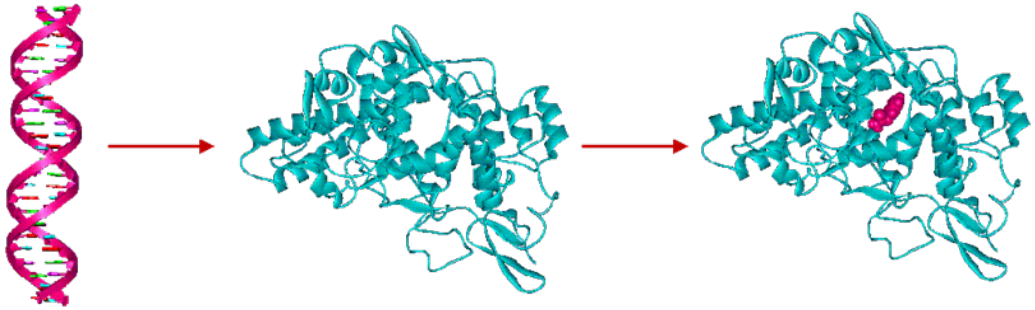


Genomes to Hit Molecules in Silico: A Country Path Today, A Highway Tomorrow

Prof. B. Jayaram

Department of Chemistry &

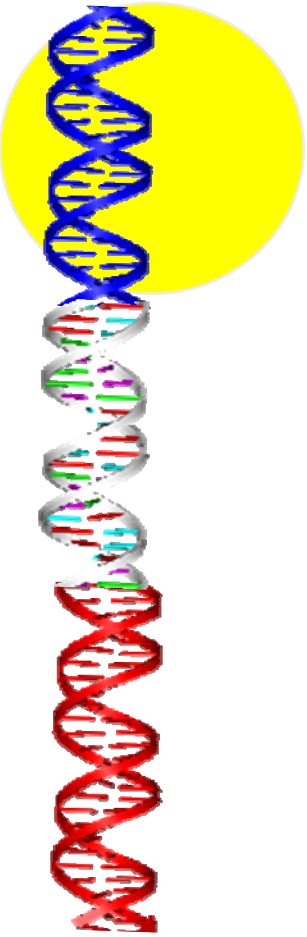
Supercomputing Facility for Bioinformatics & Computational Biology &

School of Biological Sciences

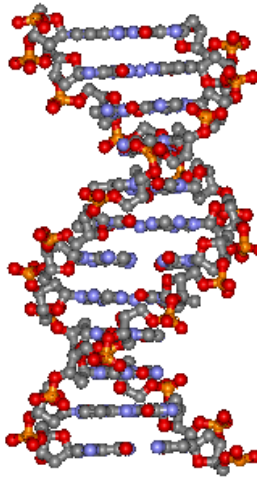
Indian Institute of Technology Delhi

The Dream @ SCFBio:

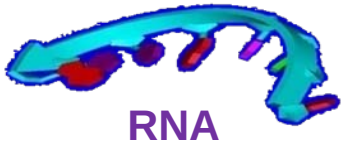
From Genome to Drug : Establishing the Central Dogma of Modern Drug Discovery



ChemGenome



DNA

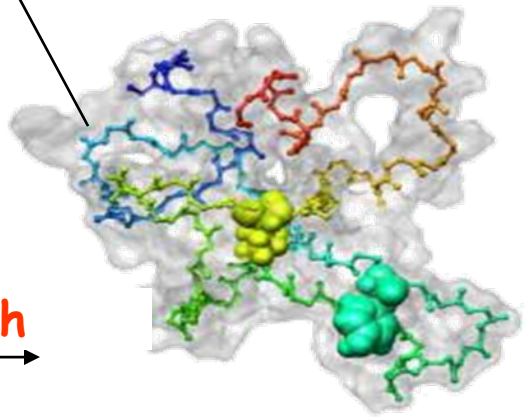


RNA

SEEARTINSCIENCE……

Primary Sequence

Bhageerath



Tertiary Structure

Drug

Sanjeevini

Genome

Protein

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.

Input the HBV Genome sequence to *ChemGenome*

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

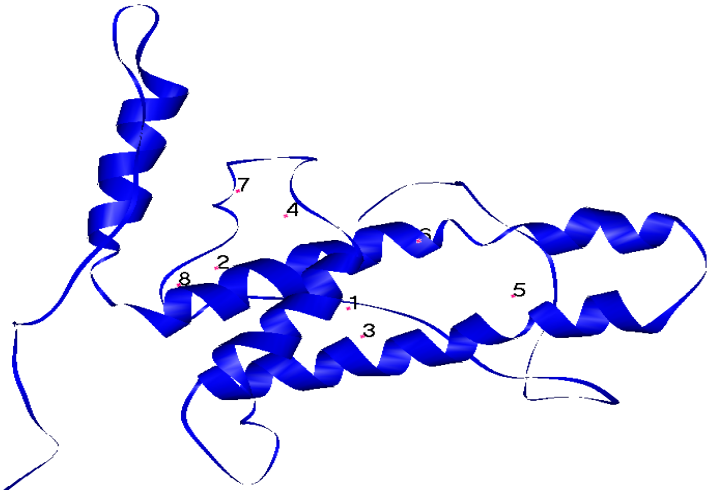
***ChemGenome 3.0* output**
Five protein coding regions identified

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 TPSPRRRRSQSPRRRRSQSRESQC

Input Amino acid sequence to *Bhageerath-H*



Input Protein Structure to Active site identifier (ASF/*Sanjeevini*)

10 potential binding sites identified

Scan a million compound library

RASPD/*Sanjeevini* calculation with an average cut off binding affinity to limit the number of candidates. (Empirical scoring function which builds in Lipisnki's rules and Wiener index)

