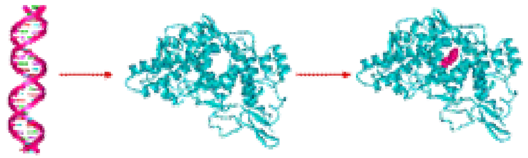
Supercomputing Facility for Bioinformatics & Computational Biology IITD

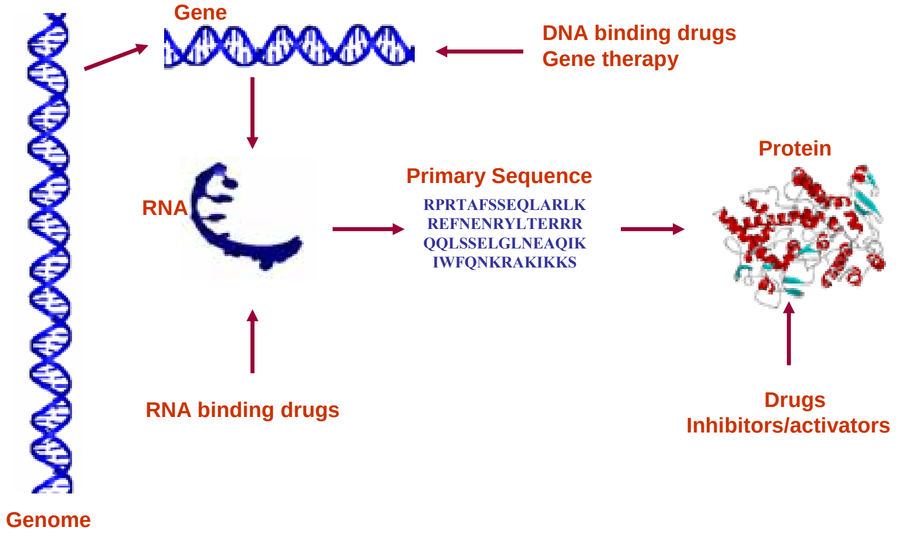


From Gene to Drug *in Silico*
Bioinformatics for A Better Tomorrow

B. Jayaram
Department of Chemistry &
Supercomputing Facility for Bioinformatics & Computational Biology
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi-110016, India
www.scfbio-iitd.res.in

Supercomputing Facility for Bioinformatics & Computational Biology IITD

The Central Dogma of Modern Drug Discovery



Gene

DNA binding drugs
Gene therapy

RNA

Primary Sequence

RPRTAFSSEQLARLK
REFNENRYLTERRR
QQLSSELGLNEAQIK
IWFQNKRAKIKKS

Protein

RNA binding drugs

Drugs
Inhibitors/activators

Genome

Bioinformatics

Bioinformatics is an emerging interdisciplinary area of Science & Technology encompassing a systematic development and application of IT solutions to biological data.

Bioinformatics addresses biological data collection and warehousing, data base searches, analyses and interpretation, modeling and product design.

Bioinformatics 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.

For *Bioinformatics* to evolve as a branch of Science, it must be practised as a Hypothesis driven endeavor with Biological Data providing information for validation, leading to newer hypotheses and discoveries.

Information → Knowledge → Products Useful to Society

Bioinformatics & Agriculture

*** Increasing the nutritional content**

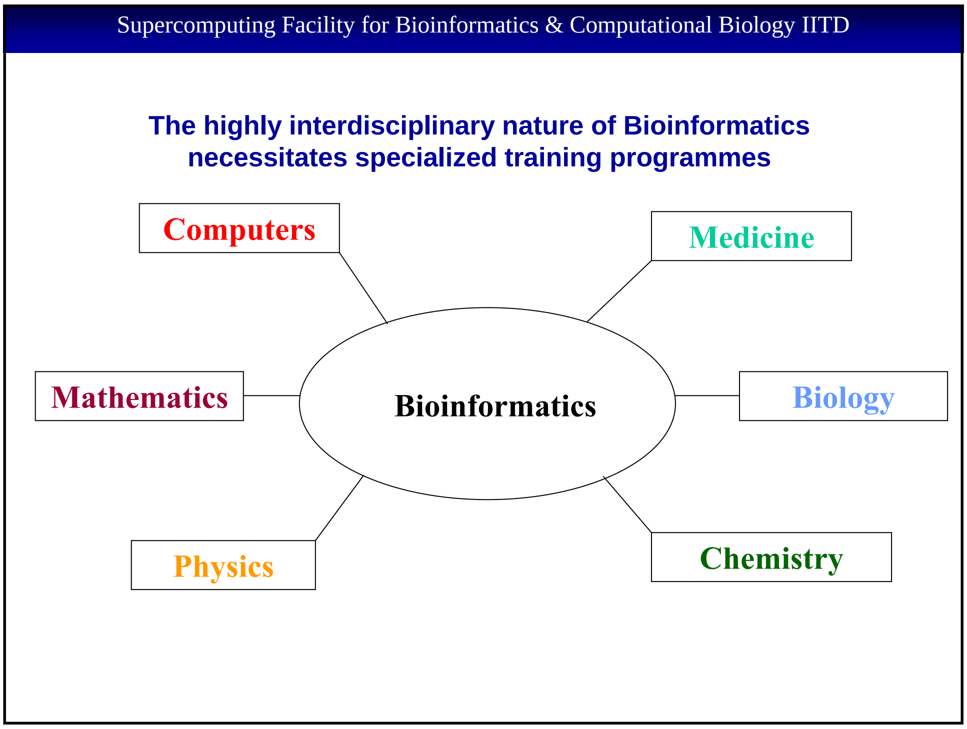
*** Increasing the volume of the agricultural produce &**

*** Implanting disease resistance etc.**

Bioinformatics & Medicine

*** Reducing the cost and time involved in drug discovery**

*** Development of personalized medicine**



Supercomputing Facility for Bioinformatics & Computational Biology IITD

The highly interdisciplinary nature of Bioinformatics necessitates specialized training programmes

Bioinformatics

Computers

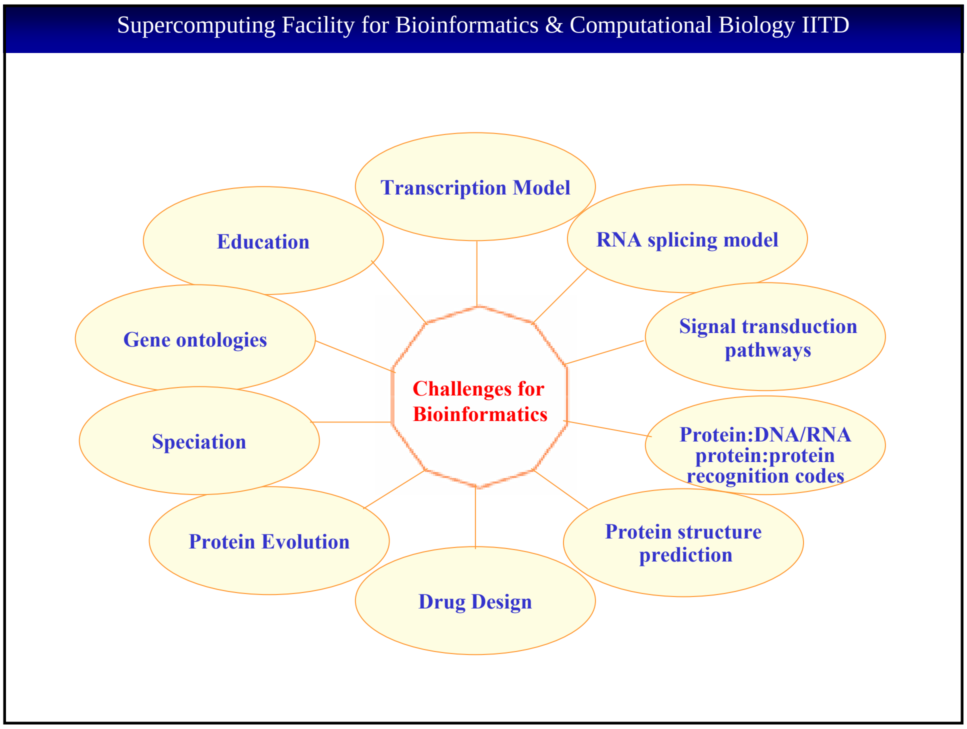
Medicine

Mathematics

Biology

Physics

Chemistry



Supercomputing Facility for Bioinformatics & Computational Biology IITD

Challenges for Bioinformatics

Transcription Model

RNA splicing model

Signal transduction pathways

Protein:DNA/RNA protein:protein recognition codes

Protein structure prediction

Drug Design

Protein Evolution

Speciation

Gene ontologies

Education

Employment Avenues in Bioinformatics

• Pharmaceutical & Biotech. Companies involved in the innovative development of drugs, agricultural products, genetically modified crops, medical and forensic tool kits…

•R&D organizations, academic institutions, software companies & product marketing companies.

•Potential opportunities as entrepreneurs, researchers, software developers, database developers, consultants and trainers.

*•Current Scenario: Supply exceeds demand but **Quality supply is far below demand.***

Bioinformatics & India

•Well-acknowledged IT Skills

• Active Governmental Initiatives, DBT, DST, CSIR, DIT, MHRD

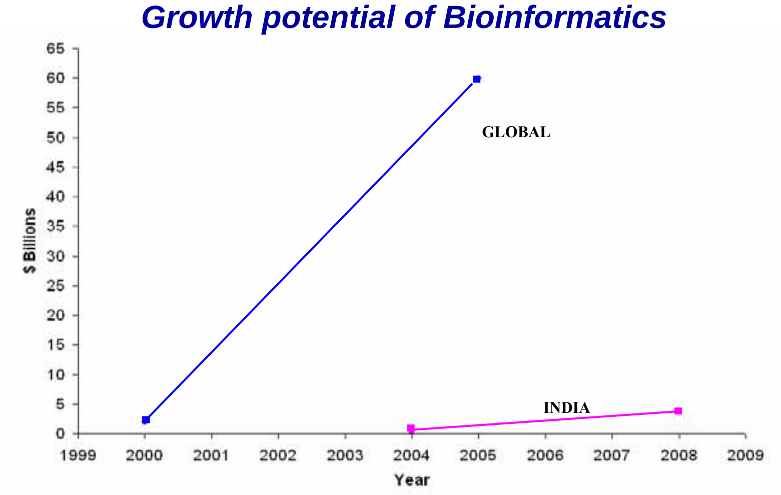
•Changing Process to Product Patent Laws. In-house R&D in Pharma sector eg. at Dabur, Ranbaxy...

• Over 200 Software & Biotech. Indian companies actively involved in related R & D and promotion eg. HCLT, TCS, Wipro, Satyam, Biocon..

•Development of non-profitable yet essential medicines for third world diseases

• Increasing agricultural output to meet the needs of increasing population.

Growth potential of Bioinformatics



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

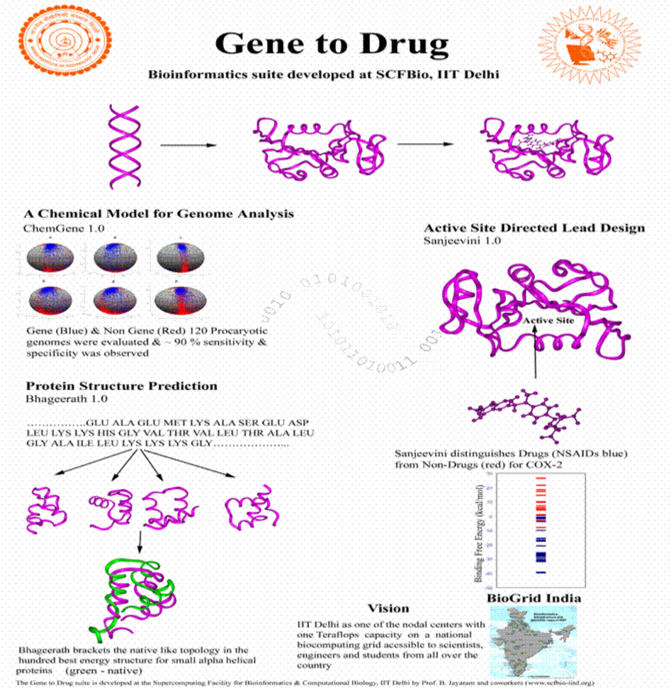
Much more is expected from the world leader in IT.

Major Research Activities in Progress & Bioinformatics Software Suites Developed at SCFBio IIT Delhi

Supercomputing Facility for Bioinformatics & Computational Biology IITD

Research @ SCFBio IIT Delhi

- Gene Evaluation (***ChemGene1.0***)
- Protein Structure Prediction (***Bhageerath1.0***)
- Active Site Directed Lead Design (***Sanjeevini1.0***)
- Biogrid-India



Gene to Drug

Bioinformatics suite developed at SCFBio, IIT Delhi

A Chemical Model for Genome Analysis
ChemGene 1.0

Gene (Blue) & Non Gene (Red) 120 Procaryotic genomes were evaluated & ~ 90 % sensitivity & specificity was observed

Protein Structure Prediction
Bhageerath 1.0

…………….GLU ALA GLU MET LYS ALA SER GLU ASP
LEU LYS LYS HIS GLY VAL THR VAL LEU THR ALA LEU
GLY ALA ILE LEU LYS LYS LYS GLY………………...

Bhageerath brackets the native like topology in the hundred best energy structure for small alpha helical proteins (green - native)

Active Site Directed Lead Design
Sanjeevini 1.0

Sanjeevini distinguishes Drugs (NSAIDs blue) from Non-Drugs (red) for COX-2

BioGrid India

Vision

IIT Delhi as one of the nodal centers with one Teraflops capacity on a national biocomputing grid acessible to scientists, engineers and students from all over the country

The Gene to Drug suite is developed at the Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi by Prof. B. Jayaram and coworkers (www.scfbio-iitd.org)

The Gene to Drug suite is developed at the Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi by Prof. B. Jayaram and coworkers (www.scfbio-iitd.org)

Genomics and Proteomics

The Nucleotide sequence and the corresponding amino acid sequence of Human Insulin (which participates in metabolism of fat and proteins).

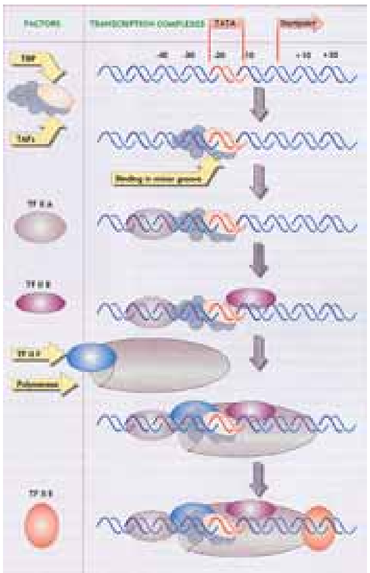
atggccctgtggatgcgcctcctgcccctgctggcgctgctggccctctggggacctgac
M A L W M R L L P L L A L L A L W G P D
ccagccgcagcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctctac
P A A A F V N Q H L C G S H L V E A L Y
ctagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagaggac
L V C G E R G F F Y T P K T R R E A E D
ctgcaggtggggcaggtggagctgggcgggggccctggtgcaggcagcctgcagcccttg
L Q V G Q V E L G G G P G A G S L Q P L
gccctggaggggtccctgcagaagcgtggcattgtggaacaatgctgtaccagcatctgc
A L E G S L Q K R G I V E Q C C T S I C
tccctctaccagctggagaactactgcaactag
S L Y Q L E N Y C N -

A base 'A' is inserted in the above nucleotide sequence as shown below. The protein sequence changes drastically.

atggccctgtggatgcgcctcctgcccctgctggcgctgctggccctctggggacctgac
M A L W M R L L P L L A L L A L W G P D
ccagccgcagAcctttgtgaaccaacacctgtgcggctcacacctggtggaagctctcta
P A A D L C E P T P V R L T P G G S S L
cctagtgtgcggggaacgaggcttcttctacacacccaagacccgccgggaggcagagga
P S V R G T R L L L H T Q D P P G G R G
cctgcaggtggggcaggtggagctgggcgggggccctggtgcaggcagcctgcagccctt
P A G G A G G A G R G P W C R Q P A A L
ggccctggaggggtccctgcagaagcgtggcattgtggaacaatgctgtaccagcatctg
G P G G V P A E A W H C G T M L Y Q H L
ctccctctaccagctggagaactactgcaactag
L P L P A G E L L Q L…….

