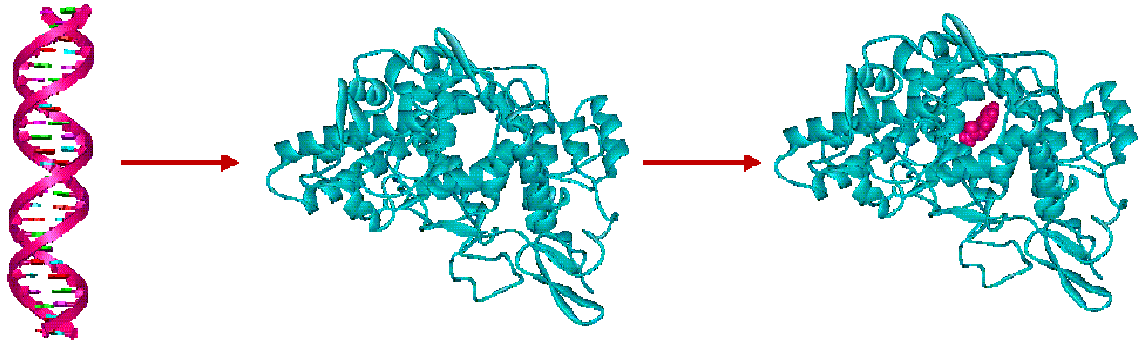


Genomes to Hits:
The Emerging Assembly Line in Silico

Prof. B. Jayaram

Department of Chemistry &

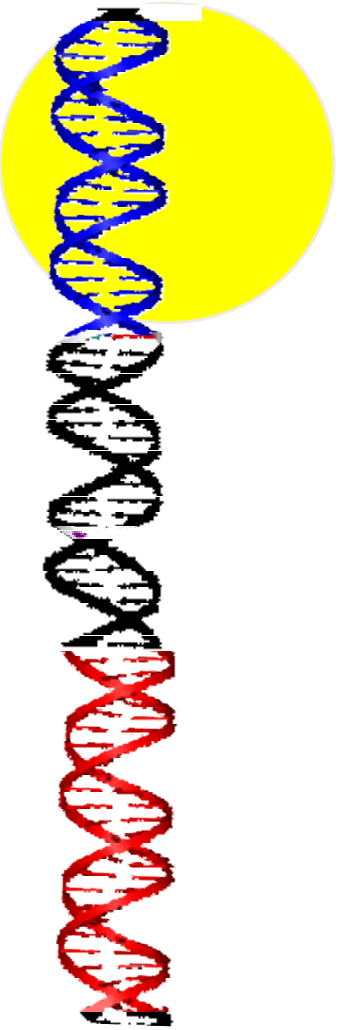
Supercomputing Facility for Bioinformatics & Computational Biology &

School of Biological Sciences

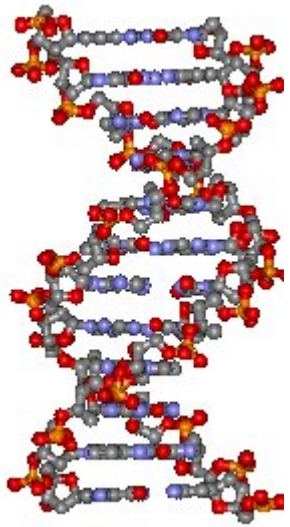
Indian Institute of Technology Delhi

The Dream @ SCFBio:

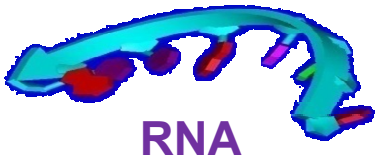
From Genome to Drug : Establishing the Central Dogma of Modern Drug Discovery



ChemGenome



DNA



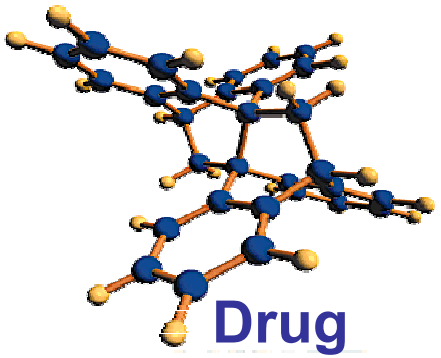
RNA

SEEARTINSCIENCE……

Primary Sequence

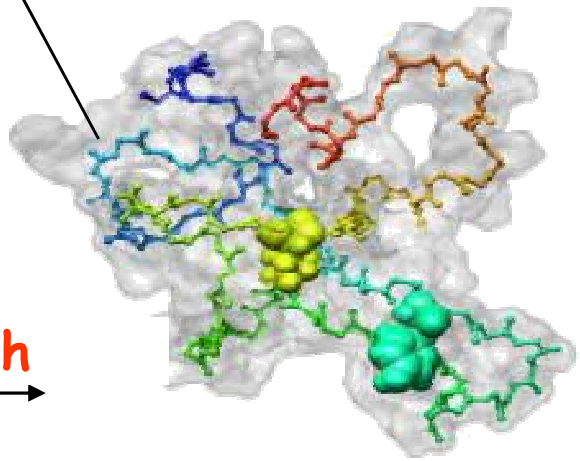
Genome

Sanjeevini



Drug

Bhageerath



Tertiary Structure

Protein

A Case Study

Hepatitis B virus (HBV) is a major blood-borne pathogen worldwide. Despite the availability of an efficacious vaccine, chronic HBV infection remains a major challenge with over 350 million carriers.

No.	HBV ORF	Protein	Function
1	ORF P	Viral polymerase	DNA polymerase, Reverse transcriptase and RNase H activity[36,48].
2	ORF S	HBV surface proteins (HBsAg, pre-S1 and pre-S2)	Envelope proteins: three in-frame start codons code for the small, middle and the large surface proteins[36,49,50]. The pre-S proteins are associated with virus attachment to the hepatocyte[51]
3	ORF C	Core protein and HBeAg	HBcAg: forms the capsid [36]. HBeAg: soluble protein and its biological function are still not understood. However, strong epidemiological associations with HBV replication[52] and risk for hepatocellular carcinoma are known[42].
4	ORF X	HBx protein	Transactivator; required to establish infection *in vivo*[53,54]. Associated with multiple steps leading to hepatocarcinogenesis[45].

United States FDA approved agents for anti-HBV therapy

Agent	Mechanism of action / class of drugs
Interferon alpha	Immune-mediated clearance
Peginterferon alpha2a	Immune-mediated clearance
Lamivudine	Nucleoside analogue
Adefovir dipivoxil	Nucleoside analogue
Tenofovir	Nucleoside analogue
Entecavir	Nucleoside analogue
Telbivudine	Nucleoside analogue

Resistance to nucleoside analogues have been reported in over 65% of patients on long-term treatment. It would be particularly interesting to target proteins other than the viral polymerase.

Wanted: New targets and new drugs

Input the HBV Genome sequence to *ChemGenome 3.0*:

Hepatitis B virus, complete genome
NCBI Reference Sequence: NC_003977.1
>gi|21326584|ref|NC_003977.1| Hepatitis B virus, complete genome

ChemGenome 3.0 output

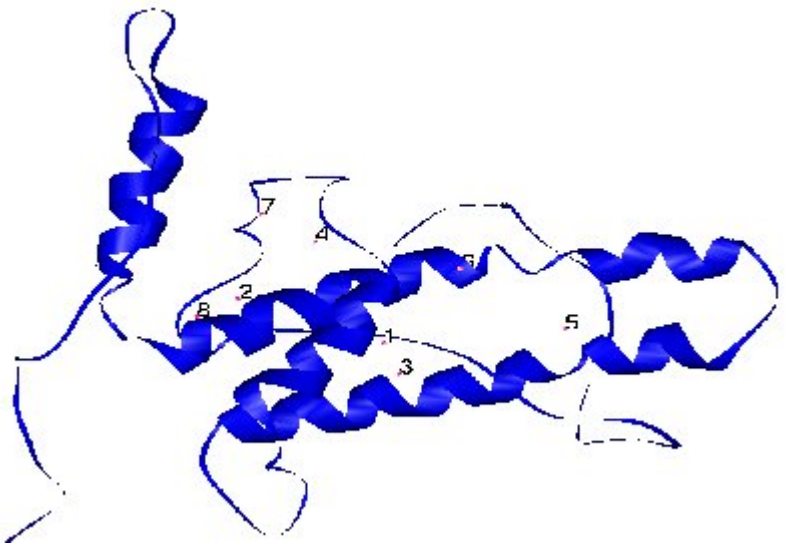
Five protein coding regions identified

Gene 2 (BP: 1814 to 2452) predicted by the *ChemGenome 3.0* software encodes for the HBV precore/ core protein (Gene Id: 944568)

>gi|77680741|ref|YP_355335.1| precore/core protein [Hepatitis B virus]

MQLFPLCLIISCSCPTVQASKLCLGWLWGMDIDPYKE FGASVELLSFLPSDFFPSIRDLLDTASALYREALESPEH CSPHHTALRQAILCWGELMNLATWVGSNLEDPASREL VVSYVNVNMGLKIRQLLWFHISCLTFGRETVLEYLVS FGVWIRTPPAYRPPNAPILSTLPETTVVRRRGRSPRRR TPSPRRRRSQSPRRRRRSQSRESQC

Input Amino acid sequence to Bhageerath



Input Protein Structure to Active site identifier (AADS)

10 potential binding sites identified

A quick scan against a million compound library

RASPD calculation with an average cut off binding affinity to limit the number of candidates.

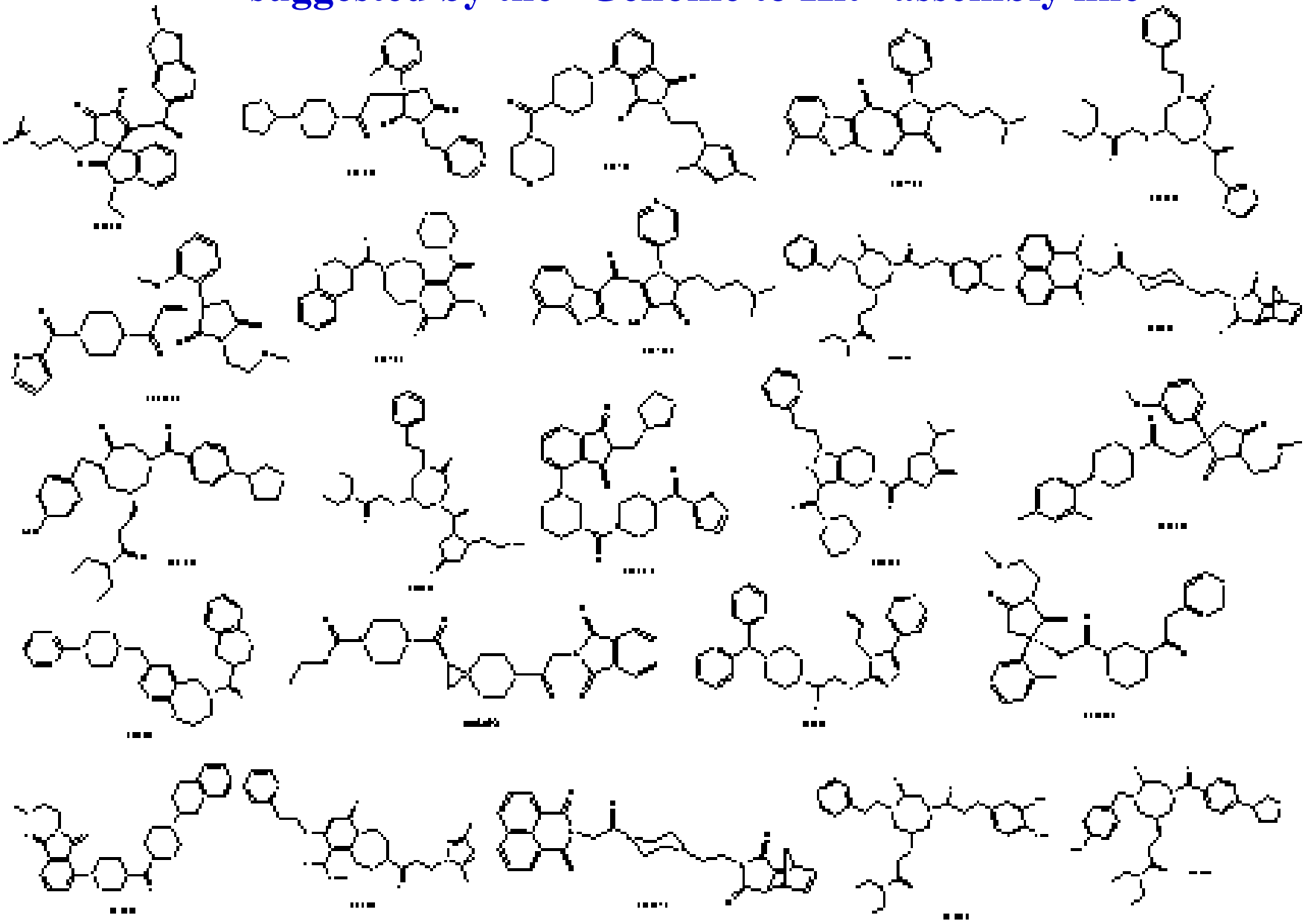
RASPD output

2057 molecules were selected with good binding energy from one million molecule database corresponding to the top 5 predicted binding sites.

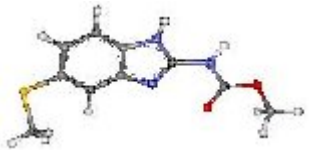
Out of the 2057 molecules, top 40 molecules are given as input to ParDOCK for atomic level binding energy calculations. Out of this 40, (with a cut off of -7.5 kcal/mol), 24 molecules are seen to bind well to precore/core protein target. These molecules could be tested in the Laboratory.

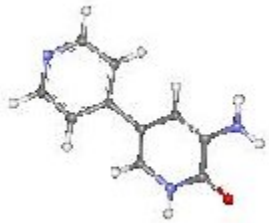
Molecule ID	Binding Energy (kcal/mol)
0001398	-10.14
0004693	-8.78
0007684	-10.05
0007795	-9.06
0008386	-8.38
0520933	-8.21
0587461	-10.22
0027252	-8.39
0036686	-8.33
0051126	-8.73
0104311	-9.3
0258280	-7.8
0000645	-7.89
0001322	-8.23
0001895	-9.49
0002386	-8.53
0003092	-8.35
0001084	-8.68
0002131	-8.07
0540853	-11.08
1043386	-10.14
0088278	-9.16
0043629	-7.5
0097895	-8.04

24 hit molecules for precore/core protein target of HBV are suggested by the "Genome to Hit" assembly line



From Genome to Hits





Genome

X Teraflops
Chemgenome
Bhageerath
Sanjeevini

Hits

www.scfbio-iitd.res.in

- **Genome Analysis - *ChemGenome***

A novel *ab initio* Physico-chemical model for whole genome analysis

- Protein Structure Prediction – *Bhageerath*

A *de novo* energy based protein structure prediction software

- Drug Design – *Sanjeevini*

A comprehensive indigenous active site directed lead molecule design protocol

Arabidopsis Thaliana

(Thale Cress)

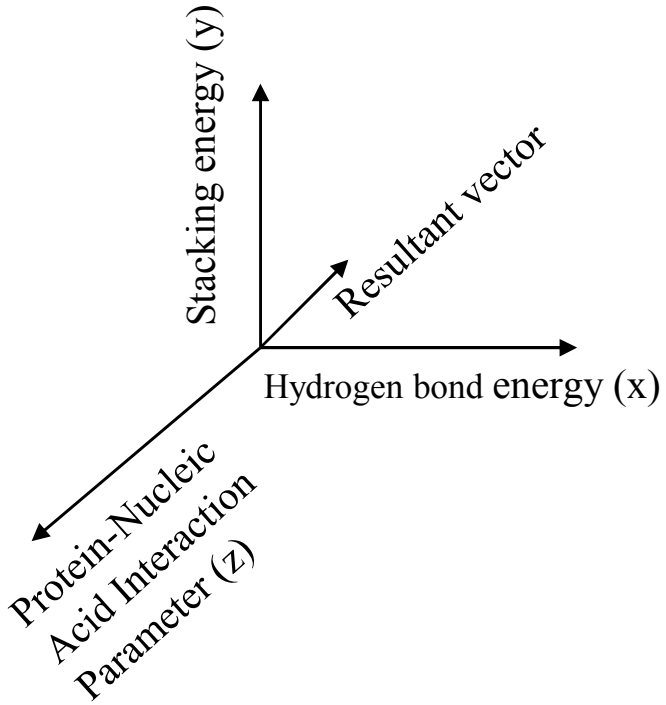
Gene Prediction Accuracies

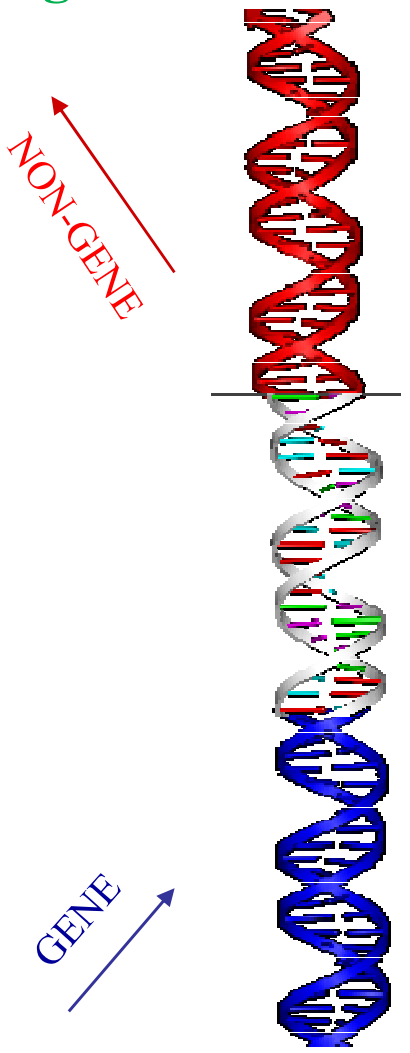
Software	Method	Sensitivity*	Specificity*
GeneMark.hmm http://www.ebi.ac.uk/genemark/	**5th-order Markov model**	**0.82**	**0.77**
GenScan http://genes.mit.edu/GENSCAN.html	**Semi Markov Model**	**0.63**	**0.70**
MZEF http://rulai.cshl.org/tools/genefinder/	**Quadratic Discriminant Analysis**	**0.48**	**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**

***Desirable: A sensitivity & specificity of unity => While it is remarkable that these methods perform so well with very limited experimental data to train on, more research, new methods and new ways of looking at DNA are required.**

ChemGenome:

Build a three dimensional physico-chemical vector which, as it walks along the genome, distinguishes genes from non-genes





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

i……l
j…..m
k…..n

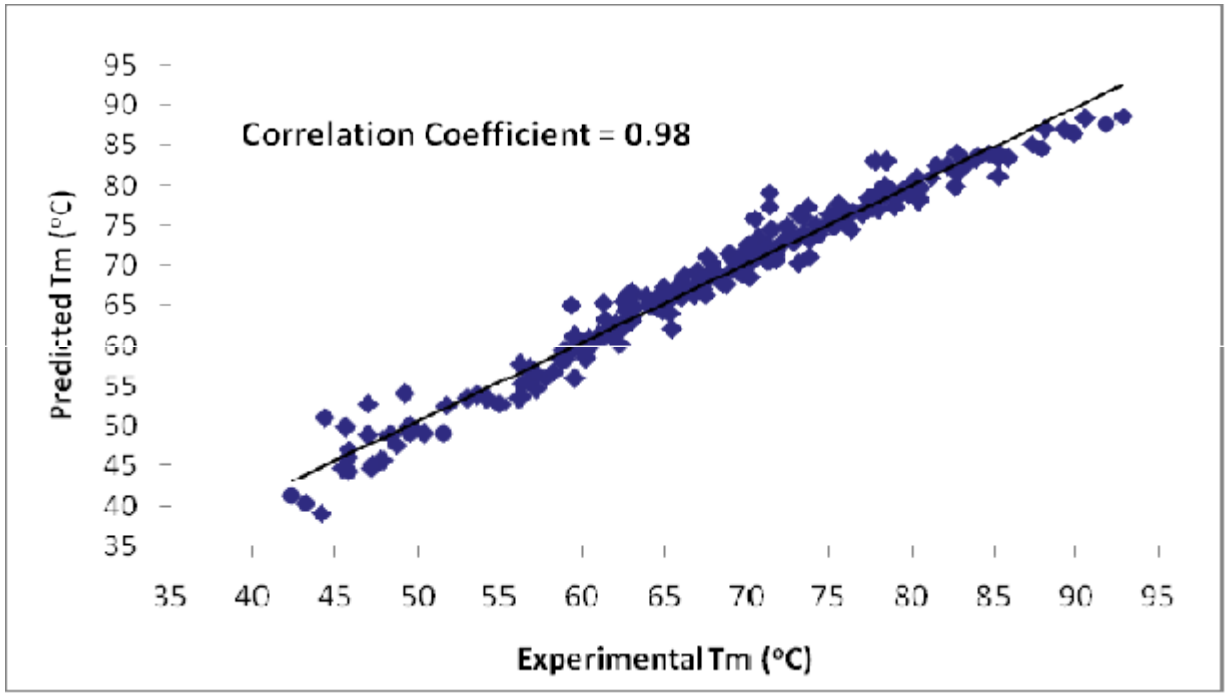
$$E_{HB} = E_{i\text{-}l} + E_{j\text{-}m} + E_{k\text{-}n}$$

$$E_{Stack} = (E_{i\text{-}m}+E_{i\text{-}n}) + (E_{j\text{-}l}+E_{j\text{-}n}) + (E_{k\text{-}l}+E_{k\text{-}m}) +(E_{i\text{-}j}+E_{i\text{-}k}+ E_{j\text{-}k}) + (E_{l\text{-}m}+E_{l\text{-}n}+ E_{m\text{-}n})$$

Hydrogen bond & Stacking energies for all 32 unique trinucleotides were calculated from 50 ns long [*] *Molecular Dynamics Simulation Trajectories on 39 sequences encompassing all possible tetranucleotides in the [#] ABC database* and the data was averaged out from the multiple copies of the same trinucleotide. The resultant energies were then linearly mapped onto the [-1, 1] interval giving the x & y coordinates for each of the 64 codons.

[*] Beveridge et al. (2004). *Biophys J* 87, 3799-813.
[#] Dixit et al. (2005). *Biophys J* 89, 3721-40.

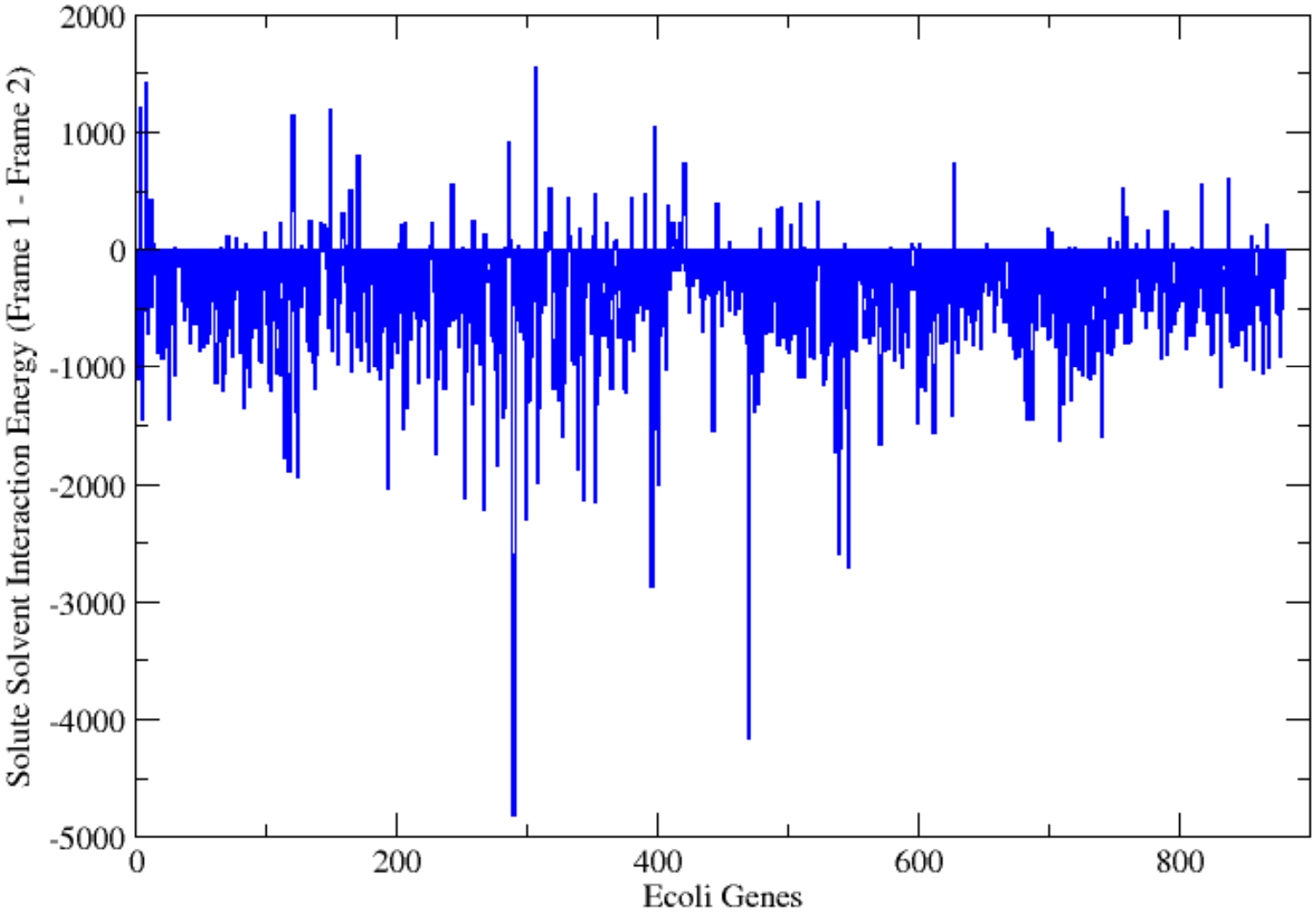
Melting temperatures of ~ 200 oligonucleotides: Prediction versus Experiment



$$Tm(°C)=(7.35 \times E) + [\ 17.34 \times \ln(Len)\ ] + [4.96 \times \ln(Conc)] + [0.89 \times \ln(DNA)] - 25.42$$

The computed (hydrogen bond + stacking) energy (E) correlates very well with experimental melting temperatures of DNA oligonucleotides

Solute-Solvent Interaction Energy for Genes/Non-genes



Coding and non-coding frames have different solvation characteristics which could be used to build the third parameter (z) besides hydrogen bonding (x) & stacking (y)

TTT Phe -1	GGT Gly +1	TAT Tyr -1	GCT Ala +1
TTC Phe -1	GGC Gly +1	TAC Tyr -1	GCC Ala +1
TTA Leu -1	GGA Gly +1	TAA Stop -1	GCA Ala +1
TTG Leu -1	GGG Gly +1	TAG Stop -1	GCG Ala +1
ATT Ile -1	CGT Arg +1	CAT His +1	ACT Thr -1
ATC Ile +1	CGC Arg -1	CAC His -1	ACC Thr +1
ATA Ile +1	CGA Arg -1	CAA Gln -1	ACA Thr +1
ATG Met -1	CGG Arg +1	CAG Gln +1	ACG Thr -1
TGT Cys -1	GTT Val +1	AAT Asn -1	CCT Pro +1
TGC Cys -1	GTC Val +1	AAC Asn +1	CCC Pro -1
TGA Stop -1	GTA Val +1	AAA Lys +1	CCA Pro -1
TGG Trp -1	GTG Val +1	AAG Lys -1	CCG Pro +1
AGT Ser -1	CTT Leu +1	GAT Asp +1	TCT Ser -1
AGC Ser +1	CTC Leu -1	GAC Asp +1	TCC Ser -1
AGA Arg +1	CTA Leu -1	GAA Glu +1	TCA Ser -1
AGG Arg -1	CTG Leu +1	GAG Glu +1	TCG Ser -1

Conjugate rule acts as a good constraint on the 'z' parameter of *Chemgenome* or one could simply use +1/-1 as in the Table for 'z'!!

