BIOINFORMATICS FOR BETTER TOMORROW

B. Jayaram, N. Latha, Pooja Narang, Pankaj Sharma, Surojit Bose, Tarun Jain, Kumkum Bhushan, Saher Afshan Shaikh, Poonam Singhal, Gandhimathi and Vidhu Pandey

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.org

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).

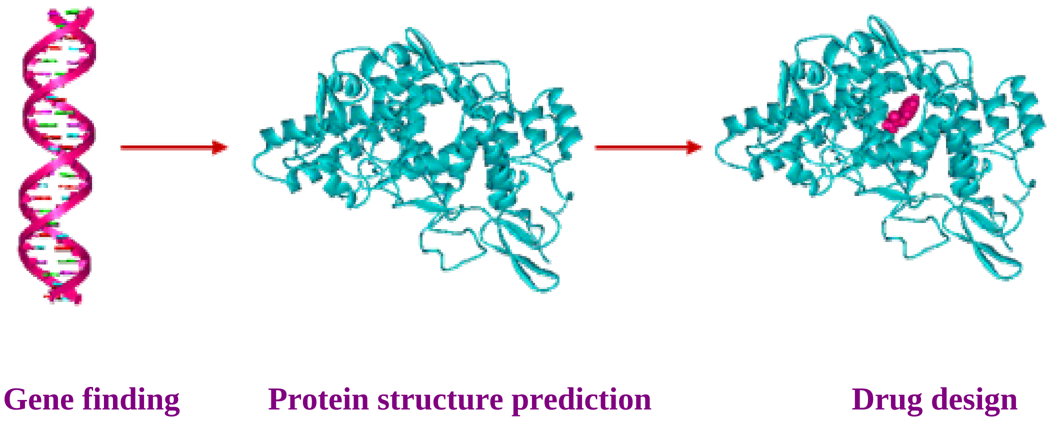


Figure 1 : Some major areas of research in bioinformatics and computational biology.

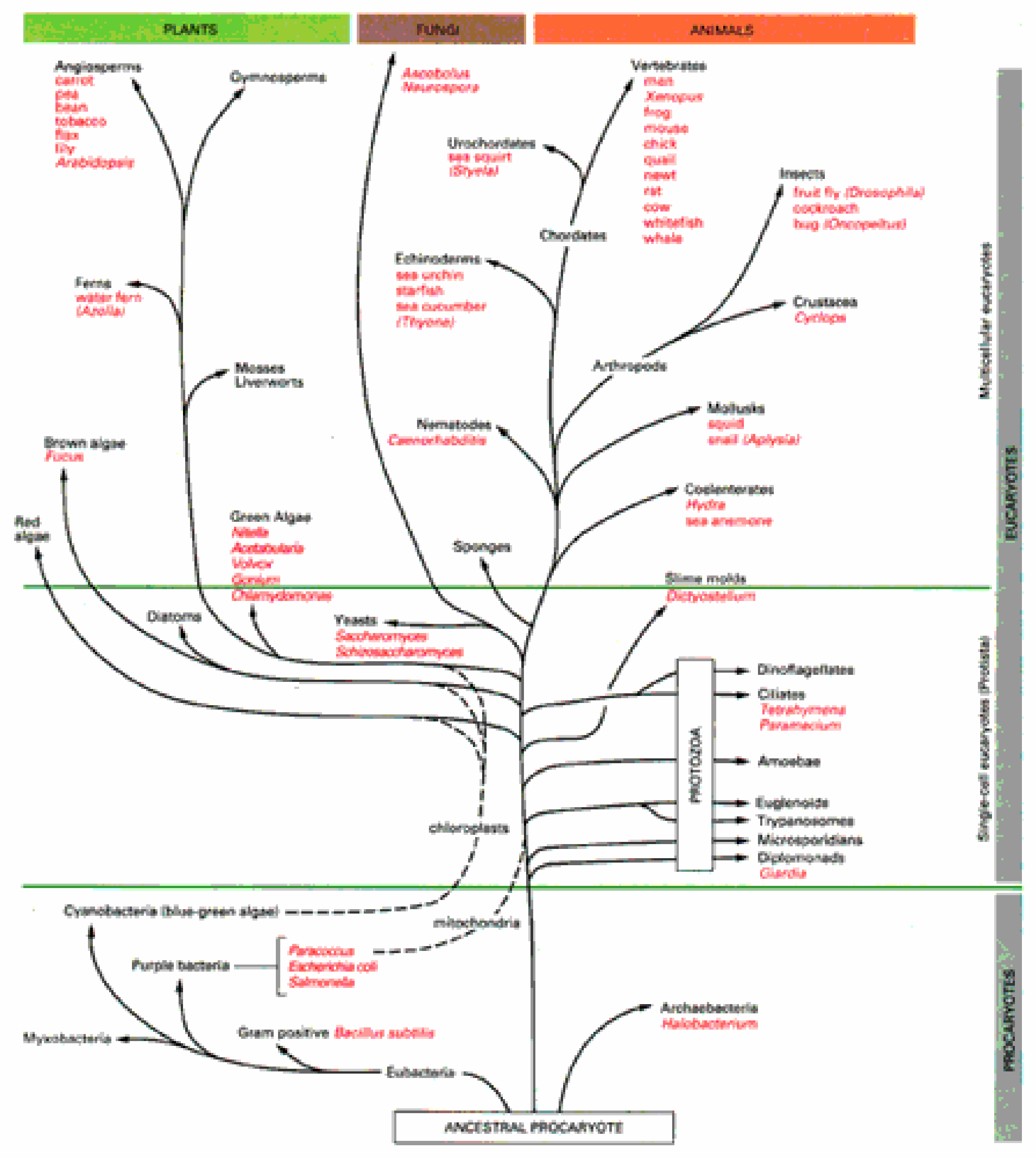
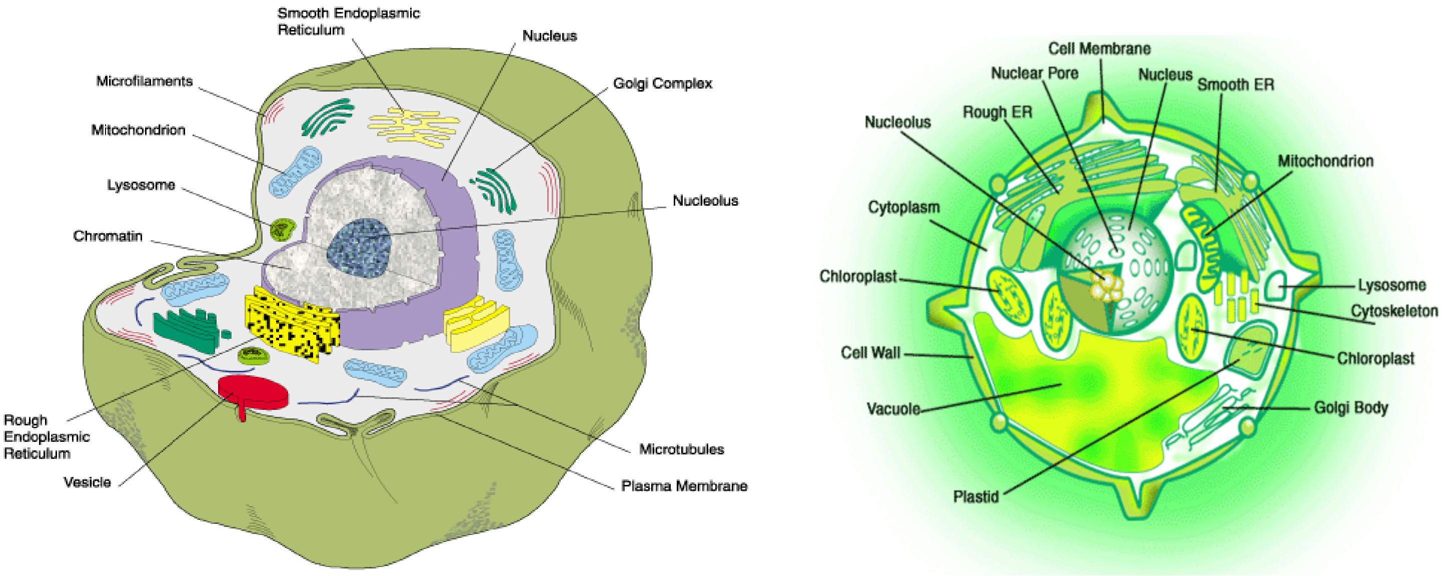