(Data from Anna Tramontano, "The Ten Most Wanted Solutions in Protein Bioinformatics", Cahpman Hall, 2005, p-2)

A Closer Look at the First Step in Gene Expression: A Complex Process in Eukaryotes



Assembly of RNA Polymerase II Preinitiation Complex.

At a molecular level, gene expression is governed by protein-DNA and protein-protein interactions – the rules of recognition are yet to be deciphered.

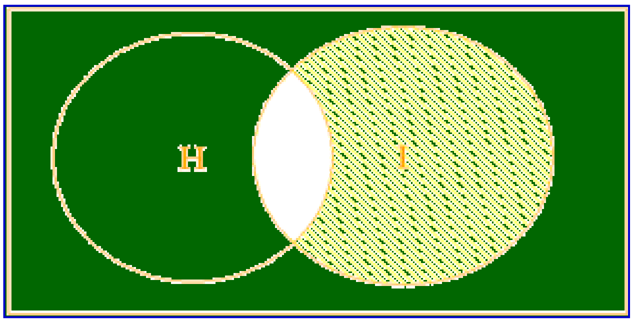
Figure from B. Lewin, "Genes", 1994, Oxford, p-861.

Genome sizes

Organism	*Genome size (Mb)*
Prokaryotes *Eschericia coli*	4.64
M tuberculosis	4.4
Bacillus Subtilisis	4.20
H.Influenza	1.83
Eukaryotes Fungi (yeast)	12.1
Invertebrates Drosophila Melanogaster	140
C Elegans	100
Bombyx Mori (silk worm)	490
Vertebrates Homo sapiens (humans)	3000
Mouse	3300
Plants Rice	565
Maize	5000
Wheat	17000
Pea	4800

Genome is the entire DNA content in a cell of an organism. The data provides a plethora of opportunities to understand creation at a molecular level (Data from : http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/G/GenomeSizes.html)

Comparative Genomics for Drug Target Identification

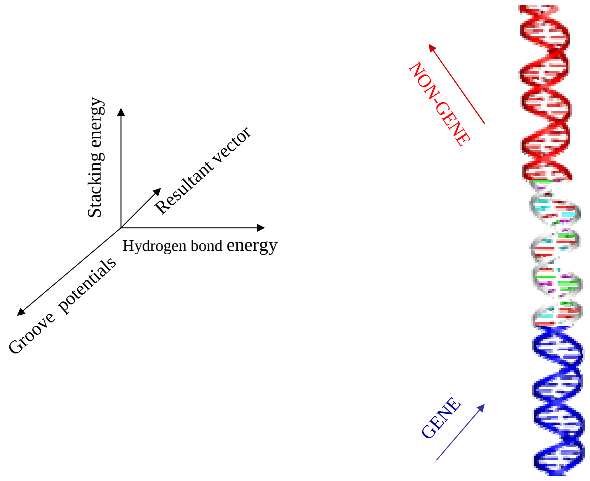


Drug Target = $H^c \cap I$

H = Human Genome / Proteome (Healthy Individual)
I = Genome / Proteome of the Invader / Pathogen

Play it on a PC. It may lead to new discoveries and help Scientists and Society

ChemGene1.0
A Chemical Model to Distinguish Genes from Non-Genes



A Physico-Chemical Model to Analyze DNA Sequences
ChemGene1.0

We constructed a 3-D vector for each codon

•X – Hydrogen bond energy

•Y – Stacking energy

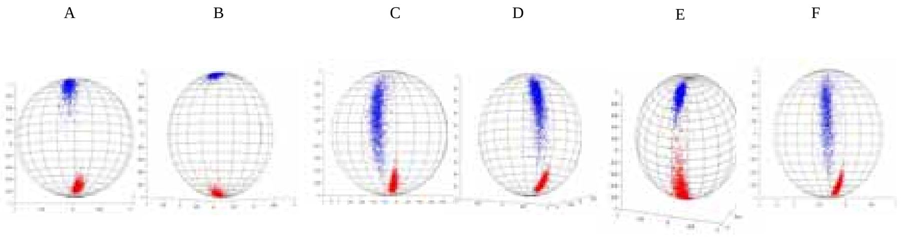
•Z – Groove potentials (Initially trained on a small data set of 1500 genes/shifted-gene pairs. Assignments made to confirm to symmetry & rule of conjugates).

As the 3D vector walks along the genome, the net orientation of the resultant vector is calculated for gene and non-gene regions

“A Physico-Chemical Model for Analyzing DNA Sequences”, Dutta S, Singhal P, Agrawal P, Tomer R, Kritee, Khurana E & Jayaram B, *J. Chem. Inf. Mod.* , 2005, *In Press.*
“Beyond the Wobble: The rule of conjugates”. Jayaram, B., *Journal of Mol Evol.* 1997, 45, 704.

Supercomputing Facility for Bioinformatics & Computational Biology IITD

ChemGene Distinguishes Genes (blue) from Non- Genes (red) in 331 Prokaryotic Genomes



Three dimensional plots of the distributions of gene and non-gene direction vectors for six best (A to F) cases calculated from the genomes of (A) *Agrobacterium tumefaciens* (NC_003304), (B) *Wolinella Succinogenes* (NC_005090), (C) *Rhodopseudomonas palustris* (NC_005296), (D) *Bordetella bronchiseptica* (NC_002927), (E) *Clostridium Acetobutylicium* (NC_003030), (F) *Bordetella Pertusis* (NC_002929)

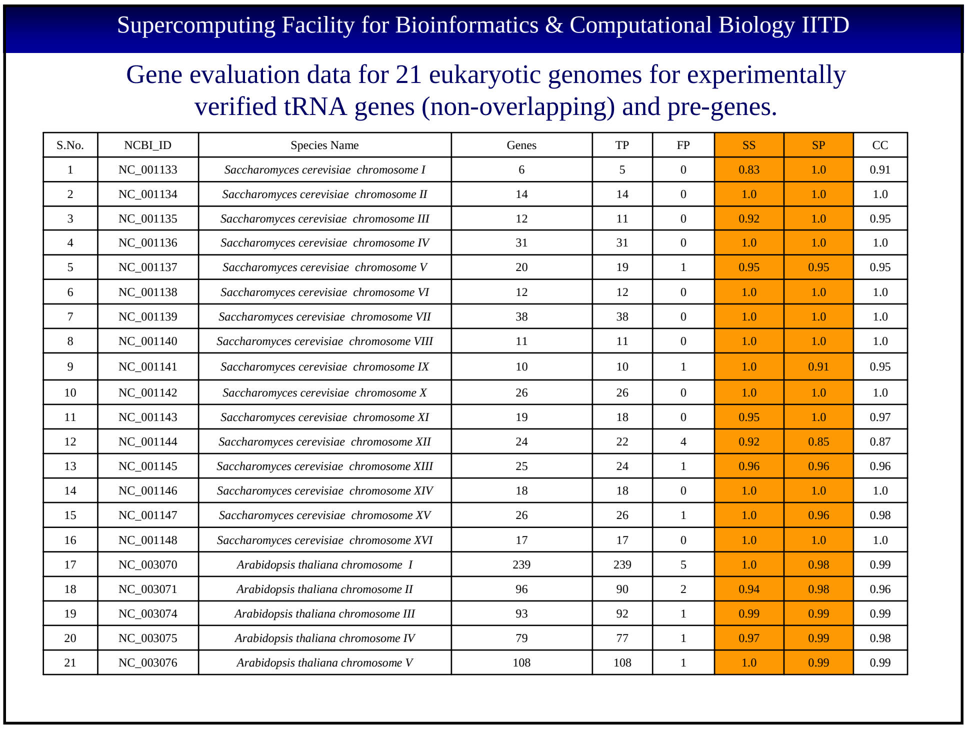
Gene vectors point to the north and the non-gene vectors to the south with >0.85 probability

Supercomputing Facility for Bioinformatics & Computational Biology IITD

Gene evaluation data for prokaryotic genomes for experimentally verified gene (non-overlapping) and non-genes

S.No.	NCBI_ID	Species Name	Genes	TP#	FP #	SS#	SP#	CC#
1	NC_000117	*Chlamydia trachomatis*	463	458	4	0.98	0.99	0.98
2	NC_000853	*Thermotoga maritima MSB8*	641	619	3	0.96	0.99	0.96
3	NC_000854	*Aeropyrum pernix K1*	561	532	7	0.94	0.98	0.93
4	NC_000868	*Pyrococcus abyssi GE5*	632	630	241	0.99	0.63	0.49
5	NC_000907	*Haemophilus influenzae*	955	953	7	0.99	0.99	0.99
6	NC_000908	*Mycoplasma genitalium G-37*	189	186	2	0.98	0.98	0.97
7	NC_000909	*Methanocaldococcus janaschii*	720	708	9	0.98	0.98	0.97
8	NC_000912	*Mycoplasma pneumoniae M129*	243	241	2	0.99	0.99	0.98
9	NC_000913	*Escherichia coli K12*	2759	175	659	0.63	0.72	0.39
10	NC_000915	*Helicobacter pylori*	731	727	4	0.99	0.99	0.98
11	NC_000916	*Methanobacterium thermoautotrophicum*	719	711	4	0.98	0.99	0.98
12	NC_000917	*Archaeoglobus fulgidus*	782	774	8	0.98	0.98	0.97
13	NC_000917	*Archaeoglobus fulgidus DSM4304*	782	774	8	0.98	0.98	0.98
14	NC_000918	*Aquifex aeolicus VF5*	584	575	3	0.98	0.99	0.97
15	NC_000921	*Helicobacter pylori strain J99*	658	648	9	0.98	0.98	0.97
16	NC_000922	*Chlamydophila pneumoniae CWL029*	597	590	9	0.98	0.98	0.97
17	NC_000948	*Borrelia burgdorferi B31 plsmids cp32-1*	11	11	0	1.0	1.0	1.0
18	NC_000949	*Borrelia burgdorferi B31 plsmids cp32-3*	11	11	0	1.0	1.0	1.0
19	NC_000950	*Borrelia burgdorferi B31 plsmids cp32-4*	11	11	0	1.0	1.0	1.0
20	NC_000951	*Borrelia burgdorferi B31 plsmids cp32-6*	10	10	0	1.0	1.0	1.0

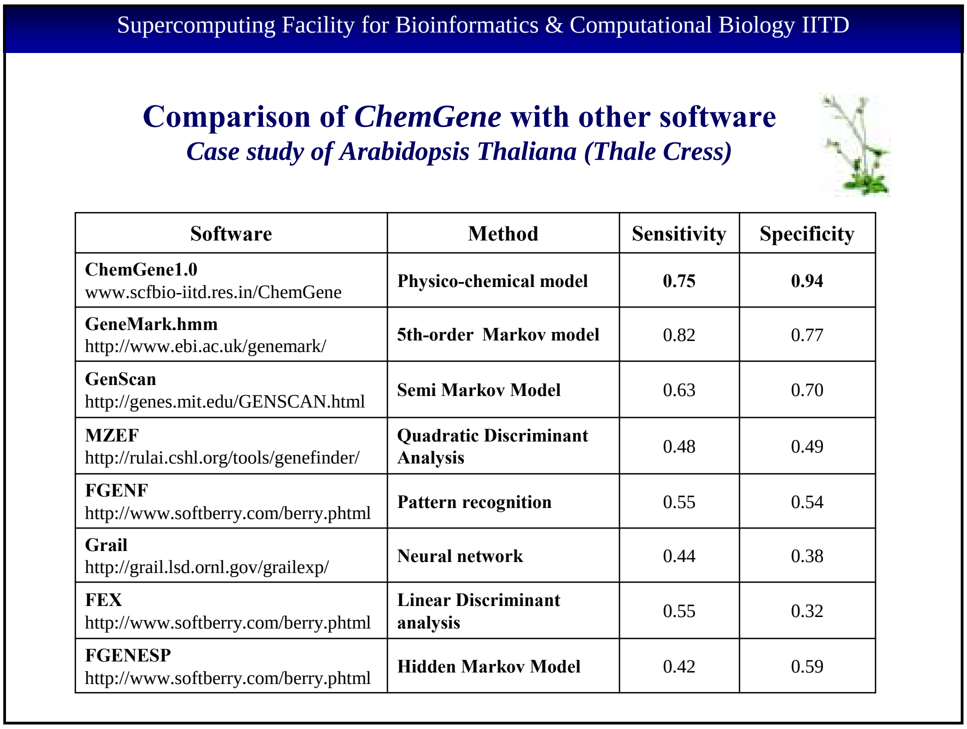
# True positives (TP): Genes evaluated as genes. False positives (FP): Non-genes evaluated as genes. True negatives (TN): Non-genes evaluated as non-genes. False negatives (FN): Genes evaluated as non-genes. Number of actual positives (AP) = TP+FN. Number of actual negatives (AN) = FP+TN. Predicted number of positives (PP) =TP+FP. Predicted number of negatives (PN) = TN+FN. Sensitivity (SS) =TP / (TP+FN). Specificity (SP) =TP / (TP+FP). $Correlation - coefficient = \left(TP \times TN - FP \times FN\right) / \sqrt{\mathrm{AN} \times \mathrm{PP} \times \mathrm{AP} \times \mathrm{PN}}$



Supercomputing Facility for Bioinformatics & Computational Biology IITD

Gene evaluation data for 21 eukaryotic genomes for experimentally verified tRNA genes (non-overlapping) and pre-genes.