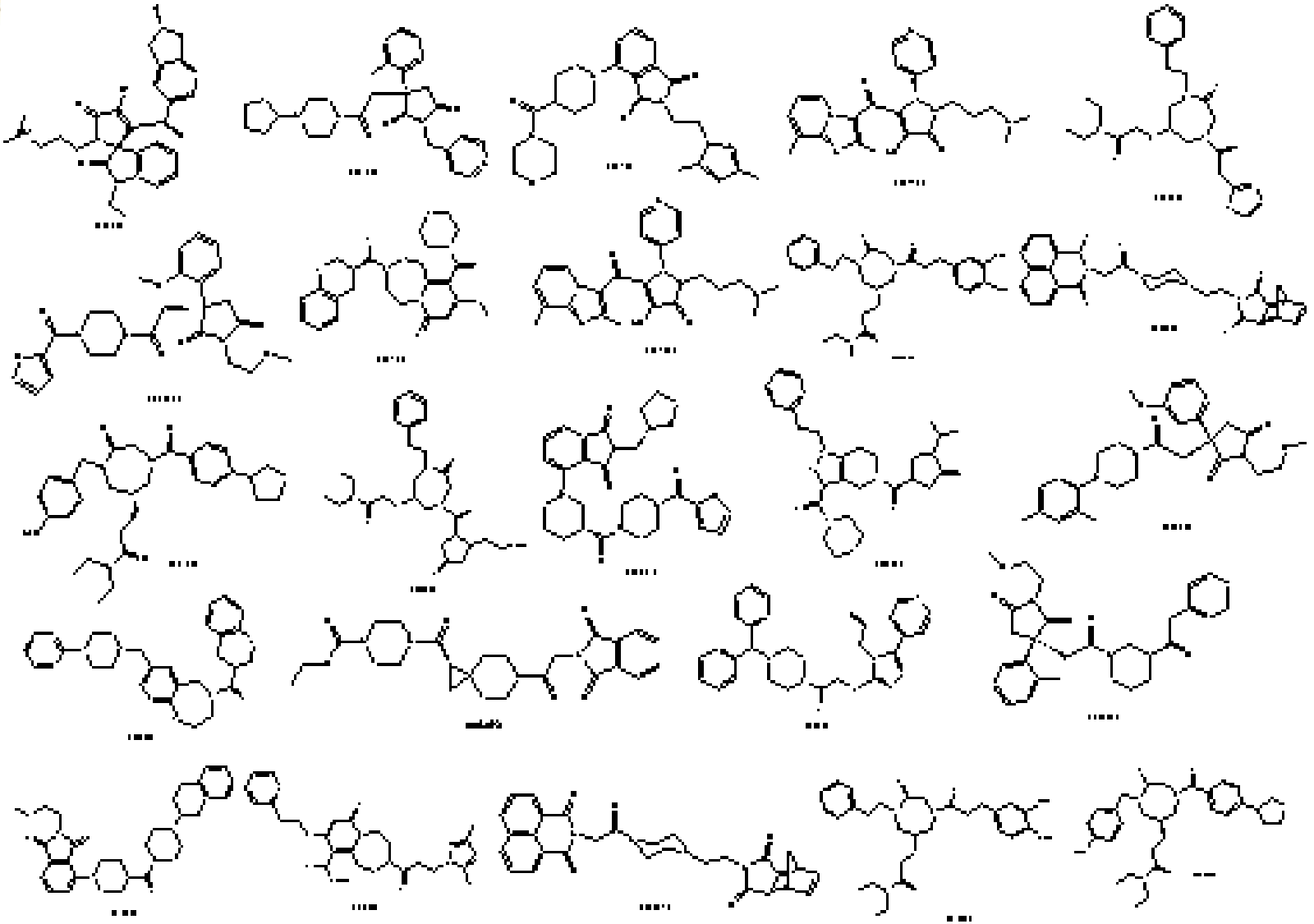
RASPD output

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/*Sanjeevini* 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.

Mol. 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



B. Jayaram, Priyanka Dhingra, Goutam Mukherjee, Vivekanandan Perumal, "Genomes to Hits: The Emerging Assembly Line", Proceedings of the Ranbaxy Science Foundation 17th Lecture Series, 2012, Ch-3, 13-35.

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 active site/target directed lead molecule design protocol

List of tools available for gene prediction

Sl. No.	Softwares	URLs	Methodology
1.	FGENESH	http://linux1.softberry.com/all.htm	*Ab initio*
2.	GeneID	http://www1.imim.es/geneid.html	*Ab initio*
3.	GeneMark	http://exon.gatech.edu/GeneMark/gmchoice.html	*Ab initio*
4.	GeneMark.hmm	http://exon.gatech.edu/hmmchoice.html	*Ab initio*
5.	GeneWise	http://www.ebi.ac.uk/Tools/Wise2/	Homology
6.	GENSCAN	http://genes.mit.edu/GENSCAN.html	*Ab initio*
7.	Glimmer	http://www.tigr.org/software/glimmer/	*Ab initio*
8.	GlimmerHMM	http://www.cbcb.umd.edu/software/glimmerhmm/	*Ab initio*
9.	GRAILEXP	http://compbio.ornl.gov/grailexp	*Ab initio*
10.	GENVIEW	http://zeus2.itb.cnr.it/~webgene/wwwgene.html	*Ab initio*
11.	GenSeqer	http://bioinformatics.iastate.edu/cgi-bin/gs.cgi	Homology
12.	PRODIGAL	http://prodigal.ornl.gov/	Homology
13.	MORGAN	http://www.cbcb.umd.edu/~salzberg/morgan.html	*Ab initio*
14.	PredictGenes	http://mendel.ethz.ch:8080/Server/subsection3_1_8.html	Homology
15.	MZEF	http://rulai.cshl.edu/software/index1.htm	*Ab initio*
16.	Rosetta	http://crossspecies.lcs.mit.edu	Homology
17.	EuGéne	http://eugene.toulouse.inra.fr/	*Ab initio*
18.	PROCRUSTES	http://www.riethoven.org/BioInformer/newsletter/archives/2/procrustes.html	Homology
19.	Xpound	http://mobyle.pasteur.fr/cgi-bin/portal.py?#forms::xpound	*Ab initio*
20.	Chemgenome	http://www.scfbio-iitd.res.in/chemgenome/chemgenome3.jsp	*Ab initio*
21.	Augustus	http://augustus.gobics.de/	*Ab initio*
22.	Genome Threader	http://www.genomethreader.org/	Homology
23.	HMMgene	http://www.cbs.dtu.dk/services/HMMgene/	*Ab initio*
24.	GeneFinder	http://people.virginia.edu/~wc9c/genefinder/	*Ab initio*
25.	EGPRED	http://www.imtech.res.in/raghava/egpred/	*Ab initio*
26.	mGene	http://mgene.org/web	*Ab initio*