Extent of Degeneracy in Genetic Code is captured by *Rule of Conjugates*:

$A_{1,2}$ is the conjugate of $C_{1,2}$ & $U_{1,2}$ is the conjugate of $G_{1,2}$:(A_2 x C_2 & G_2 x U_2)

With 6 h-bonds at positions 1 and 2 between codon and anticodon, third base is inconsequential

With 4 h-bonds at positions 1 and 2 third base is essential

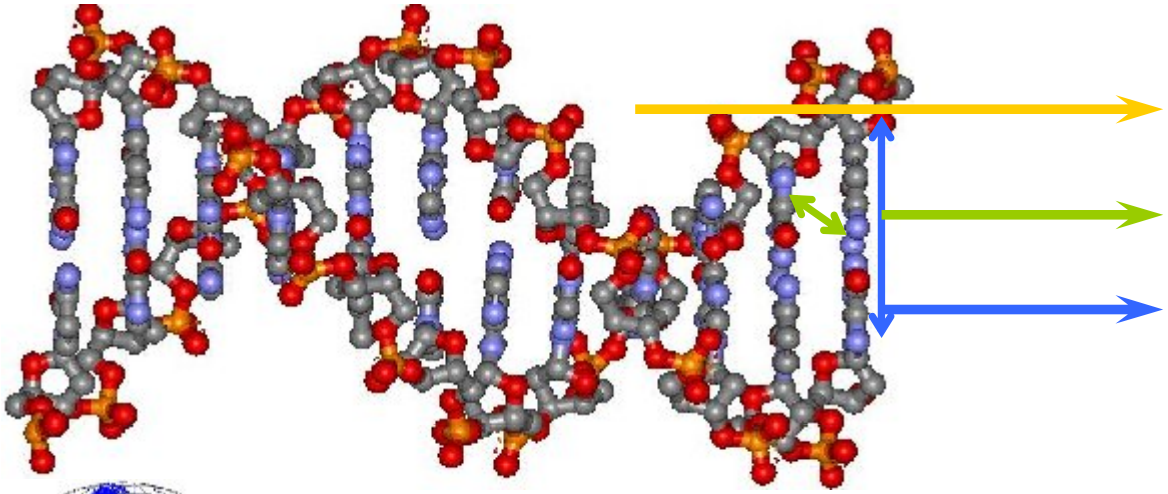
With 5 h-bonds middle pyrimidine renders third base inconsequential;
middle purine requires third base.

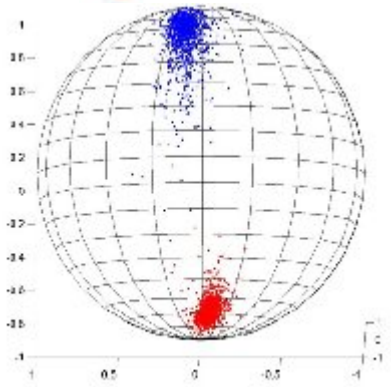
B. Jayaram, "Beyond Wobble: The Rule of Conjugates", *J. Molecular Evolution*, **1997,** *45*, 704-705.

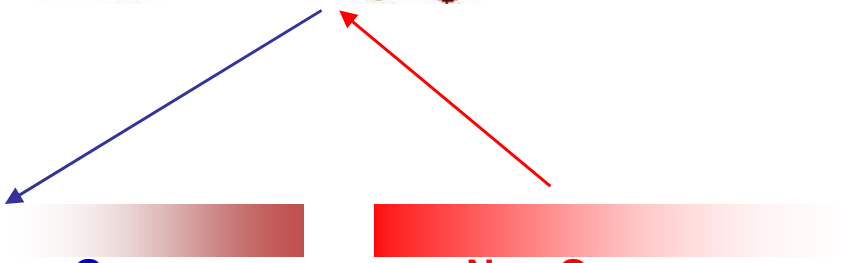
Codons with G_3 → +1; C_3G_3 or C_3T_3 → +1; C_3A_3 or C_3C_3 → -1

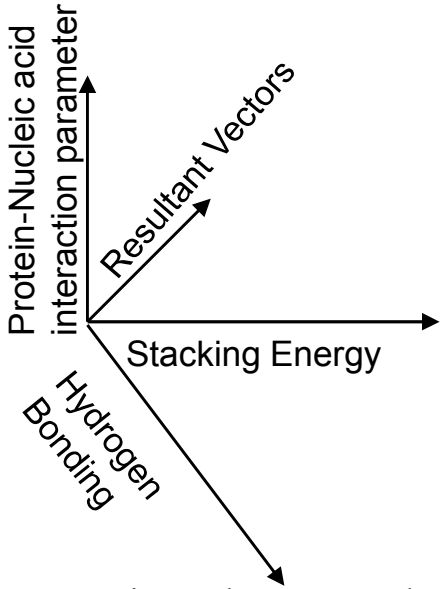
ChemGenome

A Physico-Chemical Model for identifying signatures of functional units on Genomes



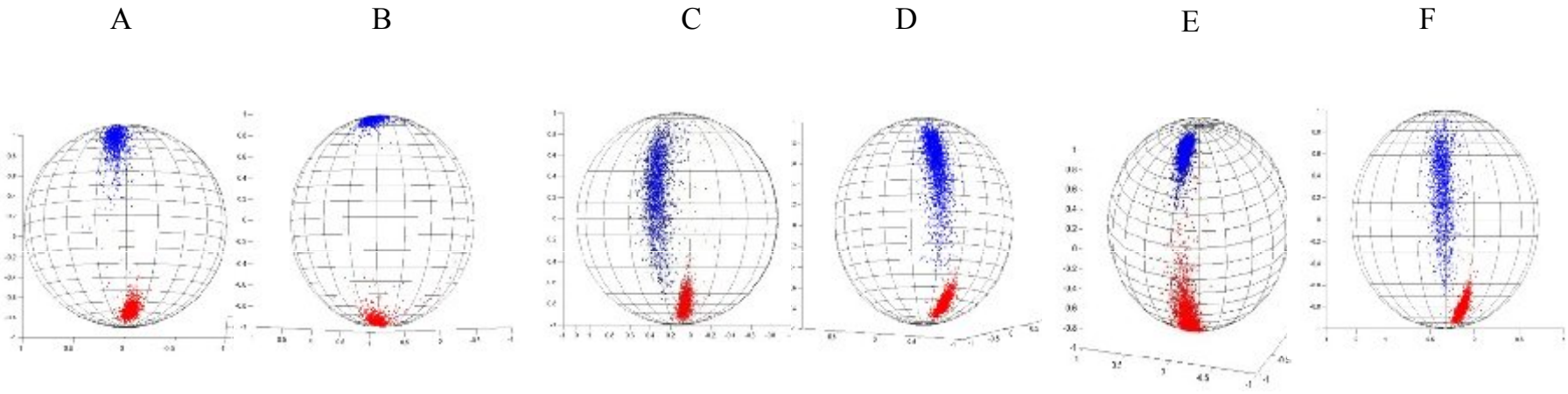






(1) "A Physico-Chemical model for analyzing DNA sequences", Dutta S, Singhal P, Agrawal P, Tomer R, Kritee, Khurana E and Jayaram B, J.Chem. Inf. Mod. , 46(1), 78-85, 2006; (2) "Molecular Dynamics Based Physicochemical Model for Gene Prediction in Prokaryotic Genomes ", P. Singhal, B. Jayaram, S. B. Dixit and D. L. Beveridge,, *Biophys. J.*, 2008, 94, 4173-4183; (3) "A phenomenological model for predicting melting temperatures of DNA sequences", G. Khandelwal and B. Jayaram, PLoS ONE, 2010, 5(8): e12433. doi:10.1371/journal.pone.0012433

Distinguishing Genes (blue) from Non-Genes (red) in ~ 900 Prokaryotic Genomes



Three dimensional plots of the distributions of gene and non-gene direction vectors for six best cases (A to F) 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)

Computational Protocol Designed for Gene Prediction

Read the complete genome sequence in the FASTA format

↓

Search for all possible ORFs in all the six reading frames

↓

Calculate resultant unit vector for each of the ORFs

↓

Classify the ORFs as genes or nongenes depending on their orientation w.r.t. universal plane (DNA space)

↓

Genes and false positives

↓

Screening of potential genes based on stereochemical properties of proteins (Protein space)

↓

Second stage screening based on amino acid frequencies in Swissprot proteins (Swissprot space)

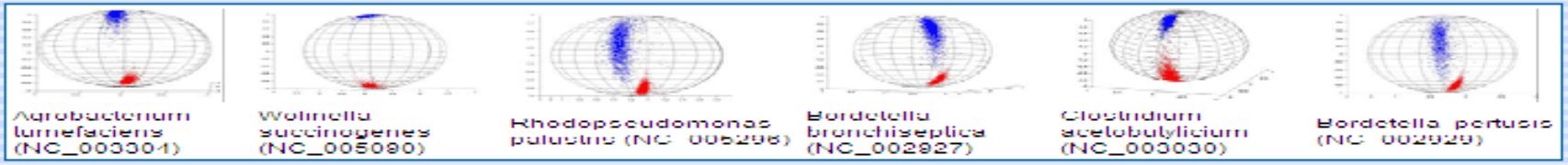
↓

Potential protein coding genes

http://www.scfbio-iitd.res.in/chemgenome/index.jsp

ChemGenome 1.1
GENE EVALUATOR

ChemGenome is a physico-chemical method [1] which accepts DNA sequence in FASTA format and characterizes it as gene or nongene based on hydrogen bonding energy, stacking energy and groove potentials for each trinucleotide (codon).



Above is a pictorial representation of the separation of genes(blue) from non genes(red).
ChemGenome is ab initio in nature and has been tested on 294706 experimentally verified genes in 331 prokaryotic genomes. The observed average sensitivity, specificity & correlation-coefficient are found to be 96.9% (min: 90%, max: 100%), 86.0% & 85.0% respectively. Preliminary studies on eukaryotic genomes show that the model successfully separates the exonic regions from the non-coding regions.A software for whole genome analysis is available at www.scfbio-iitd.res.in/chemgenome2

ChemGenome

Please specify the E-mail id : uilesh@scfbio-iitd.res.in

Insert the Nucleotide sequence (in FASTA format)* : Help

>Gene Name (This comment line is necessary)
ATGTTGGTGTCCCCCAAGGGTACACAAAACAAAAGCCGTGTTGCTTATCACGGCGAACGCCGACAGTGCTTGCTCTCCCG
TAAGG
CGTTGCCCCACCAATGTTGTTTCCCACGACAAATCTCGTCGACCAGTATCTCCCGTCAAATCAAATCCCATAACCTTTC
TAT
CAAAAGACAAGCTGCAACACTTGTGCAAAAGCACAACTGTCAAGACACGCTACACAGTCATGTCACGGGGAGACG
CTGCAC
AAATACCCTGAACTAGCAACCGAGGGTTCCCCAACCATCAAACAGAGGCTTGAGATTGCAAACGATGCGGTTGT
CCACA

SUBMIT RESET

Browse... Upload

Instructions for using the Tool

- The tool takes DNA sequence in FASTA format as input file.
- Browse to select the input file and upload.
- The input file can contain multiple sequences, each sequence being in FASTA format.
- For multiple sequences, please specify the E mail address or wait for a few minutes to get the on-line result.
- Click on Submit to get the result
- For further information, please see the Help file.

Suggestions and Comments

We will be glad to receive your suggestions and comments/feedback at scfbio@scfbio-iitd.res.in.

References

[1] "A Physico-Chemical model for analyzing DNA sequences", Dutta S, Singhal P, Agrawal P, Tomar R, Kritee, Khurana E and Jayaram B,*J.Chem. Inf. Mod.* ,46 (1), 78 - 85, 2006.[ABSTRACT].

[2] "Beyond the Wobble : The rule of conjugates", Jayaram B, *Journal of Mol. Evol.*,1997.45.701.

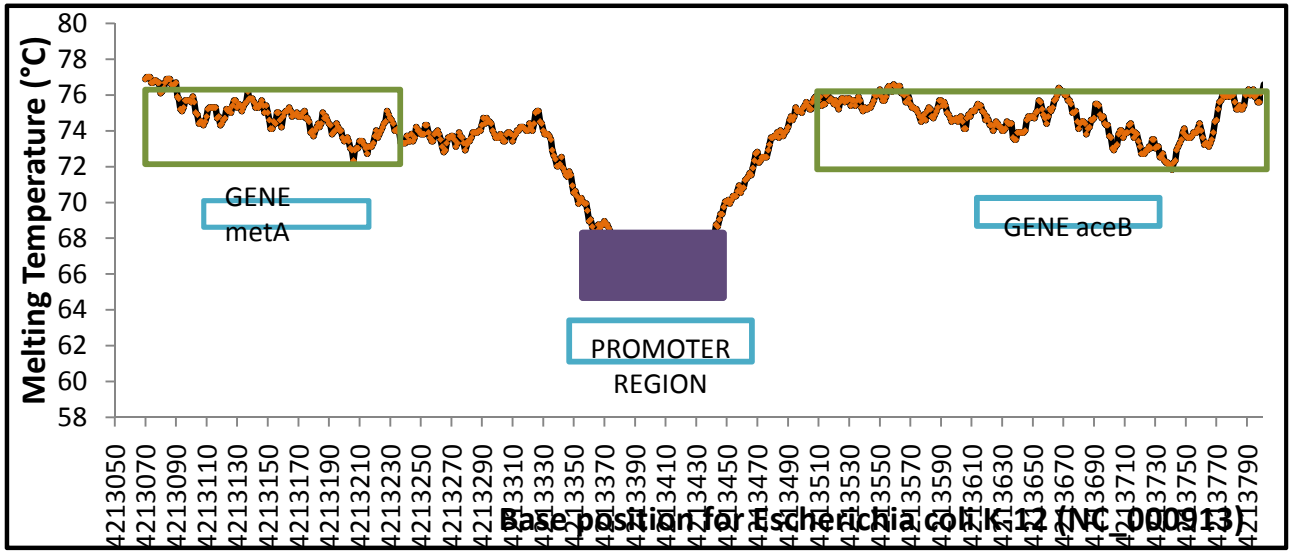
The ChemGenome2.0 WebServer

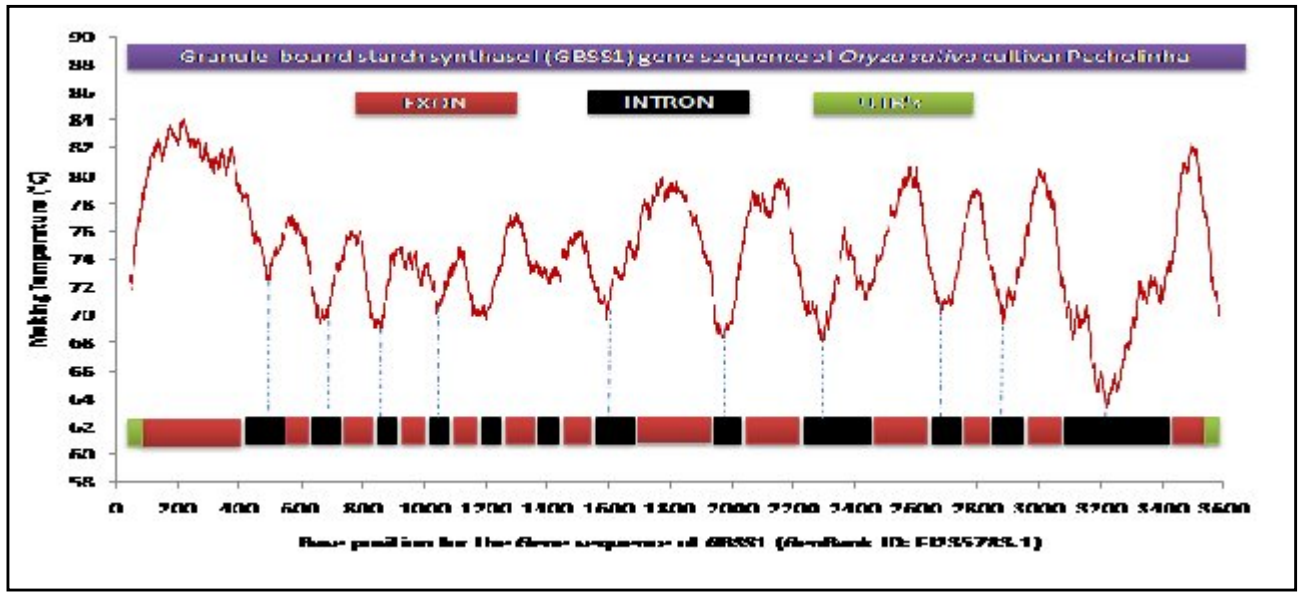
http://www.scfbio-iitd.res.in/chemgenome/chemgenomenew.jsp

Arabidopsis Thaliana
(Thale Cress)

Software	Method	Sensitivity	Specificity
ChemGenome www.scfbio-iitd.res.in/chemgenome	**Physico-chemical model**	**0.87**	**0.89**
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**

A simple physico-chemical model works just as well as any of the sophisticated knowledge base driven methods and has scope for further systematic improvements





Chemgenome **methodology enables detection of not only protein coding regions on a genome but also promoters and introns etc..**

Some day, it should be possible to read the book of Human Genome like a novel

3000 Mb

- **Gene & Gene related Sequences** — 900 Mb
  - **Coding DNA** — 90 Mb (3%) !!!
  - **Non-coding DNA** — 810 Mb
- **Extra-genic DNA** — 2100 Mb
  - **Repetitive DNA** — 420 Mb
    - **Tandemly repeated DNA**
      - Satellite, micro-satellite, mini-satellite DNA
    - **Interspersed genome wide repeats**
      - LTR elements, Lines, Sines, DNA Transposons
  - **Unique & low copy number** — 1680 Mb

www.scfbio-iitd.res.in

- Genome Analysis - *ChemGenome*

A novel *ab initio* Physico-chemical model for whole genome analysis

- **Protein Structure Prediction – *Bhageerath***

A *de novo* energy based protein structure prediction software

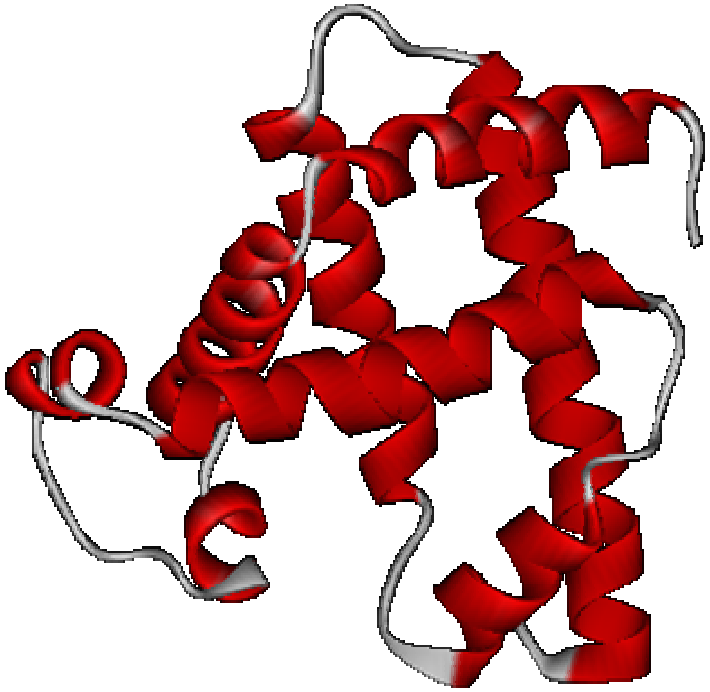
- Drug Design – *Sanjeevini*

A comprehensive indigenous active site directed lead molecule design protocol

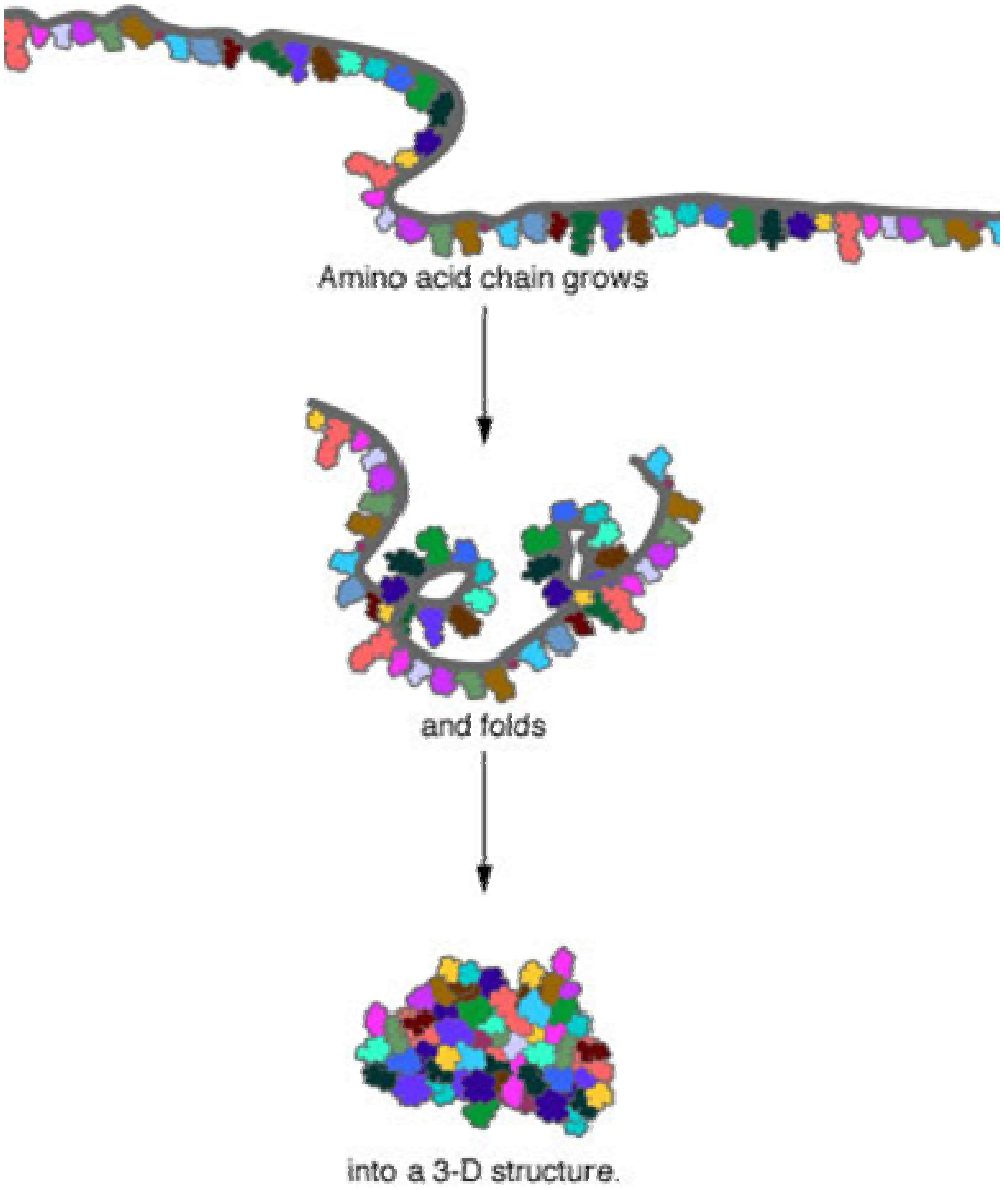
Bhageerath

Protein Tertiary 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.....

