BIOINFORMATICS FOR BETTER TOMORROW

B. Jayaram* and Kumkum Bhushan
Department of Chemistry &
Supercomputing Facility for Bioinformatics & Computational Biology,
Indian Institute of Technology, Hauz Khas, New Delhi - 110016, India.
Email: **bjayaram@chemistry.iitd.ac.in**
Web site: www.scfbio-iitd.res.in

I. What is Bioinformatics?

Bioinformatics is an emerging interdisciplinary area of Science & Technology encompassing a systematic development and application of IT solutions to handle biological information by addressing biological data collection and warehousing, data mining, database searches, analyses and interpretation, modeling and product design. Being an interface between modern biology and informatics it involves discovery, development and implementation of computational algorithms and software tools that facilitate an understanding of the biological processes with the goal to serve primarily agriculture and healthcare sectors with several spin-offs. In a developing country like India, bioinformatics has a key role to play in areas like agriculture where it can be used for increasing the nutritional content, increasing the volume of the agricultural produce and implanting disease resistance etc.. In the pharmaceutical sector, it can be used to reduce the time and cost involved in drug discovery process particularly for third world diseases, to custom design drugs and to develop personalized medicine (Fig. 1).

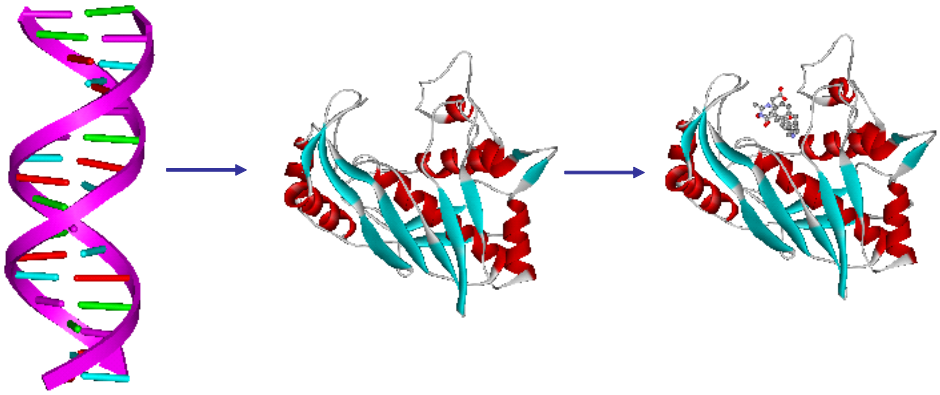
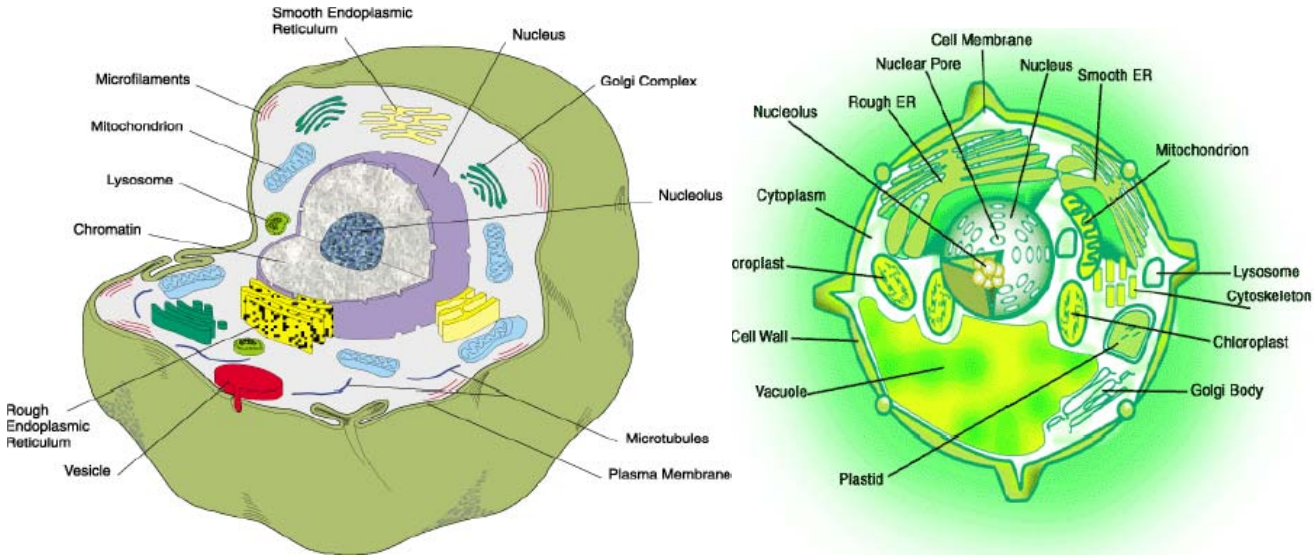


Figure 1: Some major areas of research in bioinformatics and computational biology.

Figure 2: The tree of life depicting evolutionary relationships among organisms from the major biological kingdoms. A possible evolutionary path from a common ancestral cell to the diverse species present in the modern world can be deduced from DNA sequence analysis. The branches of the evolutionary tree show paths of descent. The length of paths does not indicate the passage of time and the vertical axis shows only major categories of organisms, not evolutionary age. Dotted lines indicate the supposed incorporation of some cell types into others, transferring all of their genes and giving the tree some web-like features (adopted from: A Alberts, D Bray, J Lewis, M Raff, K Roberts & J D Watson, Molecular Biology of the Cell, p38, Garland, New York (1994)).

It is assumed that life originated from a common ancestor and all the higher organisms evolved form a common unicellular prokaryotic organism. Subsequent division of different forms of life from this makes the diversity in the morphological and genetic characters (Fig. 2).



(A) **(B)**

Figure 3: (A) An animal cell. The figure represents a rat liver cell, a typical higher animal cell in which features of animal cells are evident such as nucleus, nucleolus, mitochondria, Golgi bodies, lysosomes and endoplasmic reticulum (ER). (Source: www.probes.com/handbook/ figures/0908.html). (B) A plant cell (cell in the leaf of a higher plant). Plant cells in addition to plasma membrane have another layer called cell wall, which is made up of cellulose and other polymers where as animal cells have plasma membrane only. The cell wall, membrane, nucleus chloroplasts, mitochondria, vacuole, ER and other organelles that make up a plant cell are featured in the figure. (Source: http://www.sparknotes.com/biology/cellstructure/celldifferences/section1.html).

The common basis to all these diverse organisms is the basic unit known as the cell (Fig. 3). All cells whether they belong to a simple unicellular organism or a complex multicellular organism (human adults comprise ~ 30 trillion cells), possess a nucleus which carries the genetic material consisting of polymeric chains of DNA (deoxyribonucleic acid), holding the hereditary information and controlling the functioning. Several challenges lie ahead in deciphering how DNA, the genetic material in these cells eventually leads to the formation of organisms (Fig. 4).

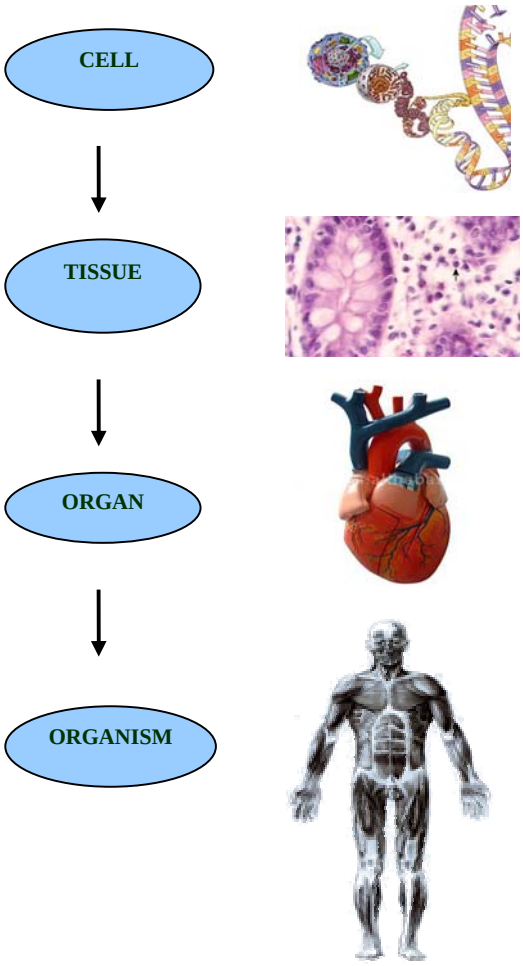


Figure 4: Levels of organization. The entire DNA content of a cell is called genome. The entire protein content in a cell is called the proteome. Cellome is the entire complement of molecules, including genome and proteome within a cell. Tissues are made of collections of cells. Tissue collections make organs. An organism is a collection of several organ systems.