Eukaryotic Gene Prediction Accuracies

Intra- and inter-species gene prediction accuracy Intra-species performance figures derived from 5-fold cross-validation are along the diagonal in bold. (Korf, 2004)

Genomic DNA									
		At		Ce		Dm		Os	
Parameters	**Measure**	**SN**	**SP**	**SN**	**SP**	**SN**	**SP**	**SN**	**SP**
At	Nuc	**97.1**	**95.2**	78.7	91.3	77.7	68.0	90.7	71.8
	Exon	**82.9**	**81.2**	44.3	52.8	38.6	24.0	57.1	42.3
	Gene	**54.3**	**46.8**	20.9	11.3	18.8	5.7	20.5	9.7
Ce	Nuc	83.5	91.5	**97.6**	**94.2**	81.3	73.6	79.7	74.5
	Exon	40.5	49.9	**85.5**	**79.3**	42.2	29.8	27.5	26.0
	Gene	25.7	18.1	**46.0**	**32.5**	21.9	8.8	13.9	7.3
Dm	Nuc	30.0	95.3	45.9	95.0	**94.3**	**86.5**	78.4	89.8
	Exon	16.5	41.3	29.9	47.2	**78.6**	**67.2**	50.0	58.4
	Gene	3.2	4.3	7.8	6.9	**50.8**	**37.5**	36.3	28.9
Os	Nuc	39.3	96.3	24.9	95.5	79.8	88.7	**86.2**	**94.0**
	Exon	30.7	47.6	11.1	36.6	47.4	44.4	**70.2**	**72.4**
	Gene	5.1	6.1	5.3	7.8	27.2	17.2	**51.2**	**37.0**

Prediction models trained on one organism do not necessarily work well on another organism, unless they incorporate molecular level language of DNA

Finding genes in Arabidopsis Thaliana

(Thale Cress)

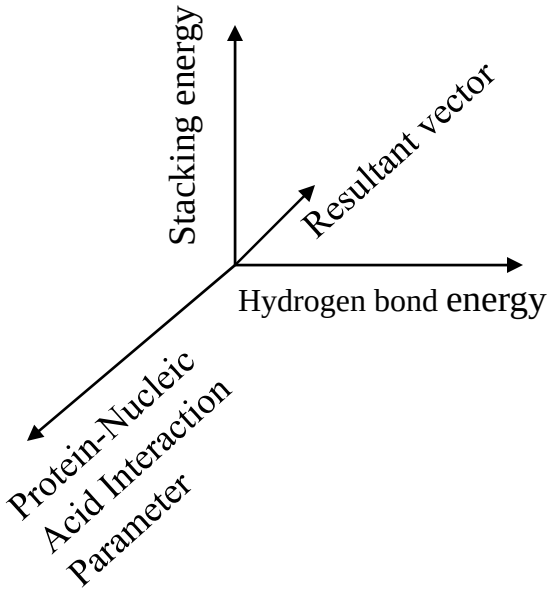
Software	Method	Sensitivity	Specificity
GeneMark.hmm http://www.ebi.ac.uk/genemark/	**5th-order Markov model**	**0.82**	**0.77**
GenScan http://genes.mit.edu/GENSCAN.html	**Semi Markov Model**	**0.63**	**0.70**
MZEF http://rulai.cshl.org/tools/genefinder/	**Quadratic Discriminant Analysis**	**0.48**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

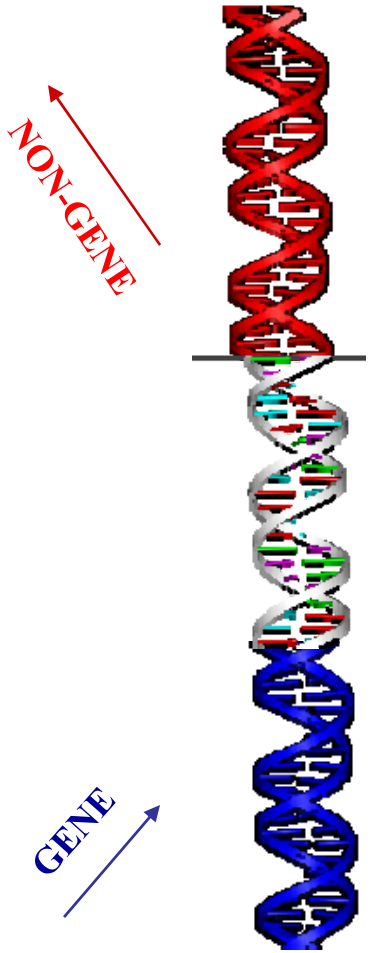
***Desired: A sensitivity & specificity of unity (all true genes are predicted with no false positives).**

While, the above methods have improved over the years and it is remarkable that they perform so well with limited experimental data to train on, more research, new methods transferable across species and new ways of looking at genomic DNA are required!