S.No.	NCBI_ID	Species Name	Genes	TP	FP	SS	SP	CC
1	NC_001133	*Saccharomyces cerevisiae chromosome I*	6	5	0	0.83	1.0	0.91
2	NC_001134	*Saccharomyces cerevisiae chromosome II*	14	14	0	1.0	1.0	1.0
3	NC_001135	*Saccharomyces cerevisiae chromosome III*	12	11	0	0.92	1.0	0.95
4	NC_001136	*Saccharomyces cerevisiae chromosome IV*	31	31	0	1.0	1.0	1.0
5	NC_001137	*Saccharomyces cerevisiae chromosome V*	20	19	1	0.95	0.95	0.95
6	NC_001138	*Saccharomyces cerevisiae chromosome VI*	12	12	0	1.0	1.0	1.0
7	NC_001139	*Saccharomyces cerevisiae chromosome VII*	38	38	0	1.0	1.0	1.0
8	NC_001140	*Saccharomyces cerevisiae chromosome VIII*	11	11	0	1.0	1.0	1.0
9	NC_001141	*Saccharomyces cerevisiae chromosome IX*	10	10	1	1.0	0.91	0.95
10	NC_001142	*Saccharomyces cerevisiae chromosome X*	26	26	0	1.0	1.0	1.0
11	NC_001143	*Saccharomyces cerevisiae chromosome XI*	19	18	0	0.95	1.0	0.97
12	NC_001144	*Saccharomyces cerevisiae chromosome XII*	24	22	4	0.92	0.85	0.87
13	NC_001145	*Saccharomyces cerevisiae chromosome XIII*	25	24	1	0.96	0.96	0.96
14	NC_001146	*Saccharomyces cerevisiae chromosome XIV*	18	18	0	1.0	1.0	1.0
15	NC_001147	*Saccharomyces cerevisiae chromosome XV*	26	26	1	1.0	0.96	0.98
16	NC_001148	*Saccharomyces cerevisiae chromosome XVI*	17	17	0	1.0	1.0	1.0
17	NC_003070	*Arabidopsis thaliana chromosome I*	239	239	5	1.0	0.98	0.99
18	NC_003071	*Arabidopsis thaliana chromosome II*	96	90	2	0.94	0.98	0.96
19	NC_003074	*Arabidopsis thaliana chromosome III*	93	92	1	0.99	0.99	0.99
20	NC_003075	*Arabidopsis thaliana chromosome IV*	79	77	1	0.97	0.99	0.98
21	NC_003076	*Arabidopsis thaliana chromosome V*	108	108	1	1.0	0.99	0.99



Supercomputing Facility for Bioinformatics & Computational Biology IITD

Comparison of *ChemGene* with other software

Case study of Arabidopsis Thaliana (Thale Cress)

Software	Method	Sensitivity	Specificity
ChemGene1.0 www.scfbio-iitd.res.in/ChemGene	**Physico-chemical model**	**0.75**	**0.94**
GeneMark.hmm http://www.ebi.ac.uk/genemark/	**5th-order Markov model**	0.82	0.77
GenScan http://genes.mit.edu/GENSCAN.html	**Semi Markov Model**	0.63	0.70
MZEF http://rulai.cshl.org/tools/genefinder/	**Quadratic Discriminant Analysis**	0.48	0.49
FGENF http://www.softberry.com/berry.phtml	**Pattern recognition**	0.55	0.54
Grail http://grail.lsd.ornl.gov/grailexp/	**Neural network**	0.44	0.38
FEX http://www.softberry.com/berry.phtml	**Linear Discriminant analysis**	0.55	0.32
FGENESP http://www.softberry.com/berry.phtml	**Hidden Markov Model**	0.42	0.59

ChemGene1.0 Summary

- An *ab-initio* physico-chemical model is proposed to analyze DNA sequences

- Analyses of 331 bacterial genomes and 21 eukaryotic genomes present a proof of concept.

- Gene and Non-gene regions separate out.

- Consequences of Frame-shift mutations are correctly predicted.

- The Sensitivities achieved are ~ 95%.

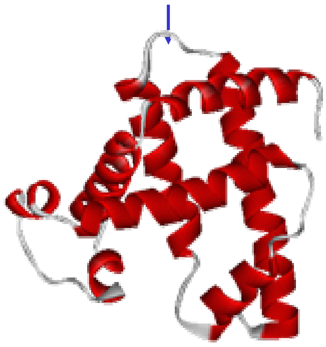
- Future work to address spatial and temporal profiles of gene expression at a molecular level and its control using *ChemGene*. *(Which gene is expressed in which cell and when?)*

- *ChemGene* [*Journal of Chemical Information & Modelling*, in press, (2005)] is web-enabled for wider usage at http://www.scfbio-iitd.res.in/ChemGene

Bhageerath 1.0

Protein Structure Prediction

................GLU ALA GLU MET LYS ALA SER GLU ASP LEU LYS
LYS HIS GLY VAL THR VAL LEU THR ALA LEU GLY ALA ILE LEU
LYS LYS LYS GLY HIS HIS GLU ALA GLU LEU LYS PRO LEU ALA
GLN SER HIS ALA THR LYS HIS LYS ILE PRO ILE LYS TYR LEU
GLU PHE ILE SER GLU ALA ILE ILE HIS LEU HIS.......................



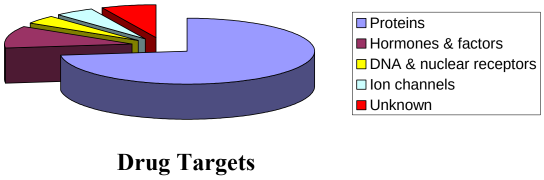
The Protein Folding Problem

Predicting the tertiary (3D) structure of a protein from the amino acid sequence and understanding the principles and pathway of folding

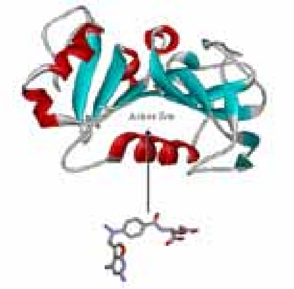
Supercomputing Facility for Bioinformatics & Computational Biology IITD

WHY FOLD PROTEINS ?

Pharmaceutical/Medical Sector



Drug Targets

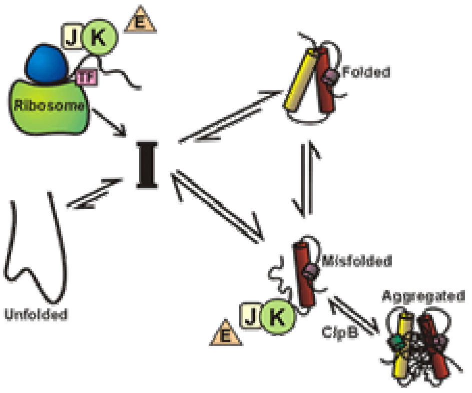


- Active site directed drug design
- Mapping the functions of proteins in metabolic pathways.

Supercomputing Facility for Bioinformatics & Computational Biology IITD

WHY FOLD PROTEINS ?

Understanding protein misfolding

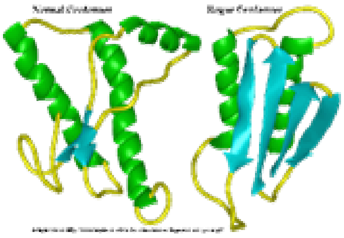


Supercomputing Facility for Bioinformatics & Computational Biology IITD

WHY FOLD PROTEINS?

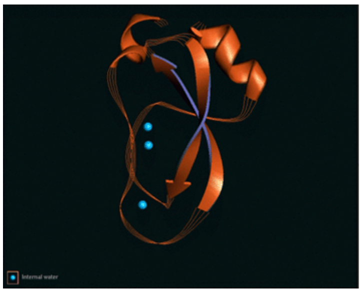
Mad cow disease

Caused due to protein misfolding of ‘prion’ protein



Alzheimer’s disease

Caused due to accumulation of beta-amyloid protein in brain cells.

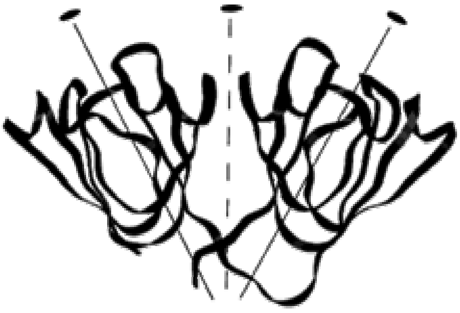


Supercomputing Facility for Bioinformatics & Computational Biology IITD

WHY FOLD PROTEINS?

Cataract

Caused due to aggregation of lens proteins



Gamma-crystallin

The protein has two similar globular domains of ‘Greek key’ motif

WHY FOLD PROTEINS?

- **Protein design:**

  Nanobiomachines: 'Self programmed' machines working as biosensors and carriers to aid in drug delivery processes. eg. ATPase in mitochondria

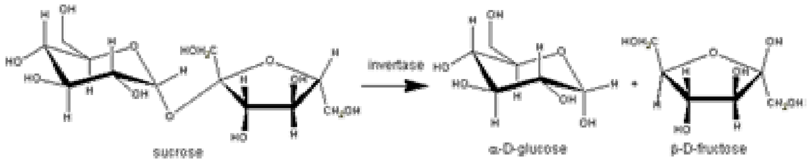
  Nanofibres: Fibers coated with extracellular matrix proteins are used as protein scaffold, reconstruction of damaged tissues

  Quantum dots: Small devices which can be used as biological probes for diagnostics.

- **Biocatalyst design**: "Catalysts of future" that will help in functions like: Making Designer Enzymes for any reaction that is thermodynamically feasible (involves inverse protein folding viz. what is the sequence to be used for obtaining an enzyme with the desired shape and function), Storing and releasing oxygen when required by the body, Controlling blood sugar level etc..

WHY FOLD PROTEINS?

- Sugar Industry: Invertase for the conversion of sucrose into glucose and fructose.

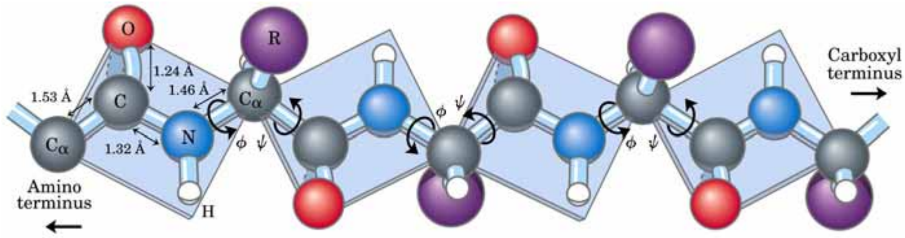


- Chocolate Industry: During cocoa beans processing, enzymes activated by fermentation process gives the characteristic chocolate flavor.
- Pulp & Paper Industry: Esterase is used to break 'stickies' into smaller components for improving paper quality.
- Textile & Leather Industry: Proteases are used in dehairing & lipases are used for degreasing, cellulase in giving smoother, glossier brighter fabrics.

RAMACHANDRAN ANGLES

Prof G.N. Ramachandran

1922-2001



A resolution to the protein folding problem entails a specification of all the Ramachandran angles along the polypeptide main chain (backbone).

Structure Determination / Prediction Methodologies

Experimental Techniques

- X-Ray diffraction
- Nuclear Magnetic Resonance (NMR)
- Electron diffraction, Neutron diffraction, Electron microscopy, Fluorescence transfer

Drawbacks of Experimental Methods

- Expensive
- Time consuming
- Don’t work well for receptors

Comparative Modeling Approaches

Homology

Similar sequences adopt similar fold is the basis.
Alignment is performed with related sequences.
(SWISS-MODEL-www.expasy.org, 3DJIGSAW-www.bmm.icnet.uk etc).

Threading

Sequence is aligned with all the available folds and scores are assigned for each alignment according to a scoring function. (Threader - bioinf.cs.ucl.ac.uk)

The above methods are fairly reliable and fast but data base dependent. Given that only (~) 8000 unique protein structures are available in structural databases (PDB) this could become a limitation, particularly with sequences with low similarity scores.

Ab initio Protein Folding Methods

Strategy A

- Generate all possible conformations and find the most stable one.
- For a protein comprising 200 AA assuming 2 degrees of freedom per AA
- 2^{200} Structures => 2^{200} Minutes to optimize and find free energy.

2^{200} Minutes = 3 x 10^{54} Years!!

Strategy B

- Start with a straight chain and solve F = ma to capture the most stable state
- A 200 AA protein evolves

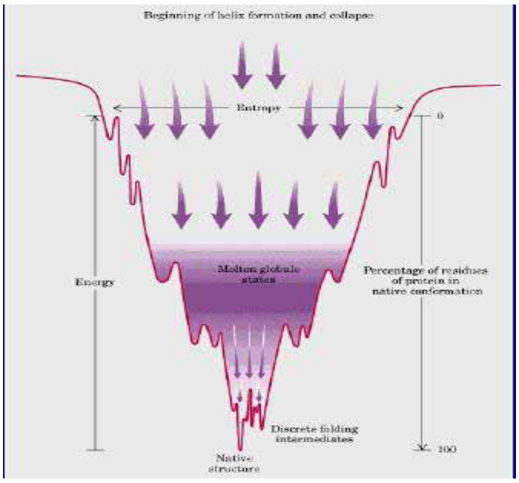
~ 10^{-11} sec / day / processor

- 10^{-3} sec (Time it takes for a protein *in vivo*) => 10^{8} days /protein / processor (to fold *in silico*) ~ 10^{6} years

With 10^{6} processors ~ 1 Year /protein

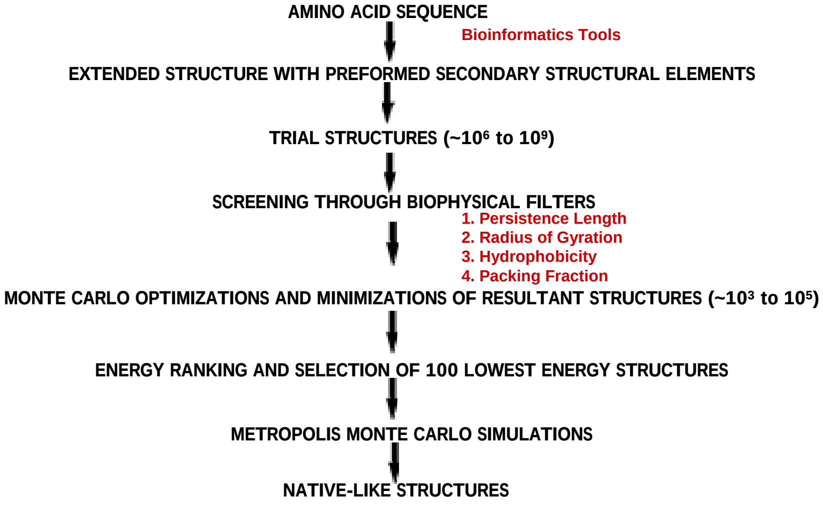
Computational requirements of *ab initio* methods are insurmountable. A smart combination of Bioinformatics tools and *ab initio* methods is required

PROTEIN FOLDING LANDSCAPE


Beginning of helix formation and collapse
Entropy
0
Energy
Molten globule states
Percentage of residues of protein in native conformation
Discrete folding intermediates
Native structure
100

Finding the global minimum on a rugged multidimensional surface is a complex unsolved problem

From Sequence to Structure: The IITD Pathway



AMINO ACID SEQUENCE

Bioinformatics Tools

EXTENDED STRUCTURE WITH PREFORMED SECONDARY STRUCTURAL ELEMENTS

TRIAL STRUCTURES (~10^6 to 10^9)

SCREENING THROUGH BIOPHYSICAL FILTERS

1. Persistence Length
2. Radius of Gyration
3. Hydrophobicity
4. Packing Fraction

MONTE CARLO OPTIMIZATIONS AND MINIMIZATIONS OF RESULTANT STRUCTURES (~10^3 to 10^5)

ENERGY RANKING AND SELECTION OF 100 LOWEST ENERGY STRUCTURES

METROPOLIS MONTE CARLO SIMULATIONS

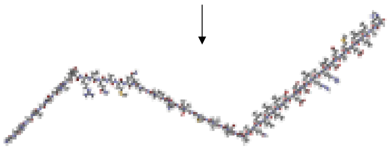
NATIVE-LIKE STRUCTURES

Narang P, Bhushan K, Bose S and Jayaram B 'A computational pathway for bracketing native-like structures for small alpha helical globular proteins.' *Phys. Chem. Chem. Phys.* 2005, 7, 2364-2375.