In spite of the complex organization, cells of all organisms possess same molecules of life for the maintenance of living state. These molecules include nucleic acids, proteins, carbohydrates and lipids (Fig. 5).

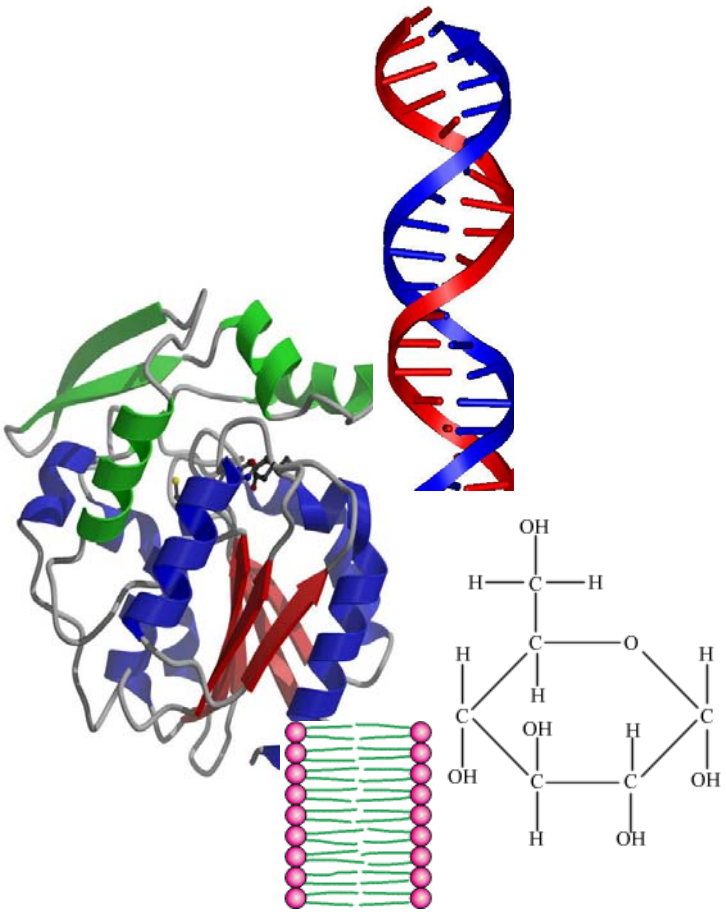


Figure 5: Molecules of life

All organisms self replicate due to the presence of genetic material DNA, the polynucleotide consisting of four bases Adenine (A), Thymine (T), Guanine (G) and Cytosine (C) (Fig. 6). The entire DNA content of the cell is what is known as the genome. The segment of genome that is transcribed into RNA is called a gene. So we can say that hereditary information is transferred in the form of genes contained on the four bases. Understanding these genes is one of the modern day challenges. Why only five percent of the entire DNA is in the form of genes [] and what is the rest of the DNA responsible for, under what conditions genes are expressed, where, when and how to regulate gene expression are some unsolved puzzles.

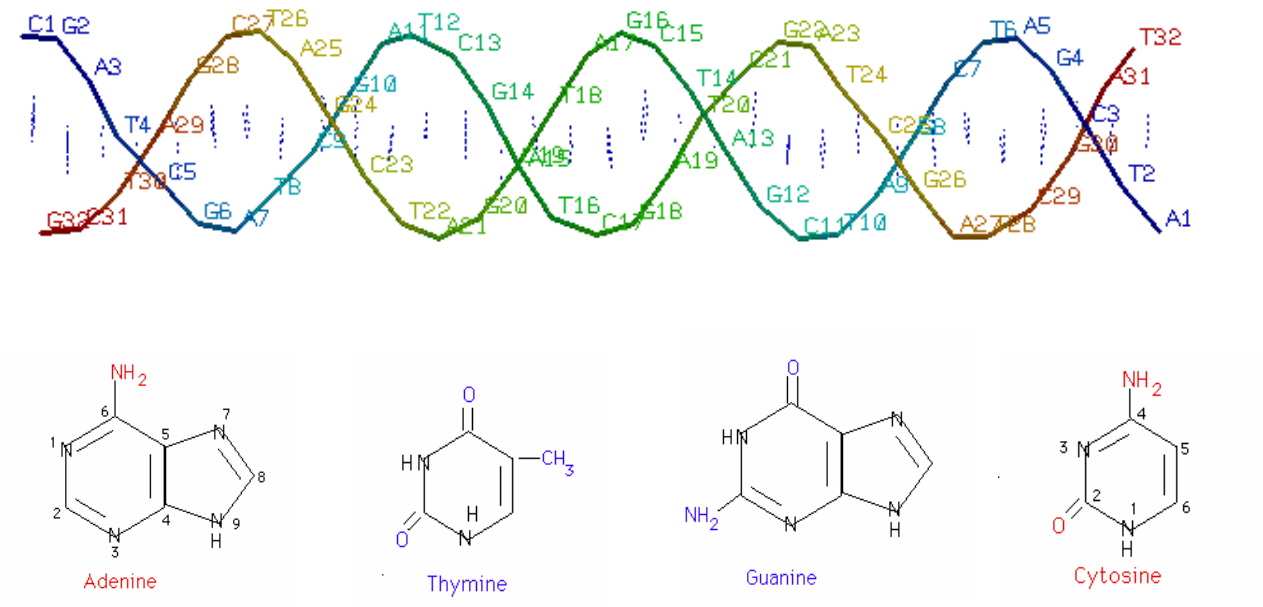


Figure 6: DNA and its alphabets – the nucleic acid bases: A, T, G and C

Some major areas of research in bioinformatics

Genome Analysis: Segments of genome coding for messenger ribonucleic acids (mRNAs), transfer ribonucleic acids (tRNAs), ribosomal ribonucleic acids (rRNAs) are called genes. Among these mRNAs determine the sequence of amino acids in proteins. The mechanism is simple for the prokaryotic cell where all the genes are converted into the corresponding mRNA (messenger ribonucleic acid) and then into proteins. The process is more complex for eukaryotic cells where rather than full DNA sequence, some parts of genes called exons are expressed in the form of mRNA interrupted at places by random DNA sequences called introns. Of the several questions posed here, one is that how some parts of the genome are expressed as proteins and yet other parts (introns as well as intergenic regions) are not expressed and which exons are combined under what conditions to make proteins necessary for the organism.

Several genome projects are being carried out world wide in order to identify all the genes in a specified organism. Human genome project [1, 2] is one such global effort to identify all the alphabets on the human genome, initiated in 1990 by the US government. A comparison of the various genome sizes of different organisms (Table 1) raises questions like what types of genetic modifications are responsible for the four times larger genome size of wheat plant and

seven times smaller size of the rice plant [3] as compared to that of humans. Mice and humans contain roughly the same number of genes – about 28K protein coding regions, about 90% of the human genome is in large blocks of homology with mouse [4]. The chimpanzee and human genomes vary by an average of just 5% [5].

Table 1: **Genome sizes of some organisms *Organism***

Organism	***Genome size ((Mb)*** *(Mb=Mega base)*
H.Influenza	1.83
M tuberculosis	4.4
Eschericia coli	4.6
Sacchromyces cerevisiae (Yeast)	12.5
Plasmodium falciparum	23
C. elegans (Nematode)	100
Drosophila melanogaster (Fruit fly)	122
Gallus gallus (Chicken)	120
Oryza sativa (Rice)	390
Canis lupis familiaris (Dog)	2400
Pan troglodytes (Chimpanzee)	2700
Mus musculus (Mouse)	3000
Homo sapiens (Humans)	3300
Triticum aestivum(Wheat)	13500

(http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html)

Several genetic disorders like Huntington's disease, Parkinson's disease, sickle cell anemia etc. are caused due to mutations in the genes or a set of genes inherited from one generation to another (Table 2). There is a need to understand the genetic origins for such disorders. Why nature carries such disorders and how to prevent these are some of the areas where extensive research is in progress.

