you enter a DNA name and coordinates for genetic elements, such as genes, to be plotted on your DNA plots.

● [Plasmid Processor](#) (Department of Biochemistry and Biotechnology, University of Kuopio, Finland) - is a simple tool for plasmid presentation for scientific and educational purposes. It features both circular and linear DNA, user defined restriction sites, genes and multiple cloning site. In addition you can manipulate plasmid by inserting and deleting fragments. Created drawings can be copied to clipboard or saved to disk for later use. Printing from within the program is also supported.

### ● *Primer design:*

● [Picky](#) is an oligo microarray design program that identifies probes that are very unique and specific to input sequences. These calculations are based on parameters inputted by the user including optimal probe length, ideal percentage of guanine and cytosine content, target-melting temperature, salt concentration and the maximum length to which a target sequence matches any non-target sequence. (Reference: H.-H. Chou et al. (2004) Bioinformatics **20**: 2893-2902). Download genome \*.ffn files from [GenBank](#) for use with this program. N.B. Unfortunately these files do not include the gene names only their coordinates.

● [GenomePrimer](#) is a program that provides a high-throughput method to select, with minimal user intervention and maximum flexibility, the most unique regions within DNA sequences and design primers that meet certain preset criteria. (Reference: S.A.F.T. van Hijum et al. (2003) *Bioinformatics* **19**: 1580-1582).

● [AutoDimer](#) - AutoDimer was packaged for installation using Visual Basic 6.0 and was developed to rapidly screen previously selected PCR primers for primer-dimer and hairpin interactions. It was originally created to assist in the development of multiplex PCR assays for probing STR and SNP markers for forensic purposes.

● [Fast PCR](#) - is based on the new approach in design PCR primers, alignment and repeat sequence searching. Program can work with several sequences simultaneously and the size of files up to 2Gb, multiplex PCR primers design and "in silico" PCR is supported. The program is convenient for search homology in a personal database is similar to BLAST and also others bioinformatics tools are included. Powerful and fast search of all types of repeats in DNA based on the new theory of search of repeats is developed. Now the program supports clustering sequences (R. Kalendar & A.H. Schulman, Institute of Biotechnology, Univ. Helsinki, Finland).

●- [Oligo Analyzer](#) is a simple tool to determine primer properties like T<sub>m</sub>, GC%, primer loops, primer dimers and primer-primer compatibility. All you have to do is to paste or type primer sequence and let Oligo Analyzer to calculate all important primer properties mentioned above. [Readme](#)

●- [Oligo Explorer](#) is a tool to search primers and primer pairs. The program analyzes all important primer properties like T<sub>m</sub>, GC%, primer loops, primer dimers and etc. [Readme](#)

●- [MeltCalc](#) is the ultimate thermodynamic modelling spreadsheet for Excel™ which allows you to analyze probes. See: Spreadsheet software for thermodynamic melting point prediction of oligonucleotide hybridization with and without mismatches (Reference: Schütz, E., von Ahsen, N. (1999) *BioTechniques* **27**:1218-1224).

● [FastPCR](#) (R. Kalendar, University of Helsinki, Finland) - is based on a new approach in the design of PCR primers for standard and long PCRs, inverse PCR, direct amino acid sequence degenerate PCR, multiplex PCR, *in silico* PCR, unique PCR primers design and group-specific PCR (common primers for multiple sequences), single primering PCR, automatically SSR loci detection and direct PCR primers design; for sequence alignments, clustering and any kind repeat sequence, MITE elements, LTR-retrotransposons, and SSR loci searching; restriction enzyme analysis.

### ● **Protein analysis:**

● [ANTHEPROT](#) (ANalyse THE PROTeins) is the result of biocomputing activity at the Institute of Biology and Chemistry of Proteins (Lyon, France)

● [WinPep](#) - WinPep ( is a versatile tool for the analysis of protein sequences (determination of amino acid composition, molecular weight, isoelectric point, & potential posttranslational modifications. It also will search for sequence motifs, display of amino acid sequences as helical wheels, hydrophathy plots, & domain structure of proteins). (Reference: L. Hennig (1999) *BioTechniques* **26**: 1170-1172 ).

### ● **Viewing three dimensional structures:**

● [Yasara](#) (Gregor Högenauer, Günther Koraimann, & Andreas Kungl [Univ. Graz, Austria]; & Gert Vriend [Univ. Nijmegen, the Netherlands]) is an awesome program for viewing an labeling 3-D structures. To visual your own pdb structure right click and chose open with (Yasara). This free program is part of a more extensive molecular modeling package.

● [ArgusLab](#) (Mark A. Thompson, Planaria Software LLC, Seattle, U.S.A.) is an incredible molecular modeling, graphics, and drug design program.