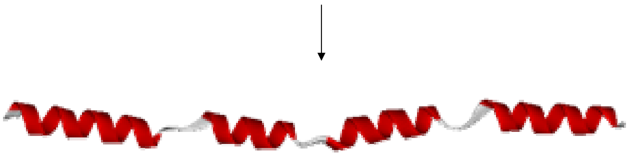
Supercomputing Facility for Bioinformatics & Computational Biology IITD

Protein Model Builder

HRQALGERLYPRVQAMQPAFASKITGMLLELSPAQLLLLLASENSLRARVNEAMELIIAHG



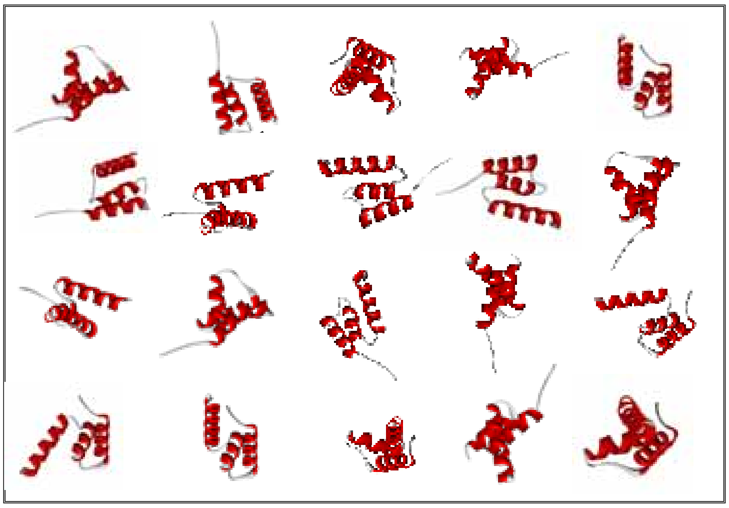
Extended Chain

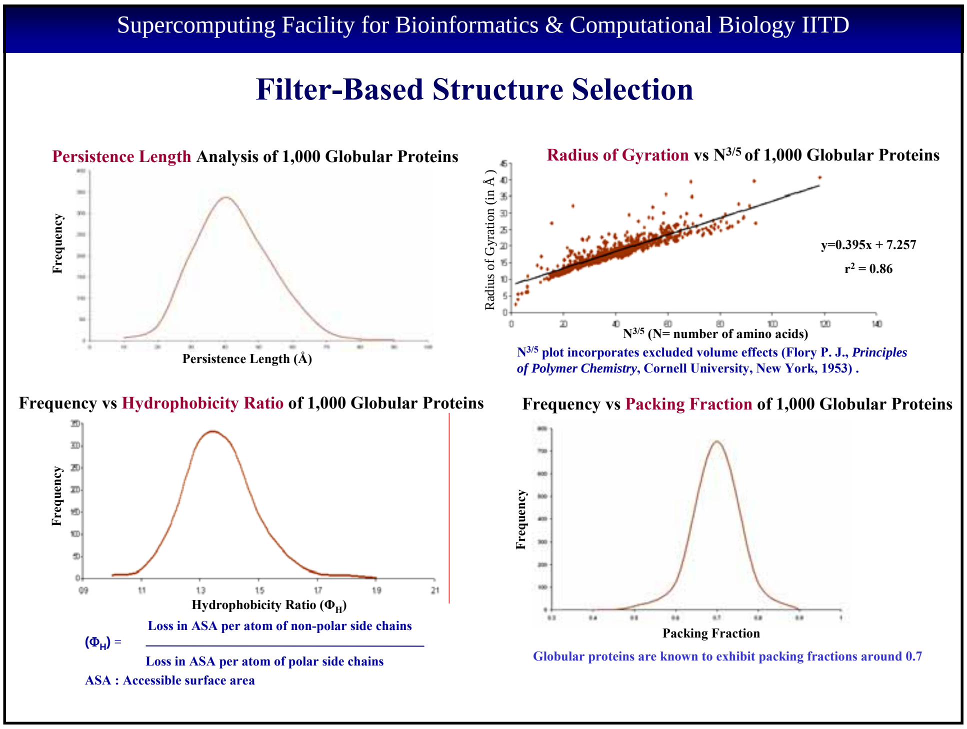


Preformed Secondary Structural Units

Supercomputing Facility for Bioinformatics & Computational Biology IITD

Trial Structure Generation





Supercomputing Facility for Bioinformatics & Computational Biology IITD

Filter-Based Structure Selection

Persistence Length Analysis of 1,000 Globular Proteins

Radius of Gyration vs $N^{3/5}$ of 1,000 Globular Proteins

$N^{3/5}$ plot incorporates excluded volume effects (Flory P. J., *Principles of Polymer Chemistry*, Cornell University, New York, 1953) .

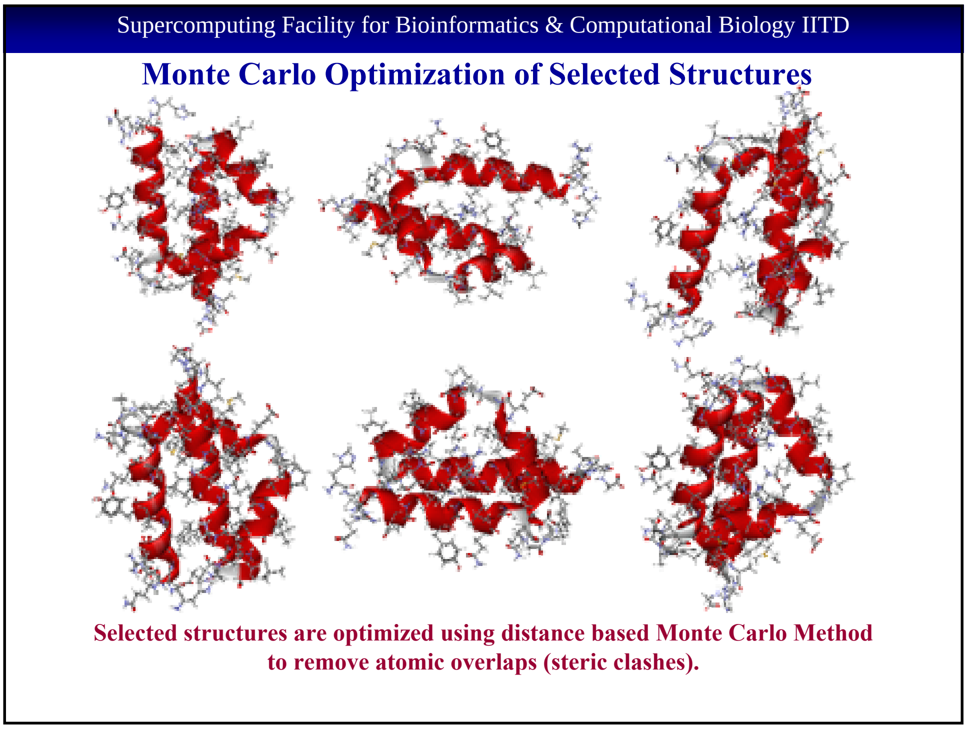
Frequency vs Hydrophobicity Ratio of 1,000 Globular Proteins

$$(\Phi_H) = \frac{\text{Loss in ASA per atom of non-polar side chains}}{\text{Loss in ASA per atom of polar side chains}}$$

ASA : Accessible surface area

Frequency vs Packing Fraction of 1,000 Globular Proteins

Globular proteins are known to exhibit packing fractions around 0.7



Supercomputing Facility for Bioinformatics & Computational Biology IITD

Monte Carlo Optimization of Selected Structures

Selected structures are optimized using distance based Monte Carlo Method to remove atomic overlaps (steric clashes).

An Empirical Scoring Function for Ranking Trial Structures

$$E = \sum E_{el} + E_{vdw} + E_{hpb}$$

Electrostatics

$$E_{el} = \frac{332 q_i q_j}{D(r) r_{ij}}$$

$$D(r) = D - \left[\left(\frac{D - D_i}{2} \right) \left(\alpha^2 + 2\alpha + 2 \right) e^{-\alpha} \right]$$

van der Waals

$$E_{vdw} = \left[\frac{C_{12}^{ij}}{r_{ij}^{12}} - \frac{C_6^{ij}}{r_{ij}^{6}} \right]$$

$$C_{12}^{ij} = \varepsilon_{ij} \left(R_{ij}^{*} \right)^{12}$$

$$C_6^{ij} = 2\varepsilon_{ij} \left(R_{ij}^{*} \right)^{6}$$

$$R_{ij}^{*} = R_i^{*} + R_j^{*}$$

$$\varepsilon_{ij} = \left(\varepsilon_i \varepsilon_j \right)^{1/2}$$

Hydrophobic

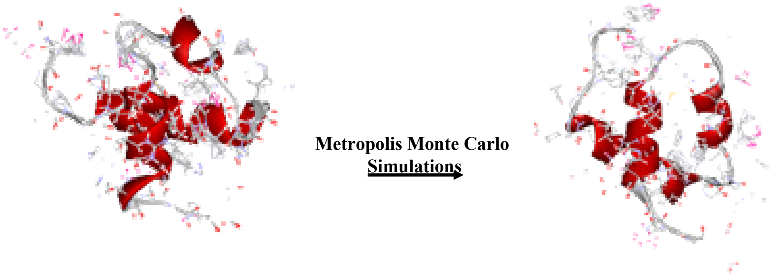
$$E_{hpb} = \begin{cases} f_{ij} \times \frac{V_{excl}}{V_w}, r_{ij} \geq \left(R_{Hi} + R_{Hj} \right); \\ 0, otherwise \end{cases}$$

$$V_{excl} = \frac{r_{ij}^3}{12} - \frac{\left(R_{Hi}^2 + R_{Hj}^2 \right)^2}{4 r_{ij}} + \frac{2}{3} \left(R_{Hi}^3 + R_{Hj}^3 \right) - \frac{r_{ij}}{2} \left(R_{Hi}^2 + R_{Hj}^2 \right.$$

The above Scoring function captures native as the lowest energy structure from among 61,640 decoys belonging to 67 different proteins and diverse decoy sets. The all-atom energy based scoring function is used to select 100 lowest energy structures.

Arora N and Jayaram B, *J. Phys. Chem.*, 1998, 102, 6139-6144.
Arora N and Jayaram B, *J. Comp. Chem.*, 1997, 18, 1245-1252.

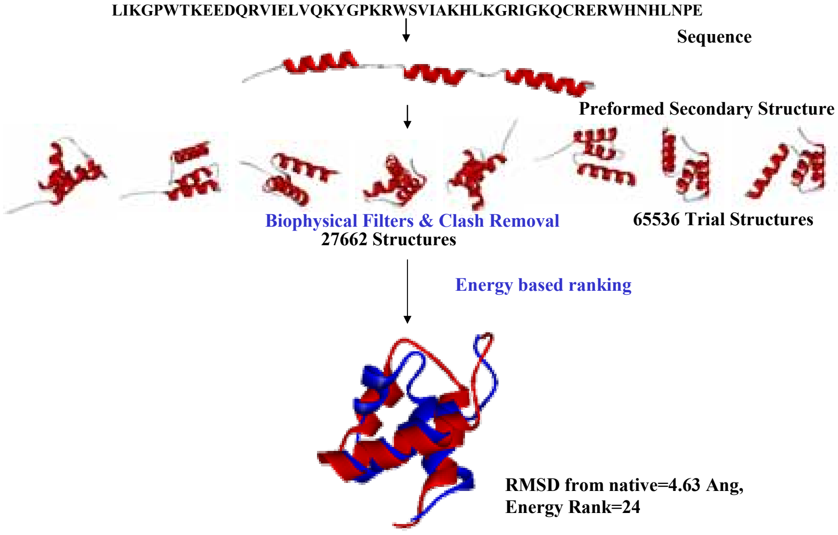
Metropolis Monte Carlo Simulations



The selected structures are optimized using Metropolis Monte Carlo Simulations

A Case Study of Mouse C-Myb

DNA Binding (52 AA)



LIKGPWTKEEDQRVIELVQKYGPKRWSVIAKHLKGRIGKQCRERWHNHLNPE

Sequence

Preformed Secondary Structure

Biophysical Filters & Clash Removal
27662 Structures

65536 Trial Structures

Energy based ranking

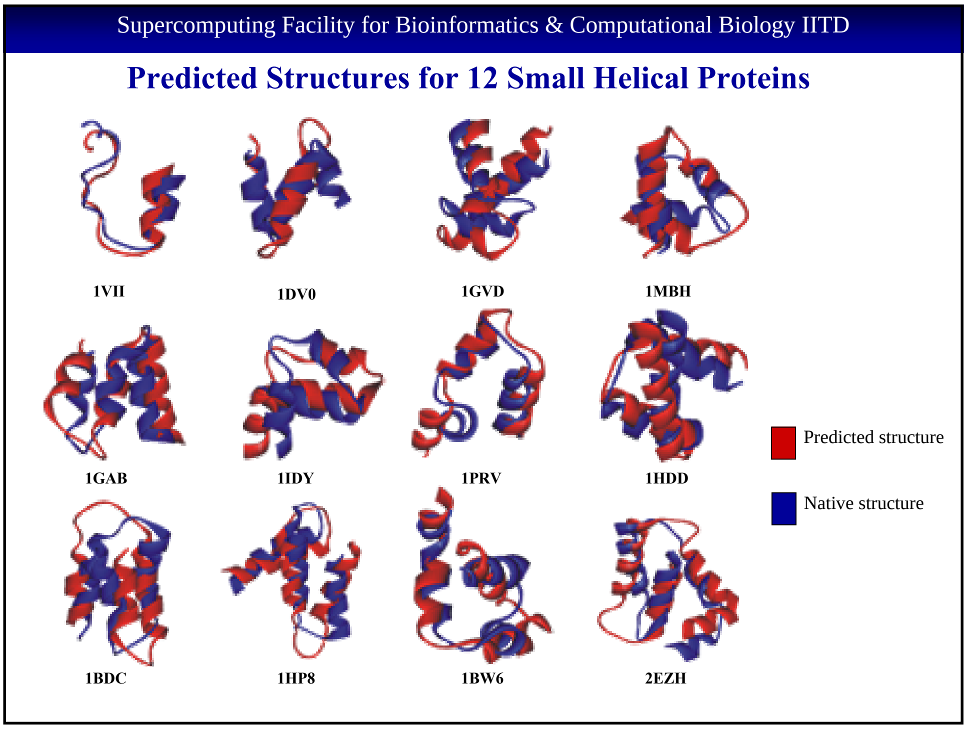
RMSD from native=4.63 Ang,
Energy Rank=24

Performance of the Protocol Devised on 12 Small Helical Proteins

No.	PDB ID (i)	No. of Residues (ii)	No. of Helices (iii)	Total No of Structures Generated (iv)	No. of Structures Accepted			RMSD without end loops (in Å) (viii)	MC Optimization & Energy Minimization		Characterization of 100 lowest energy structures			
					After Persistence Length (v)	After Radius of Gyration (vi)	Lowest RMSD (in Å) (vii)		Lowest RMSD (in Å) (ix)	Rank (Energy) (x)	Lowest RMSD (in Å) (xi)	Rank (Energy) (xii)	Metropolis Monte Carlo simulations	
													Lowest RMSD (in Å) (xiii)	Rank (Energy) (xiv)
1.	1VII	36	3	65536	65536	47976	3.29	2.63	2.35	6958	2.85	3	2.88	1
2.	1DV0	45	3	65536	65536	28606	4.23	3.72	3.78	7429	4.74	31	4.74	2
3.	1GVD	52	3	65536	65257	25980	4.97	4.08	4.23	19351	4.88	71	4.89	71
4.	1MBH	52	3	65536	65536	27662	3.64	3.24	2.87	1774	4.66	72	4.63	24
5.	1GAB	53	3	65536	65483	18941	3.89	3.37	3.16	838	4.01	50	4.08	25
6.	1IDY	54	3	65536	65536	18953	4.85	2.97	2.38	2468	3.28	66	3.36	14
7.	1PRV	56	3	65536	65515	7545	5.56	3.40	2.7	727	4.23	52	3.87	2
8.	1HDD	57	3	65536	61427	16523	4.08	3.29	2.46	1134	4.58	32	4.27	20
9.	1BDC	60	3	65536	57903	6800	6.64	4.42	4.12	5	4.12	5	4.21	2
10.	1HP8	68	3	65536	48171	5189	4.98	4.22	3.78	4610	3.89	90	4.20	41
11.	1BW6	56	4	262144	254975	44872	5.99	4.13	4.32	6826	4.68	11	4.69	5
12.	2EZH	65	4	1048576	1041303	249740	3.37	3.21	3.33	30851	4.34	11	4.40	2

Structures with native-like topology are bracketed within the 100 lowest energy structures.