ChemGenome

Build a hypothesis driven three dimensional Physico-Chemical vector for DNA sequences, which as it walks along the genome, distinguishes Genes (coding regions) 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

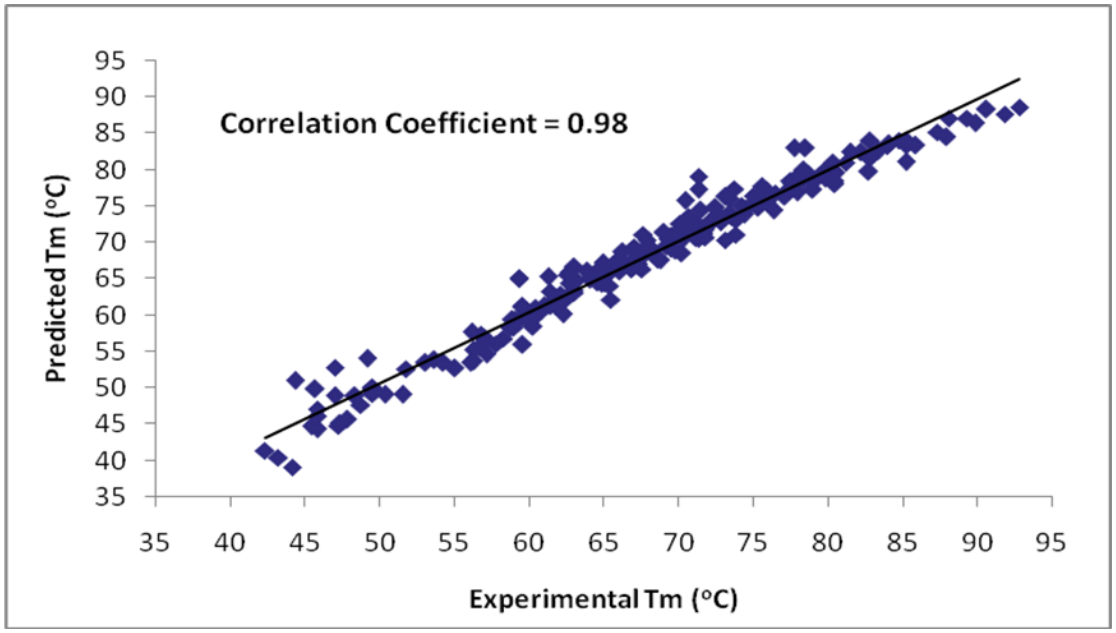
$$E_{HB} = E_{i\text{-}l} + E_{j\text{-}m} + E_{k\text{-}n}$$

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

Hydrogen bond & Stacking energies for all 32 unique trinucleotides were calculated from 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 codon (double helical trinucleotide) .

*Beveridge et al. (2004). *Biophys J,* 87, 3799-813; #Dixit et al. (2005). *Biophys J,* 89, 3721-40; #Lavery et al. (2009). Nucl. Acid Res., 38, 299-313.