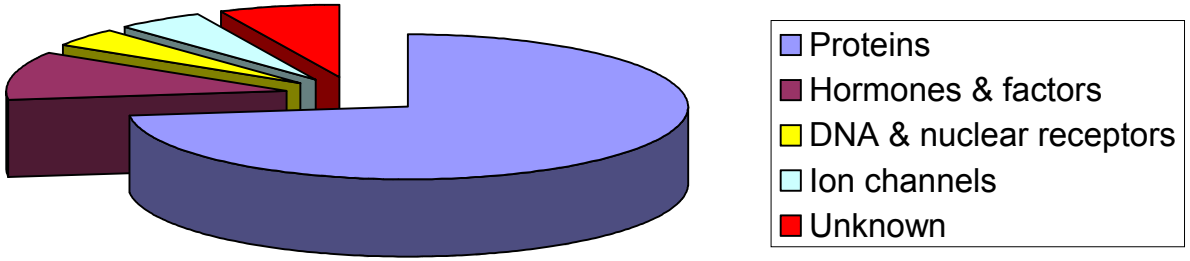


Protein Folding Problem

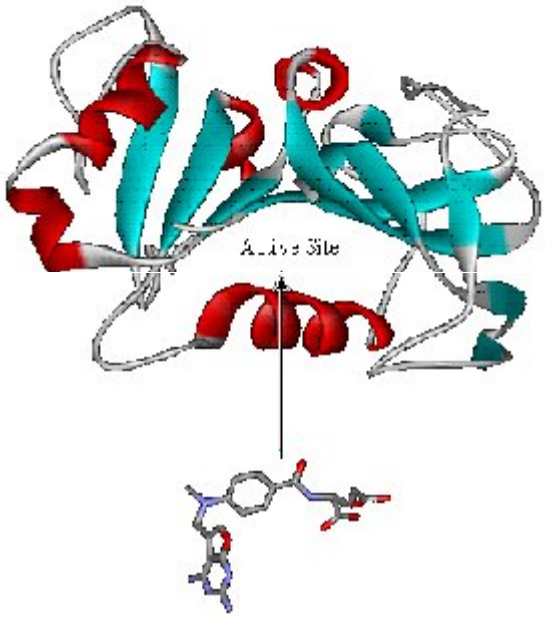


WHY FOLD PROTEINS ?

Pharmaceutical/Medical Sector



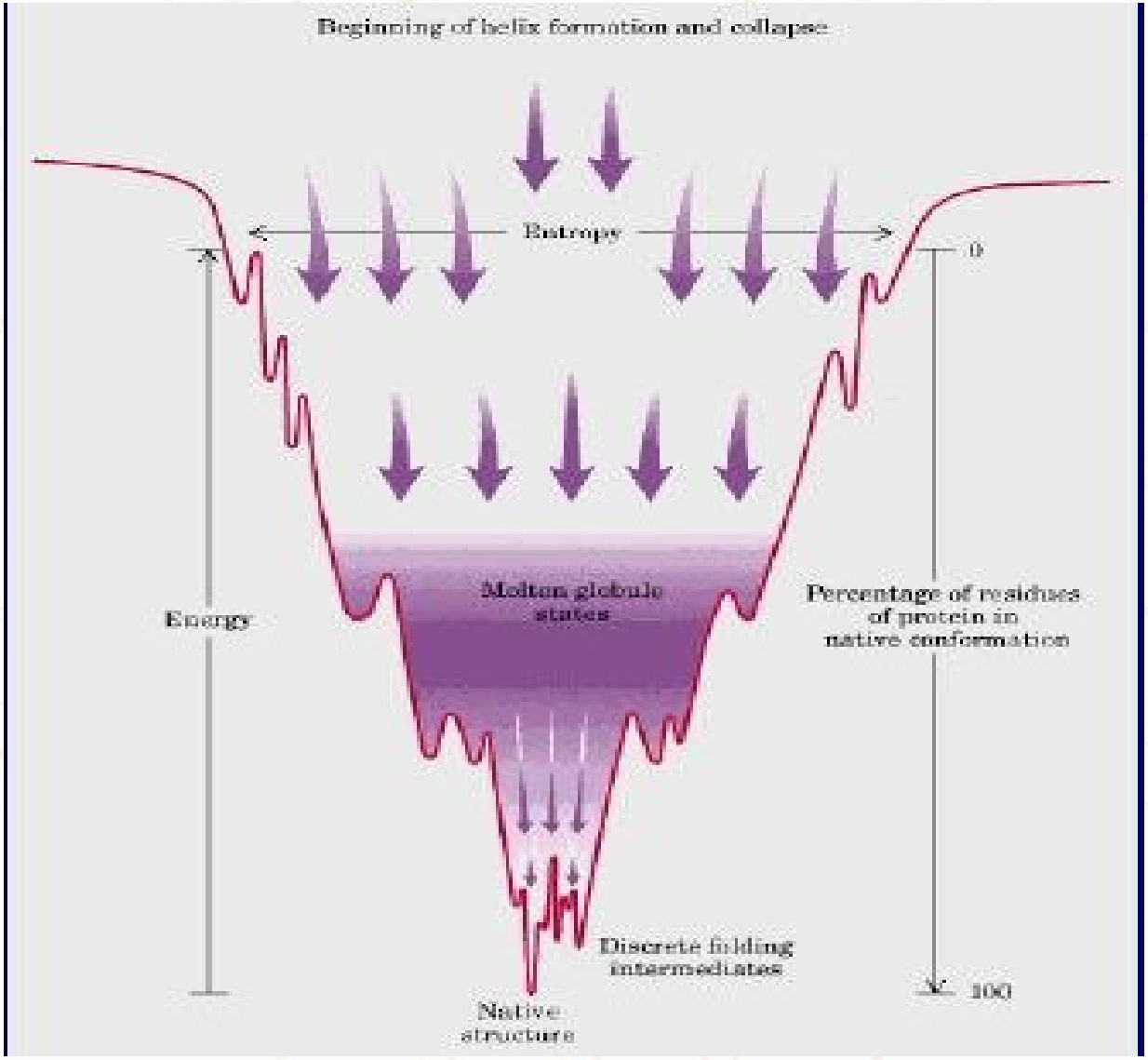
Drug Targets



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

PROTEIN FOLDING LANDSCAPE

Native structure at the bottom of the rugged free energy well is the folded protein.



Protein Folding is considered as a Grand Challenge Problem!

Protein Structure Prediction Approaches

Comparative Modeling

Homology

Similar sequences adopt similar fold is the basis.

Alignment is performed with related sequences. (SWISS-MODEL-www.expasy.org, 3D JIGSAW-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)

Computational Requirements for *ab initio* Protein Folding

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

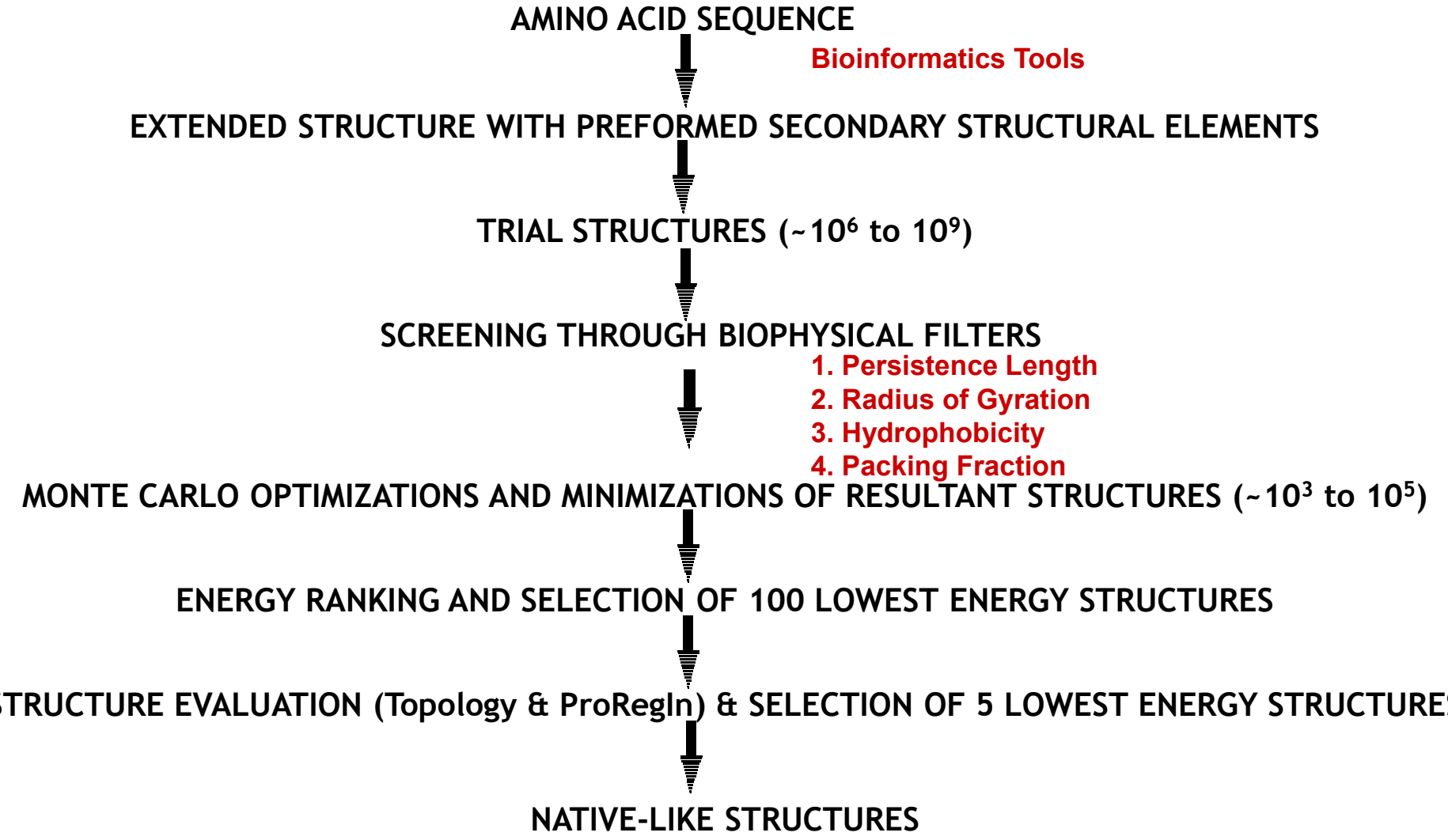
~ 10^{-10} sec / day / processor

- 10^{-2} sec => 10^{8} days

~ 10^{6} years

With 10^{6} processors ~ 1 Year

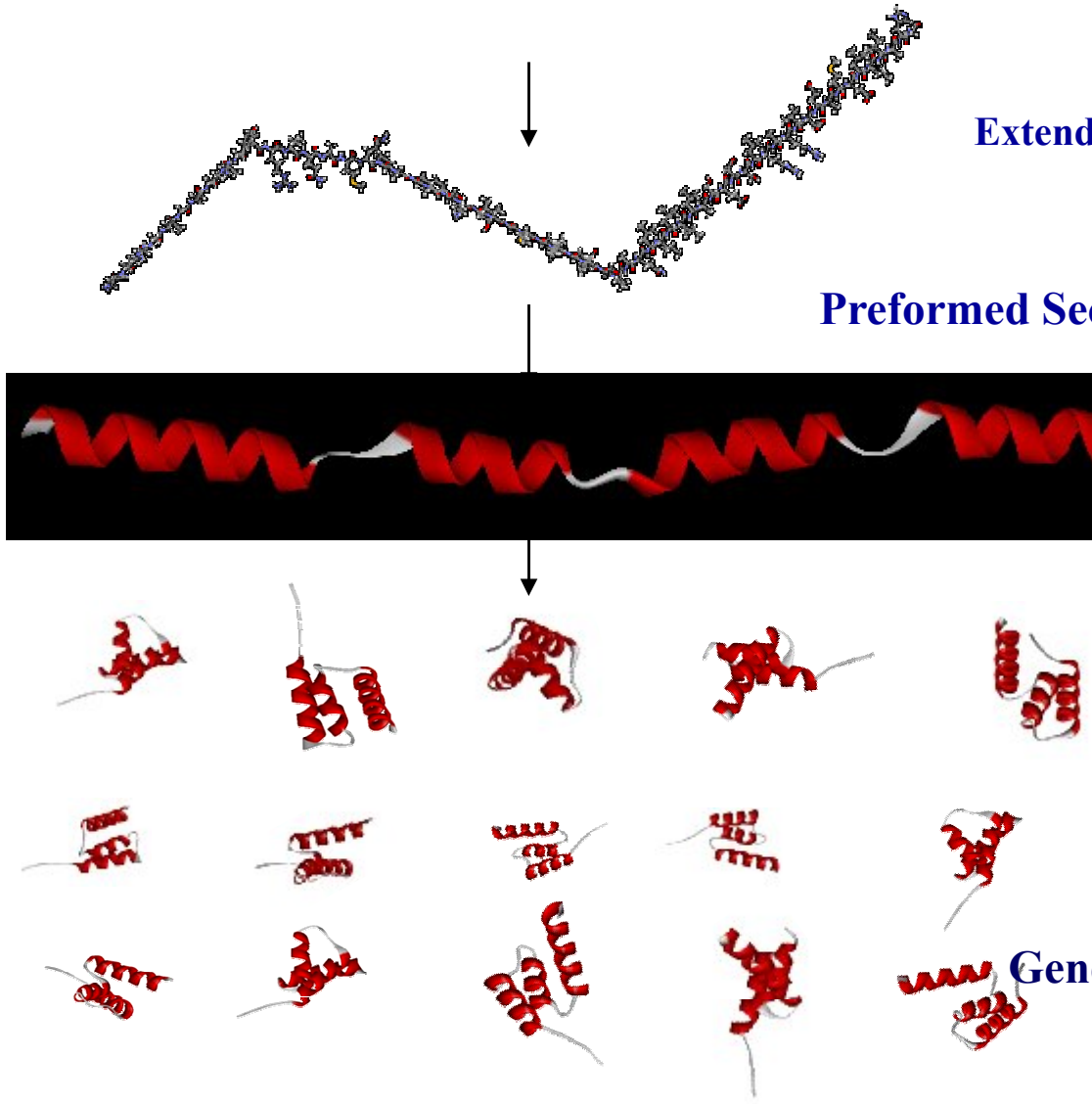
From Sequence to Structure: The *Bhageerath* Pathway



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.

Sampling 3D Space

HRQALGERLYPRVQAMQPAFASKITGMLLELSPAQLLLLLASENSLRARVNEAMELIIAHG



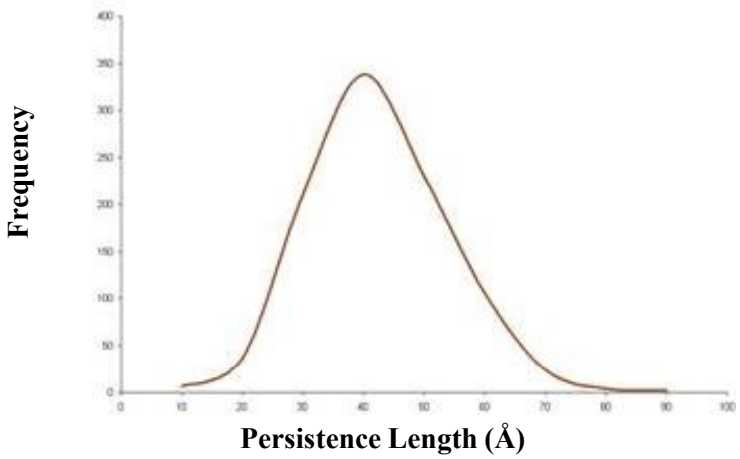
Extended Chain

Preformed Secondary Structural Units

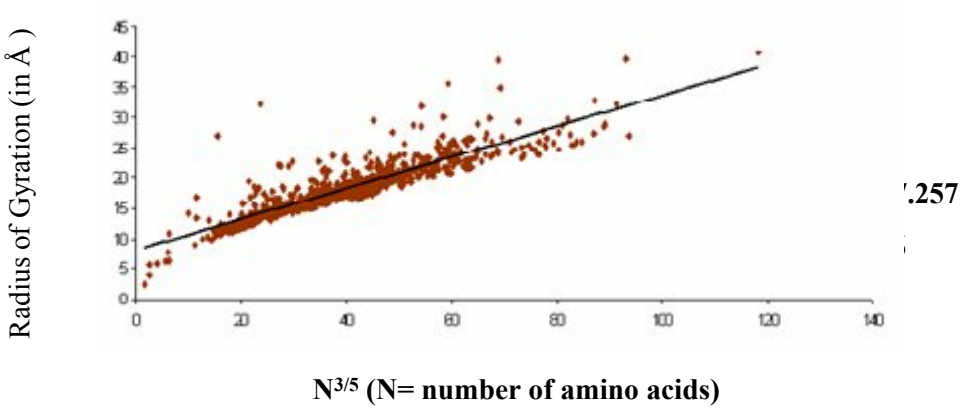
Generation of Trial Structures

Filter-Based Structure Selection

Persistence Length Analysis of 1,000 Globular Proteins



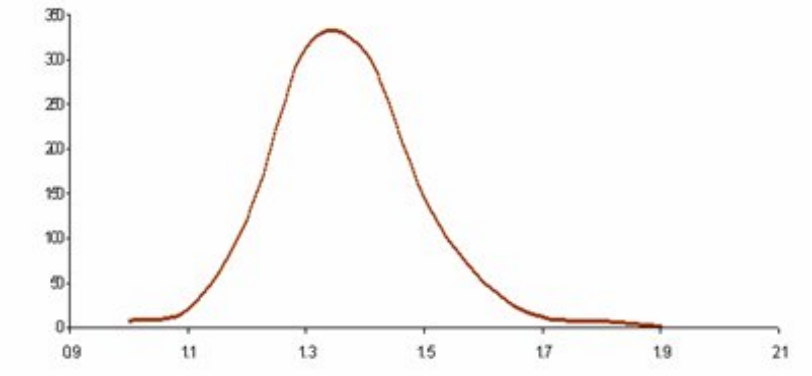
Radius of Gyration vs $N^{3/5}$ of 1,000 Globular Proteins



$N^{3/5}$ (N= number of amino acids)

$N^{3/5}$ 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

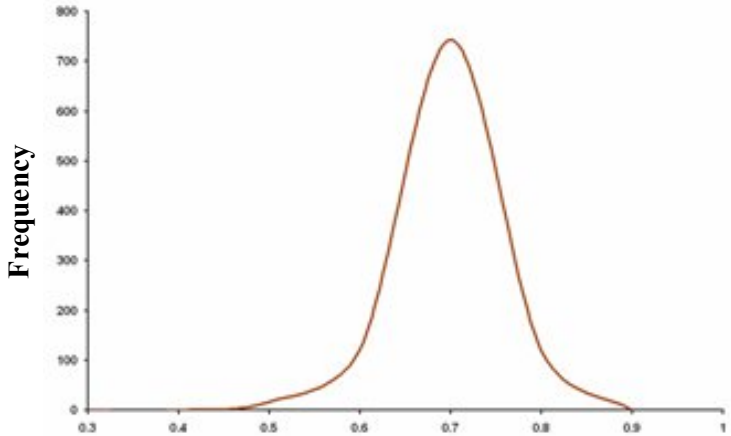


Hydrophobicity Ratio (Φ_H)

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

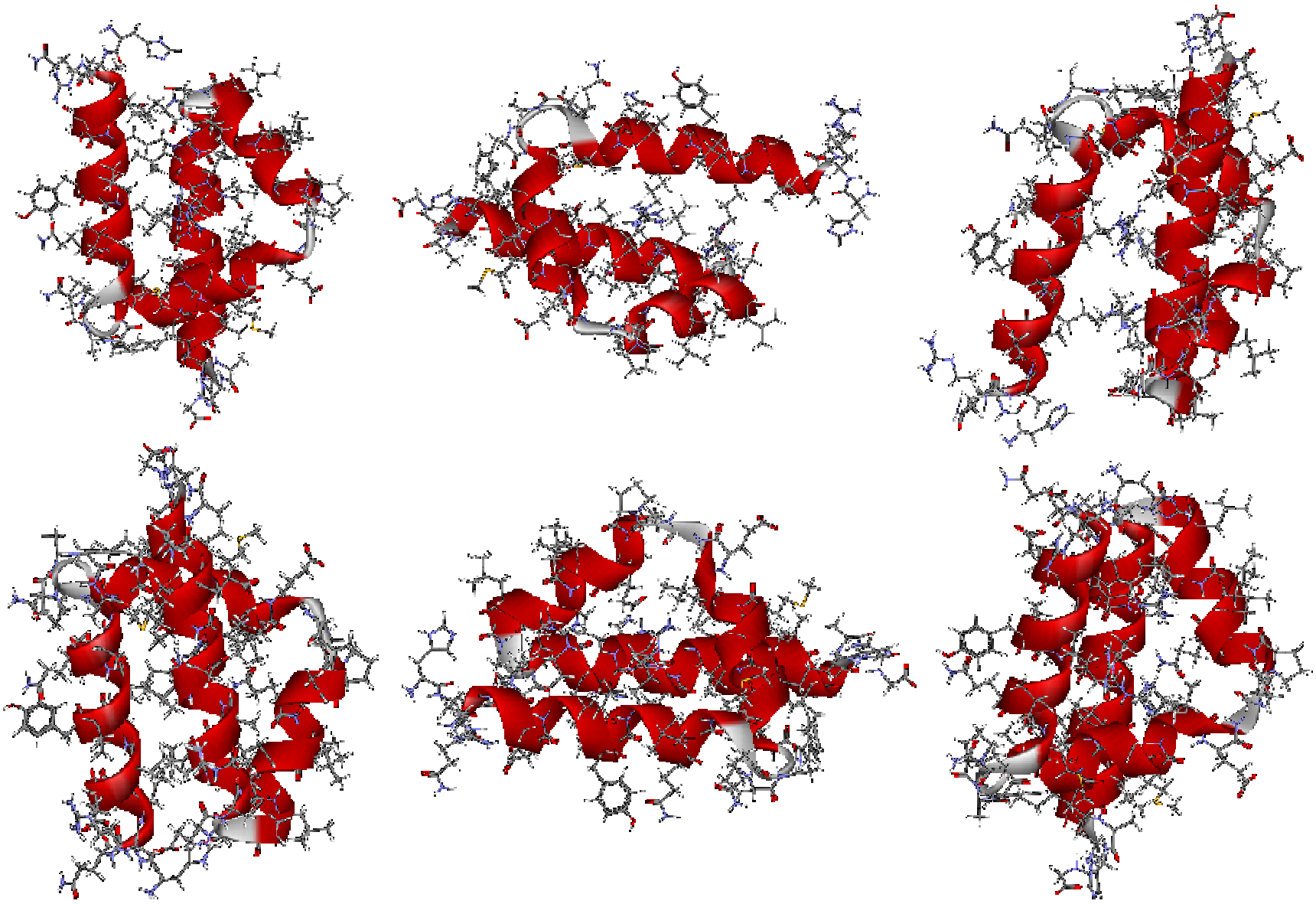
ASA : Accessible surface area

Frequency vs Packing Fraction of 1,000 Globular Proteins



Globular proteins are known to exhibit packing fractions around 0.7

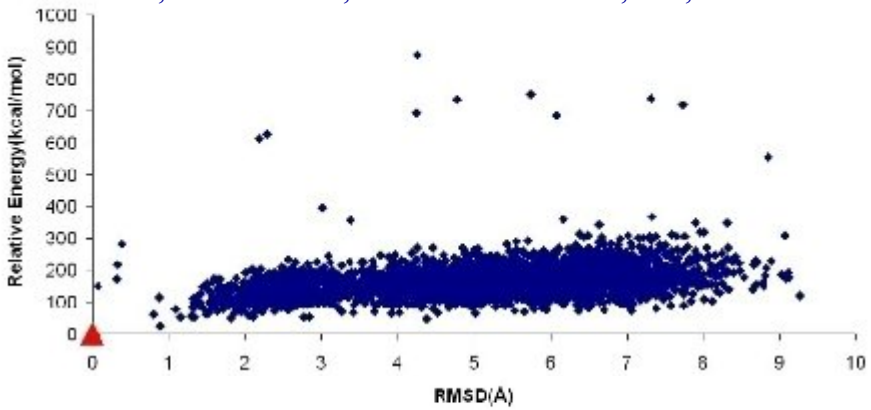
Removal of Steric Clashes in Selected Structures (Distance Based Monte Carlo)



Validation of Empirical Energy Based Scoring Function

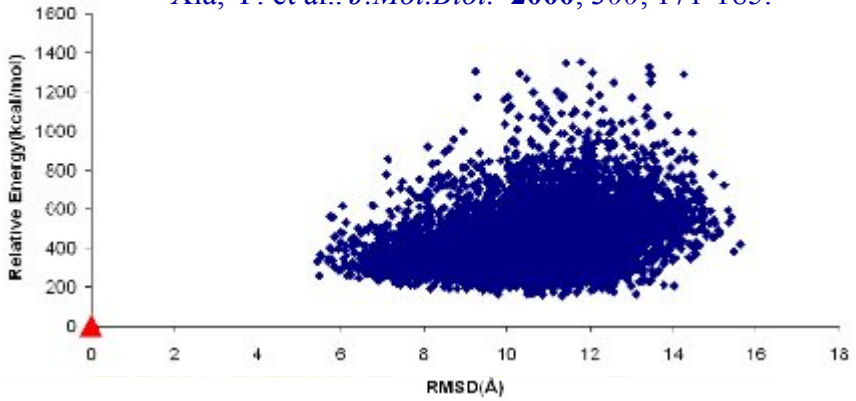
Four-state reduced decoy set

Park, B. and Levitt, M. *J.Mol.Biol.* **1996**, *258*, 367-392.