Figure 2 : The tree of life depicting evolutionary relationships among organisms from the major biological kingdoms. A possible evolutionary path from a common ancestral cell to the diverse species present in the modern world can be deduced from DNA sequence analysis. The branches of the evolutionary 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. (Source : 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). Animal cell **(B). Plant cell**

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 (deoxyribo nucleic 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).

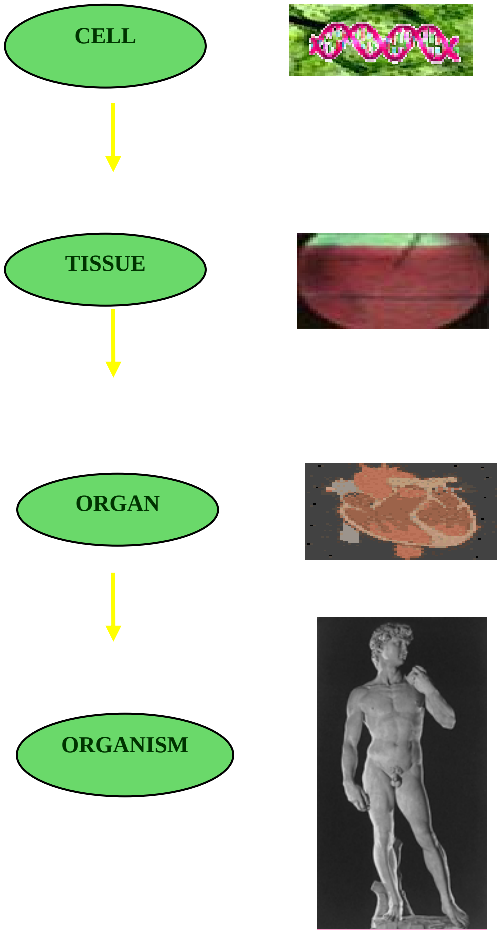


Figure 4 : Levels of organization. The entire DNA content in a cell is called the genome. The entire protein content in a cell is called the proteome. Cellome is the entire complement of molecules, including genome and proteome within a cell. Tissues are made of collections of cells. Tissue collections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