Melting temperatures of ~ 200 oligonucleotides: Prediction versus Experiment

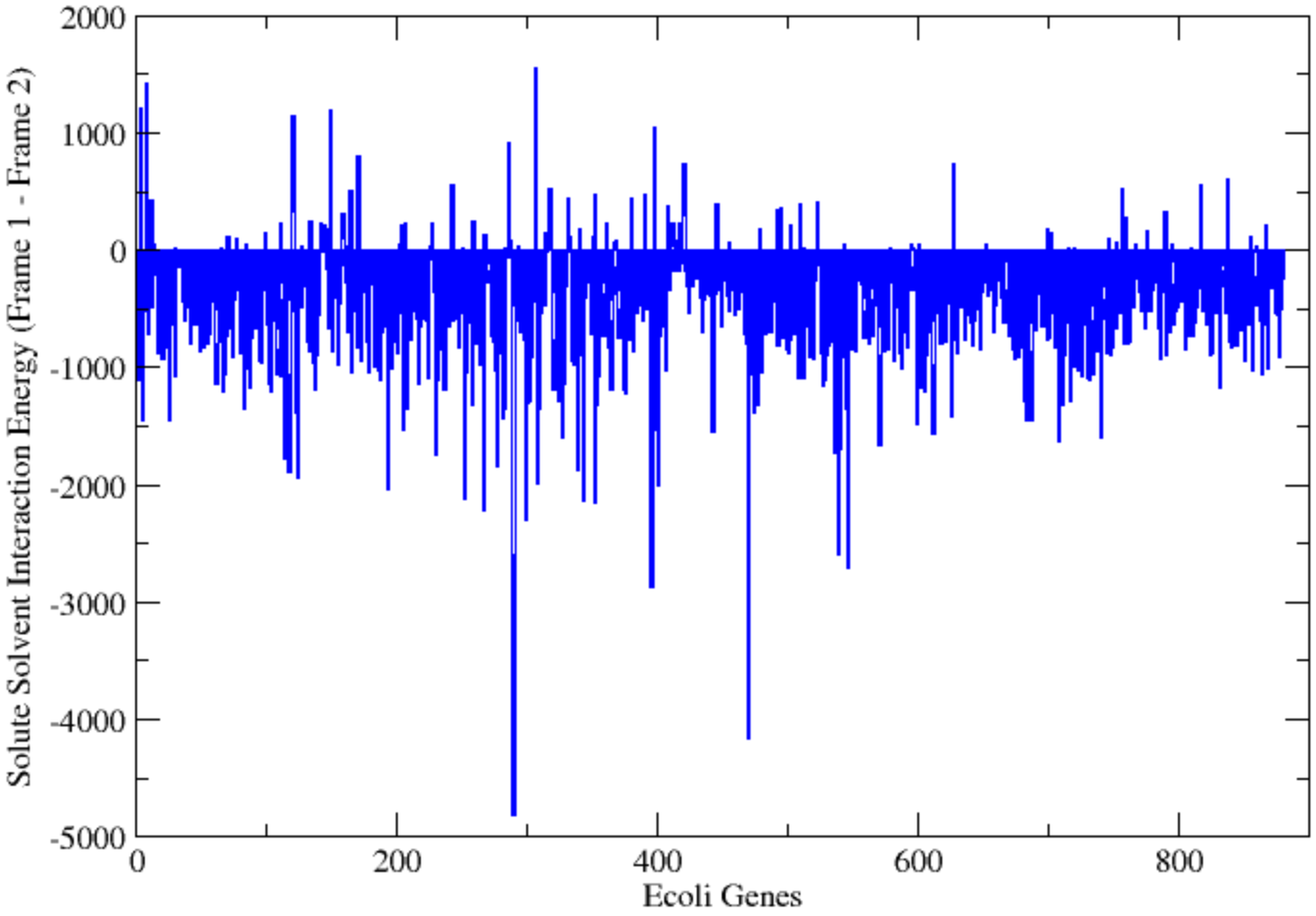


Tm(°C)=(7.35 ×E) + [17.34 ×ln(Len)] + [4.96 ×ln(Conc])+ [0.89× ln(DNA)] - 25.42

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

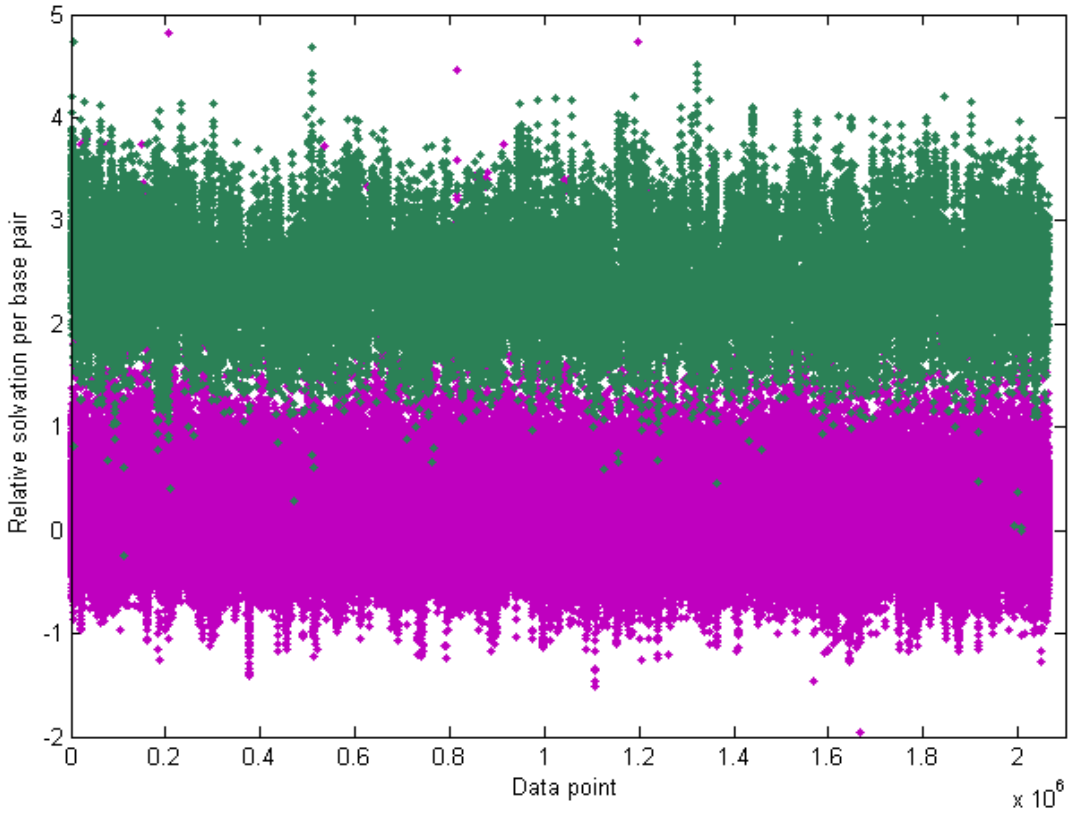
Garima Khandelwal, Jalaj Gupta and B. Jayaram, "DNA energetics based analyses suggest additional genes in prokaryotes" *J Bio Sc.*, 2012, 37, 433-444; DOI 10.1007/s12038-012-9221-7

Solute-Solvent Interaction Energy for Genes/Non-genes



Coding and noncoding frames have different solvation characteristics which can be used to build the third parameter (z), besides hydrogen bonding (x) and stacking (y).

Relative solvation energies per base pair for 2063537 mRNA (magenta) and 56251 tRNA (green) genes



Garima Khandelwal and B. Jayaram, "DNA-water interactions distinguish messenger RNA genes from transfer RNA genes", *J. Am. Chem. Soc.*, 2012, 134 (21), 8814–8816; DOI: 10.1021/ja3020956

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' coordinate of *chemgenome* or one can simply use +1/-1 as in the adjacent 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;