Table 2: Specific genetic disorders

Genetic Disorder	***Reason***
Sickle Cell Anemia	Mutation in hemoglobin-b gene on chromosome 11
Brukitt lymphoma	Translocations on chromosome 8
Hemophilia A	Mutation of the *HEMA* gene on the X chromosome
Breast Cancer	Mutation on genes found on chromosomes 13 & 17
Leukemia	Exchange of genetic material between the long arms of chromosome 6 & 22.
Colon cancer	Proteins MSH2, MSH6 on chromosome 2 & MLH1 on chromosome 3 are mutated.
Cystic Fibrosis	Mutations in a single (CFTR) gene
Best disease	Mutation in one copy of a gene located on chromosome 11.
Rett Syndrome	Disfunctioning of a gene on the X chromosome.
Pendred Syndrome	Defective gene on chromosome 7.
Asthma	Disfunctioning of genes on chromosome 5, 6, 11, 14&12.
Diastrophic dysplasia	Mutation in a gene on chromosome 5
Angelman Syndrome	Deletion of a segment on maternally derived chromosome 15.
Werner Syndrome	Mutations on genes located on chromosome 8.
Alzheimer disease	Mutations on four genes located on chromosome 1, 14, 19 & 21.
Parkinson’s Disease	Variations in genes on chromosomes 4, 6.
Huntington’s Disease	Excessive repeats of a three-base sequence, "CAG" on chromosome 4.
Tay-Sachs Disease	Controlled by a pair of genes on chromosome 15
William’s syndrome	Deletion of the gene for elastin and LIM kinase from chromosome 7.
Zellweger syndrome	Mutations in the PXR1 gene on chromosome 12.
Gaucher disease	Mutation in a gene on chromosome 1
Achondroplasia	Mutations in the gene encoding FGFR3 situated on chromosome 4
Familial mediteraanean fever	Mutation in a gene on chromosome 16

(Source:http://www.ncbi.nlm.nih.gov)

An understanding of the genome organization can lead to progresses in drug-target identification. Genome level comparisons of healthy individuals with those carrying some disorder can help identify drug targets. If the genome for humans and a pathogen, a virus causing harm is identified, comparative genomics can predict possible drug-targets for the invader without causing side effects to humans. SNPs (single nucleotide polymorphisms) are common DNA sequence variations that occur when a single nucleotide in the genome sequence changes. SNPs occur every 100 to 300 bases along the human genome. The SNP variants promise to significantly advance our ability to understand and treat human diseases. Comparative genomics is the establishment of the relation between two genes from different organisms. Comparison of series of sequences between two genomes generates intergenomic maps which help in identifying the evolutionary process responsible for divergence of two genomes / species. Functional genomics involves identification of gene function. DNA micro array [6] data analysis is another research area for quantifying the levels of gene expression in various tissues or at different stages in the development of diseases.

Over the past two decades, genetic modifications have enabled plant breeders to develop new varieties of crops like cereals, soya, maize at a faster rate. Genes are transfered from one species to another species called as transgenic varieties, engineered to possess special characteristics that make them better. Research is in progress world-wide, utilizing GM (genetically modified) crops to produce therapeutic plants [7]. Modern plant biotechnology faces a challenge of feeding an increasing world population. The emerging field of genomics has provided huge information to improve crop characteristics like size and height of the plant, seed and flower colour (phenotypes) [8].

In order to contribute to the sustainability of rural agriculture, studies are being conducted to identify medicinal substances based on indigenous knowledge and publicly available databases, to critically evaluate these products using controlled functional genomics experiments and bioinformatics and to increase awareness and assess perceptions about the technology used and to disseminate outcomes. Studies are also in progress to evaluate the effectiveness of traditional therapeutics on inflammatory and parasitic processes in livestock (cows and goats) and to establish models for comparative genomic analyses of functional consequences of exposure using cell and molecular biology, bioinformatics and micro array techniques. Neem, wormwood and garlic are some examples of plants used in traditional medicine that are known

to possess anti-helmintic and anti-inflammatory properties The main biologically active constituents of these selected agents are presented in Table 3.

Table 3: Selected medicinals and their biologically active constituents

Medicinal	Main Biologically Active Constituent
Garlic	Allicin
Tobacco	Nicotine Anabasine
Neem	Limonoids (e.g. azadirachtin, salannin, meliantriol, nimbin, nimbidin)
Wormwood	Thujone
Shitake	1-3 beta glucans and lentinan
Diatomaceous earth	Diatomites
Whey	Proteins

(Source: 9th ICABR International Conference on Agricultural Biotechnology: Ten years later, indigenous knowledge, bioinformatics and rural agriculture technology)

Comparative genomics of plant genomes has suggested that the organization of genes has been conserved during evolution. The complete genomes of many crop plants (e.g. Oryza sativa, wheat) help in providing information about the agronomically important genes which could be used for further improvement in food crops. Genes from Bacillus thuringiensis which control a number of pests are successfully transferred to crops like cotton, maize and potatoes. This helps plants to become insect resistant and the amount of insecticides being used is reduced thus improving overall economics.

Given the whole genome of an organism, finding the genes is a challenging task. Various sophisticated mathematical methods have been proposed. Most of these approaches are database driven which rely on the existing experimental information. Some of the successful strategies are based on short range correlations between bases along the genome (eg. Markov models) and some others are based on long range / global correlations (eg. Fourier transform techniques). Table 4 lists some of the common gene prediction softwares available freely over the internet.

Table 4: A list of gene prediction softwares available freely over the internet

Sl. No.	Name of the Software	URL	Remarks
1.	FGENESH	http://www.softberry.com/all.htm	HMM
2.	GeneID	http://www1.imim.es/geneid.html	*Ab initio*
3.	GeneParser	http://beagle.colorado.edu/~eesnyder/GeneParser.html	*Ab initio*
4.	GeneWise	www.sanger.ac.uk/Software/Wise2/genewiseform.shtml	Homology
5.	GeneMark	http://exon.gatech.edu/GeneMark	*Ab initio*
6.	GENSCAN	http://genes.mit.edu/GENSCAN.html	*Ab initio*
7.	Glimmer	http://www.tigr.org/software/glimmer/	*Ab initio*
8.	GlimmerHMM	http://www.tigr.org/software/GlimmerHMM/	*Ab initio*
9.	GRAILEXP	http://compbio.ornl.gov/grailexp	*Ab initio*
10.	GENVIEW	www.itba.mi.cnr.it/webgene	*Ab initio*
11.	GenSeqer	http://bioinformatics.iastate.edu/cgi-bin/gs.cgi	Homology
12.	ORFgene	http://125itba.mi.cnr.it/~webgene/wwworfgene2.html	Homology
13.	MORGAN	http://www.cs.jhu.edu/labs/compbio/morgan.html	*Ab initio*
14.	PredictGenes	http://mendel.ethz.ch:8080/Server/subsection3_1_8.html	Homology
15.	MZEF	http://rulai.cshl.edu/software/index1.htm	*Ab initio*
16.	Rosetta	http://crossspecies.lcs.mit.edu	Homology
17.	VEIL	http://www.tigr.org/~salzberg/veil.html	*Ab initio*
18.	PROCRUSTES	http://hto-13.usc.edu/software/procrustes/index.html	Homology
19.	Xpound	ftp://igs-server.cnrs-mrs.fr/pub/Banbury/xpound	*Ab initio*
20.	Chemgenome	http://scfbio-iitd.res.in/chemgenome2.0	*Ab initio*

We have recently developed a hypothesis driven physico-chemical model for gene prediction. In this model a physico-chemical vector which attempts to capture forces responsible for DNA structure and protein-nucleic acid interactions walks along the genome identifying the potential protein coding regions. As of now, the physico-chemical model *(ChemGenome 2.0)* is seen to predict genes in reasonably well with fairly high specificities and sensitivities. The methodology has been web enabled at www.scfbio-iitd.res.in/chemgenome2. Also software for distinguishing genes from non-genes is available for free access at www.scfbio-iitd.res.in/chemgenome [9].

It is hoped that bioinformatics research in genomics will eventually generate a molecular view of genome organization and genetic networks.

Protein Folding

Proteins are polymers of amino acids with unlimited potential variety of structures and metabolic activities. Proteins may be classified into structural proteins, enzymes, hormones, transport proteins, protective proteins, contractile proteins, storage proteins, toxins. Each protein possesses a characteristic three-dimensional shape and function which is defined by the sequence of amino acids constituting it. This in turn is genetically controlled by the sequences of bases in DNA of the cell through the genetic code. Substitution of a single amino acid can cause a major alteration in function as in the well known case of sickle cell anemia. Study of proteins with related folds and related sequences (homologous proteins) from different species is of interest in constructing taxonomic relationships.

Protein folding can be considered to involve changes in the polypeptide chain conformation to attain a stable conformation corresponding to the global minimum in free energy which is about 10 to 15 kcal/mol lower relative to the unfolded state [10]. How does a given sequence of amino acids fold into a specific conformation as soon as it is conceived on the ribosomal machinery using the information on mRNA in millisecond - second timescales, is the problem pending a resolution for over five decades. Linus Pauling solved the secondary structure problem in proteins and later Prof G N Ramachandran, made some significant contributions towards a deeper understanding of the secondary structure of proteins. His fundamental work in this area is remembered in the form of Ramachandran maps [11]. For a 200 amino acid protein with just two conformations per amino acid (i.e. considering only a highly restricted Ramachandran map), a systematic search for this minimum among all possible 2^{200}

conformations, will take approximately $3x10^{54}$ years which is much longer than the present age of the universe. Despite this innumerable number of conformations to search, nature does it in milliseconds to seconds (Fig. 7).