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

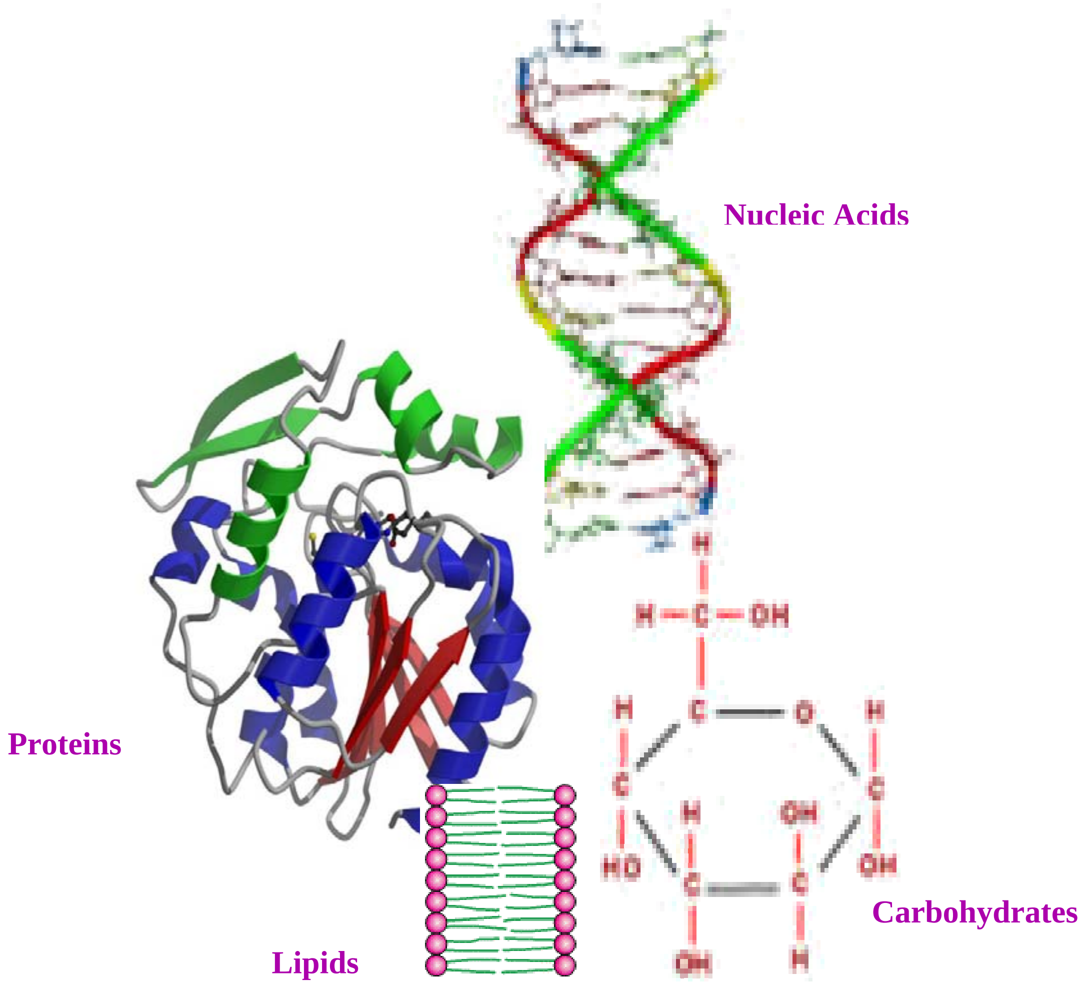


Figure 5 : Biomolecules of life

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 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.

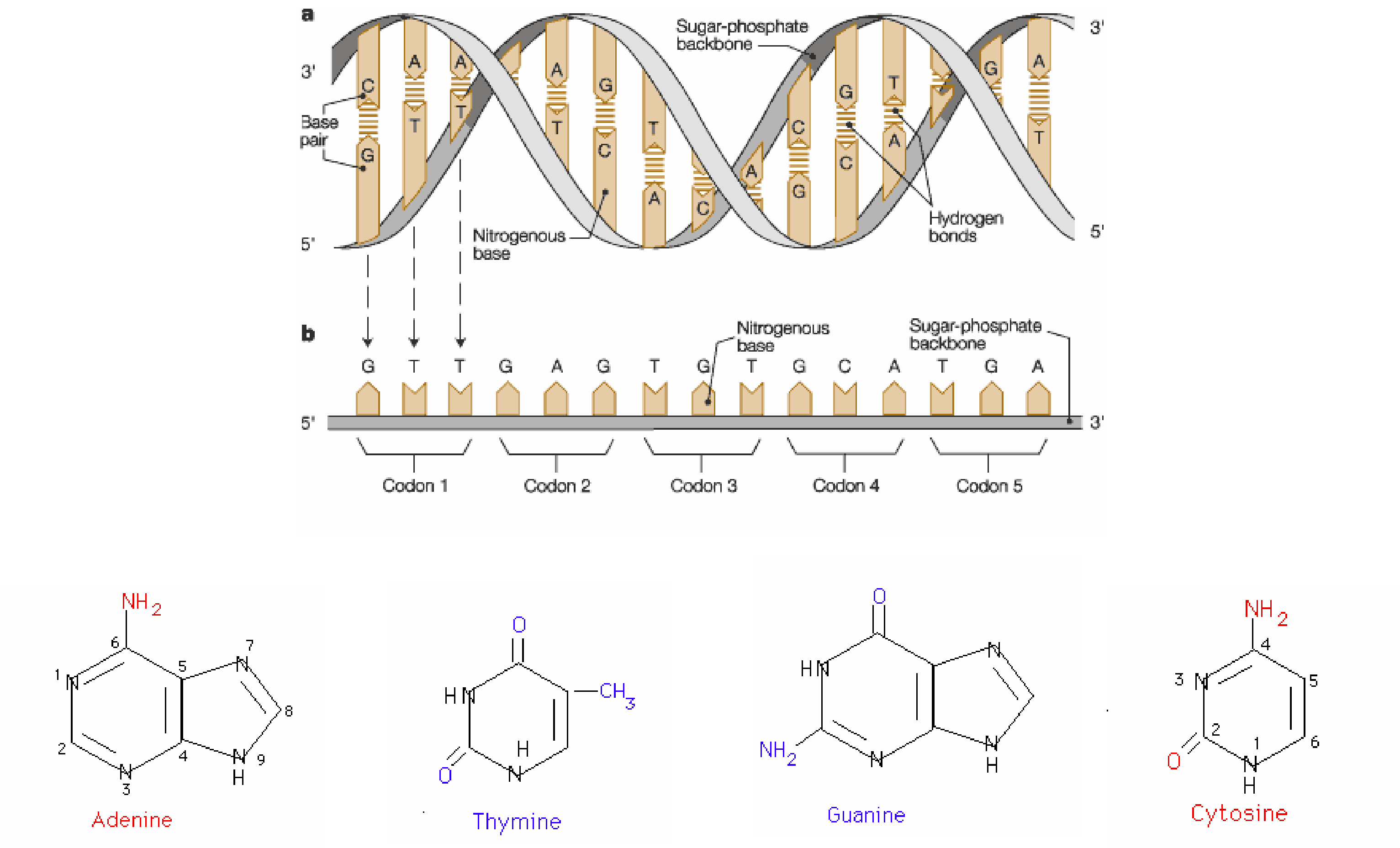


Figure 6 : DNA and its alphabets A, G, C & T the nucleic acid bases. (Source: Molecular biology of the cell, Garland Publishing. Inc. New York, 1994)

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. 7). Potential drugs can bind to DNA, RNA or proteins to suppress or enhance the action at any stage in the pathway.