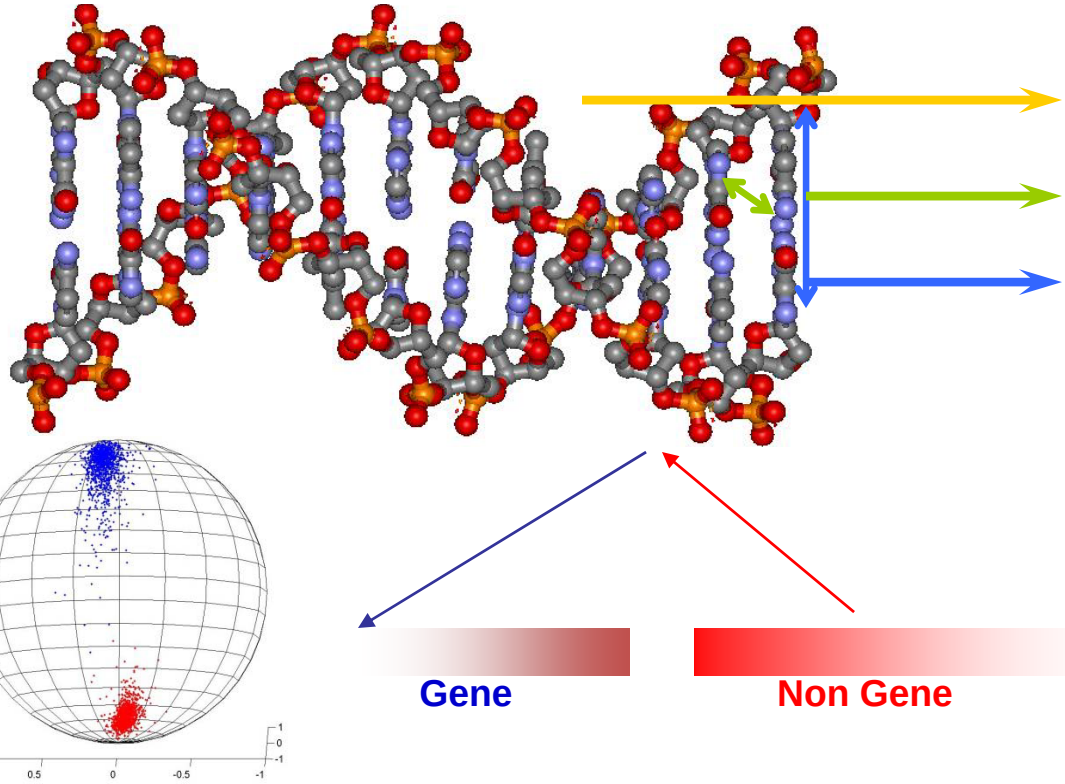
middle purine requires third base.

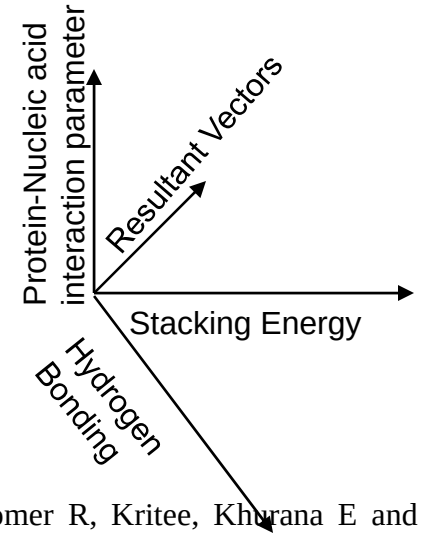
B. Jayaram, "Beyond Wobble: The Rule of Conjugates", *J. Molecular Evolution*, 1997, *45*, 704-705.

Codons with G_1 → +1; C_1G_3 or C_1T_3 → +1; C_1A_3 or C_1C_3 → -1

ChemGenome

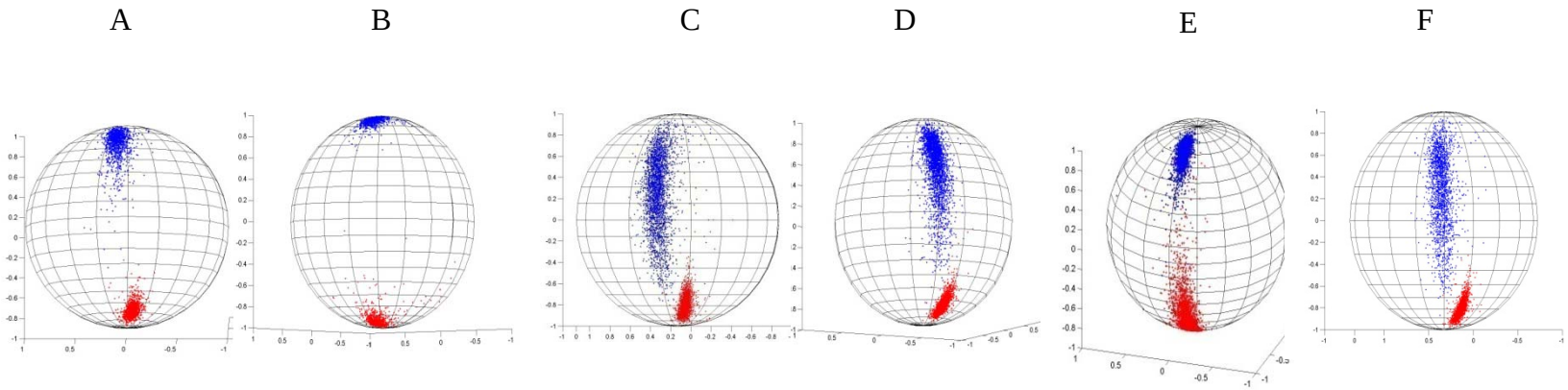
A Physico-Chemical Model for identifying signatures of functional units on Genomes





(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; (4) G. Khandelwal, J. Gupta and B. Jayaram, "DNA energetics based analyses suggest additional genes in prokaryotes" *J Bio Sc.*, 2012, 37, 433-444.

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)

Poonam Singhal, B. Jayaram, Surjit B. Dixit and David L. Beveridge, Molecular Dynamics Based Physicochemical Model for Gene Prediction in Prokaryotic Genomes, *Biophys. J.*, 2008, 94, 4173-4183.