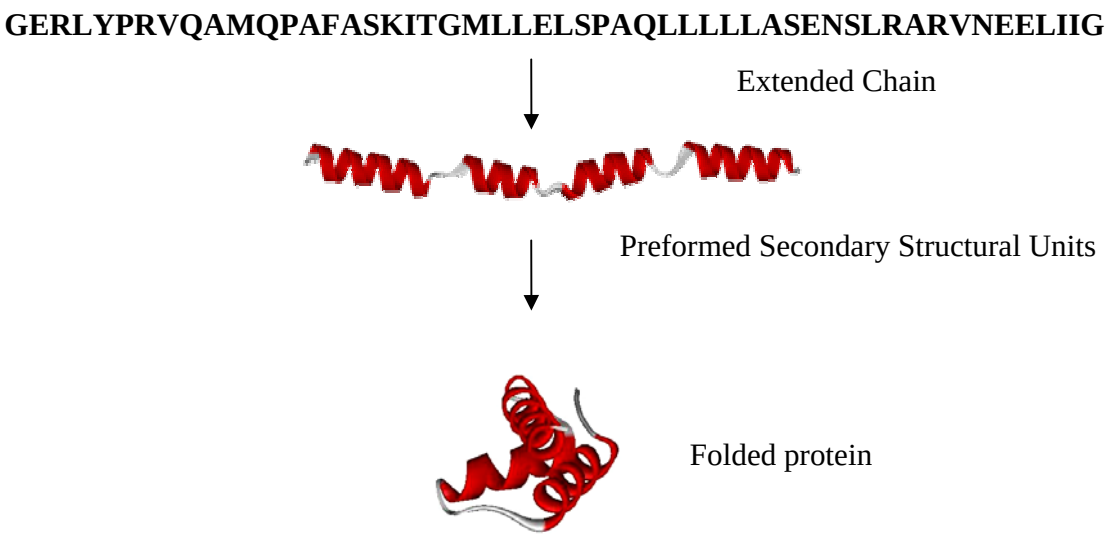


Figure 7: Sequence to structure: the protein folding problem

A computational solution to the protein folding problem – i.e. a specification of the Cartesian coordinates of all the atoms of the protein from its amino acid sequence information - has an immense immediate impact on society. In biotech industry, this can be helpful in the design of nanobiomachines and biocatalysts to carry out the required function. Pulp, paper and textile industry, food and beverages industry, leather and detergents industry are among the several potential beneficiaries. Other important implications are in structure based drug discovery wherein the three dimensional structure of the drug target is the starting point (Fig. 8). Structures of receptors – a major class of drug targets - are refractory to experimental techniques thus leaving the field open to computer modeling.

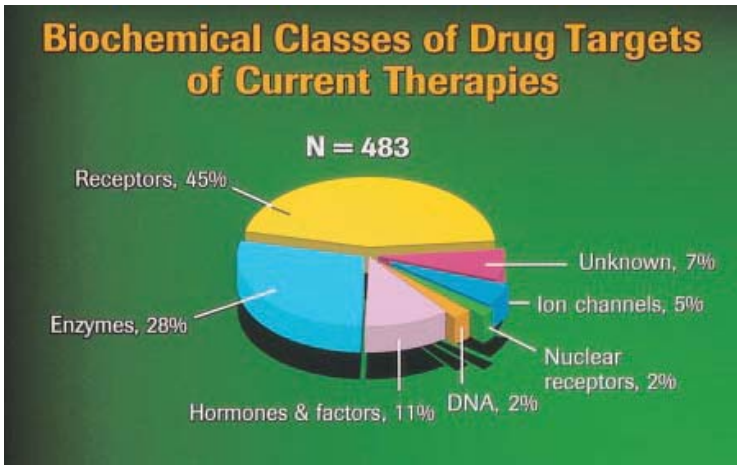


Figure 8: Number and classification of known drug targets (in year 2006) [12].

There is an urgent demand for faster and better algorithms for protein structure prediction. There are two major ways in which protein structure prediction attempts are currently progressing viz. comparative modeling, and *de novo* approaches. The former is database dependent methodology relying on known structures and the latter is independent of the databases and starts from the physical principles. Comparative modeling techniques are extremely popular, reliable and fast where sequence homologues exist in the database. With increasing structural information, these techniques should prove more useful. Although the *de novo* techniques till date are able to predict structures of only small proteins, because of their first principles approach and the concurrent computational requirements, they have the potential to predict new / novel folds and structures. The time required to fold a 200 amino acid protein which evolves ~10^{-11} sec per day per processor according to Newton's laws of motion will require approximately a million years to fold. If one can envision a million processors working together, a single mid-sized protein can be folded in one year computer time. Against this backdrop, IBM has launched a five year *Blue Gene* project (www.research.ibm.com/bluegene/) in the year 1999 to address complex biomolecular phenomena such as protein folding. The full Blue Gene/L machine was designed and built in collaboration with the Department of Energy's NNSA/Lawrence Livermore National Laboratory in California, and has a peak speed of 360 Teraflops. Blue Gene is by far the fastest supercomputing system in the world, giving scientists access to unprecedented computing power. Table 5 lists some of the freely available comparative modeling as well as *de novo* protein structure prediction softwares available over the internet.

Table 5: A list of protein structure prediction softwares available freely over the internet

Sl. No.	Name of the software	Description	URL
1.	PSI-BLAST	The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches	http://www.ncbi.nlm.nih.gov/BLAST/
2.	CPHModels2.0	An automated protein structure homology-modeling server	http://www.cbs.dtu.dk/services/CPHmodels/
3.	Swiss-Model	A fully automated protein structure homology-modeling server	http://swissmodel.expasy.org/SWISS-MODEL.html
4.	ModWeb	A web server implementation of MODELLER (comparative protein structure modeling by satisfaction of spatial restraints)	http://alto.compbio.ucsf.edu/modweb-cgi/main.cgi
5.	3DJigSaw	An automated server to build three-dimensional models for proteins based on homologues of known structure	http://www.bmm.icnet.uk/servers/3djigsaw/
6.	GenTHREADER	A combination of methods such as sequence alignment with structure based scoring functions and neural network based jury system to calculate final score for the alignment	http://bioinf.cs.ucl.ac.uk/psipred/
7.	3D PSSM	Threading approach using 1D and 3D profiles coupled with secondary structure and solvation potential	http://www.sbg.bio.ic.ac.uk/~3dpssm
8.	ROBETTA	*De novo* Automated structure prediction analysis tool used to infer protein structural information from protein sequence data	http://robetta.bakerlab.org
9.	PROTINFO	*De novo* protein structure prediction web server utilizing simulated annealing for generation and different scoring functions for selection of final five conformers	http://protinfo.compbio.washington.edu
10.	SCRATCH	Protein structure and structural features prediction server which utilizes recursive neural networks, evolutionary information, fragment libraries and energy	http://www.igb.uci.edu/servers/psss.html
11.	ROKKY	*De novo* structure prediction by the simfold energy function with the multi-canonical ensemble fragment assembly	http://www.proteinsilico.org/rokky/rokky-p/
12.	BHAGEERATH	Energy based methodology for narrowing down the search space of small globular proteins	http://www.scfbio-iitd.res.in/bhageerath

We have recently developed a computational protocol for modeling and predicting protein structures of small globular proteins. Here a combination of bioinformatics tools, physicochemical properties of proteins and *de novo* approaches are used. This suite of programs is named *Bhageerath* (http://www.scfbio-iitd.res.in/bhageerath) [13, 14]. Starting with the sequence of amino acids, for 50 small globular proteins, 10 candidate structures for each protein within 3-6 Å of the native are predicted within a day on a 96 processor cluster. Attempts are in progress to further improve the prediction accuracies of the structures to within a root mean square deviation of <3Å from the native structures via explicit solvent molecular dynamics and Metropolis Monte Carlo simulations.

Function follows form [15] and hence the need for structures. Stated alternatively, sequence to consequence [16] is the major challenge in proteomics investigations.

Drug Design

The information present in DNA is expressed via RNA molecules into proteins which are responsible for carrying out various activities. This information flow is called the central dogma of molecular biology (Fig. 9). Potential drugs can bind to DNA, RNA or proteins to suppress or enhance the action at any stage in the pathway.