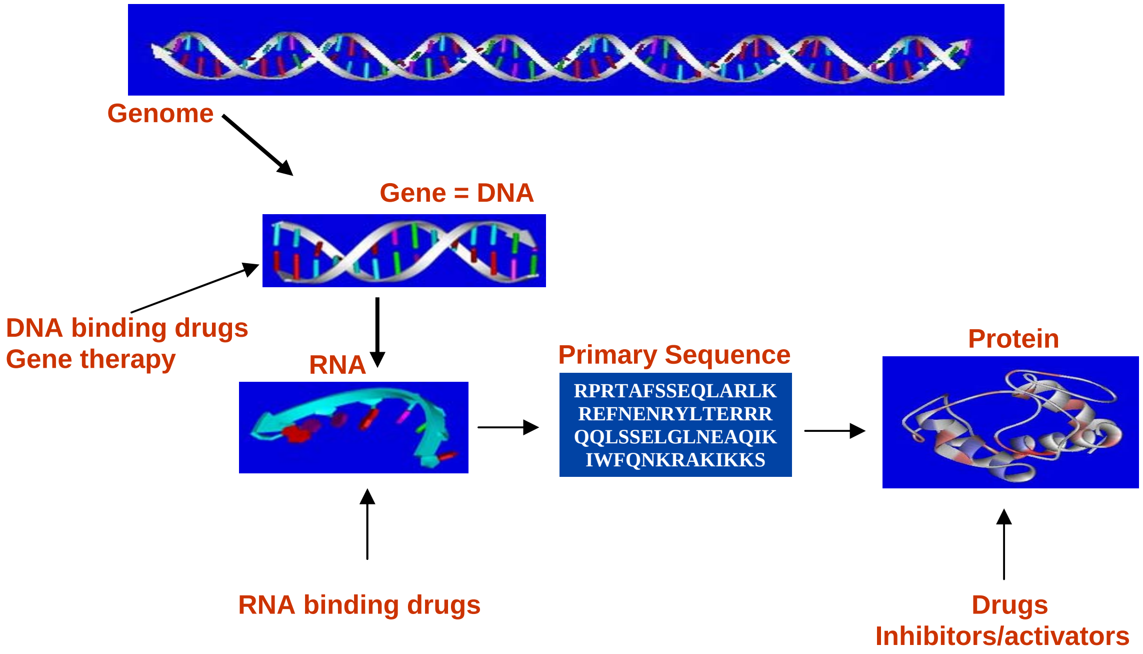


Figure 7 : Central dogma of modern biology

Genome Analysis

Segments of genome called genes 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.

Scores of genome projects are being carried out world wide in order to identify all the genes and ascertain their functions in a specified organism. **Human genome project** 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 genome sizes of different organisms (Table 1) raises questions like what types of genetic modifications are responsible for the four times large genome size of wheat plant and seven times small size of the rice plant as compared to that of humans. Mice and humans contain roughly the same number of genes – about 28K protein coding regions. The chimp and human genomes vary by an average of just 2% i.e. just about 160 enzymes.

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

Organism	***Genome size (Mb)*** *(Mb=Mega base)*
Eschericia coli	4.64
M tuberculosis	4.4
H.Influenza	1.83
Homo sapiens (humans)	3300
Mouse	3000
Rice	430
Wheat	13500

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

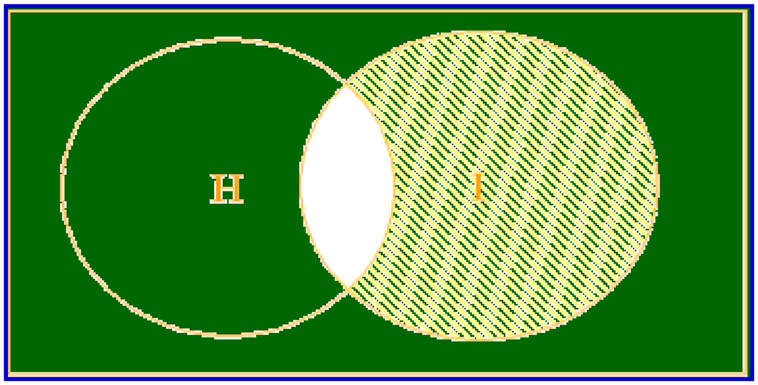
Table 2 : **Specific genetic disorders**

Genetic Disorder	***Reason***
Huntington's Disease	Excessive repeats of a three-base sequence, "CAG" on chromosome 4.
Parkinson's Disease	Variations in genes on chromosomes 4,6.
Sickle Cell Disease	Mutation in hemoglobin-b gene on chromosome 11
Tay-Sachs Disease	Controlled by a pair of genes on chromosome 15
Cystic Fibrosis	Mutations in a single (CFTR) gene
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.
Asthma	Disfunctioning of genes on chromosome 5,6,11,14&12.
Rett Syndrome	Disfunctioning of a gene on the X chromosome.
Brukitt lymphoma	Translocations on chromosome 8
Alzheimer disease	Mutations on four genes located on chromosome 1, 14, 19 & 21.
Werner Syndrome	Mutations on genes located on chromosome 8.
Angelman Syndrome	Deletion of a segment on maternally derived chromosome 15.
Best disease	Mutation in one copy of a gene located on chromosome 11.
Pendred Syndrome	Defective gene on chromosome 7.
Diastrophic dysplasia	Mutation in a gene on chromosome 5

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

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 cause for such disorders. Why nature carries such disorders and how to prevent these are some of the areas where extensive research endeavor has to be invested.

An understanding of the genome organization can lead to concomitant progresses in drug-target identification. Comparative genomics has become a very important emerging branch with tremendous scope, for the above mentioned reasons. 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 (Fig. 8).



Potential Areas for Drug Targets = $H^c \cap I$
H = Human Genome / Proteome (Healthy Individual)
I = Genome / Proteome of the Invader / Pathogen

Figure 8 : Comparative genomics for drug target identification – an application of Bioinformatics

Over the past two decades genetic modification has enabled plant breeders to develop new varieties of crops like cereals, soya, maize at a faster rate. Some of these called as transgenic varieties have been engineered to possess special characteristics that make them better. Recently efforts are on in the area of utilizing GM (genetically modified) crops to produce therapeutic plants.