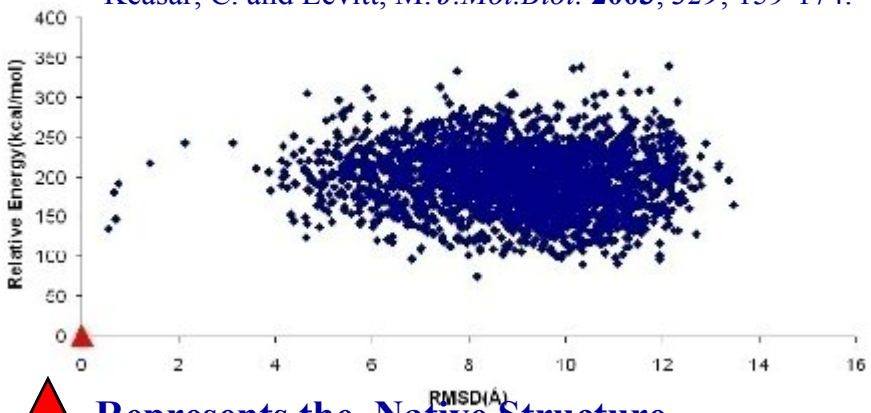
Lattice_ssfit decoy set

Xia, Y. et al.. *J.Mol.Biol.* **2000**, *300*, 171-185.



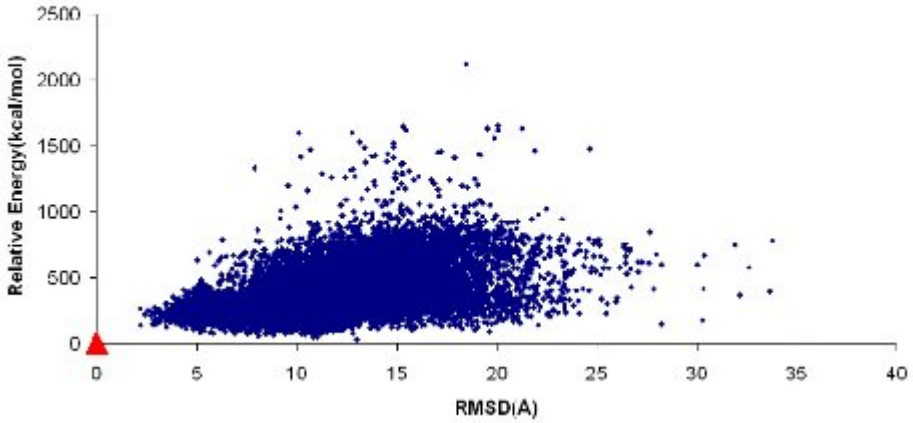
Lmds decoy set

Keasar, C. and Levitt, M. *J.Mol.Biol.* **2003**, *329*, 159-174.



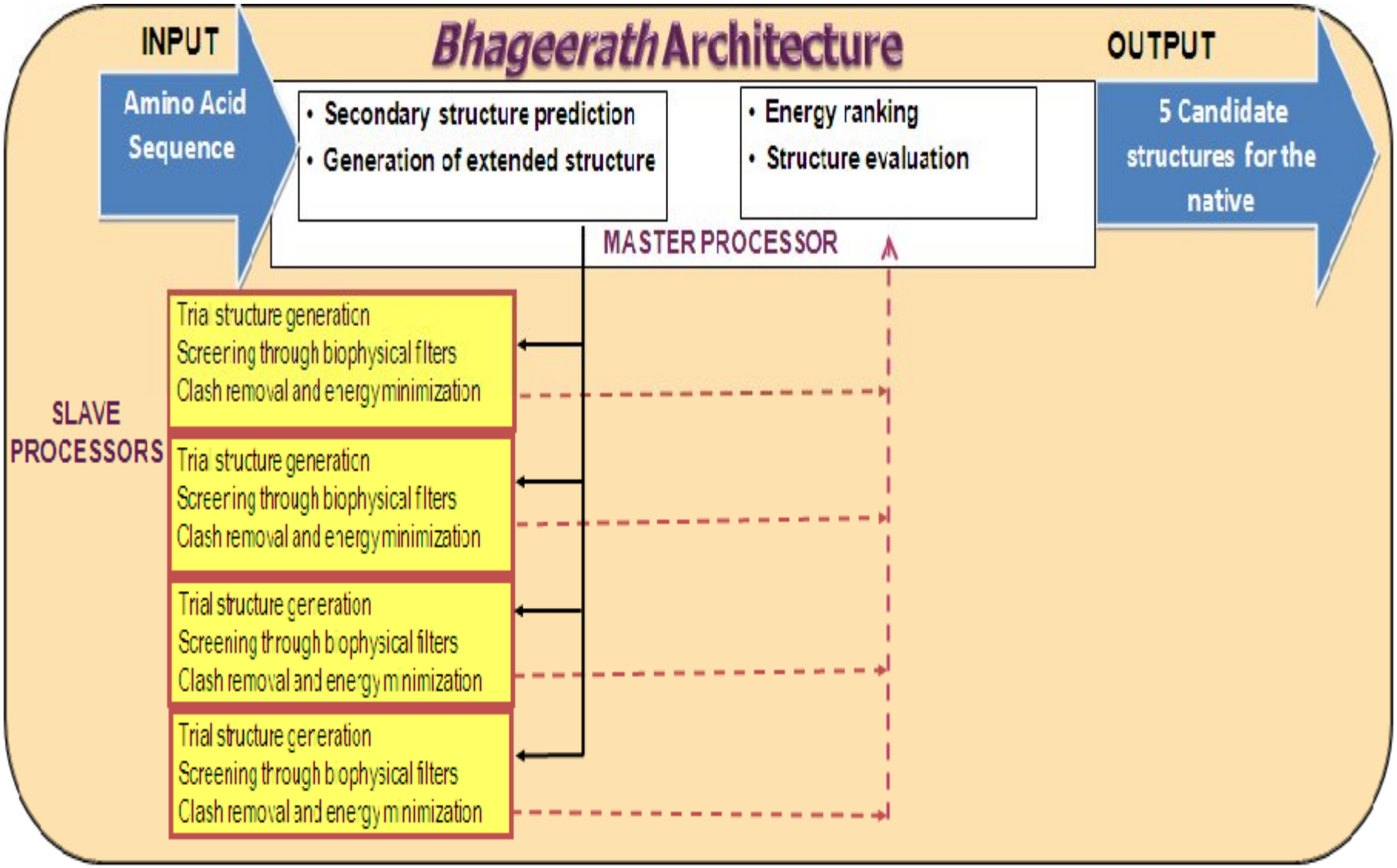
Rosetta decoy set

Simons, K.T. et al.. *Proteins* **1999**, *37* S3, 171-176.



Represents the Native Structure

Narang, P., Bhushan, K., Bose, S., and Jayaram, B. *J. Biomol.Str.Dyn,* **2006**,*23*,385-406;
Arora N.; Jayaram B.; *J. Phys. Chem. B.* **1998**, *102*, 6139-6144;
Arora N, Jayaram B, *J. Comput. Chem*., .**1997**, *18*, 1245-1252.



***Bhageerath* is currently implemented on a 280 processor (~3 teraflops) cluster**

Jayaram et al., *Bhageerath*, 2006, *Nucleic Acid Res.*, 34, 6195-6204

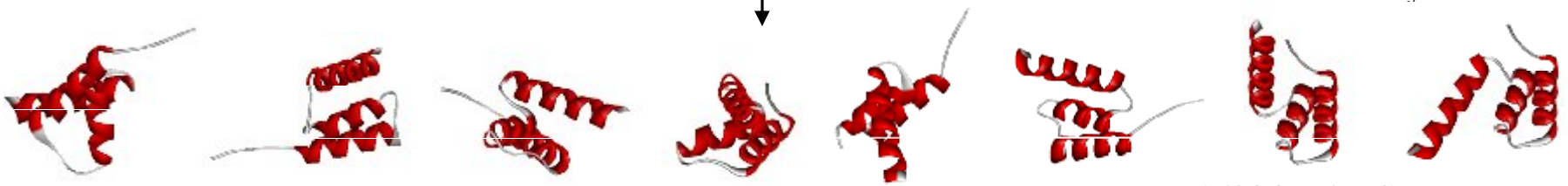
A Case Study of Mouse C-Myb

DNA Binding (52 AA)

LIKGPWTKEEDQRVIELVQKYGPKRWSVIAKHLKGRIGKQCRERWHNHLNPE

Sequence

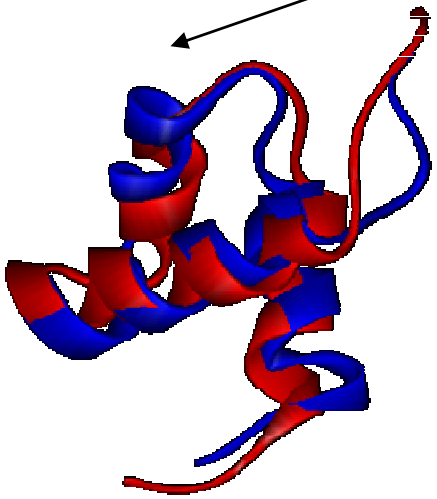
Preformed Secondary Structure



16384 Trial Structures

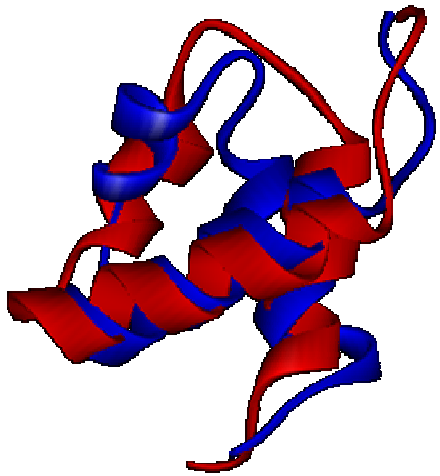
Biophysical Filters & Clash Removal

10632 Structures



RMSD=2.87, Energy Rank=1774

Energy Scans



RMSD=4.0, Energy Rank=4

Blue: Native & Red: Predicted

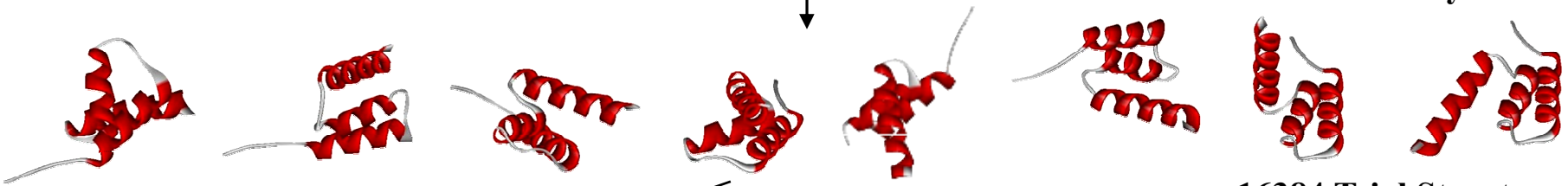
A Case Study of *S.aureus* Protein A

Immunoglobulin Binding (60 AA)

RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKS

Sequence

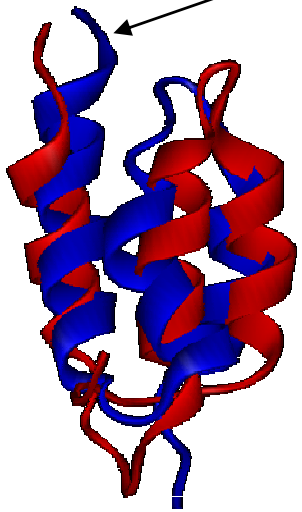
Preformed Secondary Structure



16384 Trial Structures

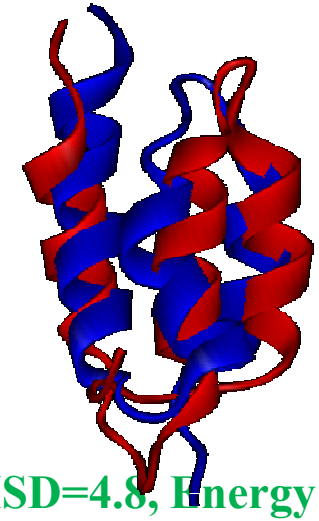
Biophysical Filters & Clash Removal

11255 Structures



RMSD=4.2, Energy Rank=44

Energy Scans



RMSD=4.8, Energy Rank=5

Blue: Native & Red: Predicted

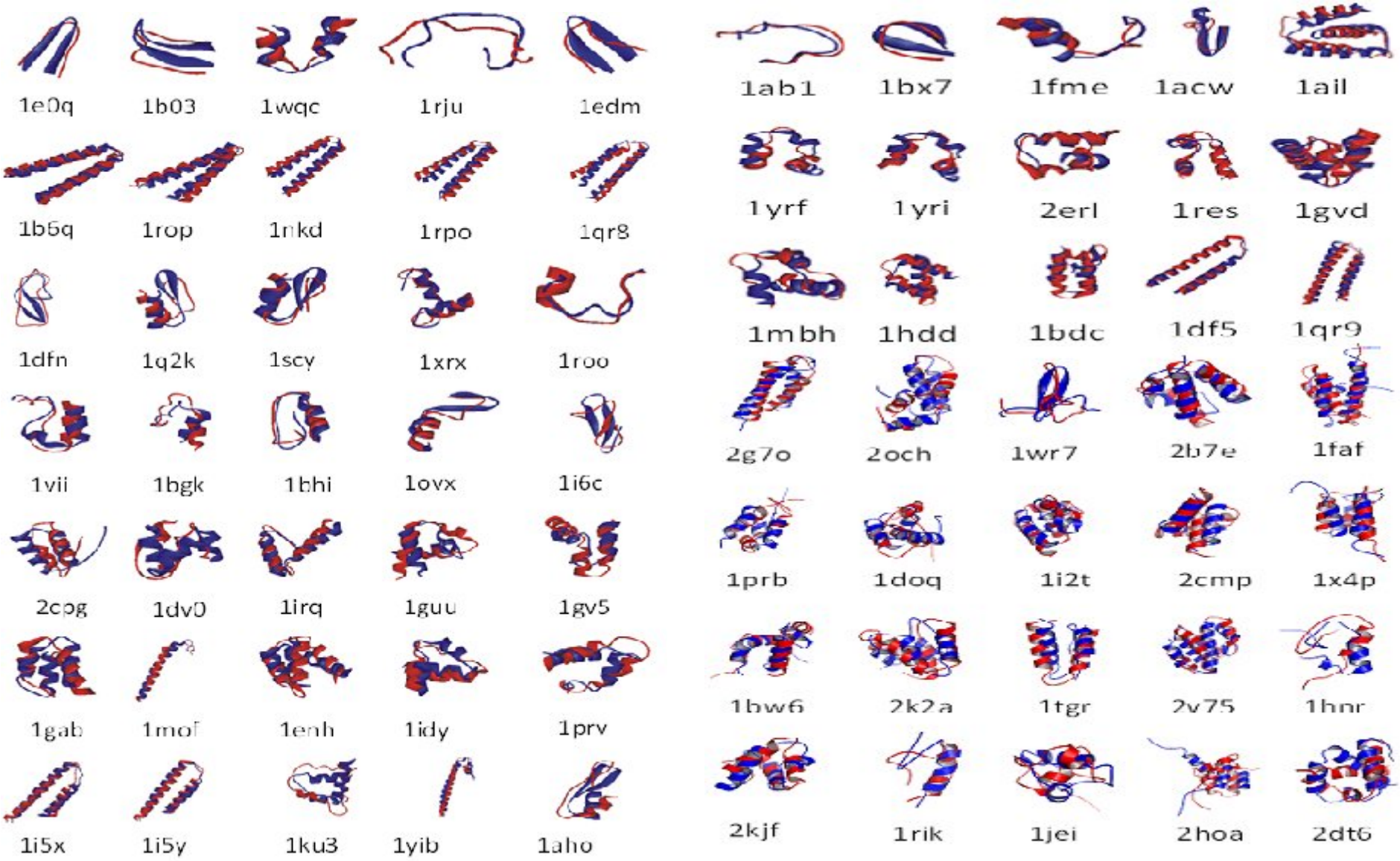
Performance of *Bhageerath* on 70 Small Globular Proteins

S.No.	PDBID	No of Amino Acids	No. of Secondary Structure elements	Lowest RMSD Å (from native)	Energy rank of lowest RMSD structure in top 5 structures
1	1E0Q	17	2E	**2.5**	**2**
2	1B03	18	2E	**4.4**	**2**
3	1WQC	26	2H	**2.5**	**3**
4	1RJU	36	2H	**5.9**	**4**
5	1EDM	39	2E	**3.5**	**2**
6	1AB1	46	2H	**4.2**	**5**
7	1BX7	51	2E	**3.2**	**4**
8	1B6Q	56	2H	**3.8**	**5**
9	1ROP	56	2H	**4.3**	**2**
10	1NKD	59	2H	**3.9**	**1**
11	1RPO	61	2H	**3.8**	**2**
12	1QR8	68	2H	**3.9**	**4**
13	1FME	28	1H,2E	**3.7**	**5**
14	1ACW	29	1H,2E	**5.3**	**3**
15	1DFN	30	3E	**5**	**1**
16	1Q2K	31	1H,2E	**4.8**	**4**
17	1SCY	31	1H,2E	**3.1**	**5**
18	1XRX	34	1E,2H	**5.6**	**1**
19	1ROO	35	3H	**2.8**	**5**
20	1YRF	35	3H	**4.8**	**4**
21	1YRI	35	3H	**4.6**	**3**
22	1VII	36	3H	**3.7**	**2**
23	1BGK	37	3H	**4.1**	**3**
24	1BHI	38	1H,2E	**5.3**	**2**

S.No.	PDBID	No of Amino Acids	No. of Secondary Structure elements	Lowest RMSD Å	Energy rank of lowest RMSD structure in top 5 structures
25	1OVX	38	1H,2E	4	1
26	1I6C	39	3E	5.1	2
27	2ERL	40	3H	4	3
28	1RES	43	3H	4.2	2
29	2CPG	43	1E,2H	5.3	2
30	1DV0	45	3H	5.1	4
31	1IRQ	48	1E,2H	5.5	3
32	1GUU	50	3H	4.6	4
33	1GV5	52	3H	4.1	2
34	1GVD	52	3H	5.1	4
35	1MBH	52	3H	4	4
36	1GAB	53	3H	4.9	1
37	1MOF	53	3H	2.9	5
38	1ENH	54	3H	4.6	3
39	1IDY	54	3H	3.6	5
40	1PRV	56	3H	5	5
41	1HDD	57	3H	5.5	4
42	1BDC	60	3H	4.8	5
43	1I5X	61	3H	3.6	3
44	1I5Y	61	3H	3.4	5
45	1KU3	61	3H	5.5	4
46	1YIB	61	3H	3.5	5
47	1AHO	64	1H,2E	4.5	4
48	1DF5	68	3H	3.4	1
49	1QR9	68	3H	3.8	2
50	1AIL	70	3H	4.4	3

S.No.	PDBID	No of Amino Acids	No. of Secondary Structure elements	Lowest RMSD Å	Energy rank of lowest RMSD structure in top 5 structures
51	2G7O	68	4H	**5.8**	2
52	2OCH	66	4H	**6.6**	3
53	1WR7	41	3E,1H	**5.2**	2
54	2B7E	59	4H	**6.8**	4
55	1FAF	79	4H	**6.4**	4
56	1PRB	53	4H	**6.9**	4
57	1DOQ	69	5H	**6.8**	3
58	1I2T	61	4H	**5.4**	4
59	2CMP	56	4H	**5.6**	1
60	1BW6	56	4H	**4.2**	1
61	1X4P	66	4H	**5.2**	3
62	2K2A	70	4H	**6.1**	1
63	1TGR	52	4H	**6.8**	2
64	2V75	90	5H	**7.0**	3
65	1HNR	47	2E,2H	**5.2**	2
66	2KJF	60	4H	**5.0**	4
67	1RIK	29	2E,2H	**4.4**	4
68	1JEI	53	4H	**5.8**	5
69	2HOA	68	4H	**6.3**	4
70	2DT6	62	4H	**5.9**	3

Predicted Structures with *Bhageerath*
for 70 Globular Proteins superposed on their corresponding experimental structures



Native structure

Predicted structure

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 5 lowest energy structures.

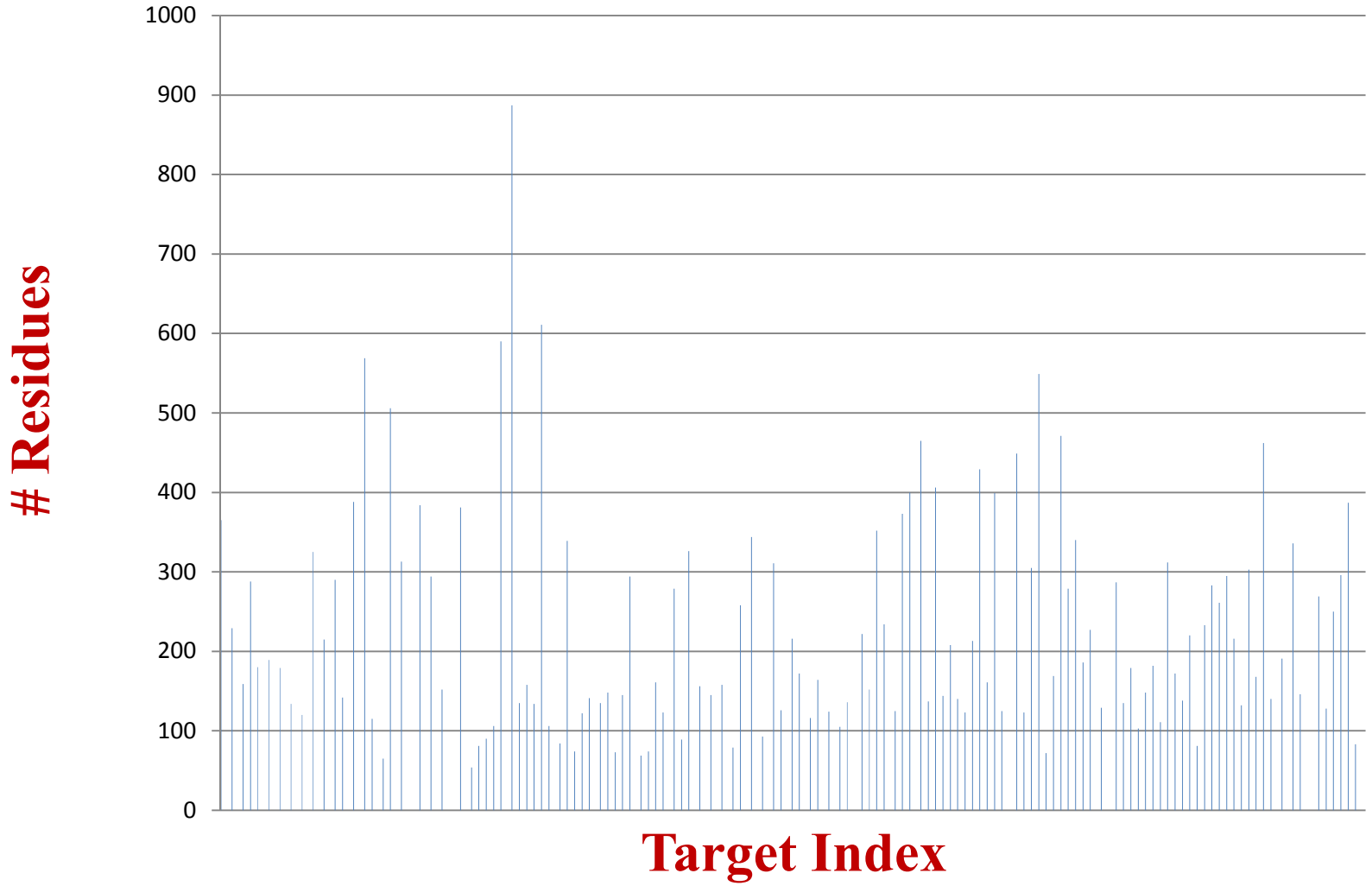
The above two proteins have maximum sequence similarity of 38% and 48% respectively.

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

Homology models are simply superb where the similarities between query sequence and template in the protein data bank are high. Where there is no match/similarity ab initio methods such as Bhageerath are the only option.