- [RasMol](#) is software for looking at molecular structures. It is very fast: rotating a protein or DNA molecule shows its 3D structure.

- [Deep View](#) (Swiss-PdbViewer) is an application that provides a user friendly interface allowing to analyze several proteins at the same time. The proteins can be superimposed in order to deduce structural alignments and compare their active sites or any other relevant parts. Amino acid mutations, H-bonds, angles and distances between atoms are easy to obtain thanks to the intuitive graphic and menu interface

- [MOLMOL](#) is a molecular graphics program for displaying, analyzing, and manipulating the three-dimensional structure of biological macromolecules, with special emphasis on the study of protein or DNA structures determined by NMR.

- [RasTop](#) - RasTop is a molecular visualization software adapted from the program RasMol by wrapping a user-friendly graphical interface around the "RasMol molecular engine". The software allows several molecules to be opened in the same window and several windows to be opened at the same time. Through an extended menu and a command panel, users can manipulate numerous molecules rapidly and learn about them. Work sessions are saved in script format and are fully regenerated with a simple mouse click.

### ● **Alignments:**

- [ClustalX](#) is a windows interface for the ClustalW multiple sequence alignment program. It provides an integrated environment for performing multiple sequence and profile alignments and analyzing the results. Online help can be found [here](#). (Reference: J.D. Thompson et al. (1997). Nucleic Acids Research **24**: 4876-4882).

- [BioEdit](#) is a mouse-driven, easy-to-use sequence alignment editor and sequence analysis program designed and written by Tom Hall (North Carolina State University). It also provides BLAST capability on local databases.

- [LalnView](#) is a graphical program for visualizing local alignments between two sequences (protein or nucleic acids). Sequences are represented by colored rectangles to give an overall picture of the similarities between the two sequences. Blocks of similarity between the two sequences are colored according to the degree of identity between segments.

### ● **Phylogeny:**

- [PHYLIP](#) (the PHYLogeny Inference Package) is a package of programs for inferring phylogenies. PHYLIP is the most widely-distributed phylogeny package, and competes with PAUP to be the one responsible for the largest number of published trees (Joe Felsenstein, University of Washington, U.S.A.).

- [HyPhy](#) - intended to perform maximum likelihood analyses of genetic sequence data and equipped with tools to test various statistical hypotheses. HYPHY was designed with maximum flexibility in mind and to that end it incorporates a simple high level programming language which enables the user to tailor the analyses precisely to his or her needs. These include relative rate and ratio tests, several methods of ML based phylogeny reconstruction, bootstrapping, model selection, positive selection, molecular clock tests and many more (Reference: S.L. Kosakovsky et al.(2005) Bioinformatics **21**:676-679).

- [TREECON](#) - is a software package developed primarily for the construction and drawing of phylogenetic trees on the basis of evolutionary distances inferred from nucleic and amino acid

sequences. It offers considerable opportunity to change the appearance of the tree. (Reference: Van de Peer, Y. & De Wachter, Y. (1994) *Comput. Applic. Biosci.* 10, 569-570).

● [TREEMAP](#) is designed as a simple tool for visually comparing host and parasite phylogenies: you can view the host and parasite trees, interactively create reconstructions of the history of the host-parasite association, perform randomization tests of tree similarity, and (if you have branch length information such as might be obtained from DNA sequence data) compare branch lengths in the two trees.

● [Treefinder](#) (Gangolf Jobb, Statistical Genetics and Bioinformatics, University of Munich) computes phylogenetic trees from nucleotide sequences. Using the widely accepted Maximum Likelihood method, it is offering a variety of evolutionary models up to the general time reversible model with Gamma and codon position rate heterogeneity among sites. The confidence of inferred relationships may be assessed by bootstrap analysis or, alternatively, by a local rearrangement paired-sites method (LRP). Linux and Mac versions also available.

● [MEGA](#) - an incredible phylogenetic analysis program. (Reference: S. Kumar et al. (2001) *Bioinformatics* 17: 1244-1245)..

● [Modeltest](#) is a program that uses hierarchical likelihood ratio tests (hLRT) to compare the fit of the nested GTR (General Time Reversible) family of nucleotide substitution models. Additionally, it calculates the Akaike Information Criterion estimate associated with the likelihood scores. (Reference: Posada, D. & K. A. Crandall. 1998. *Bioinformatics* 14: 817-818).

● [TreeView](#) provides a simple way to view the contents of a NEXUS, PHYLIP, or other format tree file. This program can be coupled with an analysis package called [FreeTree](#) written by Dr. Jaroslav Flegr (Charles University, Prague, Czech Republic).

\*\*\*\*\*