Another area where bioinformatics techniques can play an important role is in SNP (Single nucleotide polymorphisms) discovery and analysis. SNPs are common DNA sequence variations that occur when a single nucleotide in the genome sequence is changed. SNP’s 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.

Given the whole genome of an organism, finding the genes (gene annotation) is a challenging task. Various approaches have been utilized by different groups for this. These approaches are typically database driven which rely on the already known information. The Markov model based methods utilize short range correlations between bases along the genome. Other methods based on Fourier transform techniques emphasize global correlations.

At IIT Delhi, we are attempting to develop a hypothesis driven physico-chemical model for genome analysis and for distinguishing genes from non-genes. In this model, a vector which attempts to capture forces responsible for DNA structure and protein-DNA interactions walks along the length of the genome distinguishing genes from non-genes. As of now, the physico-chemical model *(ChemGene 1.0)* is able to distinguish successfully genes from non-genes in 120 prokaryotic genomes with fairly good specificities and sensitivities.

Protein Folding

Proteins are polymers of amino acids with unlimited potential variety of structures and metabolic activities. Each protein possesses a characteristic three dimensional shape called its conformation. Sequence of amino acids of protein determines its shape and function. 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. Study of homologous proteins from different species is of interest in constructing taxonomic relationships. Proteins may be classified into structural proteins, enzymes, hormones, transport proteins, protective proteins, contractile proteins, storage proteins, toxins etc..

Prof G N Ramachandran, an Indian scientist made some significant contributions towards an understanding of the secondary structure of proteins. His fundamental work in this area is remembered in the form of Ramachandran maps.

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 relative to the unfolded state. 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 timescales, is the problem pending a resolution for over five decades. For a 200 amino acid protein with just two conformations per amino acid, 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. 9).

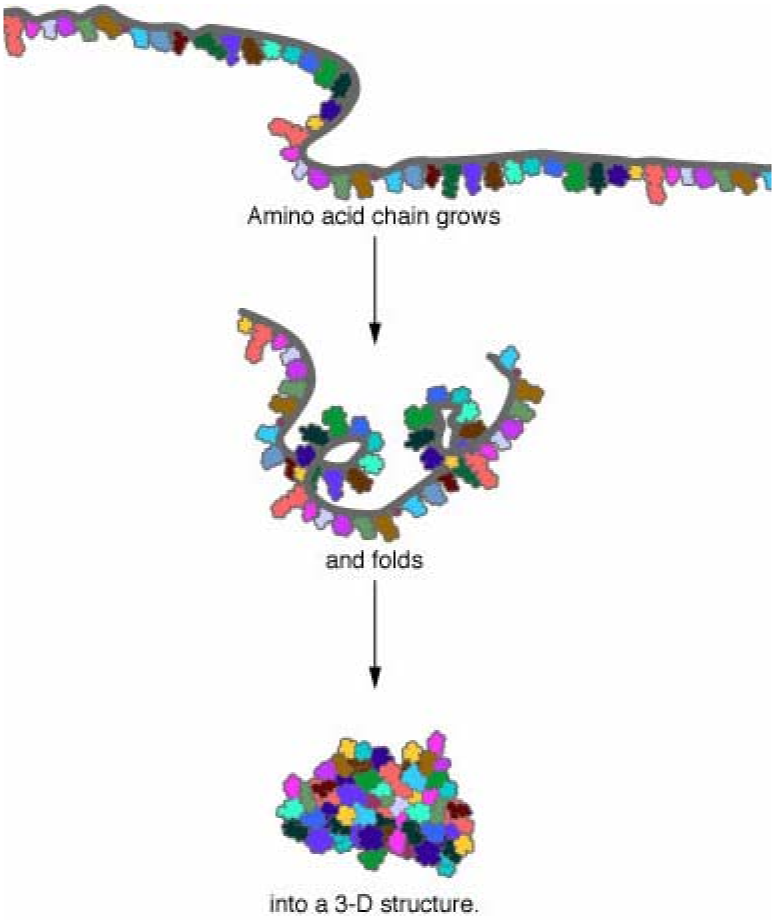


Figure 9 : The protein folding problem (Source: press2.nci.nih.gov/.../ snps_cancer21.htm)

Solution to the protein folding problem 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 detergent industry are among the several potential beneficiaries. Other important implications are in structure based drug discovery wherein the structure of the drug target is vital (Fig. 10). Structures of receptors – a major of class of drug targets - are refractory to experimental techniques thus leaving the fold open to computer modeling.

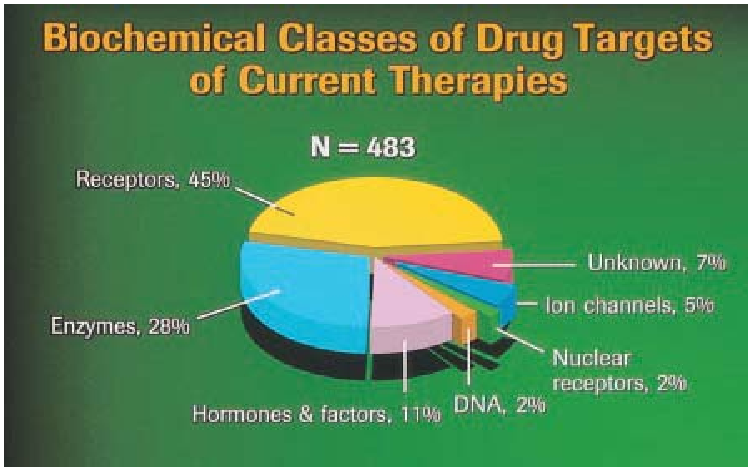


Figure 10 : Drug targets (Source: Drews J., Drug Discovery: A historical perspective, Science, 2000, 287: 1960-1964)

There is an urgent demand for faster and better algorithms for protein structure prediction. There are three major ways in which protein structure prediction attempts are currently progressing viz. homology modeling, fold recognition and *ab initio* approaches. Whereas the first two approaches are database dependent relying on known structures and folds of proteins, the third one is independent of the databases and starts from the physical principles. Although the *ab initio* techniques till date are able to predict only small proteins, because of their first principle approach, they have the potential to predict new / novel folds and structures. On account of the complexity of the problem, computational requirement for such a prediction is a major issue which can easily be appreciated by the estimates of 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. This will require approximately a million years to fold a single protein. If one can envision a million processors working together, a single protein can be folded in one year computer time. In this spirit, the IBM has launched a five year *Blue Gene* project in the year 1999 to address the complex biomolecular phenomena such as protein folding.