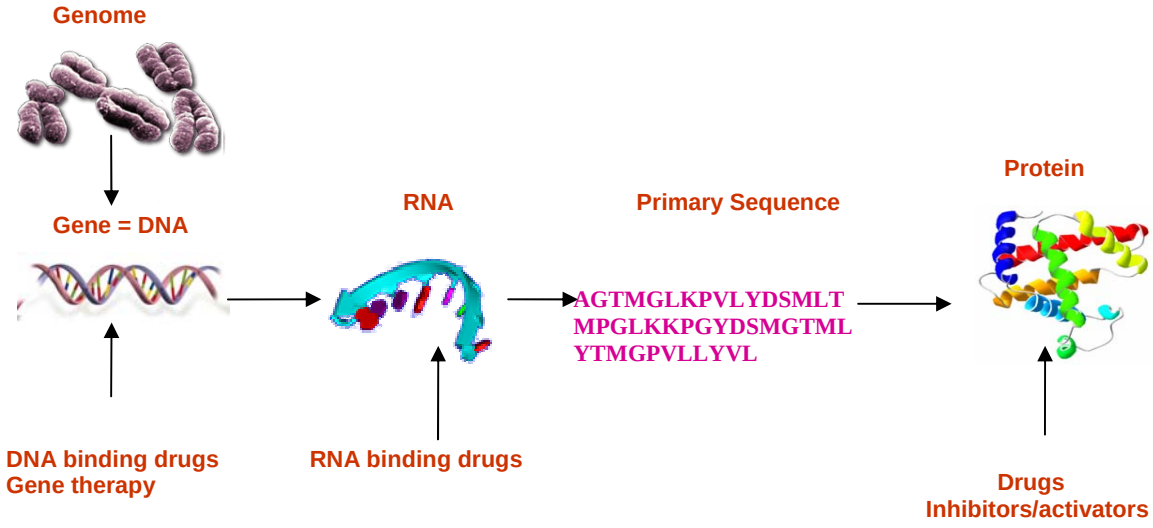


Figure 9: Central dogma of modern drug discovery

As structures of more and more protein targets become available through crystallography, NMR and bioinformatics methods, there is an increasing demand for computational tools that can identify and analyze active sites and suggest potential drug molecules that can bind to these sites specifically (Fig. 10).

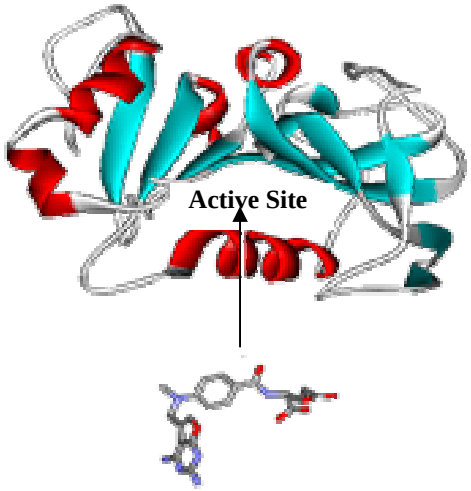


Figure 10: Active-site directed drug-design

Also to combat life-threatening diseases such as AIDS, tuberculosis, malaria etc., a global push is essential (Table 6). “Millions for Viagra and pennies for the diseases of the poor” [17] is the current situation of investment in Pharma Research & Development.

Table 6: Leading causes of death

Leading causes of death in millions per year due to infectious diseases (for the year 2002)		Percentage (%)
Lower respiratory infections	3.9	19
HIV-AIDS	2.8	3
Diarrheal diseases	1.8	17
Tuberculosis	1.6	n/a
Malaria	1.2	8
Measles	0.6	4
Neonatal Causes	N/A	37
Others(including noncommunicable diseases)	N/A	10

Source: World health report, 2004, WHO, www.globalhealth.org

Time and cost required for designing a new drug are immense and at an unacceptable level. According to some estimates it costs, on an average, about $1.2-1.4 billion and 14 years of research to develop a new drug before it is introduced in the market (Table 7).

Table 7: Cost and time involved in drug discovery

Target Discovery
2.5 yrs ↓ ***4%***
Lead Generation & Lead Optimization
3.0 yrs ↓ ***15%***
Preclinical Development
1.0 yrs ↓ ***10%***
Phase I, II & III Clinical Trials
6.0 yrs ↓ ***68%***
FDA Review & Approval
1.5 yrs ↓ ***3%***
Drug to the Market
14 yrs **$1.2-1.4 billion**

(source: PAREXEL [18])

Intervention of computers at some plausible steps is imperative to bring down the cost and time involved in the drug discovery process (Table 8). Making a drug is more like designing a key for a lock to jam or open the lock, except that both the lock and the key are dynamic and made of atoms and are susceptible to environmental effects such as solvent, salt and other small or biomolecules.

Table 8: High End Computing Needs for *In Silico* Drug Design: Estimates of current computational requirements to complete a binding affinity calculation for a given drug

Modeling complexity	**Method**	**Size of library**	**Required computing time**
Molecular Mechanics	SPECITOPE	140,000	~1 hour
Rigid ligand/target	LUDI	30,000	1-4 hours
	CLIX	30,000	33 hours
Molecular Mechanics	Hammerhead	80,000	3-4 days
Partially flexible ligand	DOCK	17,000	3-4 days
Rigid target	DOCK	53,000	14 days
Molecular Mechanics	ICM	50,000	21 days
Fully flexble ligand			
Rigid target			
Molecular Mechanics	AMBER	1	~several days
Free energy perturbation	CHARMM		
QM Active site and MM protein	Gaussian, Q-Chem	1	>several weeks

(Source: http://cbcg.lbl.gov)

In silico methods can help in identifying drug targets via bioinformatics tools. They can also be used to analyze the target structures for possible binding / active sites, generate candidate molecules, check the molecules for their drug-likeness, dock these molecules with the target, rank them according to their binding affinities, further optimize the molecules to improve binding characteristics, studies on newer drug delivery methods and design principles to cut down on toxicity. Some popular softwares for drug design are listed in Table 9.

Table 9: Comprehensive softwares for drug design

S.No	Name of the software	Description	URL
1	InsightII, Discovery studio Cerius ADME/ Tox Package	A suite for molecular modeling and *de nova* drug design Provides computational models for the prediction of ADME properties derived from chemical structures	http://www.accelrys.com/products/insight/ http://www.accelrys.com/products/cerius2/cerius2products/c2adme.html
2	Sybyl	Computational informatics software for drug discovery	http://www.tripos.com/
3	Bio-suite	The package can be used for Genomics, Protein modeling & Structural analysis, Simulation and Drug Design	http://www.atc.tcs.co.in/biosuite/
4	Molecular Operating Environment (MOE)	Bioinformatics, Cheminformatics, Protein Modeling, Structure-Based Design, High Throughput Discovery and Molecular Modeling and Simulations	http://www.chemcomp.com/
5	MDL QSAR	A suite of tools and applications for decision support, visualization, analysis, and ADME-toxicity assessment	http://www.mdl.com/products/predictive/qsar/index.jsp
6	Glide	Ligand –receptor docking	http://www.schrodinger.com/
7	Autodock	Protein –ligand docking	http://www.scripps.edu/mb/olson/doc/autodock/
8	Ligplot	Program for automatically plotting protein ligand interactions	http://www.biochem.ucl.ac.uk/bsm/ligplot/ligplot.html
9	synSPROUT	A *de novo* ligand design software system which uses synthetic rules to join fragments	http://www.simbiosys.ca/products/products.html
10	Sanjeevini	A Comprehensive software for Active site directed drug design	http://www.scfbio-iitd.res.in/research/drugdesign.htm

Pursuing the dream that once the gene target is identified and validated, drug discovery protocols could be automated using bioinformatics and computational biology tools, we have developed a computational protocol for active site directed lead molecule design. The suite of programs christened "*Sanjeevini*" (http://www.scfbio-iitd.res.in/research/drugdesign.htm)[19] has the potential to evaluate and /or generate lead-like molecules for any biological target [20-22]. The *Sanjeevini* protocols when tested on a few drug targets could successfully distinguish drugs from non-drugs (Fig. 11). Validation on other targets is in progress.