The Protein Structure Prediction Olympics

CASP9 (May 3rd to July 17th, 2010: 129 Targets)



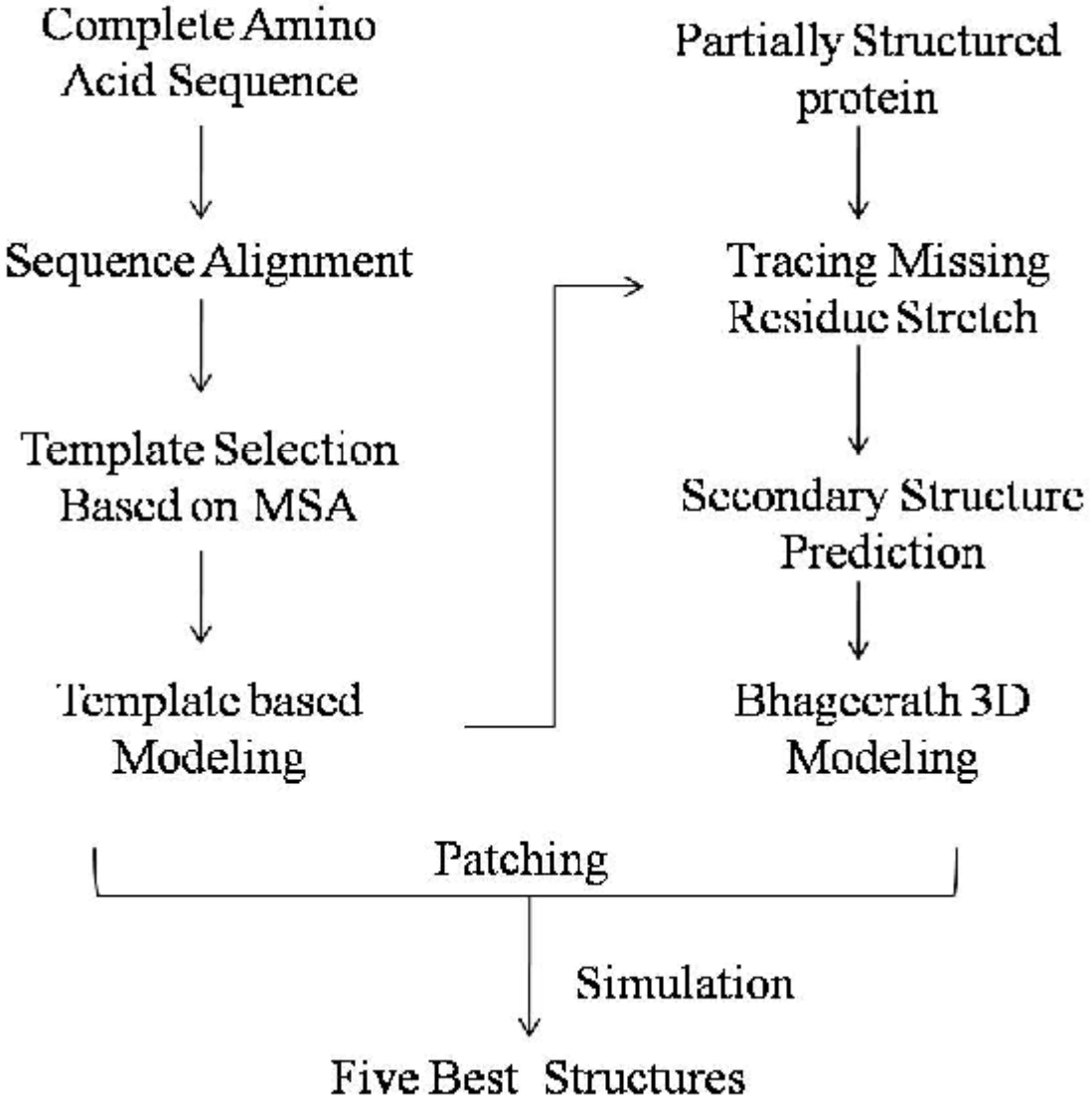
Bhageerath vs other servers for Template free prediction in CASP9

Target No.	No.of residues	PDBID	*Bhageerath* RMSD Å	TASSER RMSD Å	ROBETTA RMSD Å	SAM-T08 RMSD Å
T0531	65	2KJX	7.1	11.0	11.9	12.6
T0553	141	2KY4	9.6	6.0	11.5	8.6
T0581	136	3NPD	15.8	11.6	5.3	15.1
T0578	164	3NAT	19.2	11.6	15.5	19.1

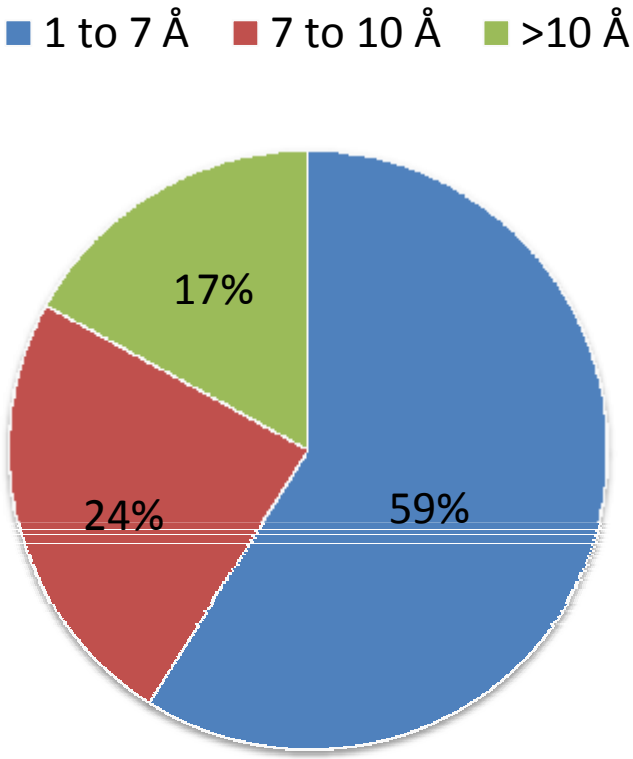
While *Bhageerath* – an *ab initio* method - works well for small proteins (<100 residues), improvements are necessary to tackle larger proteins

Development of a homology - ab initio hybrid server

Bhageerath-H Protocol

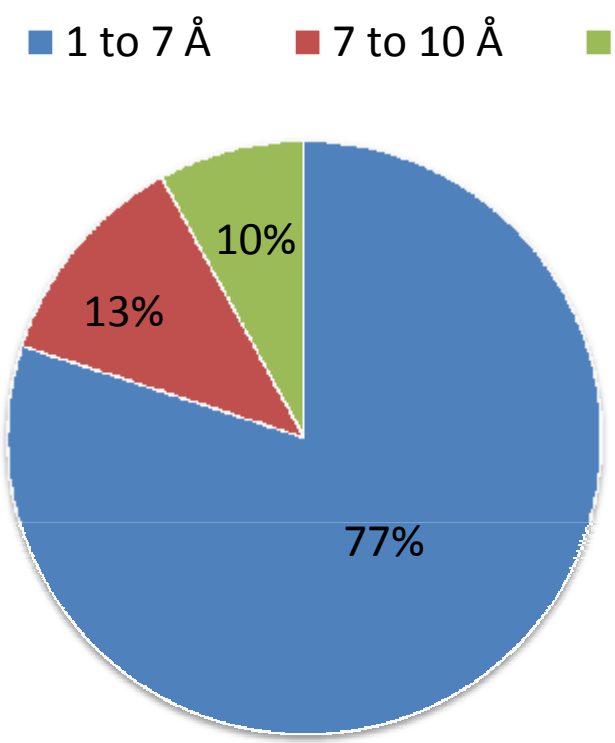


Bhageerath-H Results in CASP9



Total Number of Targets: 50
Average RMSD: 8.7 Å

Bhageerath-H Results Post CASP9



Total Number of Targets: 105
Average RMSD: 6 Å

Homology *ab initio* hybrid methods are getting better in tertiary structure prediction

BHAGEERATH : An Energy Based Protein Structure Prediction Server

The present version of"Bhageerath" accepts amino acid sequence and secondary structure information to predict 10 candidate structures for the native. It is anticipated that at least one native like structure (RMSD < 6Å without end loops) is present in the final structures. The server has been validated on 50 small globular proteins. Know about Protein Folding

Download BHAGEERATH 1.0 for Solaris 10.0 environment from here.

[Repository] [General Info] [Links] [Help] [Home]

Process ID 56703599

E-mail Address: (Optional)

Input Amino acid sequence in FASTA format **OR** Click on the Amino acid to add to the sequence

ALA	VAL	LEU	ILE	PRO
MET	PHE	TRP	GLY	SER
THR	CYS	ASN	GLN	TYR
ASP	GLU	LYS	ARG	HIS

Secondary Structure Information

Auto Secondary Structure Prediction **Enter Secondary Structure Information**

Helix Residue Range - Add Clear

SUBMIT SUBMIT RESET

Retrieve previous results

Job ID: Get Status

In case of any Suggestions/Exceptions, Please contact us at scfbio@scfbio-iitd.res.in

Bhageerath-H WebServer

http://www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp

BHAGEERATH-H: A Homology ab-intio Hybrid Web server for Protein Tertiary Structure Prediction

"Bhageerath-H" accepts amino acid sequence to predict 5 candidate structures for the native. Here user has the flexibility to mention reference PDB(s) for modeling. Method has been fielded in CASP9 Experiment and has been improved since.

[Repository] [Tutorial] [Sample File] [Links] [Help] [Home]

Process ID 1764624

E-mail Address:

Upload sequence in FASTA format Choose File No file chosen

OR Input Amino acid sequence in FASTA format

ALA	VAL	LEU	ILE	PRO
MET	PHE	TRP	GLY	SER
THR	CYS	ASN	GLN	TYR
ASP	GLU	LYS	ARG	HIS

Template Information

Auto Template Searching User Defined Template

PDB ID - Chain ID Add Clear

SUBMIT RESET

In search of rules of protein folding:

Margin of Life: Amino acid compositions in proteins have a tight distribution

The average percentage occurrence of each amino-acid for folded proteins gives the "Chargaff's rules" for protein folding and the standard deviations give the "margin of life".

Amino Acid	Folded Proteins – Margin of Life (mean ± std, n = 3718)
A	7.8 ± 3.4
V	7.1 ± 2.4
I	5.8 ± 2.4
L	9.0 ± 2.9
Y	3.4 ± 1.7
F	3.9 ± 1.8
W	1.3 ± 1.0
P	4.4 ± 2.0
M	2.2 ± 1.3
C	1.8 ± 1.5
T	5.5 ± 2.4
S	6.0 ± 2.5
Q	3.8 ± 2.0
N	4.3 ± 2.2
D	5.8 ± 2.0
E	7.0 ± 2.7
H	2.3 ± 1.4
R	5.0 ± 2.3
K	6.3 ± 2.8
G	7.2 ± 2.8

The average percentage occurrence of each amino-acid from the ExPASy Server.

Amino Acid	Protein sequences confirmed by annotation and experiments (mean ± std, n = 131855)
A	7.2 ± 3.0
V	6.3 ± 2.1
I	5.1 ± 2.2
L	9.6 ± 2.9
Y	3.0 ± 1.5
F	3.9 ± 1.8
W	1.2 ± 0.9
P	5.4 ± 2.6
M	2.2 ± 1.3
C	1.9 ± 2.3
T	5.5 ± 1.8
S	7.9 ± 2.8
Q	4.3 ± 2.0
N	4.2 ± 1.9
D	5.2 ± 1.9
E	6.8 ± 2.8
H	2.4 ± 1.3
R	5.3 ± 2.9
K	6.0 ± 2.9
G	6.6 ± 2.8

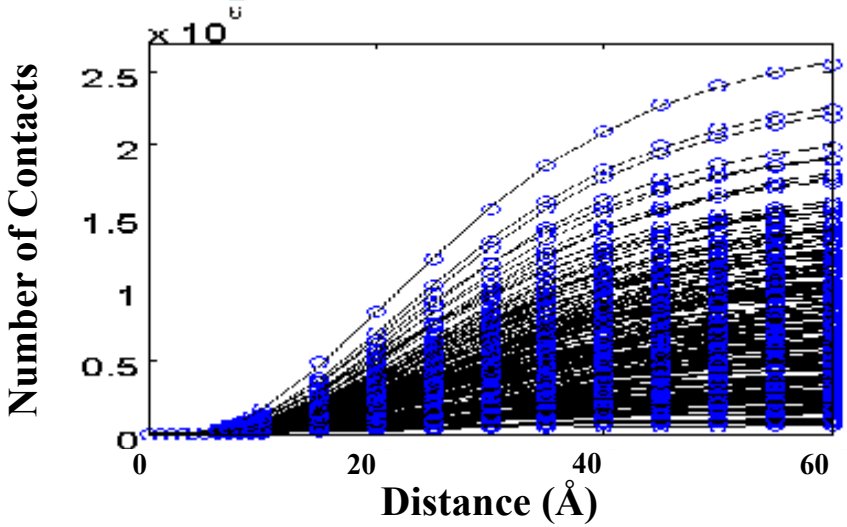
The average percentage occurrence of each amino acid, their STD as observed and as calculated from the binomial distribution.

	P (%)	STD (observed)	STD (random)
A	7.8	3.4	7.2
V	7.1	2.4	6.6
I	5.8	2.4	5.5
L	9.0	2.9	8.2
Y	3.4	1.7	3.3
F	3.9	1.8	3.7
W	1.3	1.0	1.3
P	4.4	2.0	4.2
M	2.2	1.3	2.2
C	1.8	1.5	1.8
T	5.5	2.4	5.2
S	6.0	2.5	5.6
Q	3.8	2.0	3.7
N	4.3	2.2	4.1
D	5.8	2.0	5.5
E	7.0	2.7	6.5
H	2.3	1.4	2.2
R	5.0	2.3	4.8
K	6.3	2.8	5.9
G	7.2	2.8	6.7

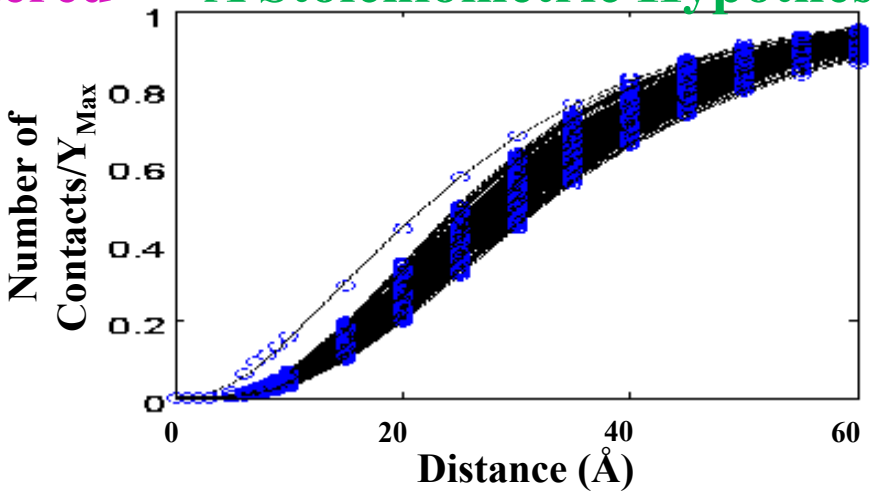
Mezei (2011), JBSD

In search of rules of protein folding:

Cα atoms of proteins of varying sequences and sizes follow a single (universal) spatial distribution



All 400 Cα spatial distributions (above) collapse into one narrow band (below) irrespective of the chemical nature of the amino acids when their percentage occurrences are considered => A Stoichiometric Hypothesis for Protein Folding.



$$Y = Y_{Max}(1-e^{-kX})^n$$

Mittal & Jayaram et al., (2010) *JBSD*, 28, 133-142; (2011), *JBSD*, 28, 443-454; (2011), *JBSD*, 28, 669-674.

While structure prediction attempts are progressing well, rules of folding are still elusive.

www.scfbio-iitd.res.in

•Genome Analysis - *ChemGenome*

A novel *ab initio* Physico-chemical model for whole genome analysis

•Protein Structure Prediction – *Bhageerath*

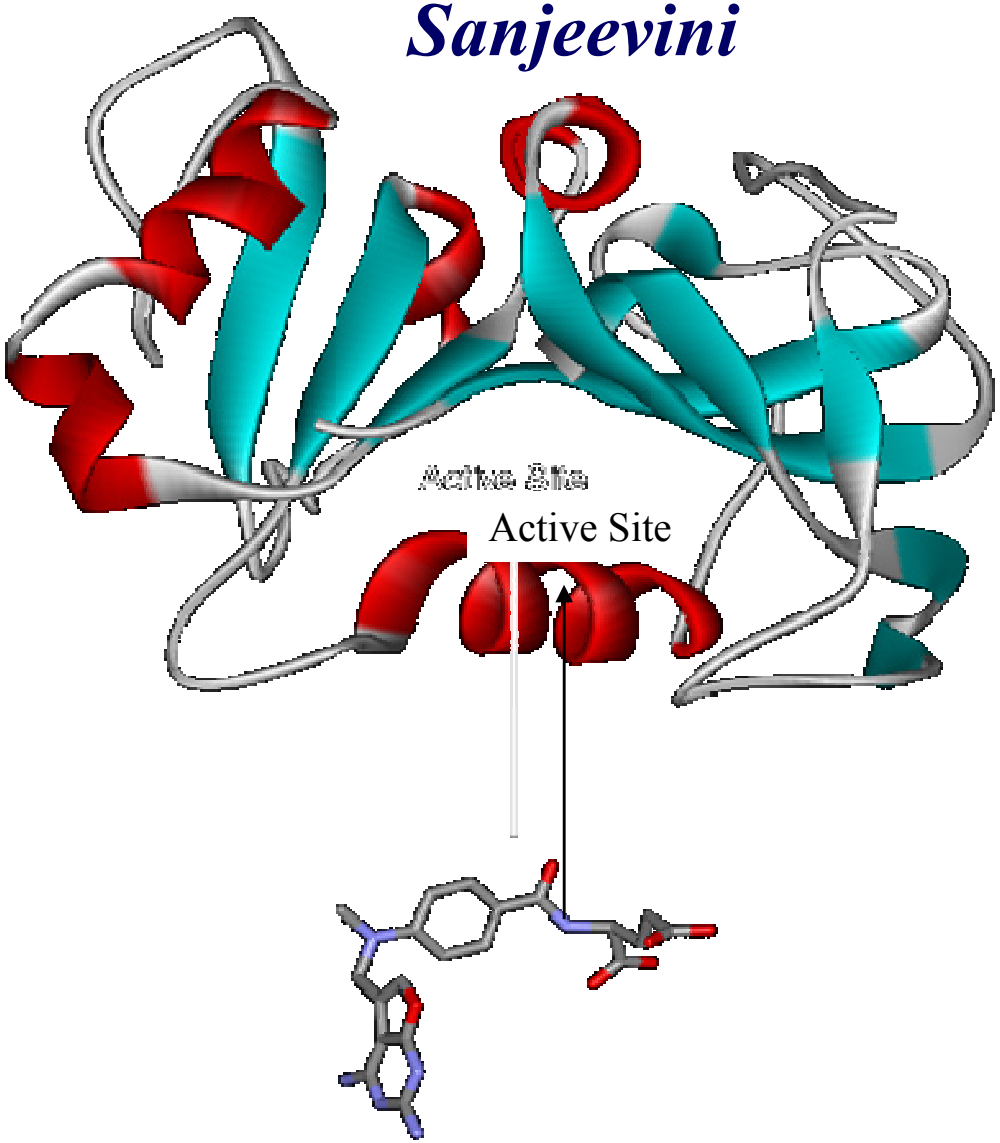
A *de novo* energy based protein structure prediction software

•Drug Design – *Sanjeevini*

A comprehensive indigenous target directed lead molecule design protocol

Target Directed Lead Design

Sanjeevini



Given the structure of the drug target, design a molecule that will bind to the target with high affinity and specificity

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

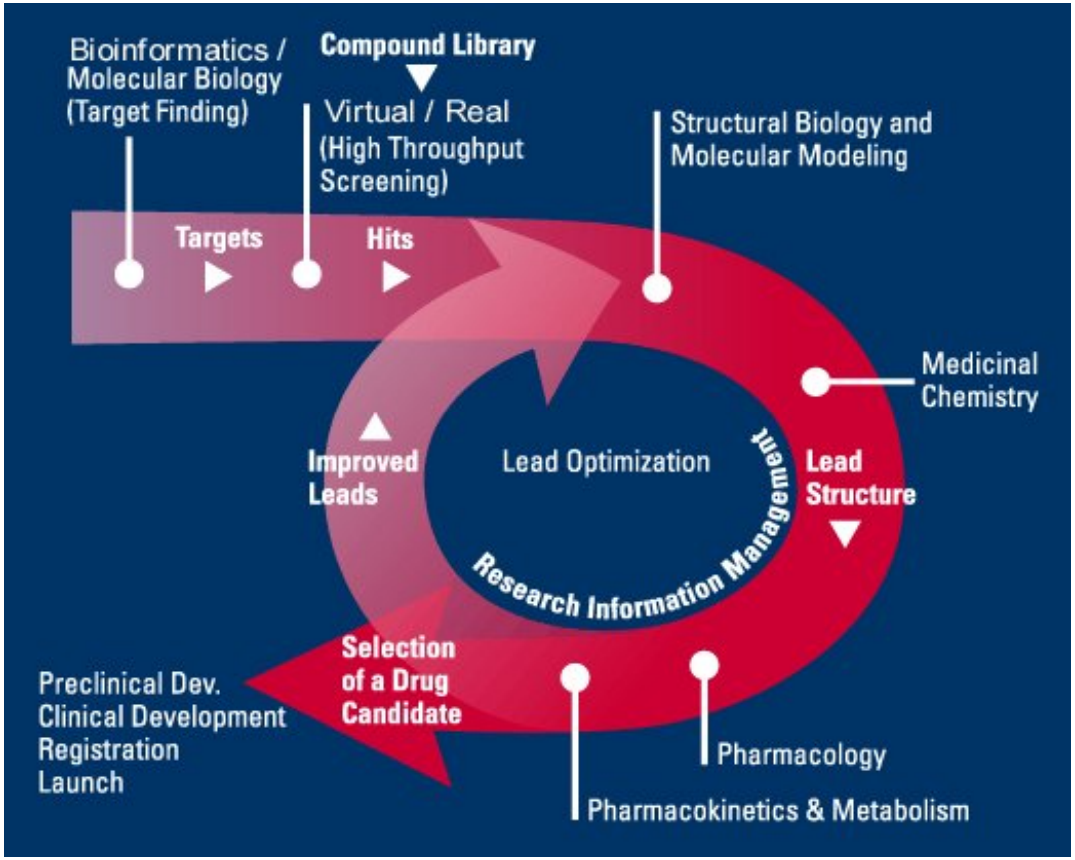
6.0yrs ↓ *68%*

FDA Review & Approval

1.5yrs ↓ *3%*

Drug to the Market

14 yrs **$1.4 billion**



Source: PAREXEL's Pharmaceutical R&D Statistical Sourcebook, 2001, p96.; Hileman, Chemical Engg. News, 2006, 84, 50-1.

Pharmaceutical R&D is Expensive

New Chemical Entities (NCEs) need to be continuously developed since income from older drugs gets gradually reduced on account of increasing competition from other products, generics as well *drug resistance*.

Drug Development is an Uphill Task

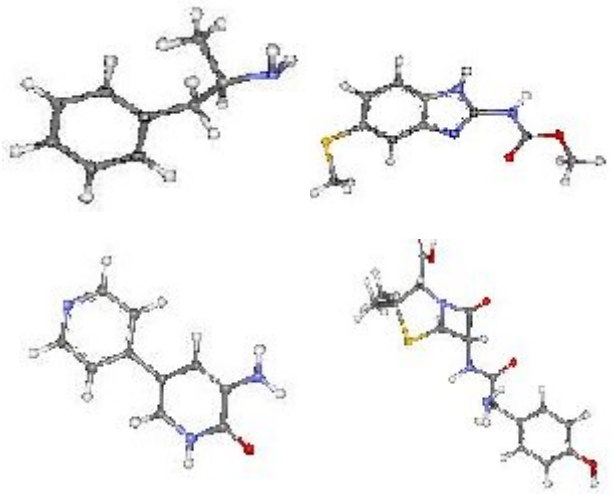
1035 new drugs approved by FDA between 1989 to 2000

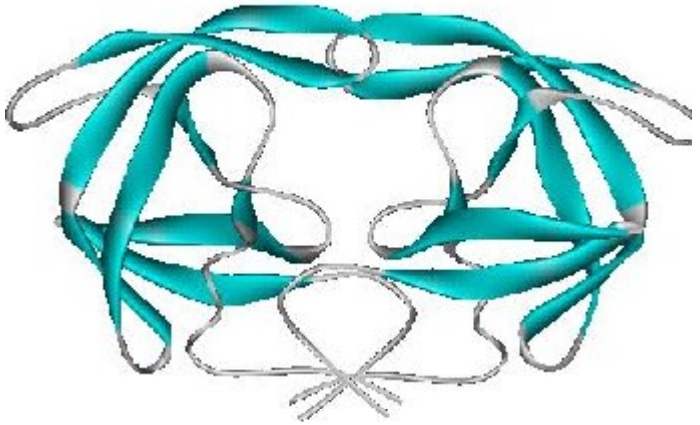
361 (35%) were New Molecular Entities (NME).

Only 15% were deemed to provide significant improvement over existing medicines.