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)

↓

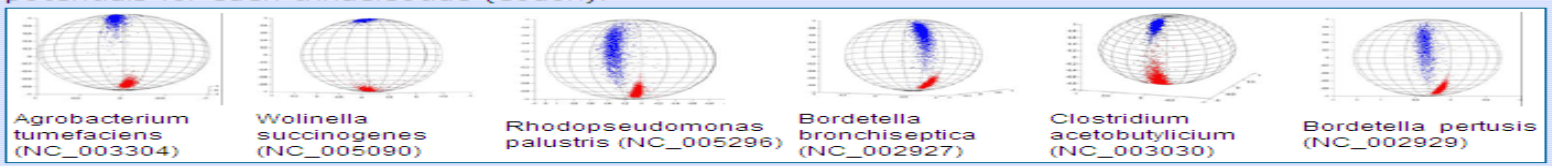
Potential protein coding genes

Poonam Singhal, B. Jayaram, Surjit B. Dixit and David L. Beveridge. Molecular Dynamics Based Physicochemical Model for Gene Prediction in Prokaryotic Genomes, **2008**, *Biophysical Journal*, 94, 4173-4183

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 294786 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 : ailesh@scfbio-iitd.res.in

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

```
>Gene Name (This comment line is necessary)
ATGTTGGTGTCCGCAAGGGTAGAGAAACAAAAGCGTGTTGCTTATCAGGGGAAGGCGACAGTGCTTGCTCTCGG
TAAGG
CCTTGCCGAGCAATGTTGTTTCCCAGGAGAATCTCGTGGAGGAGTATCTCCGTGAAATCAAATGCGATAACCTTTC
TAT
CAAAGACAAGCTGCAACACTTGTGCAAAAGCACAACTGTCAAGACACGCTACACAGTCATGTCACGGGAGACG
CTGCAC
AAATACCCTGAACTAGCAACCGAGGGTTCCCCAACCATCAAACAGAGGCTTGAGATTGCAAACGATGCGGTTGT
GCAGA
```

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, Tomer 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.*,1997.45.704.

The ChemGenome2.0 WebServer

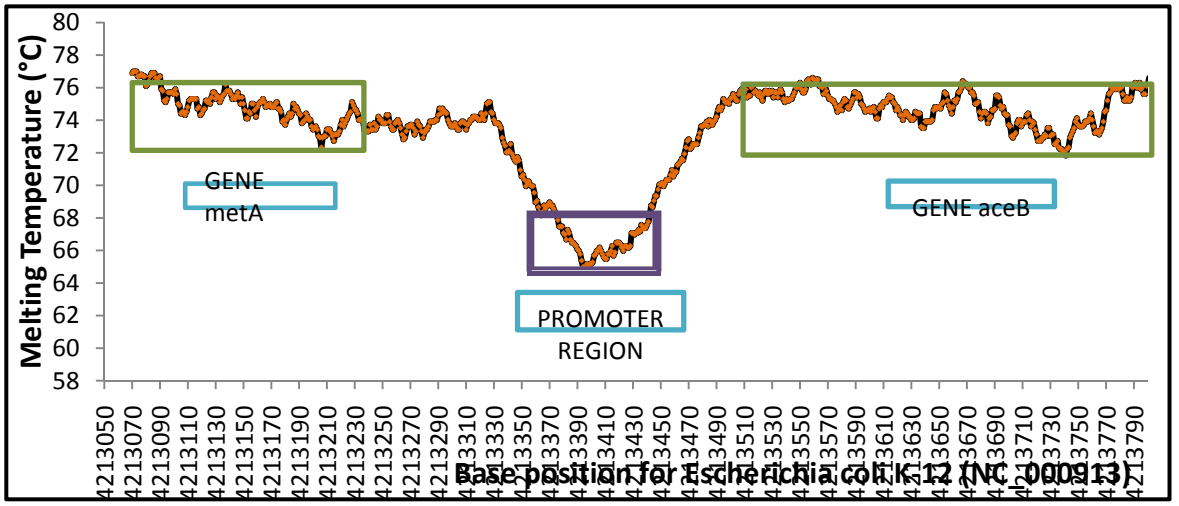
http://www.scfbio-iitd.res.in/chemgenome/chemgenomenew.jsp

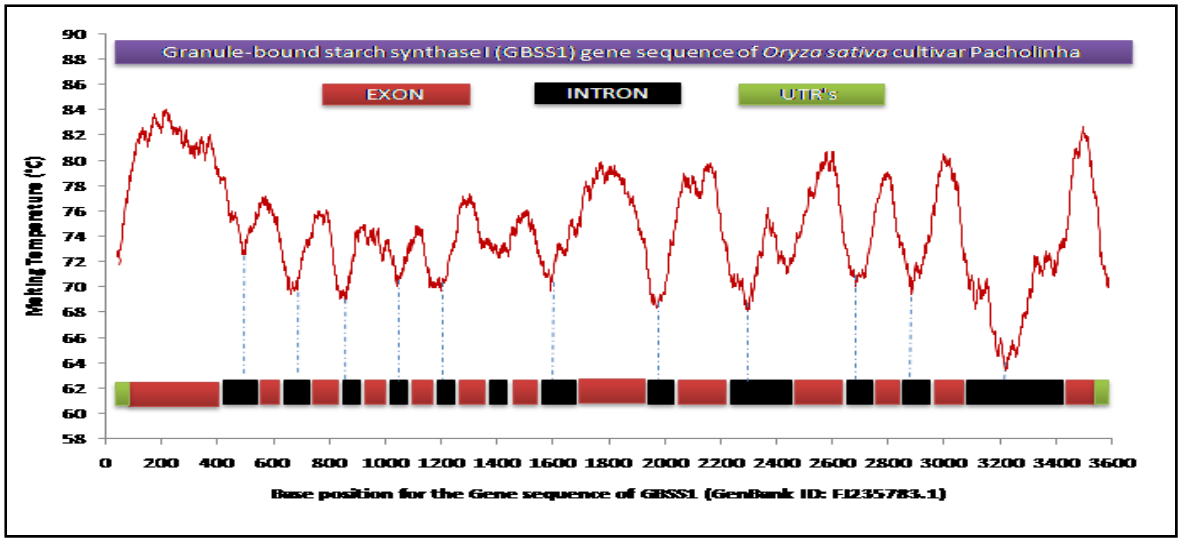
Back to Finding Genes in Arabidopsis Thaliana

(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 (Chemgenome) performs as well as any other sophisticated knowledge based methods and is amenable to further systematic improvements.