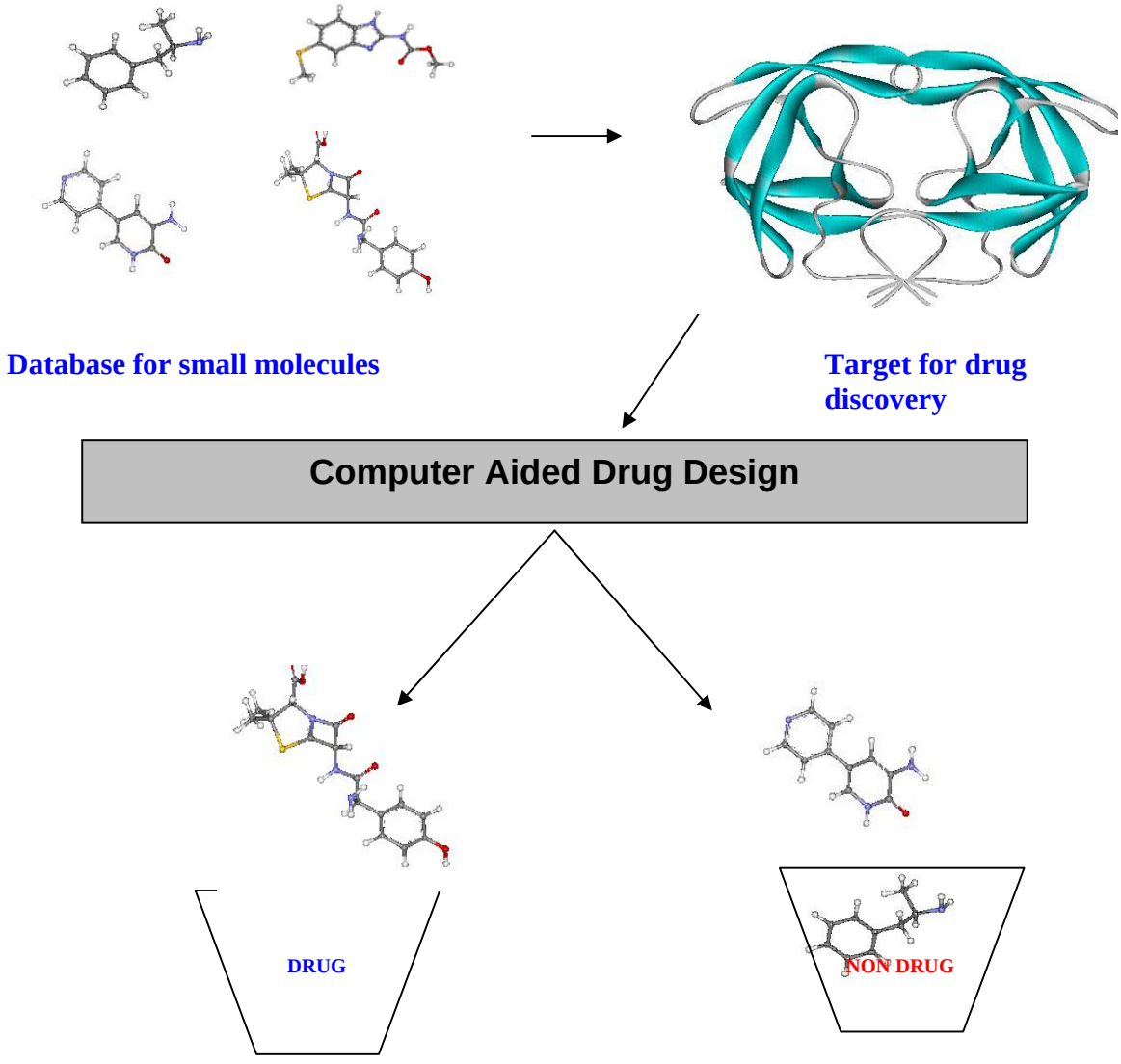


Figure 11: The active site directed lead design protocols developed from first principles and implemented on a supercomputer could segregate drugs from Non-drugs based on binding affinity estimates.

Bioinformatics applied in the form of pharmacogenomics involves developing personalized medicine for individuals based on their genetic profile. Databases of genetic profiles of patients with ailments like diabetes, cancer etc. play an important role in individual health care. The aim is to study a patient's individual genetic profile and compare it with a collection of reference profiles which may help in improving the diagnosis and prevention of the disease.

Metabolomics

Metabolomics is the "systematic study of the unique chemical fingerprints that specific cellular processes leave behind" - specifically, the study of their small-molecule metabolite profiles. The goals of Metabolomics are to catalog and quantify the myriad small molecules found in biological fluids under different conditions. The words 'Metabolomics' and 'Metabonomics' are often used interchangably, though a consensus is beginning to develop as to the specific meaning of each. The goals of Metabolomics are to catalog and quantify the myriad small molecules found in biological fluids under different conditions. Metabonomics is the study of how the metabolic profile of a complex biological system changes in response to stresses like disease, toxic exposure, or dietary change.

The metabolome represents the collection of all metabolites (such as metabolic intermediates, hormones and other signalling molecules, and secondary metabolites) in a biological organism, which are the end products of its gene expression. Thus, while mRNA gene expression data and proteomic analyses do not tell the whole story of what might be happening in a cell, metabolic profiling can give an instantaneous snapshot of the physiology of that cell. Metabolites are the intermediates and products of metabolism. The term metabolite is usually restricted to small molecules. A primary metabolite is directly involved in the normal growth, development, and reproduction. A secondary metabolite is not directly involved in those processes, but usually has important ecological function. Examples include antibiotics and pigments. Various small molecule databases have been created and a few have been listed in Table 10.

Table 10: A list of small molecule databases and metabolomics sites available over the internet

Sl. No.	Name of the database / site	Description	URL
1.	ChEBI	Chemical Entities of Biological Interest (ChEBI) is a freely available dictionary of molecular entities focused on 'small' chemical compounds.	http://www.ebi.ac.uk/chebi/
2.	ChemFinder	A free database for 2D structures and 3D models of chemical compounds.	http://chemfinder.cambridgesoft.com/
3.	CAS	CAS, a division of the American Chemical Society, provides the most comprehensive databases of publicly disclosed research in chemistry and related sciences.	http://www.cas.org/index.html
4.	ChemIDplus	This database allows users to search the NLM ChemIDplus database of over 370,000 chemicals	http://www.cas.org/index.html
5.	CSD	The world repository of small molecule crystal structures	http://www.ccdc.cam.ac.uk/products/csd/
6.	EUROCarbDB	European Carbohydrate Databases	http://www.eurocarbdb.org/
7.	Ligand Depot	Ligand Depot is a data warehouse which integrates databases, services, tools and methods related to small molecules bound to macromolecules.	http://ligand-depot.rutgers.edu/
8.	MSD	Consistent and enriched library of ligands, small molecules and monomers that are refered in any structure	http://www.ebi.ac.uk/msd-srv/chempdb/cgi-bin/cgi.pl
9.	OCA	A browser-database for protein structure/function	http://oca.ebi.ac.uk/oca-docs/oca-home.html
10.	PubChem	PubChem provides information on the biological activities of small molecules	http://pubchem.ncbi.nlm.nih.gov/
11.	Spectral Database for Organic Compounds (SDBS)	SDBS is an integrated spectral database system for organic compounds,which includes 6 different types of spectra under a directory of the compounds	http://riodb01.ibase.aist.go.jp/sdbs/cgi-bin/cre_index.cgi?lang=eng
12.	KEGG LIGAND	KEGG LIGAND contains knowledge on the universe of chemical substances and reactions that are relevant to life	http://www.genome.jp/kegg/ligand.html
13.	Human Metabolite Database	a freely available electronic database containing detailed information about small molecule metabolites found in the human body	http://www.hmdb.ca/
14.	MMCD	A resource for metabolomics research based on nuclear magnetic resonance (NMR) spectroscopy and mass spectrometry (MS)	http://mmcd.nmrfam.wisc.edu/

The comprehensive qualitative and quantitative analyses of the primary and secondary metabolites provides a holistic view of the biochemical status or biochemical phenotype of an organism. The correlations of biochemical information with genetic and molecular data are very useful in providing better insight into the functions of unknown gene or systems response to external stimuli. Metabolomic studies also offer unique opportunities to study regulation and signaling under the control of small molecules (i.e., metabolites). Quite often, signaling and regulation are transparent at the transcriptome and/or proteome level. Finally, metabolomics offers the unbiased ability to differentiate organisms or cell states based on metabolite levels that may or may not produce visible phenotypes/genotypes. Although metabolomics is quite promising, several challenges still exist that influence the implementation of a metabolomic approach, including chemical complexity, analytical and biological variance, and dynamic range.

The key application of metabolomics is in the toxicity assessment / toxicology of potential drug candidates. Metabolic profiling (especially of urine or blood plasma samples) can be used to detect the physiological changes caused by toxic insult of a chemical (or mixture of chemicals). In many cases, the observed changes can be related to specific syndromes, e.g. a specific lesion in liver or kidney. This is of particular relevance to pharmaceutical companies wanting to test the toxicity of potential drug candidates: if a compound can be eliminated before it reaches clinical trials on the grounds of adverse toxicity, it saves the enormous expense of the trials.

Metabolomics can be an excellent tool in functional genomics for determining the phenotype caused by a genetic manipulation, such as gene deletion or insertion. Sometimes this can be a sufficient goal in itself -- for instance, to detect any phenotypic changes in a genetically-modified plant intended for human or animal consumption. More exciting is the prospect of predicting the function of unknown genes by comparison with the metabolic perturbations caused by deletion/insertion of known genes. Such advances are most likely to come from model organisms such as Saccharomyces cerevisiae and Arabidopsis thaliana.

Nutrigenomics is a generalised term which links genomics, transcriptomics, proteomics and metabolomics to human nutrition. In general a metabolome in a given body fluid is influenced by endogenous factors such as age, sex, body composition and genetics as well as underlying pathologies. The large bowel microflora are also a very significant potential

confounder of metabolic profiles and could be classified as either an endogenous or exogenous factor. The main exogenous factors are diet and drugs. Diet can then be broken down to nutrients and non- nutrients. Metabolomics is one means to determine a biological endpoint, or metabolic fingerprint, which reflects the balance of all these forces on an individual's metabolism.

Bioinformatics endeavors in India

Owing to the well acknowledged IT skills and a spate of upcoming software, biotech and pharma industries and active support from Government organizations, the field of Bioinformatics in India appears promising. However, the projections of the growth potential in India in a global scenario clearly indicate (Fig. 12) that a lot more could be done.