Narang P, Bhushan K, Bose S and Jayaram B ‘A computational pathway for bracketing native-like structures for small alpha helical globular proteins.’ *Phys. Chem. Chem. Phys.* 2005, 7, 2364-2375.

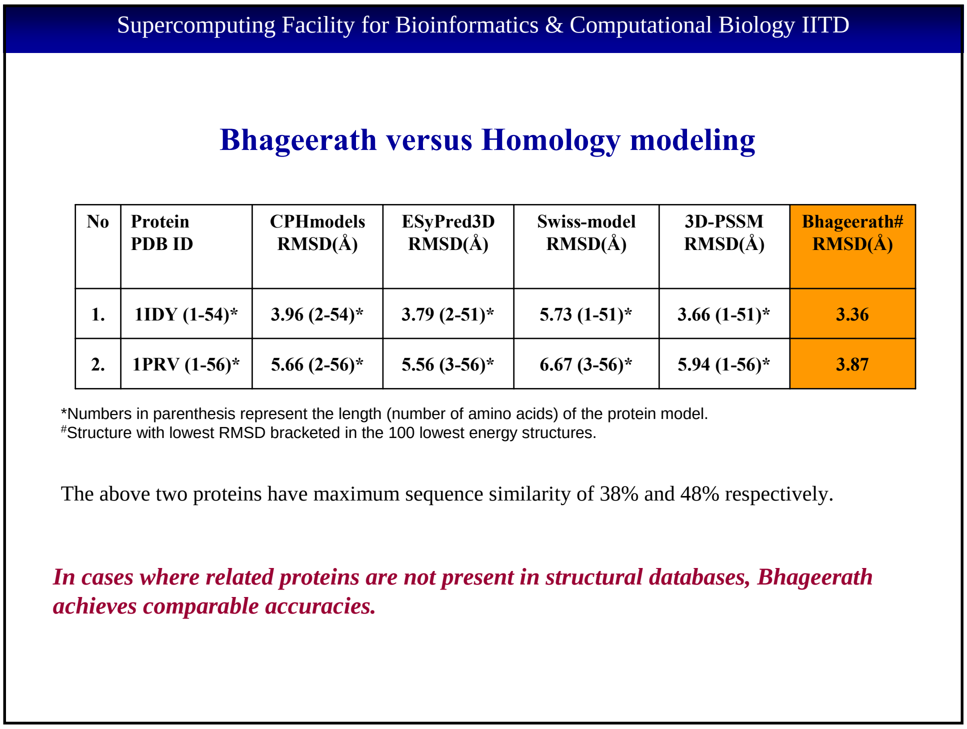


Supercomputing Facility for Bioinformatics & Computational Biology IITD

Predicted Structures for 12 Small Helical Proteins

🟥 Predicted structure

🟦 Native structure



Supercomputing Facility for Bioinformatics & Computational Biology IITD

Bhageerath versus Homology modeling

No	Protein PDB ID	CPHmodels RMSD(Å)	ESyPred3D RMSD(Å)	Swiss-model RMSD(Å)	3D-PSSM RMSD(Å)	Bhageerath# RMSD(Å)
1.	1IDY (1-54)*	3.96 (2-54)*	3.79 (2-51)*	5.73 (1-51)*	3.66 (1-51)*	3.36
2.	1PRV (1-56)*	5.66 (2-56)*	5.56 (3-56)*	6.67 (3-56)*	5.94 (1-56)*	3.87

*Numbers in parenthesis represent the length (number of amino acids) of the protein model.
#Structure with lowest RMSD bracketed in the 100 lowest energy structures.

The above two proteins have maximum sequence similarity of 38% and 48% respectively.

In cases where related proteins are not present in structural databases, Bhageerath achieves comparable accuracies.

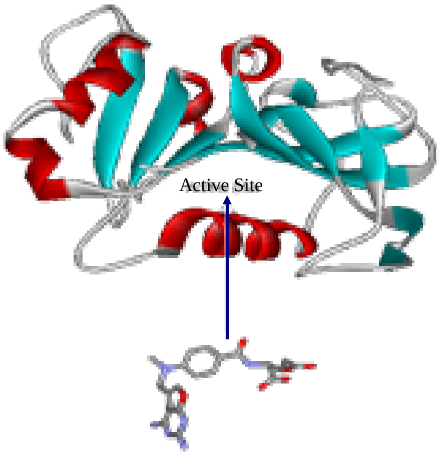
Conclusions and Future Perspectives

•Structures with native-like topology are bracketed within the 100 lowest energy structures. “Needle in a haystack problem” is thus reduced to finding best 100 energy structures at least for small proteins. The suite of programs christened “Bhageerath” is made accessible at www.scfbio-iitd.res.in/bhageerath for wider usage.

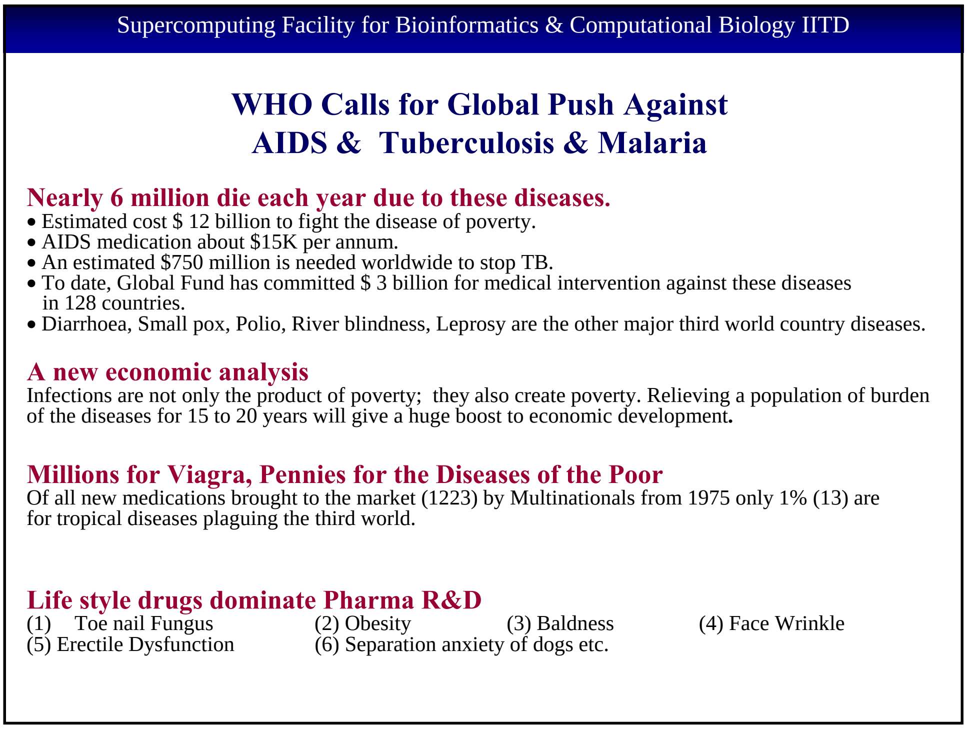
•Further improvements to the methodology such as topological equivalence have been introduced to reduce the number of candidate structures for the native.

•It is envisioned that explicit solvent molecular dynamics simulations on the selected candidate structures can aid in optimizing side chain orientations, promoting favorable packing interactions bringing the RMSD to less than 3Å.

Active Site Directed Lead Design
Sanjeevini1.0



Structure based drug design is like designing a key to open or jam a dynamic lock. The shape of the lock as well as its key hole are known.



WHO Calls for Global Push Against AIDS & Tuberculosis & Malaria

Nearly 6 million die each year due to these diseases.

- Estimated cost $ 12 billion to fight the disease of poverty.
- AIDS medication about $15K per annum.
- An estimated $750 million is needed worldwide to stop TB.
- To date, Global Fund has committed $ 3 billion for medical intervention against these diseases in 128 countries.
- Diarrhoea, Small pox, Polio, River blindness, Leprosy are the other major third world country diseases.

A new economic analysis

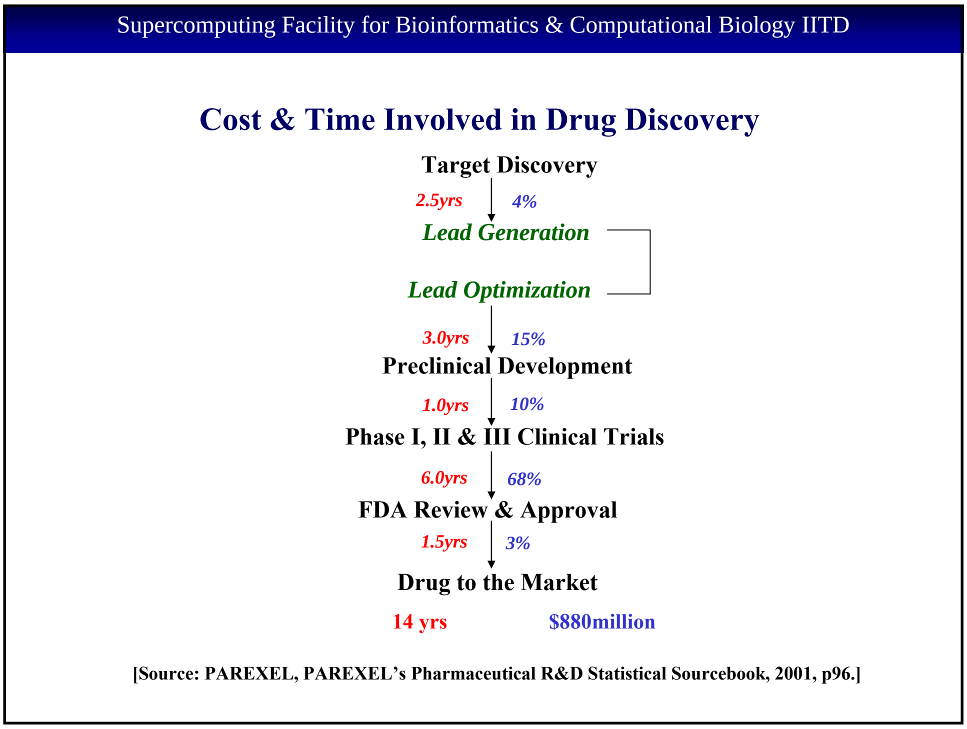
Infections are not only the product of poverty; they also create poverty. Relieving a population of burden of the diseases for 15 to 20 years will give a huge boost to economic development**.**

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.



Cost & Time Involved in Drug Discovery

Target Discovery

2.5yrs *4%*

Lead Generation

Lead Optimization

3.0yrs *15%*

Preclinical Development

1.0yrs *10%*

Phase I, II & III Clinical Trials

6.0yrs *68%*

FDA Review & Approval

1.5yrs *3%*

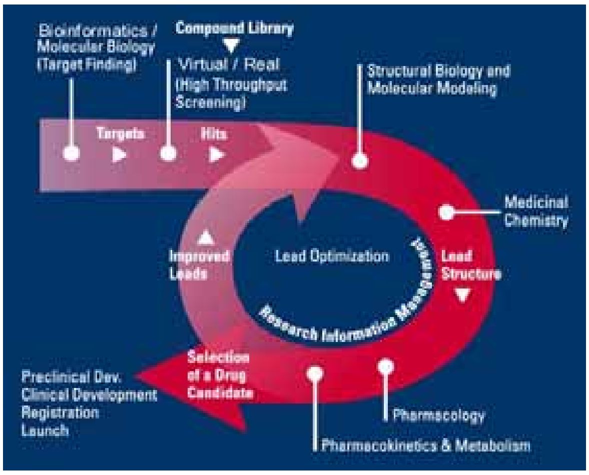
Drug to the Market

14 yrs **$880million**

[Source: PAREXEL, PAREXEL’s Pharmaceutical R&D Statistical Sourcebook, 2001, p96.]

Supercomputing Facility for Bioinformatics & Computational Biology IITD

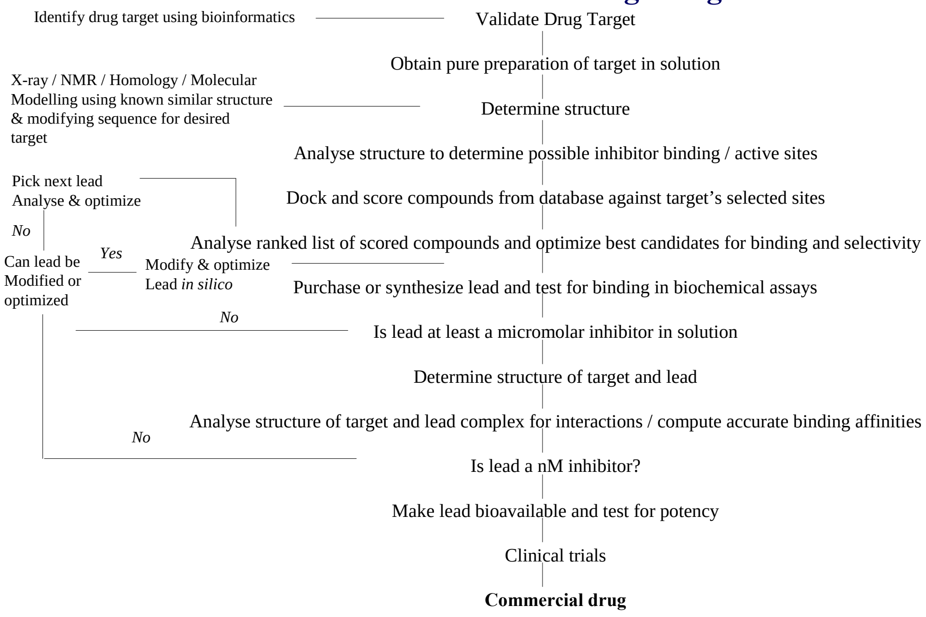
In silico Intervention in the Drug Discovery Process to Reduce Cost & Time


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

***In silico* intervention in drug discovery can save up to ~ 15% of time and cost which could be significant for life threatening diseases.**

Supercomputing Facility for Bioinformatics & Computational Biology IITD

Details of Structure Based Drug Design


Identify drug target using bioinformatics

Validate Drug Target

Obtain pure preparation of target in solution

X-ray / NMR / Homology / Molecular Modelling using known similar structure & modifying sequence for desired target

Determine structure

Analyse structure to determine possible inhibitor binding / active sites

Pick next lead
Analyse & optimize

Dock and score compounds from database against target's selected sites

No

Analyse ranked list of scored compounds and optimize best candidates for binding and selectivity

Can lead be Modified or optimized

Yes

Modify & optimize Lead *in silico*

Purchase or synthesize lead and test for binding in biochemical assays

No

Is lead at least a micromolar inhibitor in solution

Determine structure of target and lead

Analyse structure of target and lead complex for interactions / compute accurate binding affinities

No

Is lead a nM inhibitor?