At IIT Delhi, we are developing a computational protocol for modeling and predicting protein structures of small globular proteins. Here a combination of bioinformatics tools, physicochemical properties of proteins and *ab initio* approaches are used. Starting with the sequence of amino acids, for ten small alpha helical proteins, structures to within 3-5Å of the native are predicted within a day on a 50 ultra sparc III Cu 900 MHz processor cluster. Attempts are on to further bring the structures to within

<3Å of the native structures via molecular dynamics and Metropolis Monte Carlo simulations.

Drug Design

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. 11).

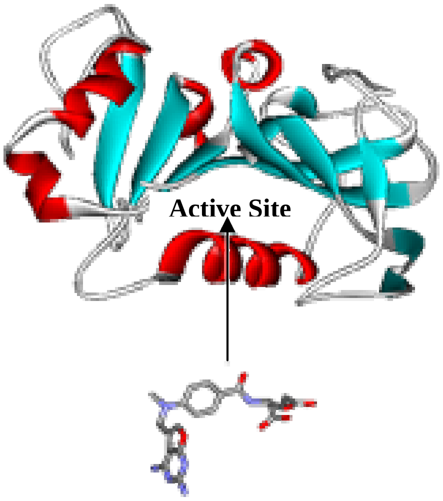


Figure 11 : Active-site directed drug-design

Also to combat life-threatening diseases such as AIDS, Tuberculosis, Malaria etc., a global push is essential (Table 3). Millions for Viagra and pennies for the diseases of the poor is the current situation of investment in Pharma R&D.

Table 3 : WHO Calls for Global Push Against AIDS & Tuberculosis & Malaria

Nearly 6 million die each year due to diseases like diarrhoea, small pox, polio, river blindness, leprosy.
Estimated cost $ 12 billion to fight the disease of poverty
(AIDS medication about $15K per annum)

A new economic analysis
Healthy people can get themselves out of poverty.
Relieving a population of burden of the diseases for 15 to 20 years will give a huge boost to economic development.

Millions for Viagra, Pennies for the Diseases of the Poor
Of all new medications brought to the market (1223) by Multinationals from 1975 only 1% (13) are for tropical diseases plaguing the third world.

Life style drugs dominate Pharma R&D

(1) Toe nail Fungus	**(2) Obesity**
(3) Baldness	**(4) Face Wrinkle**
(5) Erectile Dysfunction	**(6) Separation anxiety of dogs etc.**

(Source : www.globalhealth.org, www.thenation.com)

Time and cost required for designing a new drug are immense and at an unacceptable level. According to some estimates it costs about $880 million and 14 years of research to develop a new drug before it is introduced in the market (Table 4).

Table 4 : **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 **$880million**

(Source : PAREXEL, PAREXEL's Pharmaceutical R&D Stastical Sourcebook, 2001, p96)

Intervention of computers at some plausible steps is imperative to bring down the cost and time required in the drug discovery process (Fig. 12, Table 5).

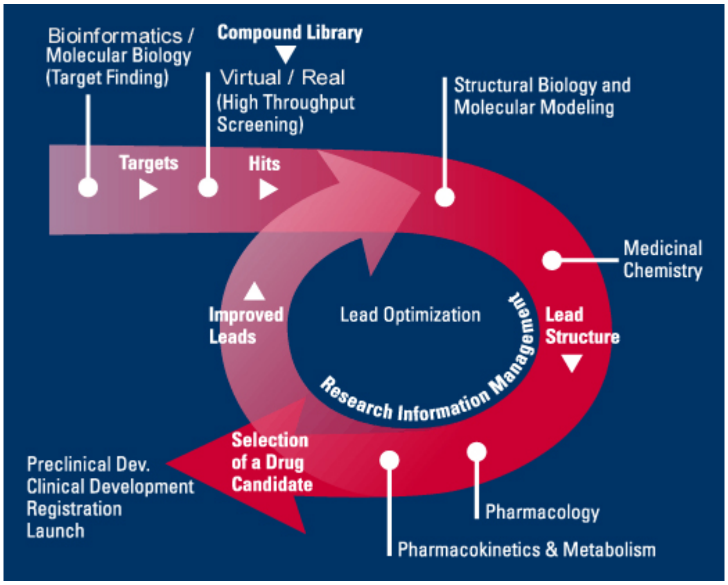

Bioinformatics / Molecular Biology (Target Finding)
Compound Library
Virtual / Real (High Throughput Screening)
Structural Biology and Molecular Modeling
Targets
Hits
Medicinal Chemistry
Improved Leads
Lead Optimization
Lead Structure
Research Information Management
Selection of a Drug Candidate
Preclinical Dev.
Clinical Development
Registration
Launch
Pharmacology
Pharmacokinetics & Metabolism

Figure 12 : Potential areas for *In silico* intervention in drug-discovery process

Table 5 : **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 Rigid ligand/target	SPECITOPE LUDI CLIX	140,000 30,000 30,000	~1 hour 1-4 hours 33 hours
Molecular Mechanics Partially flexible ligand Rigid target	Hammerhead DOCK DOCK	80,000 17,000 53,000	3-4 days 3-4 days 14 days
Molecular Mechanics Fully flexible ligand Rigid target	ICM	100,000	~1 year (extrapolated)
Molecular Mechanics Free energy perturbation	AMBER CHARMM	1	~several days
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 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 etc..