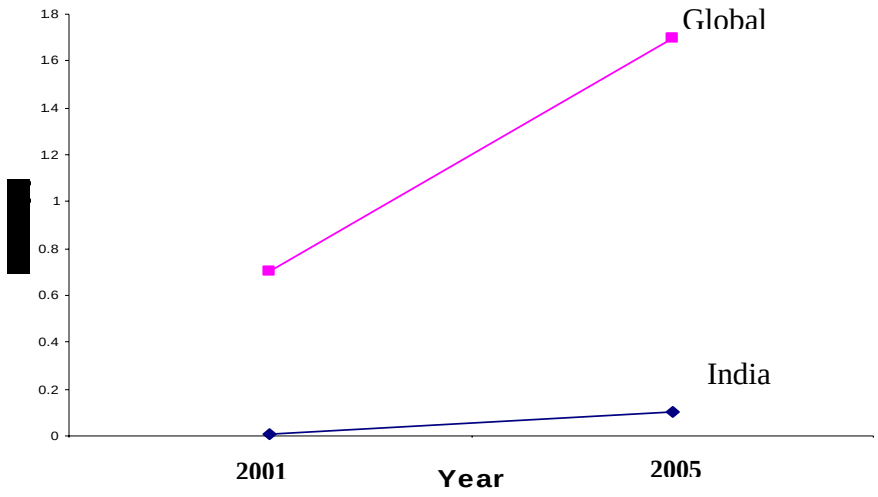


Figure 12: Growth potential for Bioinformatics based business opportunities in India according to IDC (International Data Corporation), India

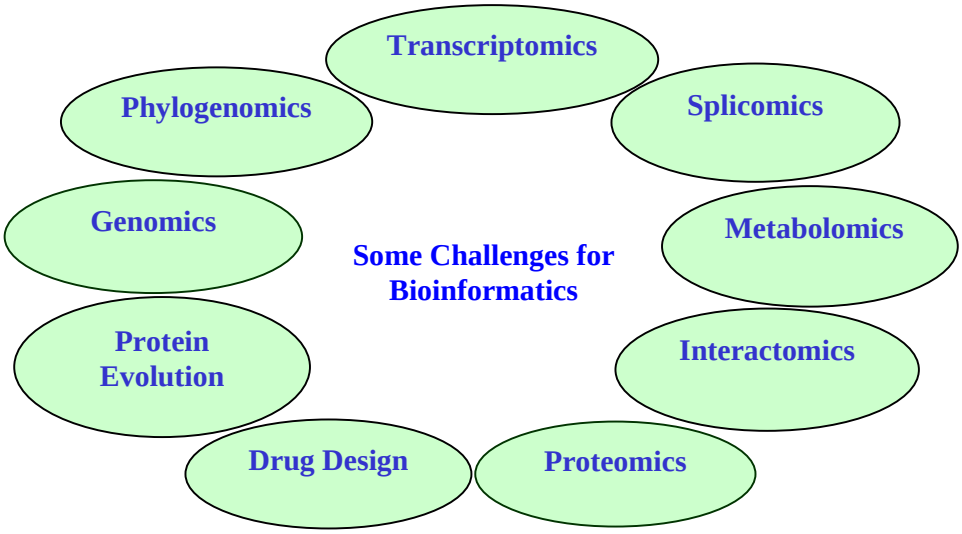


Figure 13: Challenges for bioinformatics

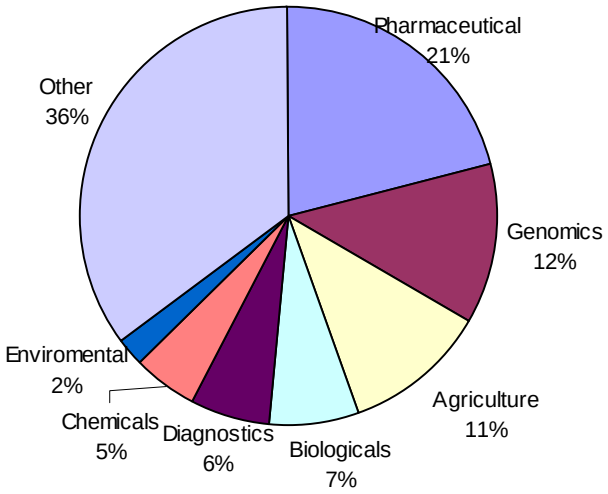
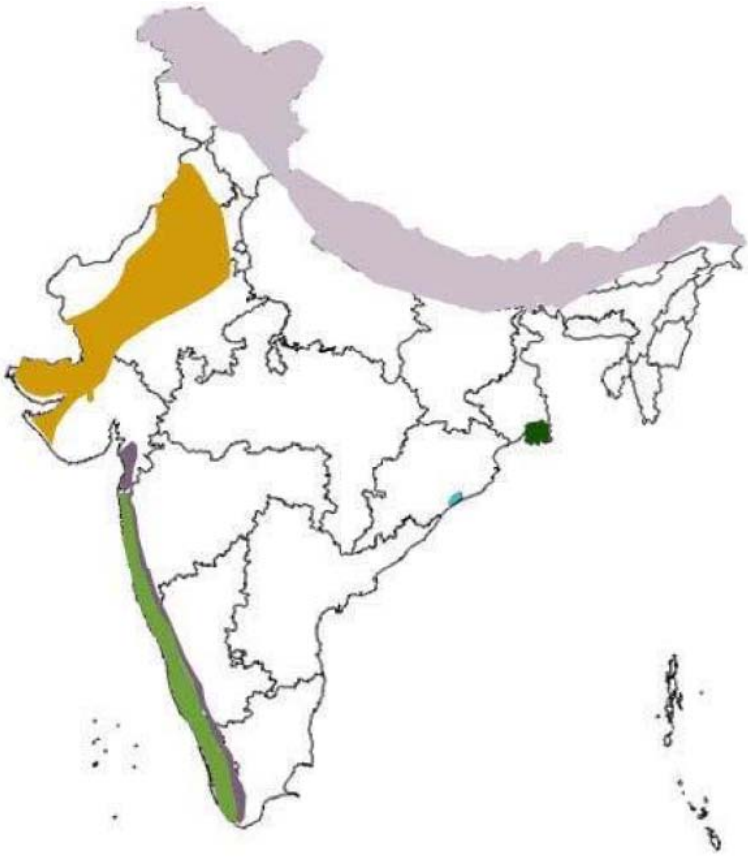


Figure 14. Major areas of Applications [23]

Besides, pharmaceutical and agriculture sectors, the challenges facing bioinformatics and areas of potential applications are captured in Figures 13 and 14.

Bioinformatics and Biodiversity

Biodiversity informatics harnesses the power of computational and information technologies to organize and analyze data on plants and animals at the macro and at genome levels. India ranks among the top twelve nations of the world in terms of biological diversity (Fig. 15 and Fig. 16).



Himalayas - This majestic range of mountains is the home of a diverse range of flora and fauna. Eastern Himalayas is one of the two biodiversity hotspots in India.

Chilika - This wetland area is protected under the Ramsar convention.

Sunderbans - The largest mangrove forest in India.

Western Ghats - One of the two biodiversity hotspots in India.

Thar desert - The climate and vegetation in this area is a contrast to the Himalayan region.

Figure 15: Biodiversity in India (Source: http://edugreen.teri.res.in/explore/maps/biodivin.htm)

Figure 16: Biodiversity bioinformatics is essential to preserve the natural balance of flora and fauna on the planet and to prevent the extinction of species (Source: www.hku.hk/ecology/envsci.htm)

Bioinformatics and environment

Deinococcus radiodurans is known for radiation resistance and being used for cleaning up the waste sites that contain toxic chemicals. Bioinformatics is also helping in climate change studies. There are many organisms which use carbon dioxide as their sole carbon source and increasing levels of carbon dioxide emission is one of the major causes of the global climate change. The study of genomes of these microbial organisms, which is possible through bioinformatics, helps in proposing ways to decrease the carbon dioxide content. The program launched by Department of Energy, USA (DOE) started Microbial Genome Project aimed at sequencing the genomes of bacteria useful in environmental cleanup. This project started in the year 1994 has brought a revolution in the field of microbiology. According to NCBI, about 100 genomes have been sequenced so far. According to some estimates, microbes constitute about 60% of the earth's biomass and play an important role in natural biogeochemical cycles. Scientists have now started realizing their potential and role in global climate processes. Several applications of microbes have been conceived, such as in cleaning up toxic waste-sites worldwide, energy generation and development of renewable energy sources, management of environmental carbon dioxide related to climate change, detection of disease-causing organisms and monitoring of the safety of food and water supplies, use of genetically altered bacteria as living sensors (biosensors) to detect harmful chemicals in soil, air or water and understanding of

specialized systems used by microbial cells to live in natural environments with other cells.