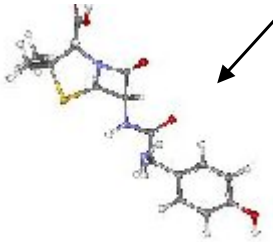
http://www.seniors.gov/articles/0502/medicine-study.htm

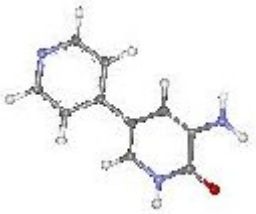
Structure Based Lead Molecule Design





Computer Aided Drug Design

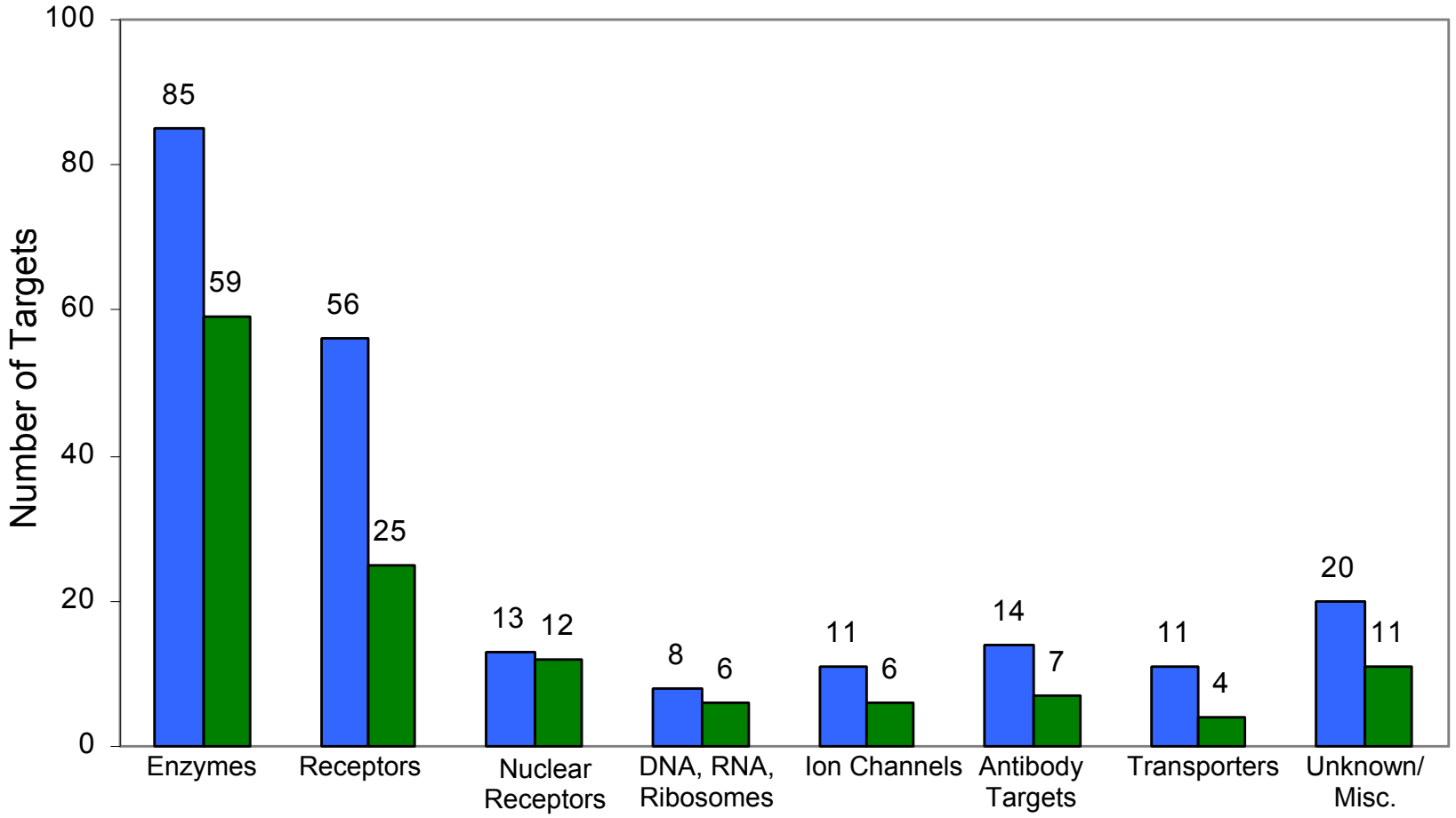




DRUG

NON-DRUG

Present Scenario of Drug Targets



BLUE: Number of targets in each class. (Imming P, Sinning C, Meyer A. *Nature Rev Drug Discov* **2006**;5: 821)
(Total 218 targets & 8 classes)
GREEN: Number of 3D structures available in each class (Total: 130) (Protein Data Bank)

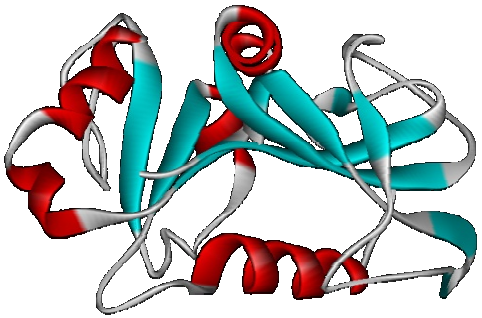
Shaikh SA, Jain T, Sandhu G, Latha N, Jayaram, B. *Current Pharmaceutical Design,* **2007**

Some Concerns in Lead Design *In Silico*

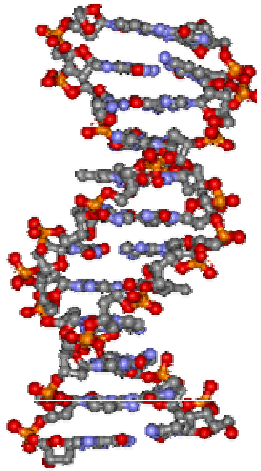
- Novelty and Geometry of the Ligands
- Accurate charges and other Force field parameters
- Ligand Binding Sites
- Flexibility of the Ligand and the Target
- Solvent and salt effects in Binding
- Internal energy versus Free energy of Binding
- Druggability
- Computational Tractability

De novo LEAD-LIKE MOLECULE DESIGN: THE *SANJEEVINI* PATHWAY

Drug Target Identification



3-Dimensional Structure of Target



Database → **Candidate molecules** ← **User**

↓

Drug-like filters ← **Mutate / Optimize**

↓

Geometry Optimization
Quantum Mechanical Derivation of Charges
Assignment of Force Field Parameters

↓

Active Site Identification on Target &
Ligand Docking

↓

Energy Minimization of Complex
Binding free energy estimates - Scoring

↓

Molecular dynamics &
***post-facto* free energy component analysis (Optional)**

↓

Lead-like compound

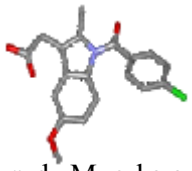
Jayaram, B., Latha, N.,Jain, T., Sharma, P., Gandhimathi, A., Pandey, V.S., Indian Journal of Chemistry-A. 2006, 45A, 1834-1837.
Tanya Singh, Goutam Mukherjee, Abhinav Mathur, B. Jayaram, *Sanjeevini* – A User Manual, 2011, manuscript in preparation

***Sanjeevini* Pathway**

NRDBM/Million Molecule Library/Natural Products and Their Derivatives

Molecular Database

or,



Ligand Molecule

or

Target Protein/DNA

Upload

Bioavalibality Check (Lipinski Compliance)

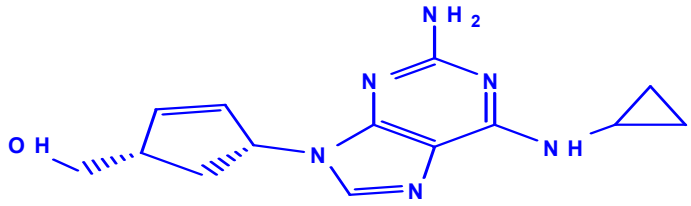
Binding EnergyEstimation by RASPD protocol

Geometry Optimization TPACM4/Quantum Mechanical Derivation of Charges

Prediction of all possible active sites (for protein only and if binding site is not known).

Assignment of Force Field Parameters

Ligand Molecule ready for Docking



+

Protein/DNA ready for Docking

or

Dock & Score

Molecular dynamics & *post-facto* free energy component analysis (Optional)

Molecular Descriptors / Drug-like Filters

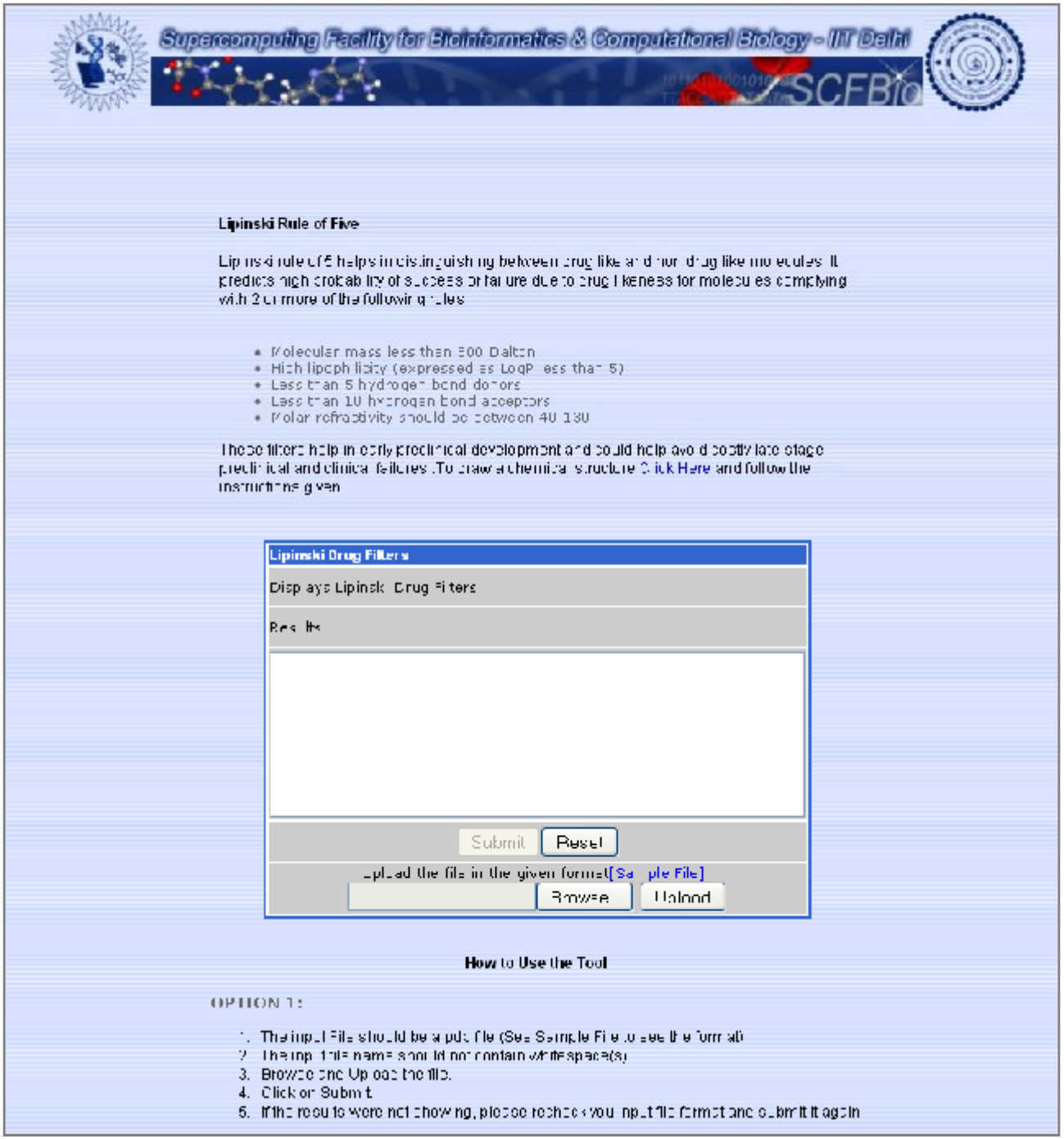
Lipinski's rule of five

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

http://www.scfbio-iitd.res.in/utility/LipinskiFilters.jsp



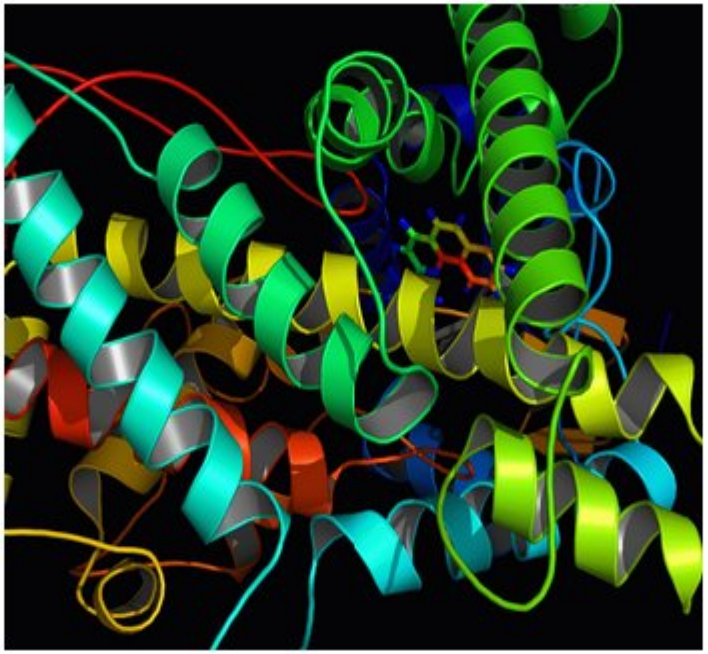
http://www.scfbio-iitd.res.in/dock/ActiveSite_new.jsp

Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi

SCFBio

Home | Group | Publications | Resources | Contact Us

ACTIVE SITE PREDICTION



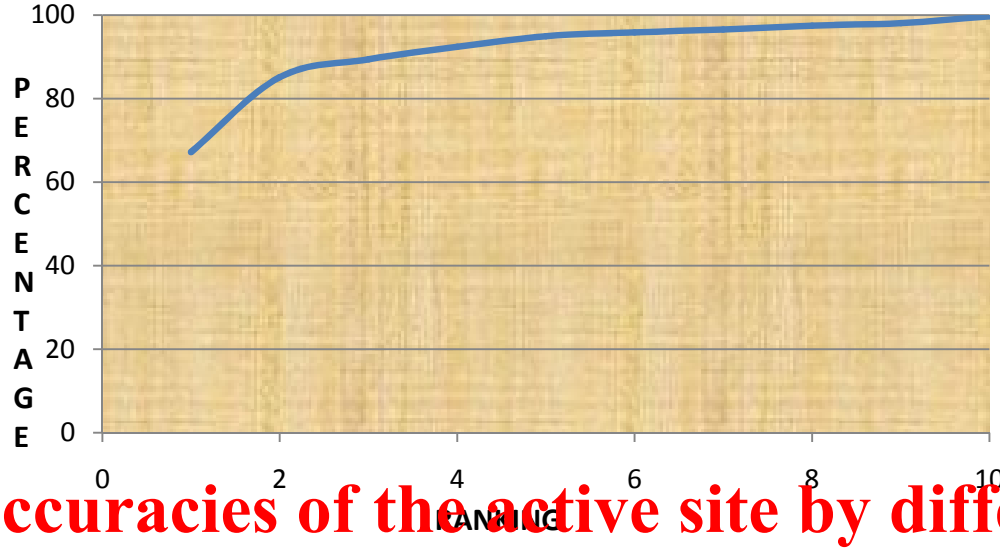
Welcome to the Act ve Site prediction

Active Site Predicticn of Protein server computes the cavities in a given protein.

Click here to see 'How to Use Tool'. [Sample Protein File] [Sample Drug File]

Rank of the cavity points vs. cumulative percentage prediction

Top ten cavity points capture the active site 100 % of time in 640 protein targets



Prediction accuracies of the active site by different softwares

Sl. No	Softwares	Top1	Top3	Top5	Top10
1	SCFBIO(Active Site Finder)	73	92	95	100
2	Fpocket	83	92	-	
3	PocketPicker	72	85	-	
4	LiGSITE^{cs}	69	87	-	
5	LIGSITE	69	87	-	
6	CAST	67	83	-	
7	PASS	63	81	-	
8	SURFNET	54	78	-	
9	LIGSITE^{csc}	79	-	-	

http://www.scfbio-iitd.res.in/software/drugdesign/raspd.jsp

Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi

Home | Drug Design Software

RASPD for Preliminary Screening of Drugs

The challenge for computer aided drug discovery is to achieve this specificity - with small molecule inhibitors - in binding to target proteins, at reduced cost and time while ensuring synthesizability, novelty of the scaffolds and proper ADMET profiles. RASPD is a computationally fast protocol for identifying good candidates for any target protein. The binding pocket of the input target protein is scanned for the number of hydrogen bond donors, acceptors, number of hydrophobic groups and number of rings. A QSAR type equation combines the afore-mentioned properties of the target protein and the candidate molecule and an estimate of the binding free energy is generated if the target protein were to complex with the candidate. The most interesting feature of this methodology is that it takes fraction of a second for calculating the binding affinities of the protein-candidate molecule complexes as opposed to several minutes in known art today for regular docking and scoring method, whereas the accuracy of this method in sorting good candidates is comparable with the conventional techniques. We have also created million molecules database. This database is prepared to include chemical formula, structure, topological index, number of hydrogen bond donors and acceptors, number of hydrophobic groups, number of rings, logP values for each of the million molecules. Scoring of 1 million small molecule database by RASPD method to identify hits for a particular protein target is also web enabled for free access at the same site.

Know more **about *RASPD Screeing*.** Click here **to see 'How to Use Tool'.** Click here **to see 'Computational Flow Chart'.**

Method A: Protein-Ligand Complex **Method B:**Only Protein3D Structure

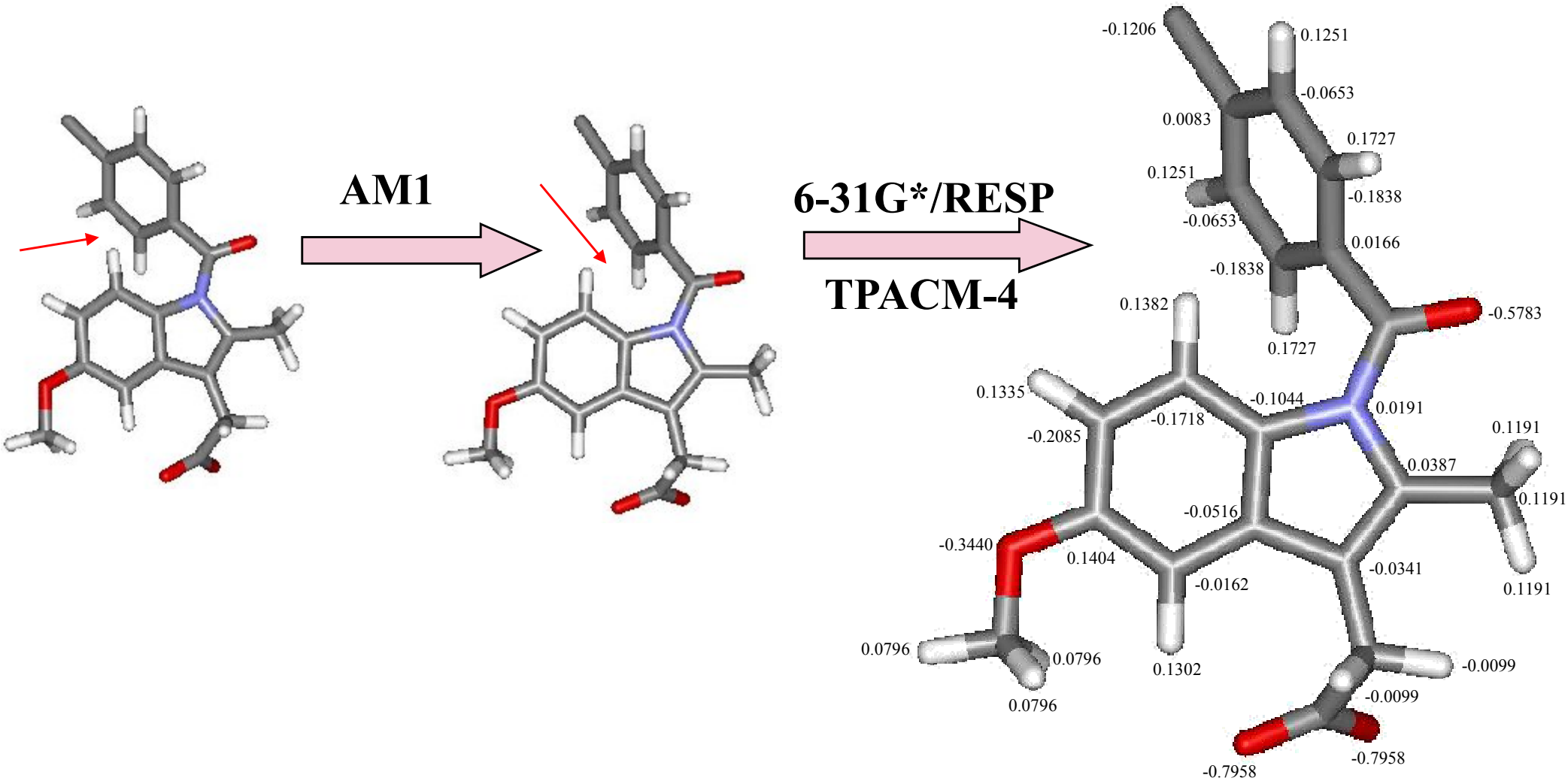
Browse...

Enter Drug Id: DRG

Step 2: Click on 'Submit' to submit your job

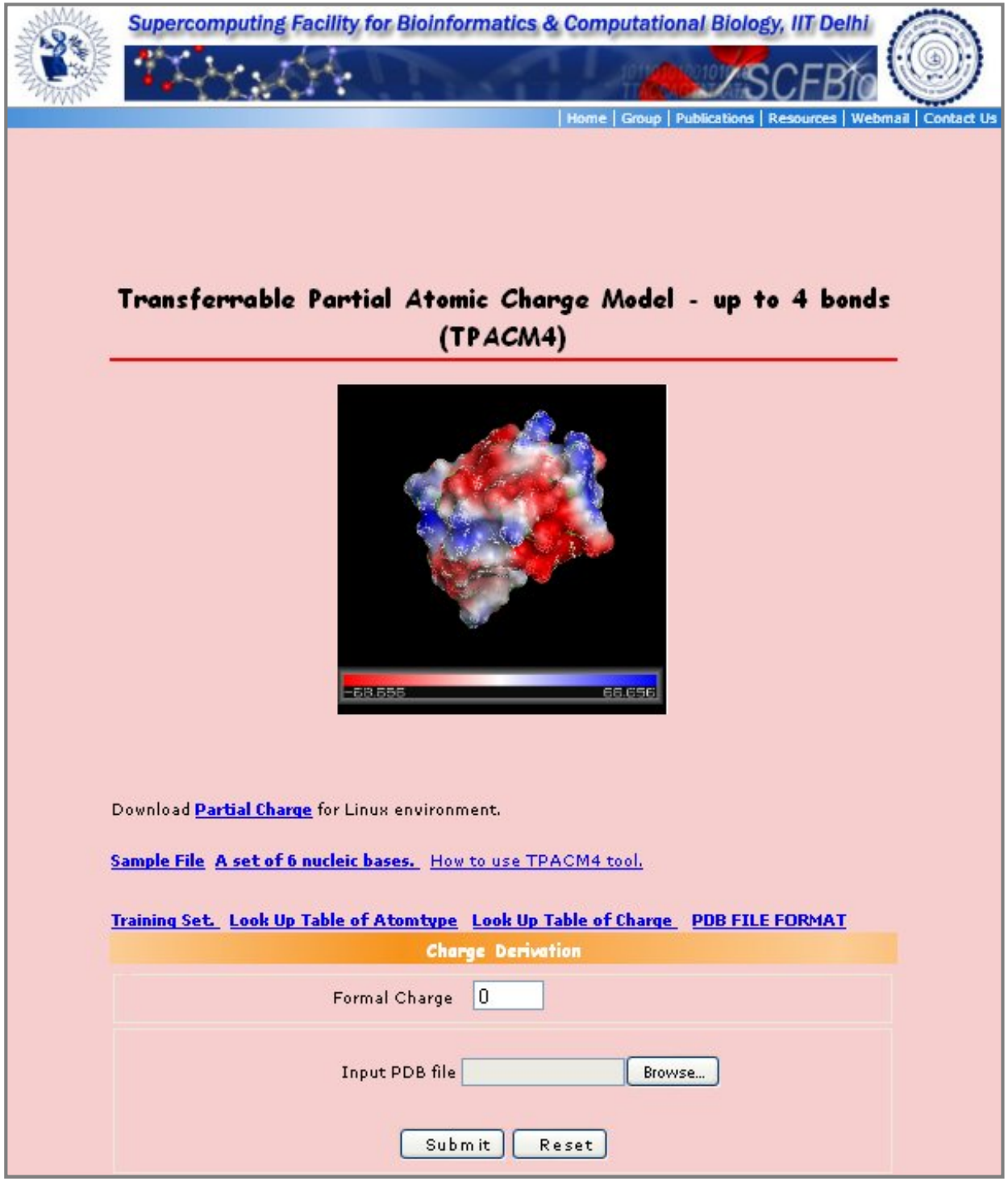
Submit

Quantum Chemistry on Candidate drugs for Assignment of Force Field Parameters



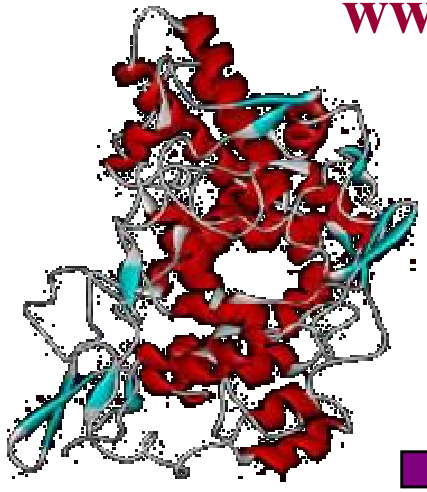
G. Mukherjee, N. Patra, P. Barua and B. Jayaram, (2011), *JCC*, 32,893-907.

http://www.scfbio-iitd.res.in/software/drugdesign/charge.jsp

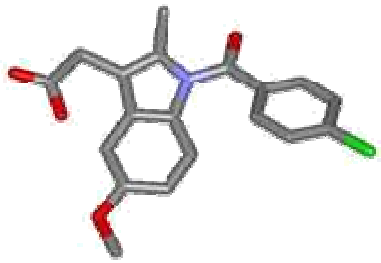


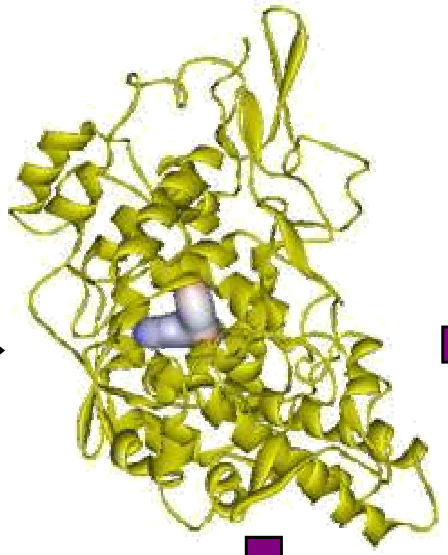
MONTE CARLO DOCKING OF THE CANDIDATE DRUG IN THE ACTIVE - SITE OF THE TARGET

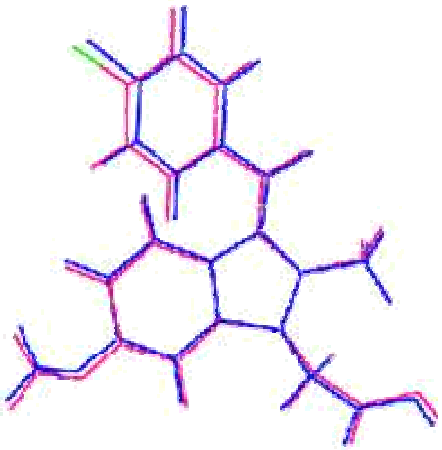
www.scfbio-iitd.res.in/dock/pardock.jsp



+







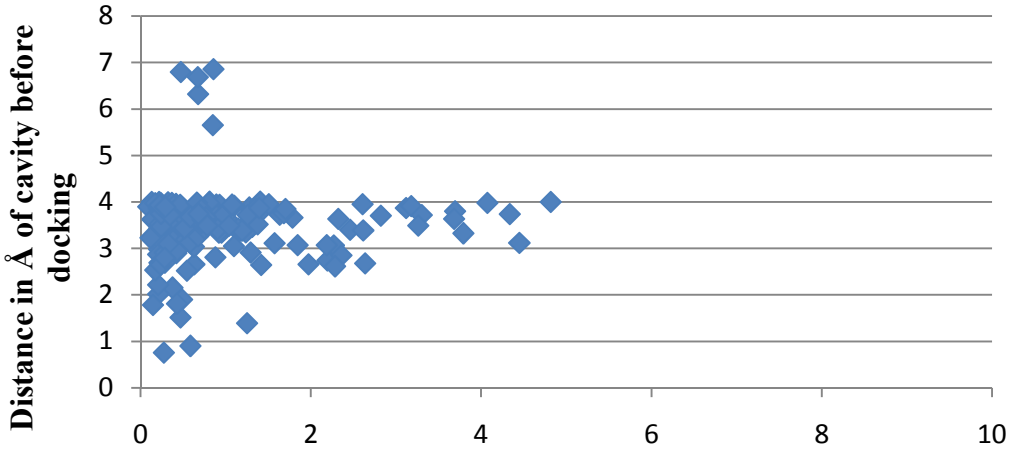
RMSD between the docked & the crystal structure is 0.2Å