Pursuing the dream that once the gene target is identified and validated drug discovery protocols could be automated using Bioinformatics & Computational Biology tools, at IIT Delhi we have developed a computational protocol for active site directed drug design. The suite of programs (christened "*Sanjeevini*") has the potential to evaluate and /or generate lead-like molecules for any biological target. The various modules of this suite are designed to ensure reliability and generality. The software is currently being optimized on Linux and Sun clusters for faster and better results. Making a drug is more like designing an adaptable key for a dynamic lock. The *Sanjeevini* methodology consists of design of a library of templates, generation of candidate inhibitors, screening candidates via drug-like filters, parameter derivation via quantum mechanical calculations for energy evaluations, Monte Carlo docking and binding affinity estimates based on *post facto* analyses of all atom molecular dynamics trajectories. The protocols tested on Cyclooxygenase-2 as a target could successfully distinguish NSAIDs (Nonsteroidal Anti-inflammatory drugs) from non-drugs (Fig 13). Validation on other targets is in progress.

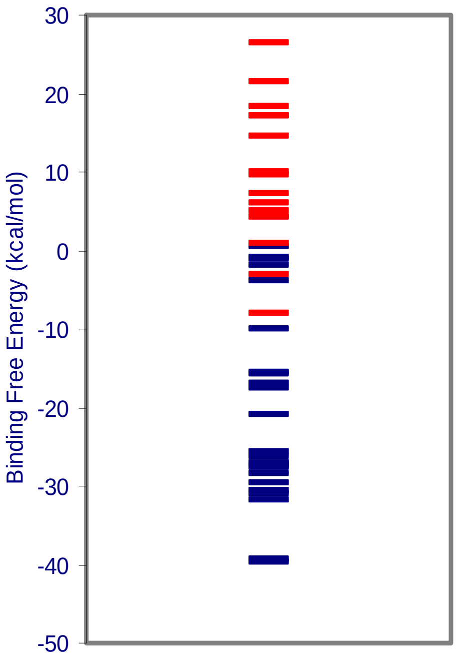
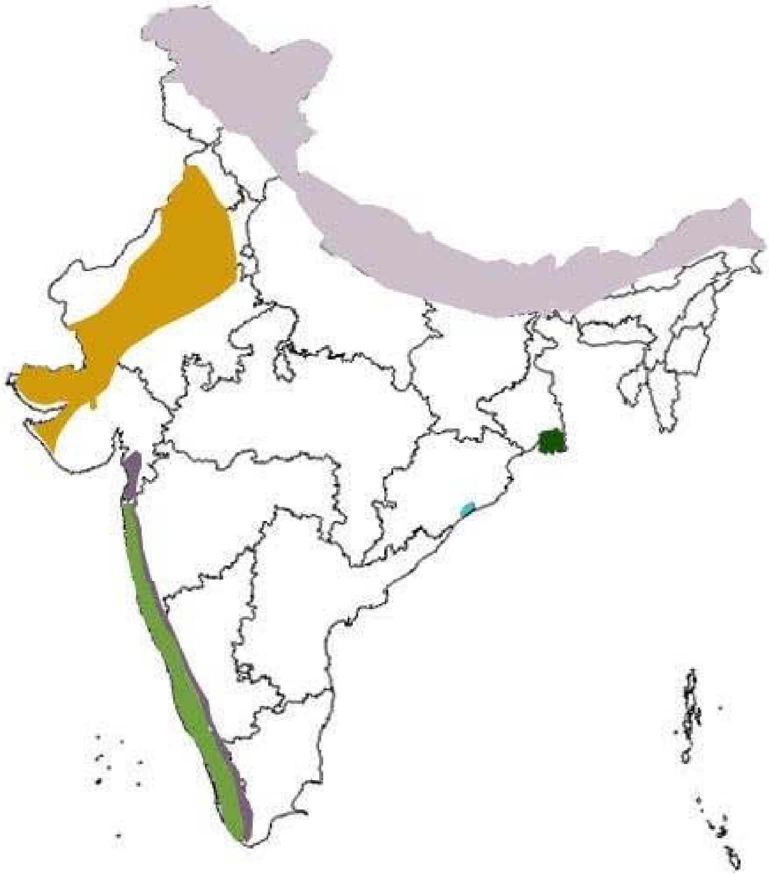


Figure 13 : The active site directed lead design protocols developed from first principles and implemented on a supercomputer could segregate NSAIDS (blue) from Non-drugs (red).

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. 14 and Fig15).



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

Figure 15: Biodiversity bioinformatics is essential to preserve the natural balance of flora and fauna on the planet and to prevent extinction of species (Source: www.hku.hk/ecology/envsci.htm).

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 appears promising. The projections of the growth potential in India in a global scenario however belie these expectations (Fig. 16).

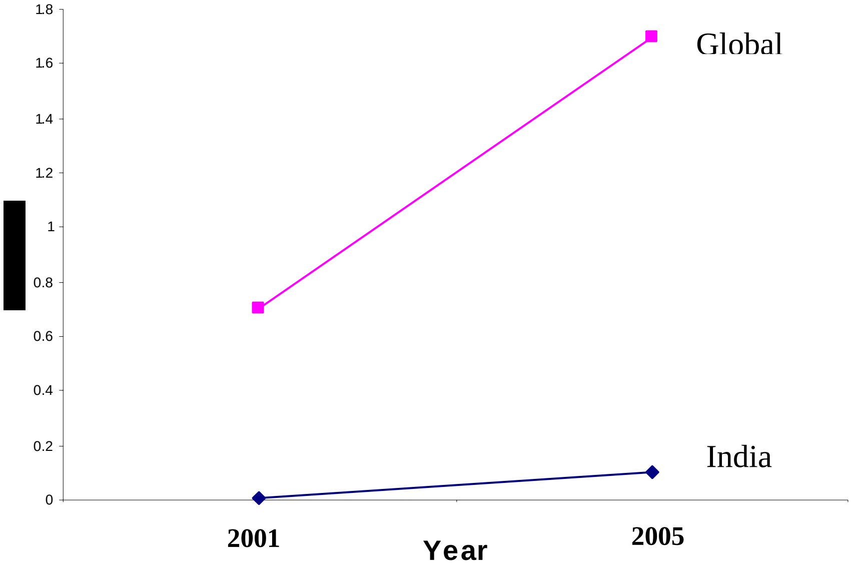


Figure 16 : Growth potential for Bioinformatics based business opportunities in India according to IDC (International Data Corporation, India.)