Make lead bioavailable and test for potency

Clinical trials

Commercial drug

Some Concerns in Lead Design *In Silico*

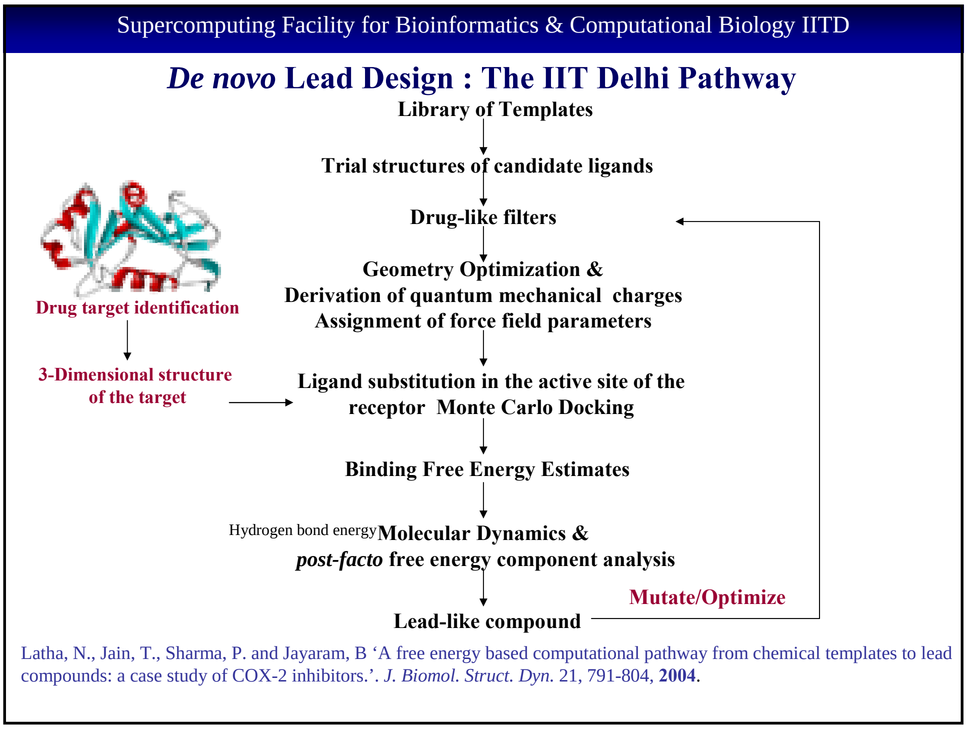
Why computers and drug design softwares don't predict new leads routinely?

- Novelty and Geometry of the Ligands
- Accurate charges and other Force field parameters
- Ligand Binding Sites
- Flexibility of the Ligand and the Target
- Solvent and salt effects in Binding
- Internal energy versus Free energy of Binding
- Computational Tractability
- Druggability (ADMET characteristics)

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



Supercomputing Facility for Bioinformatics & Computational Biology IITD

De novo Lead Design : The IIT Delhi Pathway

Library of Templates

Trial structures of candidate ligands

Drug-like filters

Geometry Optimization &
Derivation of quantum mechanical charges
Assignment of force field parameters

Drug target identification

3-Dimensional structure
of the target

Ligand substitution in the active site of the
receptor Monte Carlo Docking

Binding Free Energy Estimates

Hydrogen bond energy Molecular Dynamics &
post-facto free energy component analysis

Mutate/Optimize

Lead-like compound

Latha, N., Jain, T., Sharma, P. and Jayaram, B 'A free energy based computational pathway from chemical templates to lead compounds: a case study of COX-2 inhibitors.'. *J. Biomol. Struct. Dyn.* 21, 791-804, **2004**.

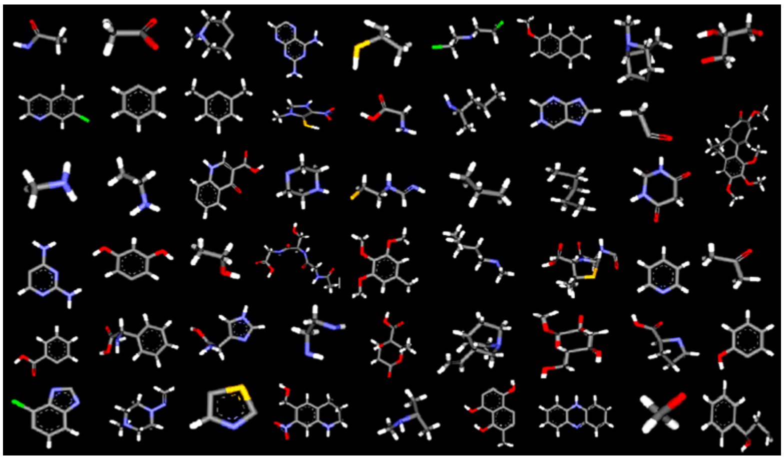
Supercomputing Facility for Bioinformatics & Computational Biology IITD

Main Modules in *Sanjeevini*

1. Template library
2. Molecule generator
3. Molecular descriptors / drug-like filters
4. Molecular docking
5. Structural analysis of the receptor-candidate complex
6. Energy analysis of the receptor-candidate complex
7. Binding affinity analysis

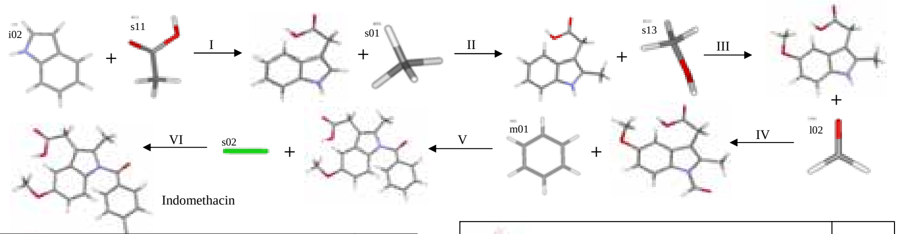
Jayaram, B., Latha, N.,Jain, T., Sharma, P., Gandhimathi, A., Pandey, V.S., '*Sanjeevini:* A Comprehensive Active-Site Directed Lead Design Software.' *Indian Journal of Chemistry-A.* **2005** (In Press)

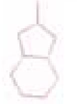
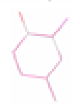
Template Library



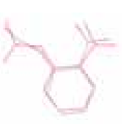
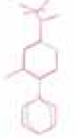
The substructure-based template library currently has ~ 160 chemical moieties consisting of unique rings, side chains and linkers, prepared in a force field compatible manner. Templates are joined to make molecules known or new.

Candidate Molecule Generation & Structure Validation

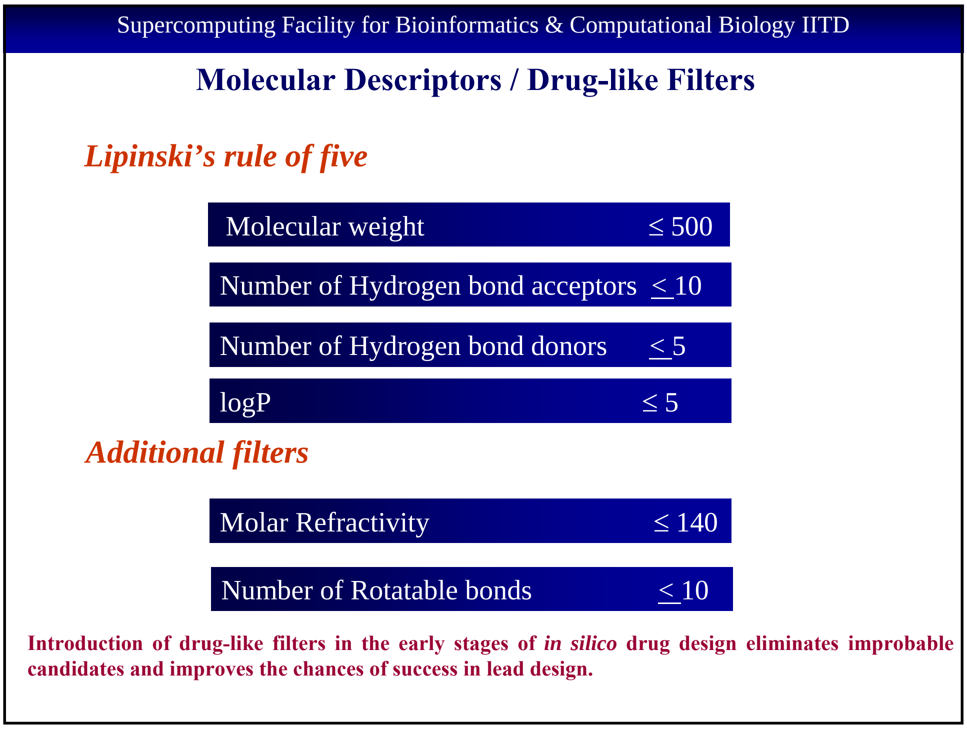


				Average RMSD
0.07Å	0.05Å	0.08Å	0.38Å	0.15Å

			Average RMSD
0.21Å	0.50Å	0.83Å	0.51Å

			Average RMSD
0.91Å	0.54Å	0.94Å	0.80Å

The *in silico* methods have come of age to predict the structures of small molecules accurately.



Supercomputing Facility for Bioinformatics & Computational Biology IITD

Molecular Descriptors / Drug-like Filters

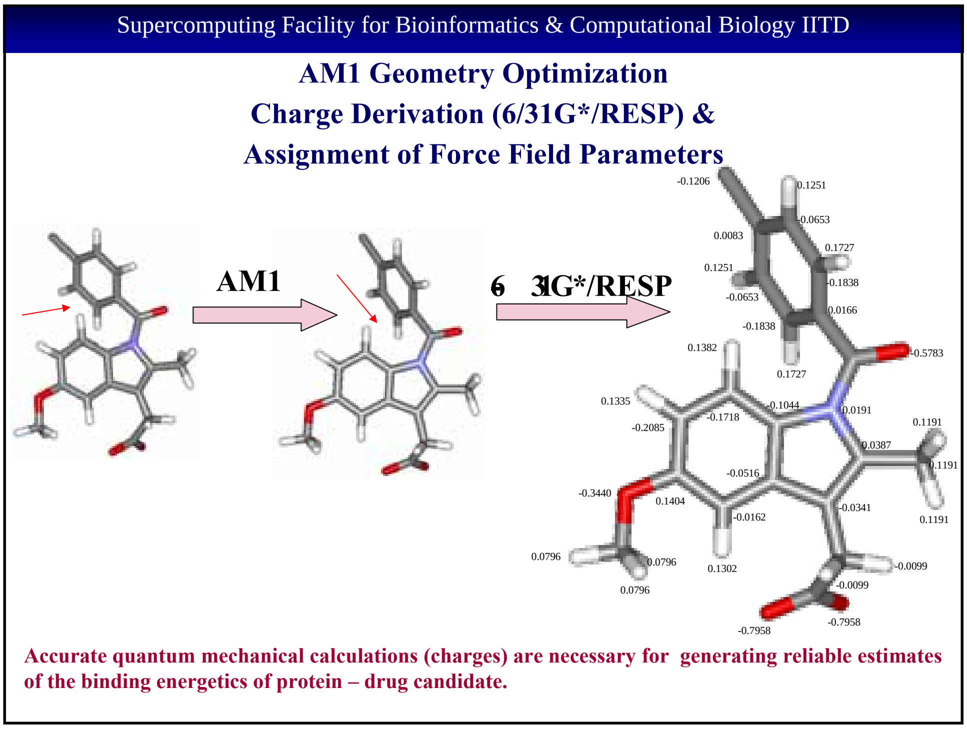
Lipinski's rule of five

Molecular weight	≤ 500
Number of Hydrogen bond acceptors	< 10
Number of Hydrogen bond donors	< 5
logP	≤ 5

Additional filters

Molar Refractivity	≤ 140
Number of Rotatable bonds	< 10

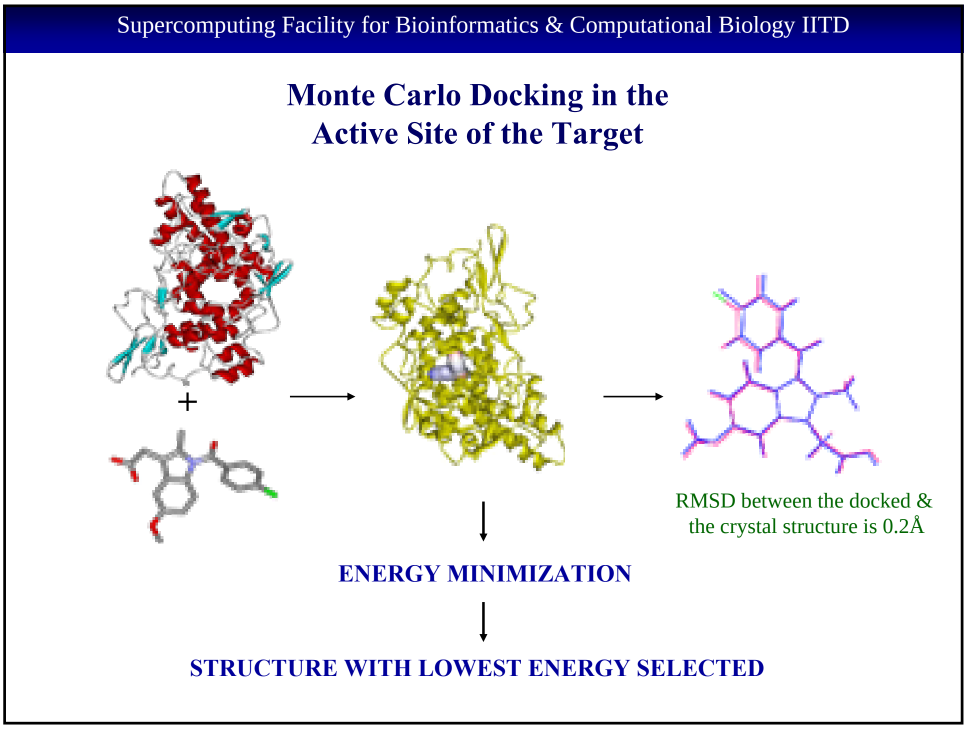
Introduction of drug-like filters in the early stages of *in silico* drug design eliminates improbable candidates and improves the chances of success in lead design.