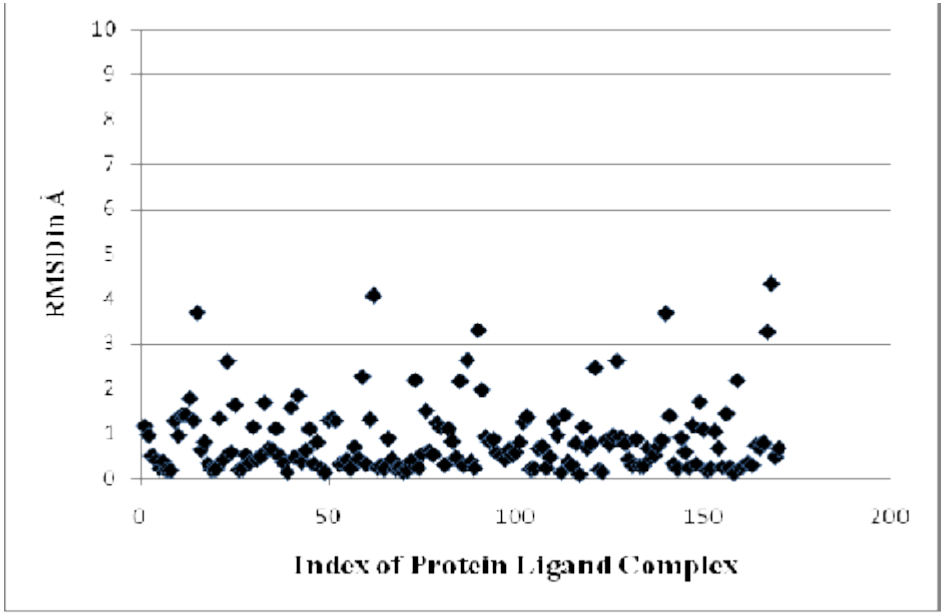
ENERGY MINIMIZATION

5 STRUCTURES WITH LOWEST ENERGY SELECTED

A. Gupta, A. Gandhimathi, P. Sharma, B. Jayaram , (2007), *PPL, 14,* 632-646.

Docking Accuracies



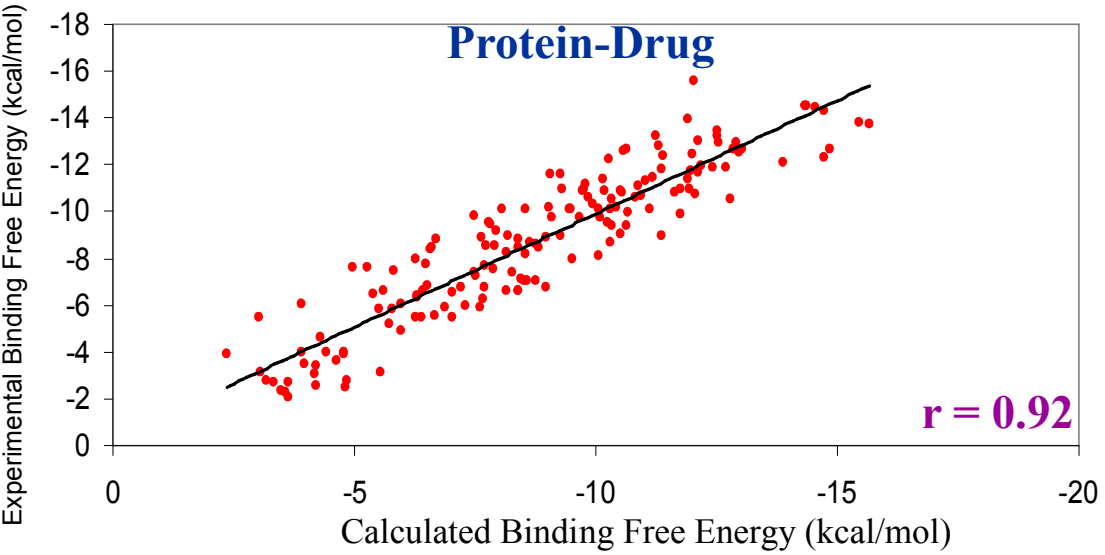


RMSD between the crystal structure and one of the top five docked structures

T. Singh, D. Biswas and B. Jayaram, *AADS - An automated active site identification, docking and scoring protocol for protein targets based on physico-chemical descriptors,* (2011), *JCIM*, 51 (10), 2515-2527

ENERGY BASED SCORING FUNCTION

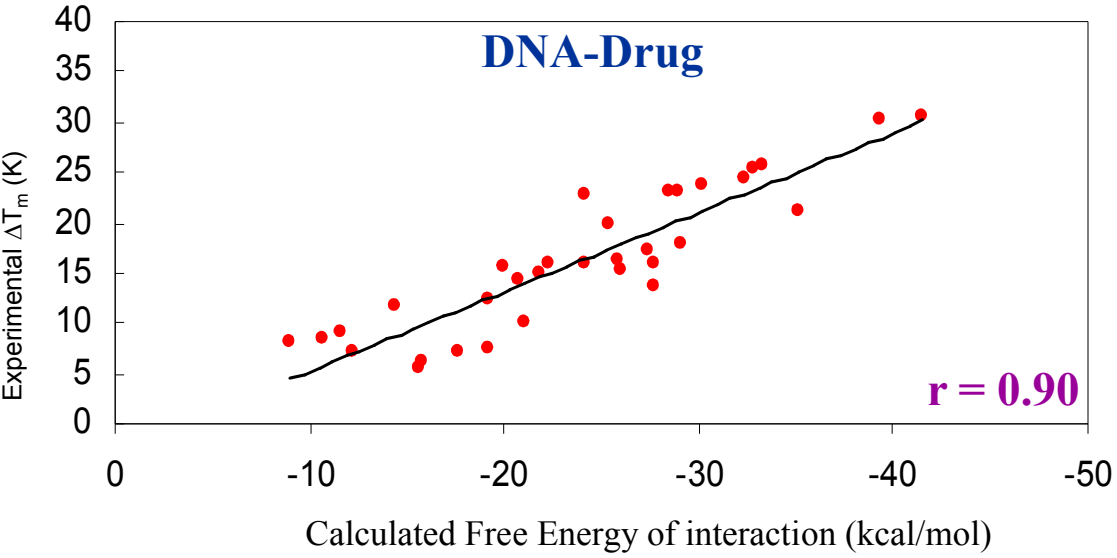
$$\Delta G^\circ_{bind} = \Delta H^\circ_{el} + \Delta H^\circ_{vdw} - T\Delta S^\circ_{rtvc} + \Delta G^\circ_{hpb}$$



Correlation between experimental & calculated binding free energy for 161 protein-ligand complexes (comprising 55 unique proteins)

Jain, T & Jayaram, B, *FEBS Letters*, **2005**, 579, 6659-6666

www.scfbio-iitd.res.in/software/drugdesign/bappl.jsp



Correlation between experimental ΔT_m and calculated free energy of interaction for DNA-Drug Complexes

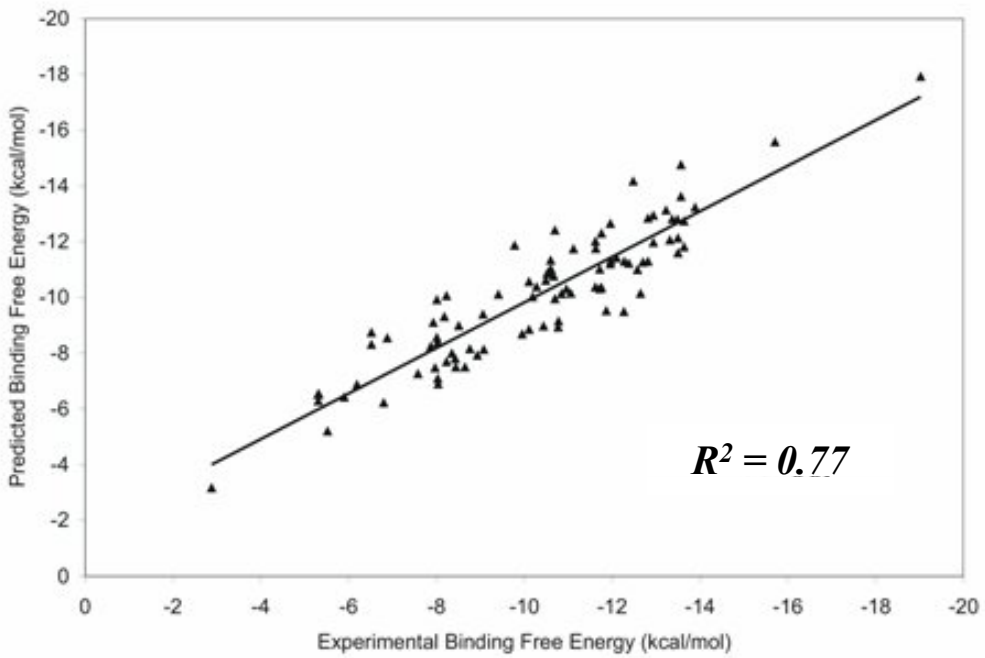
S.A Shaikh and B.Jayaram, *J. Med.Chem.* , **2007**, 50, 2240-2244

www.scfbio-iitd.res.in/software/drugdesign/preddicta.jsp

Comparative Evaluation of Scoring Functions

S. No.	Scoring Function	Method	Dataset		Correlation Coefficient (r)	Reference
			Training	Test		
1.	Present Work(BAPPL*)	Force field / Empirical	61	100	r = 0.92	*FEBS Letters*, 2005, 579, 6659
2.	DOCK	Force field	-	-	-	J. Comput.-Aided Mol. Des. 2001, 15, 411
3.	EUDOC	Force field	-	-	-	J. Comp. Chem. 2001, 22, 1750
4.	CHARMm	Force field	-	-	-	J. Comp. Chem. 1992, 13, 888
5.	AutoDock	Force field	-	-	-	J. Comp. Chem. 1998, 19, 1639
6.	DrugScore	Knowledge	-	-	-	J. Mol. Biol. 2000, 295, 337
7.	SMoG	Knowledge	-	36	r = 0.79	J. Am. Chem. Soc. 1996, 118, 11733
8.	BLEEP	Knowledge	-	90	r = 0.74	J. Comp. Chem. 1999, 202, 1177
9.	PMF	Knowledge	-	77	r = 0.78	J. Med. Chem. 1999, 42, 791
10.	DFIRE	Knowledge	-	100	r = 0.63	J. Med. Chem. 2005, 48, 2325
11.	SCORE	Empirical	170	11	r = 0.81	J. Mol. Model. 1998, 4, 379
12.	GOLD	Empirical	-	-	-	J. Mol. Biol. 1997, 267, 727
13.	LUDI	Empirical	82	12	r = 0.83	J. Comput.-Aided Mol. Des. 1994, 8, 243 & 1998, 12, 309
14.	FlexX	Empirical	-	-	-	J. Mol. Biol. 1996, 261, 470
15.	ChemScore	Empirical	82	20	r = 0.84	J. Comput.-Aided Mol. Des. 1997, 11, 425
16.	VALIDATE	Empirical	51	14	r = 0.90	J. Am. Chem. Soc. 1996, 118, 3959
17.	Ligscore	Empirical	50	32	r = 0.87	J. Mol. Graph. Model. 2005, 23, 395
18.	X-CSCORE	Empirical (consensus)	200	30	r = 0.77	J. Comput.-Aided Mol. Des. 2002, 16, 11
19.	GLIDE	Force field / Empirical	-	-	-	J. Med. Chem. 2004, 47, 1739

Binding Affinity Analysis on Zinc Containing Metalloprotein-Ligand Complexes



Correlation between the predicted and experimental binding free energies for 90 zinc containing metalloprotein-ligand complexes comprising 5 unique targets

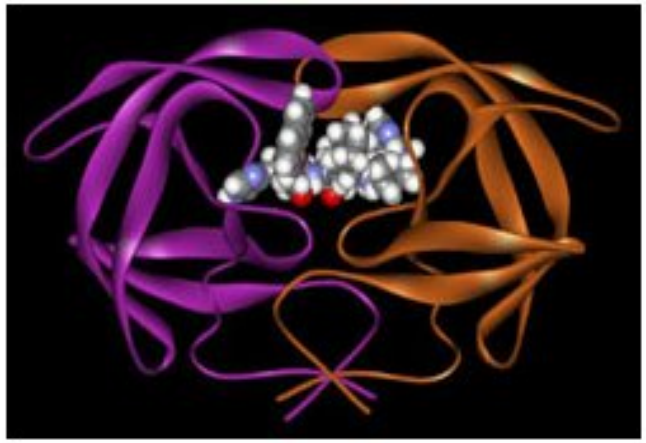
T. Jain & B. Jayaram, *Proteins: Struct. Funct. Bioinfo.* 2007, 67, 1167-1178.

www.scfbio-iitd.res.in/software/drugdesign/bapplz.jsp

Comparative evaluation of some methodologies reported for estimating binding affinities of zinc containing metalloprotein-ligand complexes

S. No.	Contributing Group	Method	Protein Studied	Training Set	Test Set	R^2
1.	Donini *et al*	MM-PBSA	MMP	-	6	
2.	Raha *et al*	QM	CA & CPA	-	23	0.69
3.	Toba *et al*	FEP	MMP	-	2	-
4.	Hou, *et al*	LIE	MMP	-	15	0.85
5.	Hu *et al*	Force Field	MMP	-	14	0.50
6.	Rizzo *et al*	MM-GBSA	MMP	-	6	0.74
7.	Khandelwal *et al*	QM/MM	MMP	-	28	0.76
8.	*Present Work*	*Force Field / Empirical*	*CA, CPA, MMP, AD & TL*	*40*	*50*	*0.77*

BAPPL server

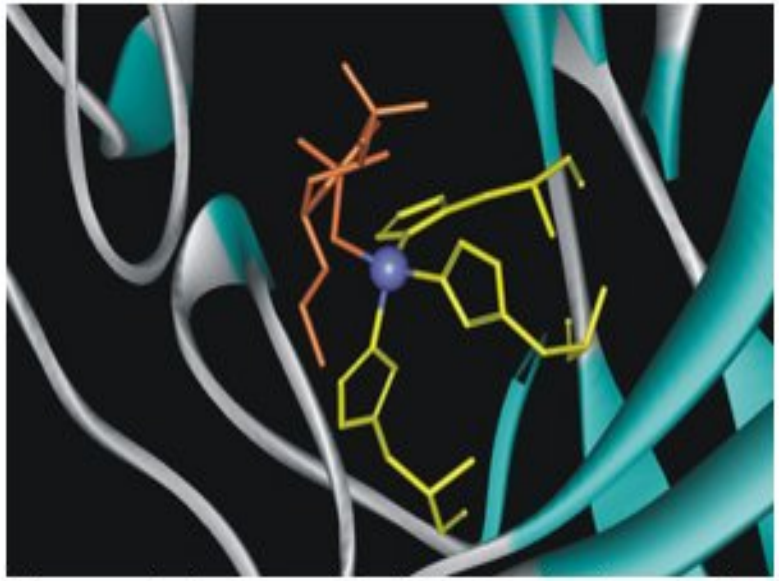


HIV-I Protease complexed with U75875 (1hiv.pdb)

Welcome to the BAPPL server

Binding Affinity Prediction of Protein-Ligand (BAPPL) server computes the binding free energy of a non-metallo protein-ligand complex using an all atom energy based empirical scoring function [1] & [2].

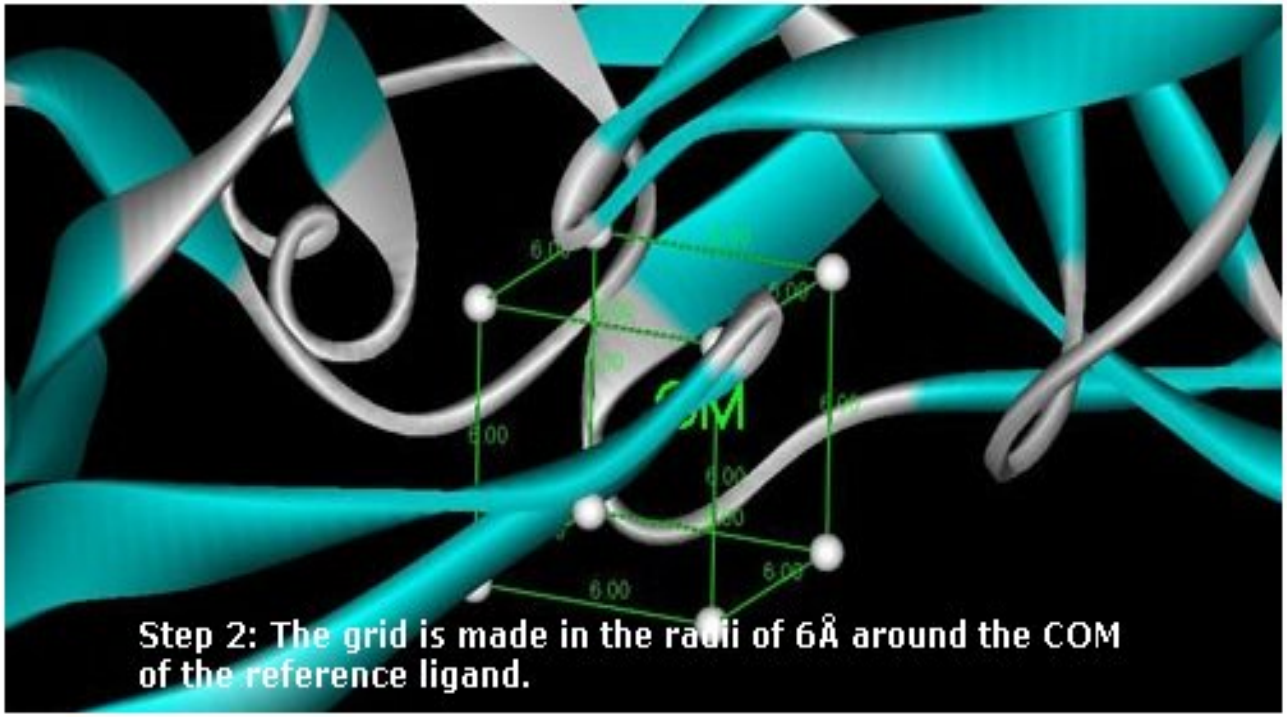
BAPPL-Z server



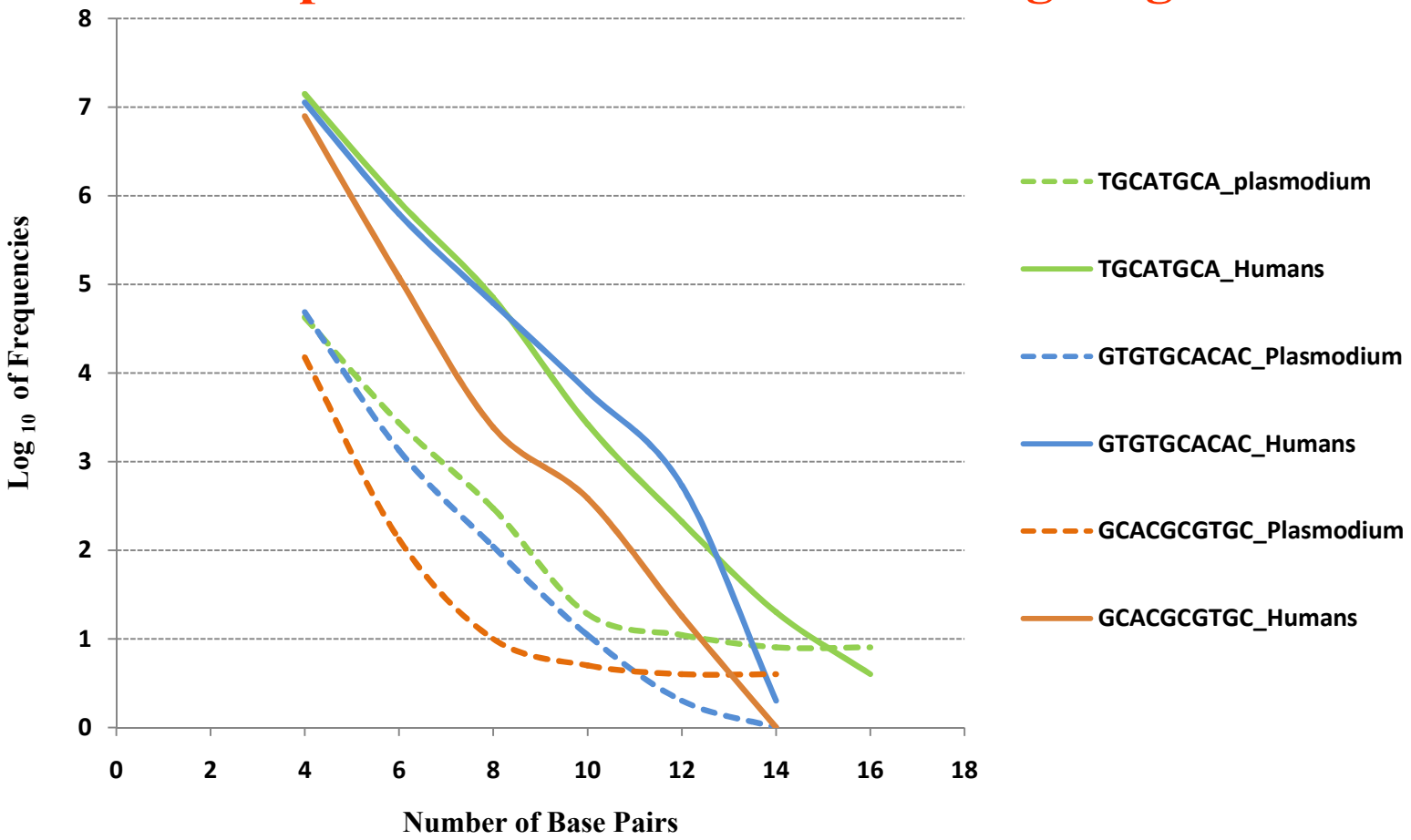
Carbonic Anhydrase complexed with Ligand and Zinc ion (1cil)

ParDOCK

Automated Server for Protein Ligand Docking



Optimum size of DNA as a drug target



Logarithm of the frequencies of the occurrence of base sequences of lengths 4 to 18 base pairs in *Plasmodium falciparum* and in humans embedding a regulatory sequence TGCATGCA (shown in green), GTGTGCACAC (blue) and GCACGCGTGC (orange) or parts thereof, of the plasmodium. The solid lines and the dashed lines correspond to humans and plasmodium, respectively. Curves lying between 0 and 1 on the log scale indicate occurrences in single digits.

One needs to cover at least 18 bp for uniqueness of the drug target

Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi

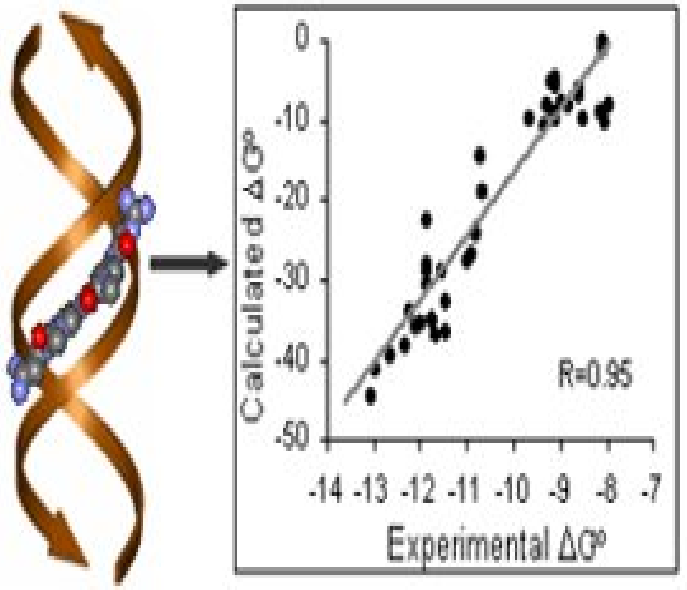
PreDDICTA

Predict **D**NA-**D**rug **I**nteraction strength by **C**omputing **A**Tm and **A**ffinity of binding.