Creating the necessary infrastructure around specialists to meet the high computational power required for these activities (Table 6, Fig. 17) is one step in the right direction.

Table 6 : Genome to drug discovery research: A rough estimate of computational requirements

1. Gene Prediction	
Homology/string comparison.	300 Giga flop ~ $3*10^{9}$ bp
2. Protein Structure Prediction	
- Threading	100 Giga flop
- Statistical Models	
- Filters to reduce plausible structures	
Molecular Dynamics	
100 structures	30 Peta flop
1-ns simulation for structure refinement	
Total Compute Time 5000ns	
Number of atoms per simulation 25000	
3. Active site directed drug design	
Scan 1000 drug molecules/protein	18 Peta flop
3ns simulation per drug molecule	
(Active site searches, docking, rate and affinity determinations etc.)	
Total Compute Time 3000ns	
25000 atoms per simulation	
Summary	
Total Computational requirement to design one lead compound from genome	
~ 50 Peta flop ($5.x10^{16}$ floating point operations)	
To design ten lead compounds per day (on a dedicated machine)	
the requirement is	**5.8 tera flops capacity.**
(Out of every 100 lead compounds, only one may become a drug, which further increases the computer requirements)	

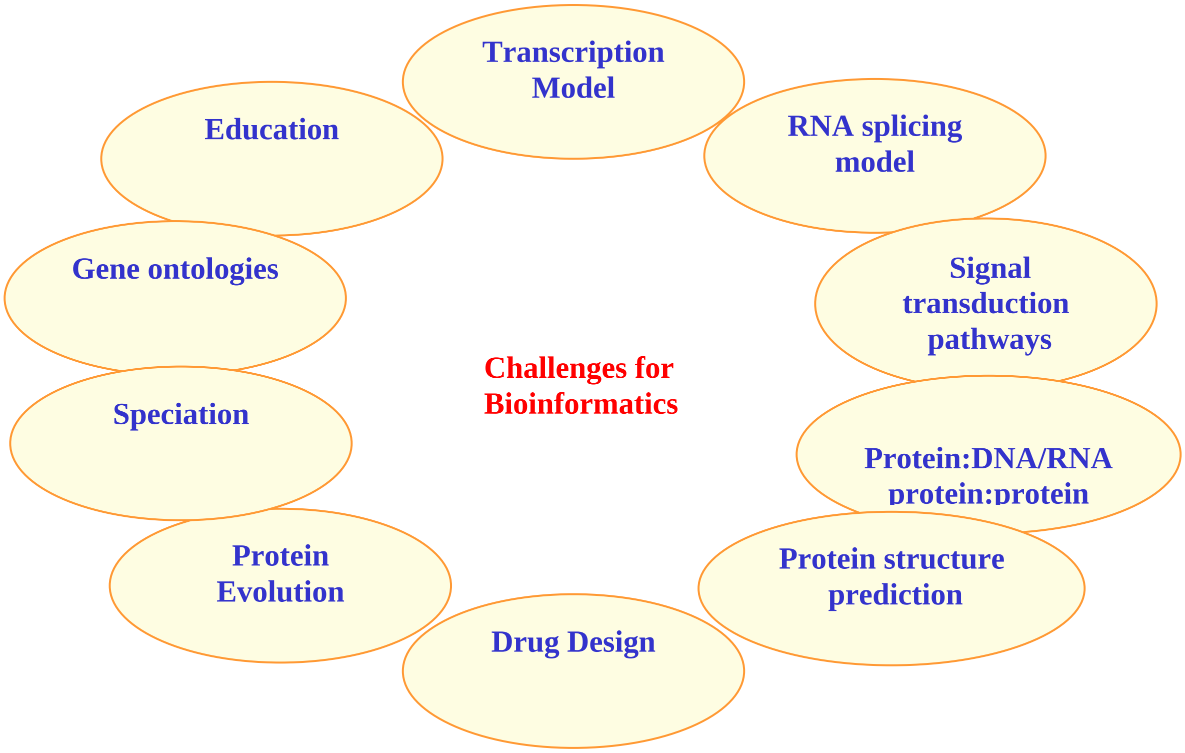

Transcription Model
Education
RNA splicing model
Gene ontologies
Signal transduction pathways
Challenges for Bioinformatics
Speciation
Protein:DNA/RNA protein:protein
Protein Evolution
Protein structure prediction
Drug Design

Figure 17 : Challenges for bioinformatics

To enable access by the student and scientific community from across the country, the Supercomputing Facility at IIT Delhi is connected to Biogrid-India (a Virtual Private Network of the Department of Biotechnology, Govt of India). It is envisioned that the Biogrid will span all the 60+ Bioinformatics Centres of the Department of Biotechnology with GBPS bandwidth and at least 10 teraflops of compute capacity soon (Figure 18).

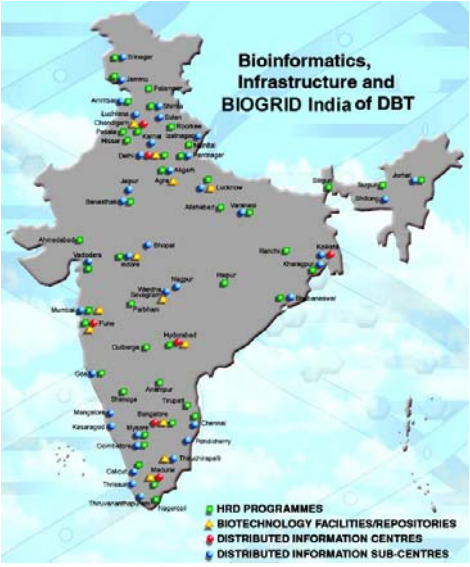


Figure 18 : Biogrid-India

Pro-active policies and conducive environment to researchers, entrepreneurs and Industry will go a long way in steering India towards leadership in the area of Bioinformatics and Computational Biology.

Acknowledgements.

BJ is thankful to the Department of Biotechnology, the Department of Science & Technology, Council of Scientific & Industrial Research, Indo-French Centre for the Promotion of Advanced Research and IIT Delhi administration for support over the years.