Supercomputing Facility for Bioinformatics & Computational Biology IITD

AM1 Geometry Optimization
Charge Derivation (6/31G*/RESP) &
Assignment of Force Field Parameters

Accurate quantum mechanical calculations (charges) are necessary for generating reliable estimates of the binding energetics of protein – drug candidate.



Supercomputing Facility for Bioinformatics & Computational Biology IITD

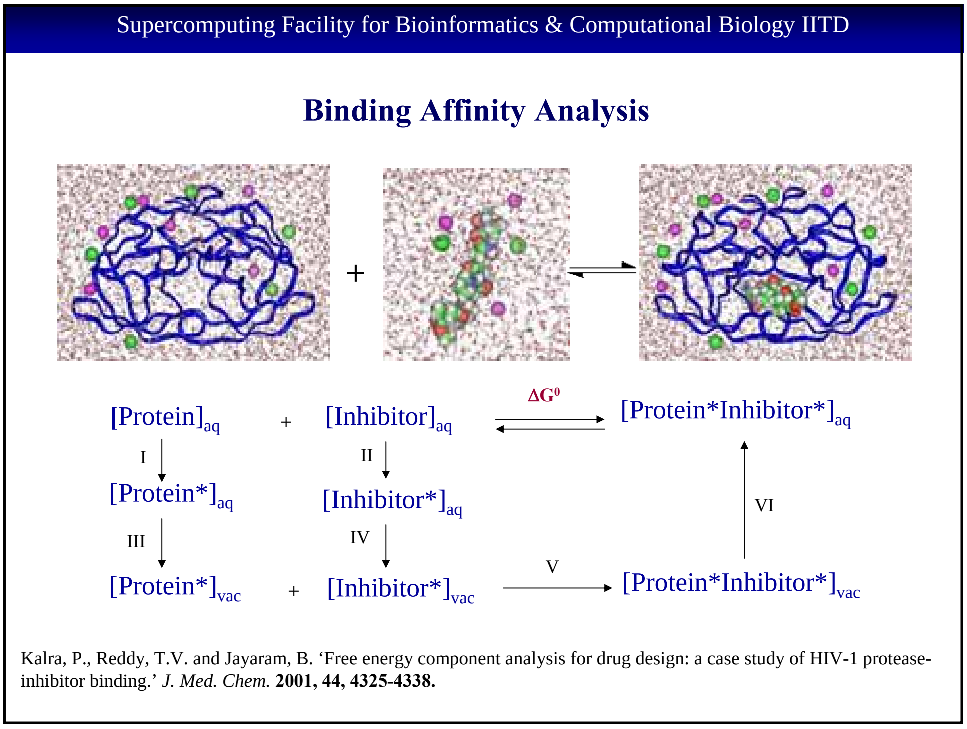
Monte Carlo Docking in the Active Site of the Target

+

RMSD between the docked & the crystal structure is 0.2Å

ENERGY MINIMIZATION

STRUCTURE WITH LOWEST ENERGY SELECTED



Supercomputing Facility for Bioinformatics & Computational Biology IITD

Binding Affinity Analysis

+

$$\begin{tikzcd}
{[\text{Protein}]_{aq}} \arrow[d, "I"'] & + & {[\text{Inhibitor}]_{aq}} \arrow[d, "II"'] \arrow[r, shift left, "\Delta G^0"] & {[\text{Protein*Inhibitor*}]_{aq}} \arrow[l, shift left] \\
{[\text{Protein*}]_{aq}} \arrow[d, "III"'] & & {[\text{Inhibitor*}]_{aq}} \arrow[d, "IV"'] & \\
{[\text{Protein*}]_{vac}} & + & {[\text{Inhibitor*}]_{vac}} \arrow[r, "V"] & {[\text{Protein*Inhibitor*}]_{vac}} \arrow[uu, "VI"']
\end{tikzcd}$$

Kalra, P., Reddy, T.V. and Jayaram, B. ‘Free energy component analysis for drug design: a case study of HIV-1 protease-inhibitor binding.’ *J. Med. Chem.* **2001, 44, 4325-4338.**

Statistical Mechanics of Binding

$\Delta G^o = - RT \ln K_{eq.} = - RT \ln [\{Q_{P*D*.}/(N_A Q_w)\}/\{(Q_{P.}/(N_A Q_w))(Q_D/(N_A Q_w))\}] + P\Delta V^o$

$Q_{p.aq} \simeq Q^{tr}_{p\,.} Q^{rot}_{p.}\ Z_{p.aq}/V^N$

$Z_{P.aq} = \int\ldots\ldots\int \exp \{-E(X^N_p,X^M_W)/k_BT\}\ dX^N_p\ dX^M_W = \langle \exp (E(X^N_p,X^M_W)/k_BT\rangle$

$\Delta G^o \simeq \Delta G^o_{tr} + \Delta G^o_{rot} + \Delta G^o_{(intra\ +solvn.)}$ **Free Energy Simulations**

$Z_{P.aq} \simeq Z_{P.aq}^{vib.config} . Z_{P.aq}^{solvn}$

$\Delta G^o \simeq \Delta G^o_{tr} + \Delta G^o_{rot} + \Delta G^o_{intra} + \Delta G^o_{solvn.}$ **Master Equation**

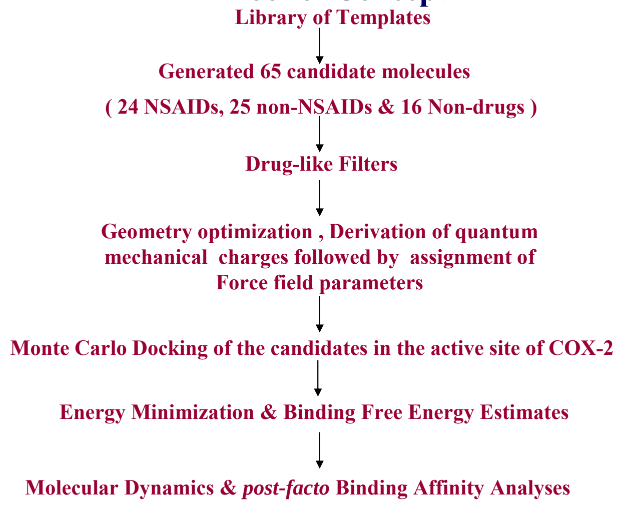
$\Delta G^o \simeq \Delta G^o_{tr} + \Delta G^o_{rot} + \Delta E^o_{vac} + \Delta G^o_{solvn.}$ **Energy Minimized Structure Analysis**

$\Delta G^o \simeq \Delta G^o_{tr} + \Delta G^o_{rot} + \Delta H^o_{intra} - T\Delta S^o_{intra\ (vib+config)} + \Delta G^o_{solvn}$

***post facto* Analysis of MD Trajectories**

For details please see www.scfbio-iitd.res.in/training/lecturenotes.html

A CASE STUDY OF COX-2 INHIBITORS – A Proof of Concept



Library of Templates

Generated 65 candidate molecules

(24 NSAIDs, 25 non-NSAIDs & 16 Non-drugs)

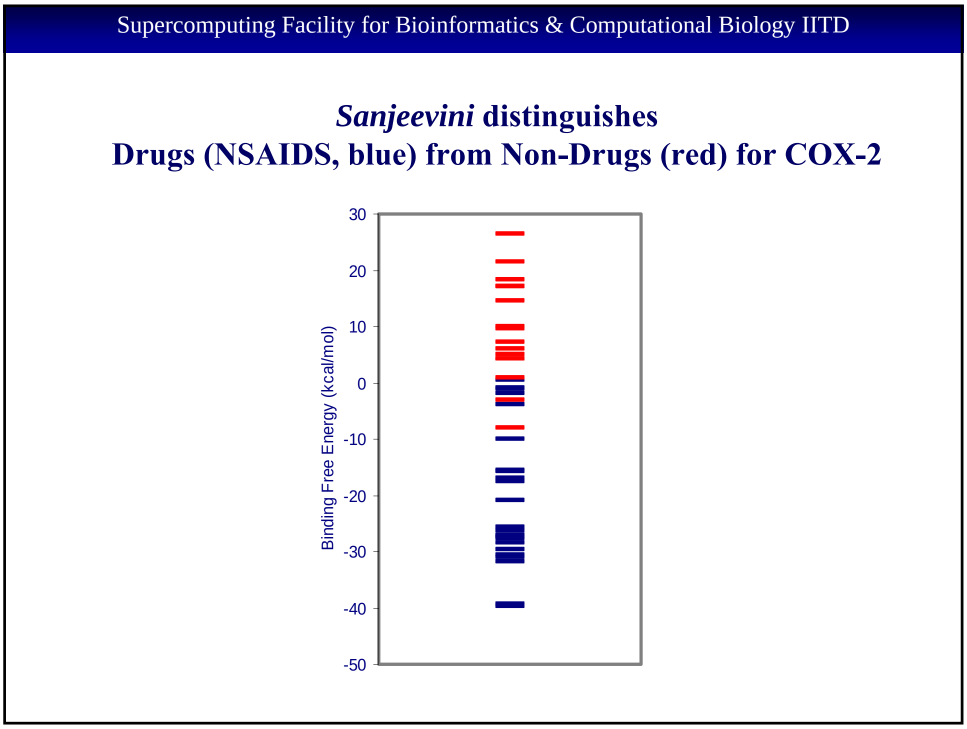
Drug-like Filters

Geometry optimization , Derivation of quantum mechanical charges followed by assignment of Force field parameters

Monte Carlo Docking of the candidates in the active site of COX-2

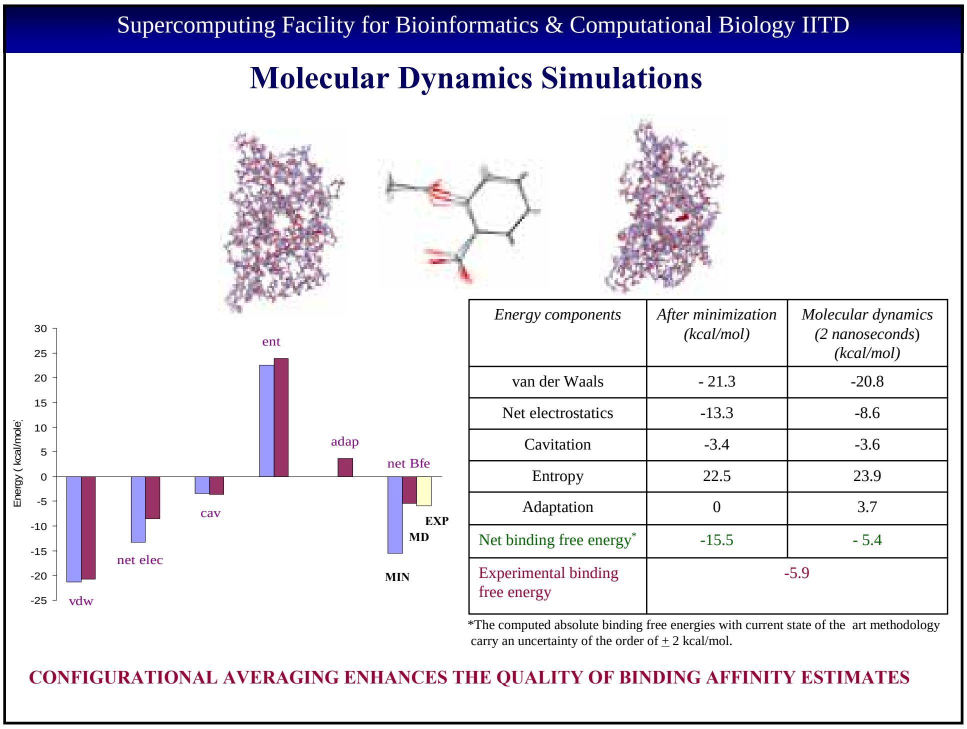
Energy Minimization & Binding Free Energy Estimates

Molecular Dynamics & *post-facto* Binding Affinity Analyses



Supercomputing Facility for Bioinformatics & Computational Biology IITD

Sanjeevini distinguishes Drugs (NSAIDS, blue) from Non-Drugs (red) for COX-2



Supercomputing Facility for Bioinformatics & Computational Biology IITD

Molecular Dynamics Simulations

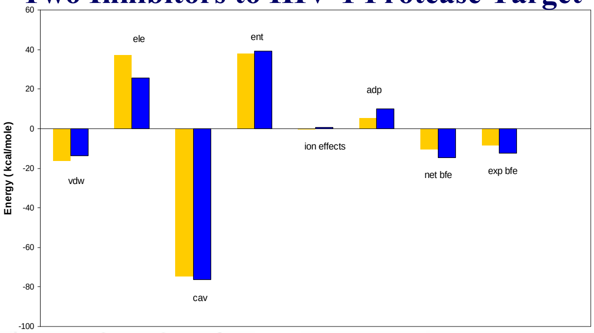
Energy (kcal/mole)
30
25
20
15
10
5
0
-5
-10
-15
-20
-25
vdw
net elec
cav
ent
adap
net Bfe
EXP
MD
MIN

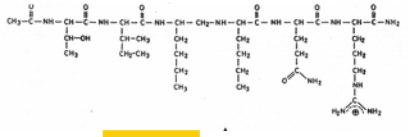
Energy components	*After minimization (kcal/mol)*	*Molecular dynamics (2 nanoseconds) (kcal/mol)*
van der Waals	- 21.3	-20.8
Net electrostatics	-13.3	-8.6
Cavitation	-3.4	-3.6
Entropy	22.5	23.9
Adaptation	0	3.7
Net binding free energy*	-15.5	- 5.4
Experimental binding free energy	-5.9	

*The computed absolute binding free energies with current state of the art methodology carry an uncertainty of the order of ± 2 kcal/mol.