Bioinformatics and Diagnostics

Microarray experiments generate the sort of data where the number of measurements of each sample is much greater than the number of samples. Bioinformatics helps in building new statistical techniques specifically for microarray data to cope up with the multivariate nature of microarray data and to extract meaningful information from it. These tools enable identification of diagnostic markers that are based on a very small number of genes. The fewer genes that are required to diagnose a disease, the simpler and cheaper the diagnostic tests can be.

Bioinformatics and biotechnology

The microbes Archaeoglobus fulgidus, Thermotoga maritima and Corynebacterium glutamicum have the potential for practical applications in industry and environmental projects. These microorganisms thrive in water temperatures above the boiling point and therefore may provide heat-stable enzymes suitable for use in industrial processes. Corynebacterium glutamicum is used by the chemical industry for the biotechnological production of the amino acid lysine. The substance is employed as a source of protein in animal nutrition. Lysine is one of the essential amino acids in animal nutrition. Biotechnologically produced lysine is an alternative to soybeans and bonemeal.

Bioinformatics and veterinary sciences

Sequencing projects of farm animals like pig, cow and others are aimed at understanding the biology of these animals which thus helps in improving their health and therefore benefits in human nutrition. Conservation of extinct species is another area where bioinformatics finds applications.

Bioinformatics and systems biology

It is anticipated that many more computational innovations will ensue in going from genome to the organism and systems biology is that all encompassing field. It is a multidisciplinary approach to integrate different levels of information to understand how biological systems function. By studying the relationships and interactions between various parts

of a biological system (e.g., gene and protein networks involved in cell signalling, metabolic pathways, organelles, cells, physiological systems, organisms, etc.) [24], this nascent field is expected to provide a prior knowledge about the whole system including response of the system to external perturbations at both individual and collective levels.

Conclusion

With the increasingly large amounts of biological data, integration with information technology has become essential. Originally started as a speciality for storage of data and as a tool kit for analyzing data, bioinformatics now encompasses many emerging areas like, evolutionary studies, protein structure-function prediction, gene expression studies etc.. It may not be long before bioinformatics becomes a hypothesis driven molecular science bridging the gap between the genome and the organism, with data providing a platform for validation and new product development.

References

1. Venter, J.C. et al. The sequence of the human genome. *Science,* **2001**, 291, 1304–1351.

2. Schmutz, J. et al. Human genome: Quality assessment of the human genome sequence. *Nature,* **2004**, 429, 365-368.

3. Vij, S., Gupta, V., Kumar, D., Vydianathan, R., Raghuvanshi, S., Khurana, P., Khurana, J.P. and Tyagi, A.K. Decoding the rice genome. *Bioessays*, **2006**, 28, 421-432.

4. Ross, C. Hardison. Compartivs genomics. *PLoS Biology*, **2003**, 1, 156-160.

5. Roy, J. Britten. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proceedings of the National Academy of Sciences,* **2002**, 99, 13633-13635.

6. Hanash, S. Disease Proteomics. *Nature,* **2003**, 422, 226-232.

7. Chevre, A., Eber,F., Baranger, A. and Renard, M. Gene flow from transgenic crops. *Nature,* **1997,** 389, 924.

8. Cutanda, M.C. Hernandez-Acosta P. Culianez-Macia F.A. Bioinformatics for crop improvement. *Proceedings of the World Congress of Computers in Agriculture and Natural Resources,* **2002**, 773-778.

9. Dutta, S., Singhal, P., Agrawal, P., Tomer, R., Kritee, Khurana, E. and Jayaram, B. A Physico-Chemical Model for Analyzing DNA sequences. *J. Chem. Inf. Model*., **2006**, 46(1), 78-85.

10. Anfinsen, C.B. Principles that govern the folding of protein chains. *Science*, **1973**,181,223.

11. Nelson, D.L., Cox, M. M. *Lehninger- Principles of Biochemistry* 4th Edition, W.H. Freeman, **2005**.

12. Peter, Lmming.,Christian, Sinning. and Achim, Meyer. Drugs,their targets and the nature and number of drug targets. *Nature*, **2006**, 5,821-834.

13. Jayaram, B., Bhushan, K., Shenoy, S. R., Narang, P., Bose, S., Agrawal, P., Sahu, D., Pandey, V.S. *Bhageerath :* An Energy Based Web Enabled Computer Software Suite for Limiting the Search Space of Tertiary Structures of Small Globular Proteins. *Nucleic Acids Res.*, **2006**, 34, 6195-6204.

14. Thukral,L., shenoy, S.R., Bhushan, K. and Jayaram, B. ProRegIn: A regularity index for the selection of native-like tertiary structures of proteins. *J. Biosci*., **2007**, 32, 71-81.

15. Dickerson,R.E. and Geis,I. *The Structure and Action of Proteins,* Menlo Park **1969**.

16. Petsko, G.A. From sequence to consequence. *Genome Biol*. **2000**, 1, 406.

17. Silvenstein,K. Millions for Viagra, Pennies for Diseases of the Poor. *The Nation*, **1999**, 269, 3, 13.

18. PAREXEL's Pharmaceutical R&D Statistical Sourcebook, **2001**, p96.

19. Jayaram, B., Latha, N., Jain, T., Sharma, P., Gandhimathi, A. and Pandey, V.S., S*anjeevini:* A Comprehensive Active-Site Directed Lead Design Software. *Ind. J. Chem.-A,* **2006**, 45A, 1834-1837.

20. Shaikh, S.A. and Jayaram, B., A Swift All-atom Energy based Computational Protocol to Predict DNA-Drug Binding Affinity and ΔTm. *J. Med. Chem*., **2007**, 50, 2240-2244.

21. Jain, T. and Jayaram, B. An all atom energy based computational protocol for predicting binding affinities of protein-ligand complexes. *FEBS Letters*, **2005**, 579, 6659-6666.

22. Jain,T. and Jayaram,B. A computational protocol for predicting the binding affinities of zinc containing metalloprotein-ligand complexes. *Proteins*, **2007**, 67, 1167-1178.

23. Paolo, Saviotti, P., Marie-Angele deLooze, Michelland, S., and Catherine,D., *Nature Biotechnology,* **2000**, 18, 1247 - 1249.

24. Chong L et al., Whole-istic Biology. *Science*, **2002**, 295, 1661.

25. Philip E. Bourne, *Structural Bioinformatics***,** John Wiley & Sons; 2002.

26. Westhead, D.R., Parish, J. H., Twyman, R.M. *Instant Notes in Bioinformatics***,** Bios Scientific Pub Ltd, 2002.

27. Des Higgins, Willie Taylor (eds.), Bioinformatics: Sequence, structure and databanks, Oxford University Press, 2000.

28. Hooman H. Rashidi, Lukas K. Buehler, Bioinformatics Basics: Applications in Biological Science and Medicine, CRC Press, 2000.

29. Jin Xiong, Essential Bioinformatics, Cambridge University Press, 2006.

30. Krane, D.E., Raymer, M.L., Marieb, E.N *Fundamental Concepts of Bioinformatics*, Benjamin/Cummings, 2002.

31. Andreas D. Baxevanis and B. F. F. Ouellette (eds.), Bioinformatics: a Practical Guide to the Analysis of Genes and Proteins, Wiley Interscience, 1998.

32. Arthur, M. Lesk, *Introduction to Bioinformatics***,** Oxford University Press, 2002.

33. Stephen, A.K., and David, D.W. *Introduction to Bioinformatics: A Theoretical and Practical Approach*, Humana Press, 2002.

34. Pierre Baldi and Soren Brunak, Bioinformatics The Machine Learning Approach, The MIT Press, 2001.

35. Hans-Werner Mewes, H., Seidel, B., Weiss, U., Karrenberg, A., Bioinformatics *and Genome Analysis*, Springer Verlag, 2002.

36. Sensen, C. W. Essentials *of Genomics and Bioinformatics***,** John Wiley & Sons, 2002.

37. Thomas, Lengauer., *Bioinformatics: From Genomes to Drugs*, John Wiley & Sons, 2001.

38. Stephen Misener and Stephen A. Krawetz (Eds.), Bioinformatics Methods and Protocols, Humana Press, 2001.

39. Mount, D.W. *Bioinformatics: Sequence and Genome Analysis*, Cold Spring Harbor Laboratory, 2001.

40. Cynthia Gibas, Per Jambeck, *Developing Bioinformatics Computer Skills***,** O'Reilly & Associates, 2001.

41. Zoe Lacroix and T. Critchlow, Bioinformatics: Managing Scientific Data, Morgan Kauffman Publishers, 2003.

42. Gary Benson, Roderic Page (Eds.), *Algorithms in Bioinformatics*, Springer International Edition, 2004.

43. Rastogi S.C, Mendiratta N, Rastogi P, *Bioinformatics Methods and Applications*, Prentice Hall of India Pvt Ltd, 2004.
44. Gautham, N. *Databases and Algorithms*, Narosa Publishing House, 2006.