About Preddicta

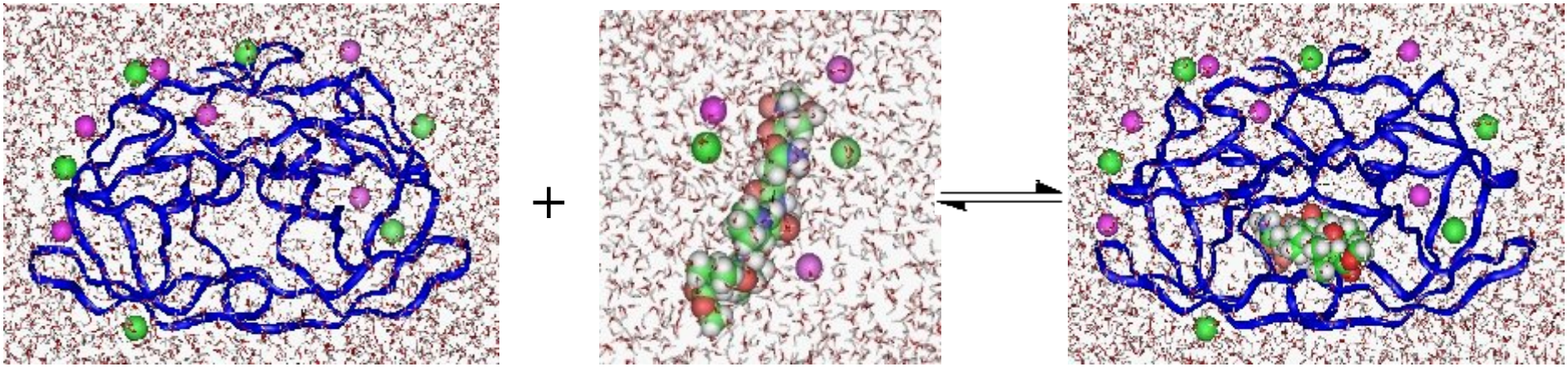
DNA Drug Interaction

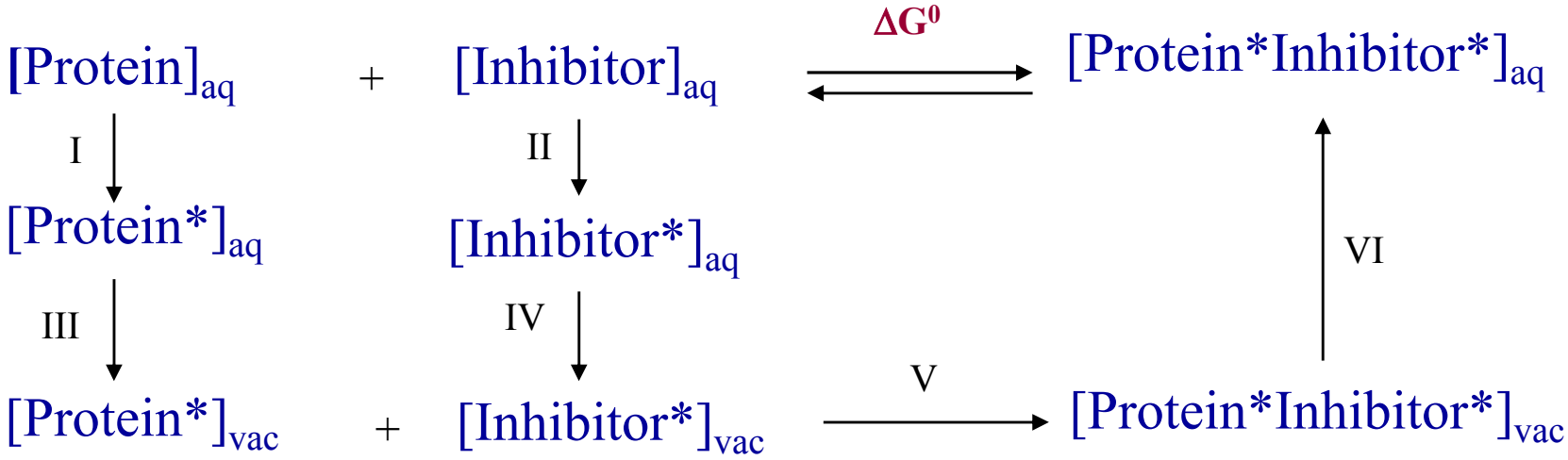
DNA Drug Complex Data Set



S.A Shaikh and B.Jayaram, *J. Med.Chem.* , **2007**, 50, 2240-2244

Binding Affinity Analysis

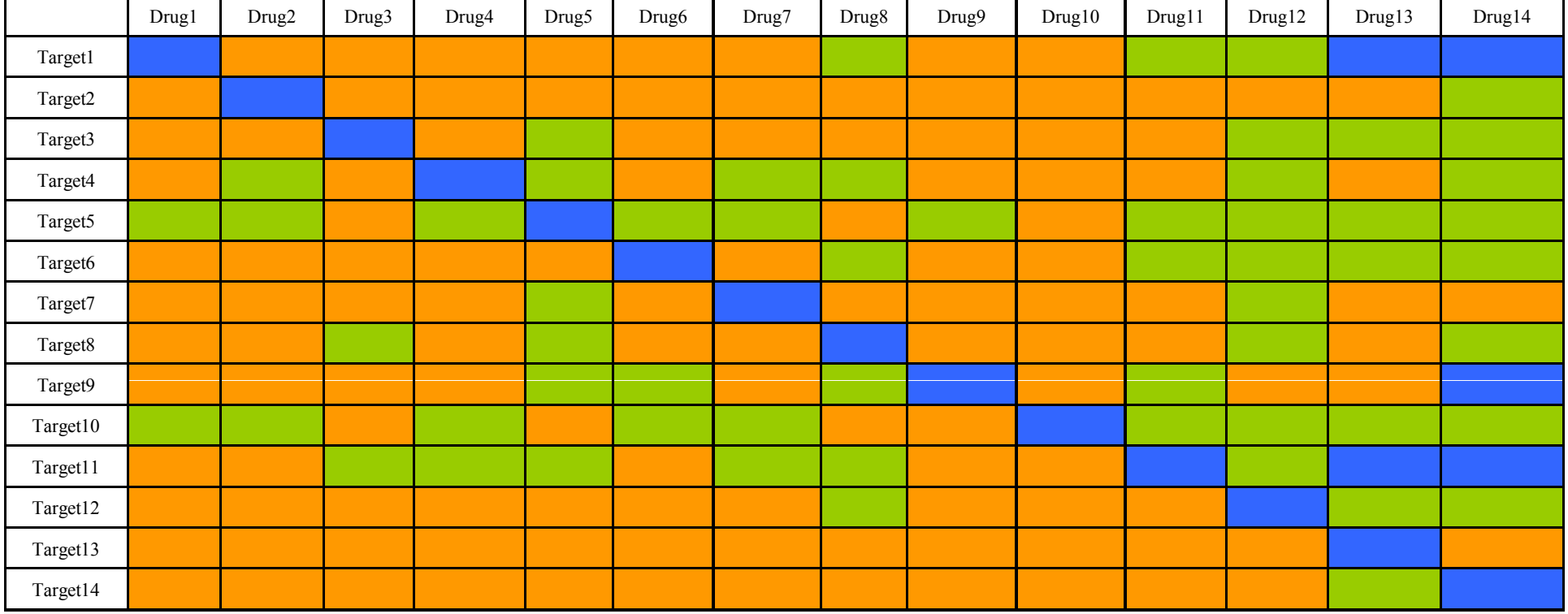




P. Kalra, T. V. Reddy, and B. Jayaram, "Free energy component analysis for drug design: A case study of HIV-1 protease-inhibitor binding", *J. Med. Chem.*, 2001, *44*, 4325-4338.

Affinity / Specificity Matrix for Drugs and Their Targets/Non-Targets

Shaikh, S., Jain. T., Sandhu, G., Latha, N., Jayaram., B., *A physico-chemical pathway from targets to leads*, 2007, *Current Pharmaceutical Design*, 13, 3454-3470.

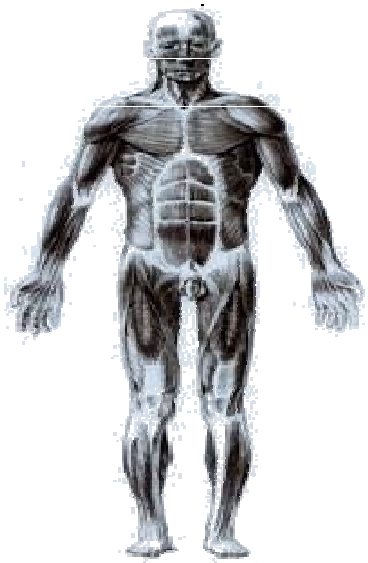


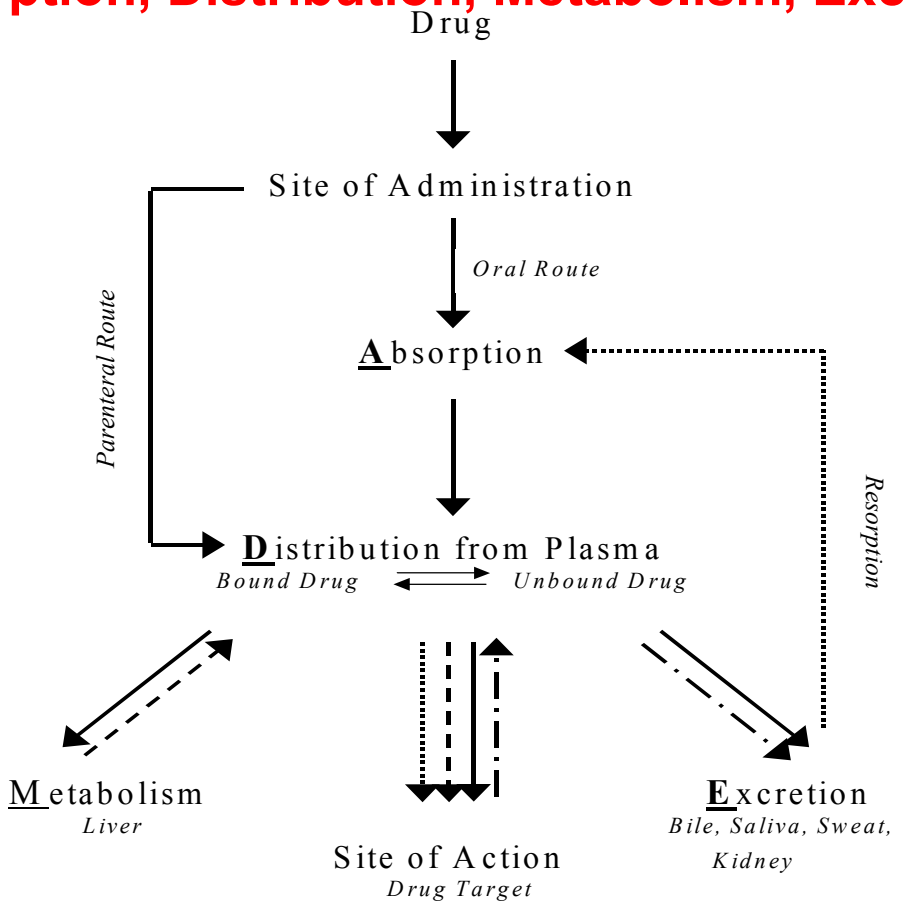
	Drug1	Drug2	Drug3	Drug4	Drug5	Drug6	Drug7	Drug8	Drug9	Drug10	Drug11	Drug12	Drug13	Drug14
Target1														
Target2														
Target3														
Target4														
Target5														
Target6														
Target7														
Target8														
Target9														
Target10														
Target11														
Target12														
Target13														
Target14														

BLUE: HIGH BINDING AFFINITY **GREEN: MODERATE AFFINITY** **ORANGE: POOR AFFINITY**

Diagonal elements represent drug-target binding affinity and off-diagonal elements show drug-non target binding affinity. Drug 1 is specific to Target 1, Drug 2 to Target 2 and so on. Target 1 is lymphocyte function-associated antigen LFA-1 (CD11A) (1CQP; Immune system adhesion receptor) and Drug 1 is lovastatin.Target 2 is Human Coagulation Factor (1CVW; Hormones & Factors) and Drug 2 is 5-dimethyl amino 1-naphthalene sulfonic acid (dansyl acid). Target 3 is retinol-binding protein (1FEL; Transport protein) and Drug 3 is n-(4-hydroxyphenyl)all-trans retinamide (fenretinide). Target 4 is human cardiac troponin C (1LXF; metal binding protein) and Drug 4 is 1-isobutoxy-2-pyrrolidino-3[n-benzylanilino] propane (Bepridil). Target 5 is DNA {1PRP; d(CGCGAATTCGCG)} and Drug 5 is propamidine. Target 6 is progesterone receptor (1SR7; Nuclear receptor) and Drug 6 is mometasone furoate. Target 7 is platelet receptor for fibrinogen (Integrin Alpha-11B) (1TY5; Receptor) and Drug 7 is n-(butylsulfonyl)-o-[4-(4-piperidinyl)butyl]-l-tyrosine (Tirofiban). Target 8 is human phosphodiesterase 4B (1XMU; Enzyme) and Drug 8 is 3-(cyclopropylmethoxy)-n-(3,5-dichloropyridin-4-yl)-4-(difluoromethoxy)benzamide (Roflumilast). Target 9 is Potassium Channel (2BOB; Ion Channel) and Drug 9 is tetrabutylammonium. Target 10 is {2DBE; d(CGCGAATTCGCG)} and Drug 10 is Diminazene aceturate (Berenil). Target 11 is Cyclooxygenase-2 enzyme (4COX; Enzymes) and Drug 11 is indomethacin. Target 12 is Estrogen Receptor (3ERT; Nuclear Receptors) and Drug 12 is 4-hydroxytamoxifen. Target 13 is ADP/ATP Translocase-1 (1OKC; Transport protein) and Drug 13 is carboxyatractyloside. Target 14 is Glutamate Receptor-2 (2CMO; Ion channel) and Drug 14 is 2-({[(3e)-5-{4-[(dimethylamino)(dihydroxy)-lambda~4~-sulfanyl]phenyl}-8-methyl-2-oxo-6,7,8,9-tetrahydro-1H-pyrrolo[3,2-H]isoquinolin-3(2H)-ylidene]amino}oxy)-4-hydroxybutanoic acid. The binding affinities are calculated using the software made available at http://www.scfbio-iitd.res.in/software/drugdesign/bappl.jsp and http://www.scfbio-iitd.res.in/preddicta.

Future of Drug Discovery: Towards a Molecular View of ADMET (Absorption, Distribution, Metabolism, Excretion & Toxicity)





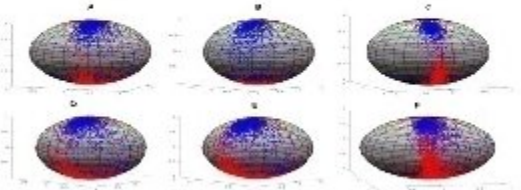
The distribution path of an orally administered drug molecule inside the body is depicted. Black solid arrows: Complete path of drug starting from absorption at site of administration to distribution to the various compartments in the body, like sites of metabolism, drug action and excretion. Dashed arrows: Path of the drug after metabolism. Dash-dot arrows: Path of drug after eliciting its required action on the target. Dot arrows: Path of the drug after being reabsorbed into circulation from the site of excretion.

Gene to Drug

Bioinformatics suite developed at SCFBio, IIT Delhi

A Chemical Model for Genome Analysis

ChemGene 1.0

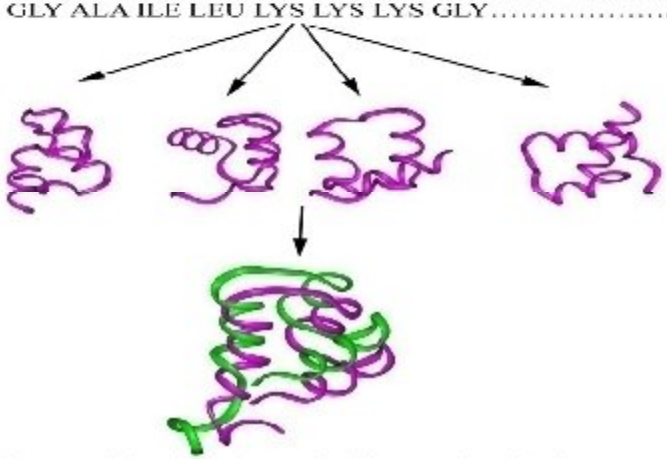


Gene (Blue) & Non Gene (Red) 120 Procaryotic genomes were evaluated & ~ 90 % sensitivity & specificity was observed

Protein Structure Prediction

Bhageerath 1.0

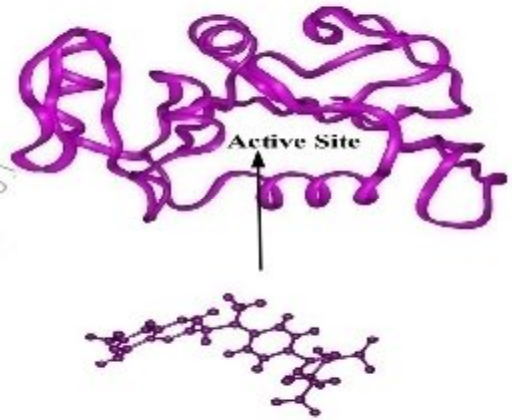
……GLU ALA GLU MET LYS ALA SER GLU ASP
LEU LYS LYS HIS GLY VAL THR VAL LEU THR ALA LEU
GLY ALA ILE LEU LYS LYS LYS GLY……



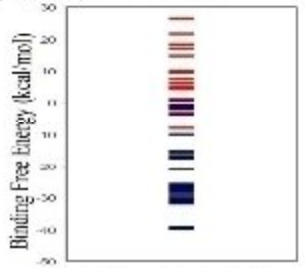
Bhageerath brackets the native like topology in the hundred best energy structure for small alpha helical proteins (green - native)

Active Site Directed Lead Design

Sanjeevini 1.0



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



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

BioGrid India

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)

SCFBio Team

16 processor Linux Cluster

Storage Area Network

~ 6 teraflops of computing; 20 terabytes of storage + huge brain power

BioComputing Group, IIT Delhi (PI : Prof. B. Jayaram)

Present

Shashank Shekhar
Tanya Singh
Avinash Mishra
Abhilash Jayaraj
Sahil Kapoor
Preeti Bisht

Garima Khandelwal
Priyanka Dhingra
Ashutosh Shandilya
Anjali Soni
Pooja Khurana
Sanjeev Kumar

Goutam Mukherjee
Vandana
Satyanarayan Rao
Navneet Tomar
Nagarajan

Former

Dr. Achintya Das
Dr. Tarun Jain
Dr. Kumkum Bhushan
Dr. Nidhi Arora
Pankaj Sharma
A.Gandhimathi
Neelam Singh
Dr. Sandhya Shenoy

Dr. N. Latha
Dr. Saher Shaikh
Dr. Poonam Singhal
Dr. E. Rajasekaran
Praveen Agrawal
Gurvisha Sandhu
Shailesh Tripathi
Rebecca Lee

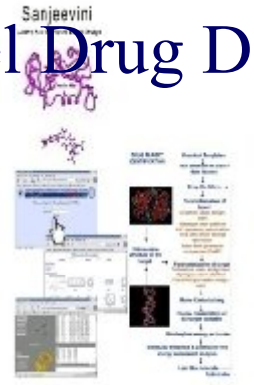
Dr. Pooja Narang
Dr. Parul Kalra
Dr. Surjit Dixit
Surojit Bose
Vidhu Pandey
Anuj Gupta
Dhrubajyoti Biswas
Bharat Lakhani

Collaborators: Dr. Aditya Mittal & Prof. D.L. Beveridge

LeadInvent

Technologies

Novel Drug Discovery



Drug Design Solutions

Biospectrum Award 2011

Asia Pacific Emerging Company of the Year

Mr. Pankaj Sharma
Mr. Surojit Bose
Mr. Praveen Aggarwal
Ms. Gurvisha Sandhu

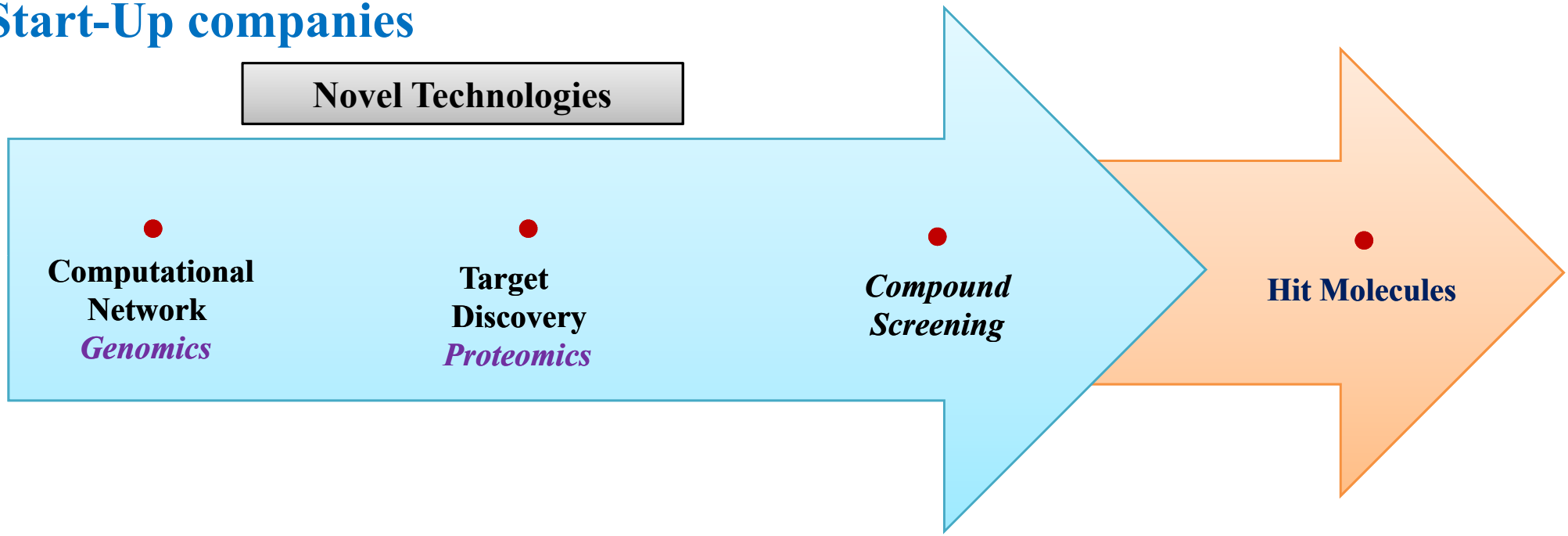
Incubated at IIT Delhi (2007-2009)

www.leadinvent.com

Novo Informatics
Smart Minds. Smart Technologies.

Under Incubation at IITD (since April, 2011)

Received TATA NEN Award 2012 for being one of the best Upcoming Start-Up companies



NI research pipeline

Sahil Kapoor
Avinash Mishra
Shashank Shekhar

Acknowledgements

Department of Biotechnology

Department of Science & Technology

Ministry of Information Technology

Council of Scientific & Industrial Research

Indo-French Centre for the Promotion of Advanced Research (CEFIPRA)

HCL Life Science Technologies

Dabur Research Foundation

Indian Institute of Technology, Delhi

A Few Key References

Genome Annotation

1.(a) S. Dutta, P. Singhal, P. Agrawal, R. Tomer, Kritee, E. Khurana and B. Jayaram. *A Physico-Chemical Model for Analyzing DNA sequences*, 2006, *Journal of Chemical Information & Modelling,* 46(1), 78-85. (b) P. Singhal, B. Jayaram, S. B. Dixit and D. L. Beveridge. *Molecular Dynamics Based Physicochemical Model for Gene Prediction in Prokaryotic Genomes*, 2008, *Biophysical Journal*, 94, 4173-4183; (c) G. Khandelwal and B. Jayaram. *A Phenomenological Model for Predicting Melting Temperatures of DNA Sequences*, 2010, *PLoS One*, 5(8): e12433. doi:10.1371/journal.pone.0012433.

Protein Structure Prediction

2.(a) P. Narang, K. Bhushan, S. Bose and B. Jayaram. *A computational pathway for bracketing native-like structures for small alpha helical globular proteins*. 2005, *Phys. Chem. Chem. Phys.,* 7, 2364-2375; (b) B. Jayaram et al., *Bhageerath*, 2006, *Nucleic Acid Res.*, 34, 6195-6204; (c) S. R. Shenoy and B. Jayaram, *Proteins: Sequence to Structure and Function – Current Status,* 2010, *Curr. Prot. Pep. Sci.,* 11, 498-514; (d) A. Mittal, B. Jayaram, S. R. Shenoy and T. S. Bawa, *A Stoichiometry driven universal spatial organization of backbones of folded proteins: Are there Chargaff's rules for protein folding ?* 2010, *J. Biomol. Struc. Dyn.*, 28, 133-142; 2011, *JBSD*, 28, 443-454; 2011, *JBSD*, 28, 669-674.

Drug Design

3.(a) T. Jain and B. Jayaram. *An all atom energy based computational protocol for predicting binding affinities of protein-ligand complexes*. 2005, *FEBS Letters*, 579, 6659-6666; (b) T. Jain and B. Jayaram. *A computational protocol for predicting the binding affinities of zinc containing metalloprotein-ligand complexes*. 2007, *Proteins: Structure, function & Bioinformatics*, 67, 1167-1178; (c) S. Shaikh and B. Jayaram. *A swift all atom energy based computational protocol to predict DNA-Drug binding affinity and* DT_m, 2007, *J. Med. Chem.*, 50, 2240-2244; (d) S. Shaikh et al.. *A physico-chemical pathway from targets to leads*, 2007, *Current Pharmaceutical Design*, 13, 3454-3470. (e) G. Mukherjee, N. Patra, P. Barua and B. Jayaram, *A fast empirical GAFF compatible partial atomic charge assignment scheme for modeling interactions of small molecules with biomolecular targets*, 2011, *J. Computational Chemistry*, 32,893-907. (f) T. Singh, D. Biswas and B. Jayaram, *AADS - An automated active site identification, docking and scoring protocol for protein targets based on physico-chemical descriptors,* 2011, *Journal of Chemical Information & Modeling*, 51 (10), 2515-2527, DOI: 10.1021/ci200193z

Visit Us at www.scfbio-iitd.res.in

Thank You