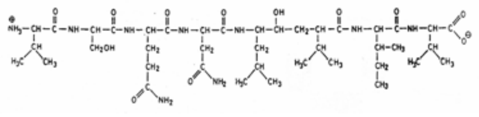
CONFIGURATIONAL AVERAGING ENHANCES THE QUALITY OF BINDING AFFINITY ESTIMATES

Free Energy Component Analysis of Binding of Two Inhibitors to HIV-1 Protease Target





4hvp



8hvp

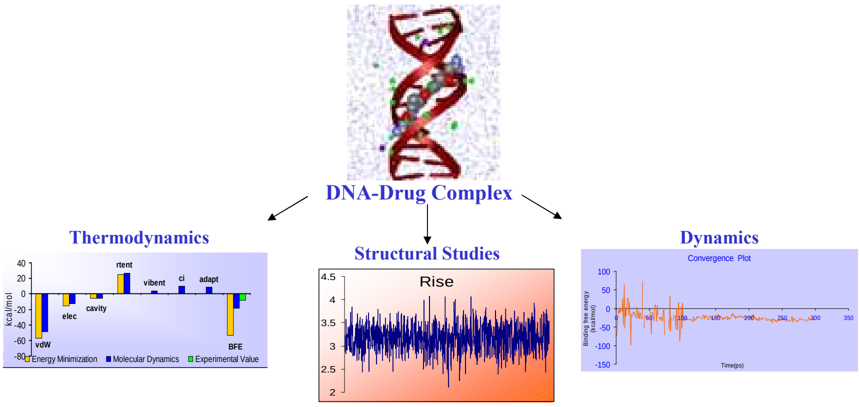
Kalra, P., Reddy, T.V. and Jayaram, B. ‘Free energy component analysis for drug design: a case study of HIV-1 protease-inhibitor binding.’ *J. Med. Chem.* **2001, 44, 4325-4338.**

CPU Times for Various Modules in *Sanjeevini*

MODULE	CPU times*	
	ULTRA SPARCIII	PIV
1.Template library	**Pre-generated database**	
2. Molecule generator	**0m0.024s**	**0m0.002s**
3. Molecular descriptors / drug-like filters	**0m0.084s**	**0m0.016s**
A. Molecular weight	*0m0.008s*	*0m0.001s*
B. Molecular volume	*0m0.020s*	*0m0.006s*
C. Hydrogen bond donors and acceptors	*0m0.016s*	*0m0.002s*
D. log P	*0m0.014s*	*0m0.001s*
E. Molar refractivity	*0m0.014s*	*0m0.001s*
F. Rotatable bonds	*0m0.012s*	*0m0.005s*
4. Molecular docking (@ Nine processors)	**21m15.338s**	**17m40.997s**
5. Structural analysis of the receptor-candidate complex	**0m0.779s**	**0m0.450s**
A. Clash identification	*0m0.573s*	*0m0.434s*
B. RMSD calculation	*0m0.070s*	*0m0.006s*
C. Charge alignment identification	*0m0.068s*	*0m0.005s*
D. Donor / acceptor alignment identification	*0m0.068s*	*0m0.005s*
6. Energy analysis of the receptor-candidate complex	**0m7.621s**	**0m3.775s**
7. Binding affinity analysis	**4m90.254s**	

The time factors are given in minutes (m) and seconds (s). CPU times for all the modules are for single processor, except for Molecular docking (Module 4) which is implemented in parallel mode over nine processors. GAMESS14 and AMBER13 for quantum mechanical and molecular mechanics calculations respectively have been implemented. CPU time for AM1 geometry optimization is 2m7.000s, for HF/6-31G/RESP calculations is 74m2.000s for energy minimization is 16m13.507s and **for a 2 nanosecond molecular dynamics simulation on COX-2 aspirin complex containing 22,442 atoms, with explicit solvent took 210 days..**

DNA-Drug Interaction



Based on detailed thermodynamic, dynamic and structural studies on a series of DNA-minor groove binder complexes, design principles are being incorporated in *Sanjeevini* for DNA-directed lead design

Shaikh, S.A., Ahmed, S.R. and Jayaram, B. 'A molecular thermodynamic view of DNA-drug interaction: A case study of 25 minor groove binders.' *Arch. Biochem. Biophys.* 429, 81, 2004.

SUMMARY

- *Sanjeevini1.0* sorts out drugs from non-drugs for enzyme and receptor targets.

- Predicts relative affinities of drugs in conformity with experiment (COX-2, HIV-1 protease, Estrogen receptor).

- Known specificity of COX inhibitors reproduced.

- An efficient Scoring Function is developed for a rapid assay of candidates to any target

- A small molecule database comprising over 3 million molecules prepared in force-field dependent manner is being developed for high throughput lead discovery

- Work on other systems including diverse targets such as hormone receptors and nucleic acids is in progress

- Several utilities of use in computer aided drug design are made freely accessible at www.scfbio-iitd.res.in/utility.

Genome to drug discovery research
A rough estimate of computational requirements

1. Gene Prediction

Homology/string comparison. — 300 Giga flop

~ $3*10^9$ bp

Time complexity of algorithm [order N] — [100 flops per bp]

2. Protein Structure Prediction

- Threading (time complexity: Exponential) — 100 Giga flop
- Statistical Models
- Filters to reduce guess structures

Molecular Dynamics — 30 Peta flop
100 structures
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)

Supercomputer at SCFBio
2003

A 70 processor machine (over 100 GFlops) with 4.5 terabytes of storage space
Several utilities along with computational resources are freely accessible at www.scfbio-iitd.res.in

Supercomputing Facility for Bioinformatics & Computational Biology IITD

Bioinformatics,
Infrastructure and
BIOGRID India of DBT
🟩 HRD PROGRAMMES
🟨 BIOTECHNOLOGY FACILITIES/REPOSITORIES
🟥 DISTRIBUTED INFORMATION CENTRES
🟦 DISTRIBUTED INFORMATION SUB-CENTRES

SCFBio is currently connected on a VPN to

1) JNU Bioinformatics center

2) University of Delhi (south campus)

3) Madurai Kamaraj University

4) Indian Institute of Science

5) National Institute of Immunology

6) Institute of Microbial Technology Chandigarh

7) DBT CGO Complex

8) University of Pune

9) IGIB Mall Road New Delhi

10) NBRC Gurgaon

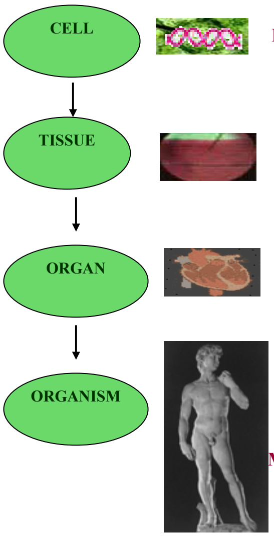
11) CDFD Hyderabad

12) IIT Delhi

***Vision*: SCFBio IIT Delhi as one of the nodal centers with multi Teraflops capacity on a national biocomputing grid with both hardware and bioinformatics software(s) accessible freely, round the clock, to scientists, engineers and students.**

Supercomputing Facility for Bioinformatics & Computational Biology IITD

Projections into the Future of Bioinformatics



CELL

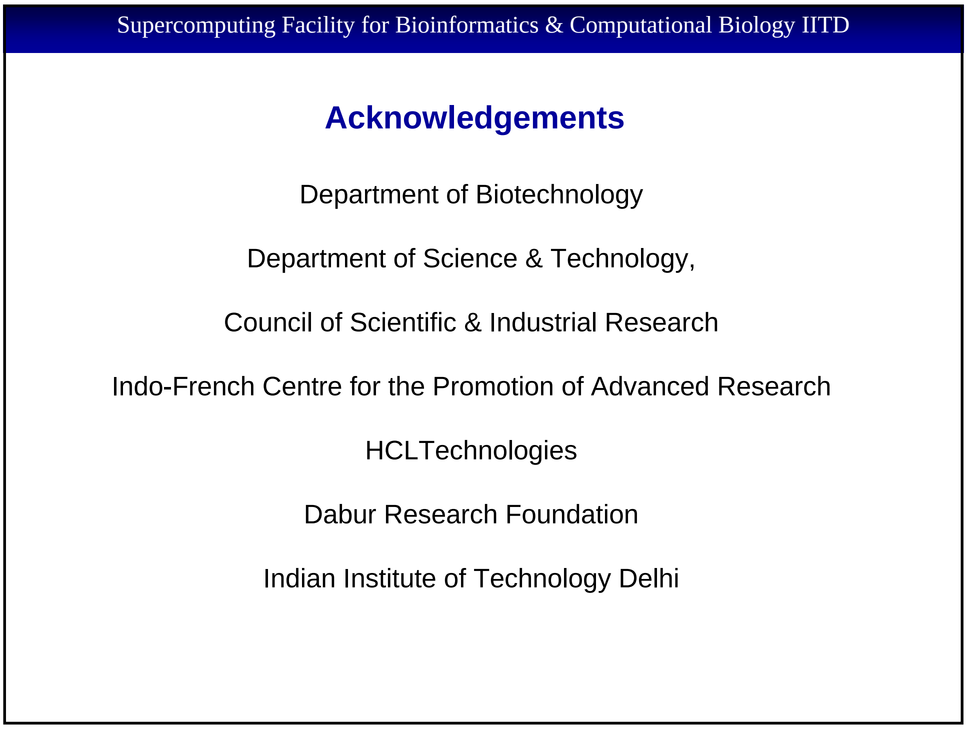
TISSUE

ORGAN

ORGANISM

**Entire DNA Content in Cell of an Organism
2000**

**Molecular Level Movie of an Entire Organism
2025**



Acknowledgements

Department of Biotechnology

Department of Science & Technology,

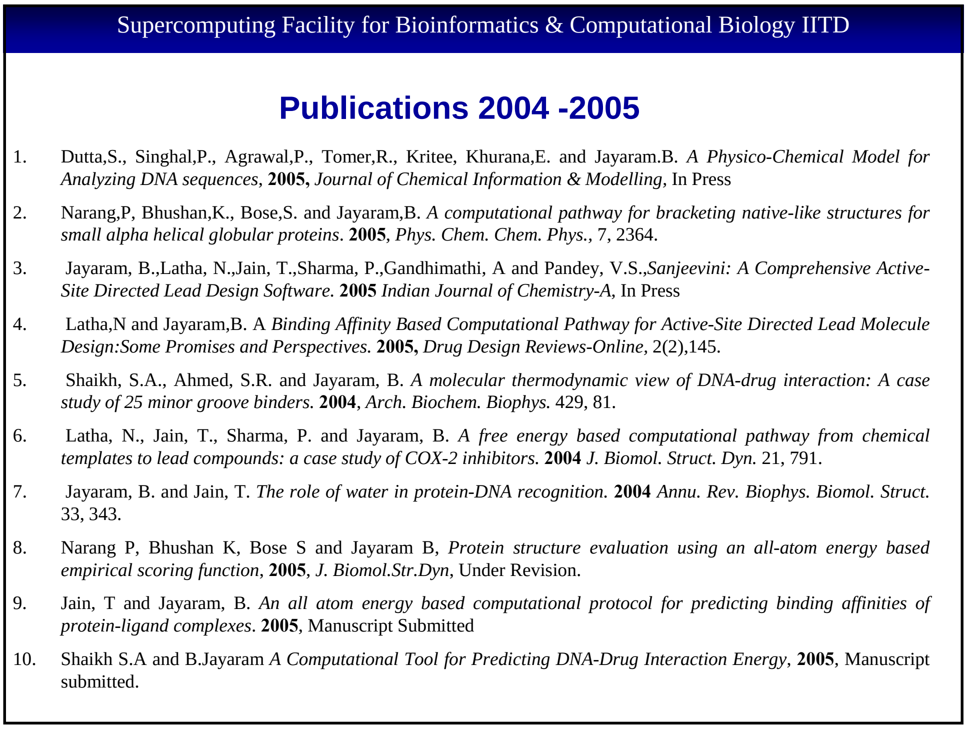
Council of Scientific & Industrial Research

Indo-French Centre for the Promotion of Advanced Research

HCLTechnologies

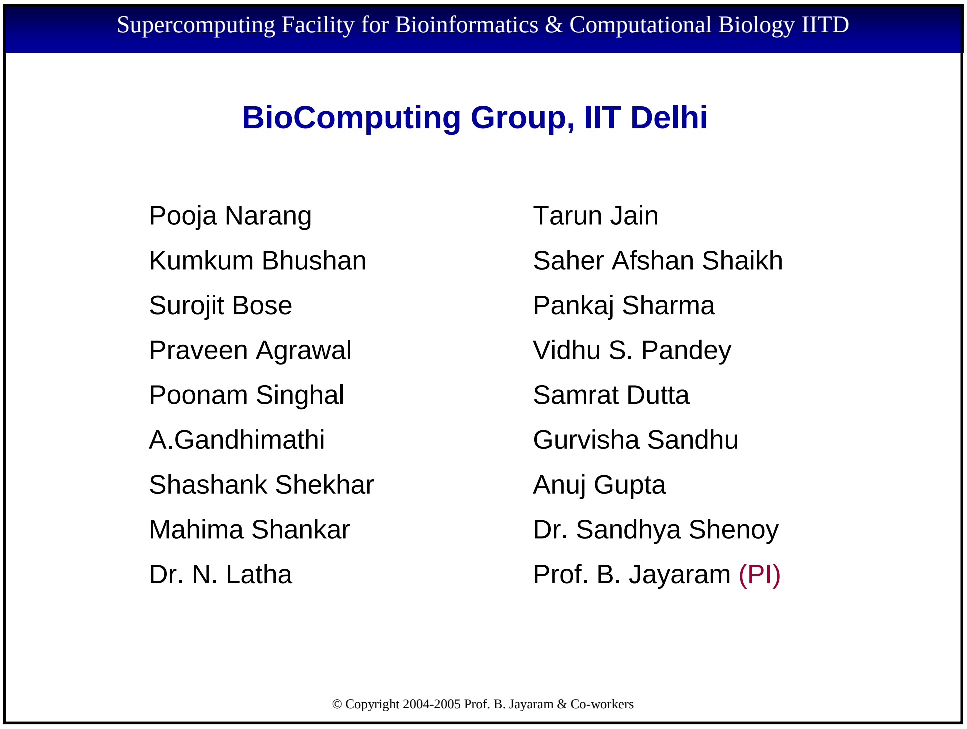
Dabur Research Foundation

Indian Institute of Technology Delhi



Publications 2004 -2005

1. Dutta,S., Singhal,P., Agrawal,P., Tomer,R., Kritee, Khurana,E. and Jayaram.B. *A Physico-Chemical Model for Analyzing DNA sequences*, **2005,** *Journal of Chemical Information & Modelling*, In Press
2. Narang,P, Bhushan,K., Bose,S. and Jayaram,B. *A computational pathway for bracketing native-like structures for small alpha helical globular proteins*. **2005**, *Phys. Chem. Chem. Phys.*, 7, 2364.
3. Jayaram, B.,Latha, N.,Jain, T.,Sharma, P.,Gandhimathi, A and Pandey, V.S.,*Sanjeevini: A Comprehensive Active-Site Directed Lead Design Software.* **2005** *Indian Journal of Chemistry-A*, In Press
4. Latha,N and Jayaram,B. A *Binding Affinity Based Computational Pathway for Active-Site Directed Lead Molecule Design:Some Promises and Perspectives.* **2005,** *Drug Design Reviews-Online*, 2(2),145.
5. Shaikh, S.A., Ahmed, S.R. and Jayaram, B. *A molecular thermodynamic view of DNA-drug interaction: A case study of 25 minor groove binders.* **2004**, *Arch. Biochem. Biophys.* 429, 81.
6. Latha, N., Jain, T., Sharma, P. and Jayaram, B. *A free energy based computational pathway from chemical templates to lead compounds: a case study of COX-2 inhibitors.* **2004** *J. Biomol. Struct. Dyn.* 21, 791.
7. Jayaram, B. and Jain, T. *The role of water in protein-DNA recognition.* **2004** *Annu. Rev. Biophys. Biomol. Struct.* 33, 343.
8. Narang P, Bhushan K, Bose S and Jayaram B, *Protein structure evaluation using an all-atom energy based empirical scoring function*, **2005**, *J. Biomol.Str.Dyn*, Under Revision.
9. Jain, T and Jayaram, B. *An all atom energy based computational protocol for predicting binding affinities of protein-ligand complexes*. **2005**, Manuscript Submitted
10. Shaikh S.A and B.Jayaram *A Computational Tool for Predicting DNA-Drug Interaction Energy*, **2005**, Manuscript submitted.



BioComputing Group, IIT Delhi

Pooja Narang
Kumkum Bhushan
Surojit Bose
Praveen Agrawal
Poonam Singhal
A.Gandhimathi
Shashank Shekhar
Mahima Shankar
Dr. N. Latha

Tarun Jain
Saher Afshan Shaikh
Pankaj Sharma
Vidhu S. Pandey
Samrat Dutta
Gurvisha Sandhu
Anuj Gupta
Dr. Sandhya Shenoy
Prof. B. Jayaram (PI)