Chemgenome methodology enables detection of not only coding regions but also promoters, introns & exons etc.. G. Khandelwal, B. Jayaram, *PLoS One* , 2010, *5(8)*, e12433

Let us read the book of Human Genome soon like a Harry Potter novel !

Human Genome
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

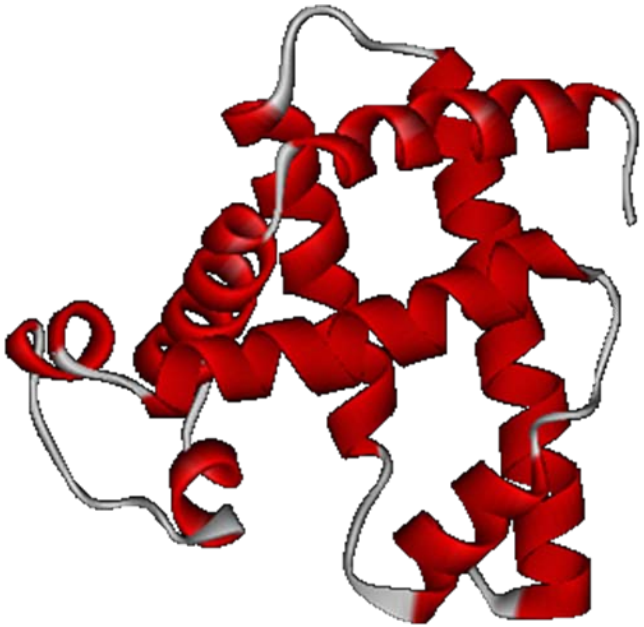
•Drug Design – *Sanjeevini*

A comprehensive active site/target directed lead molecule design protocol

Bhageerath

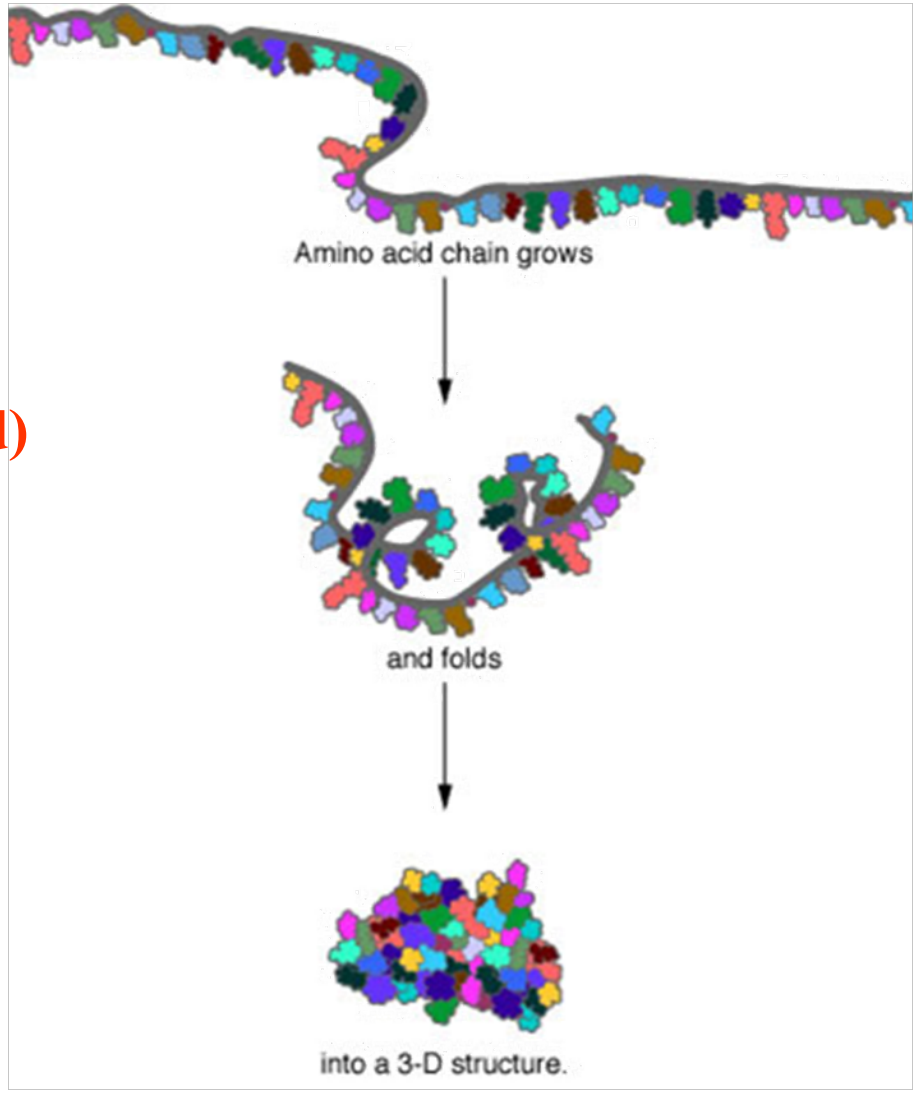
Protein Tertiary Structure Prediction

.....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.....



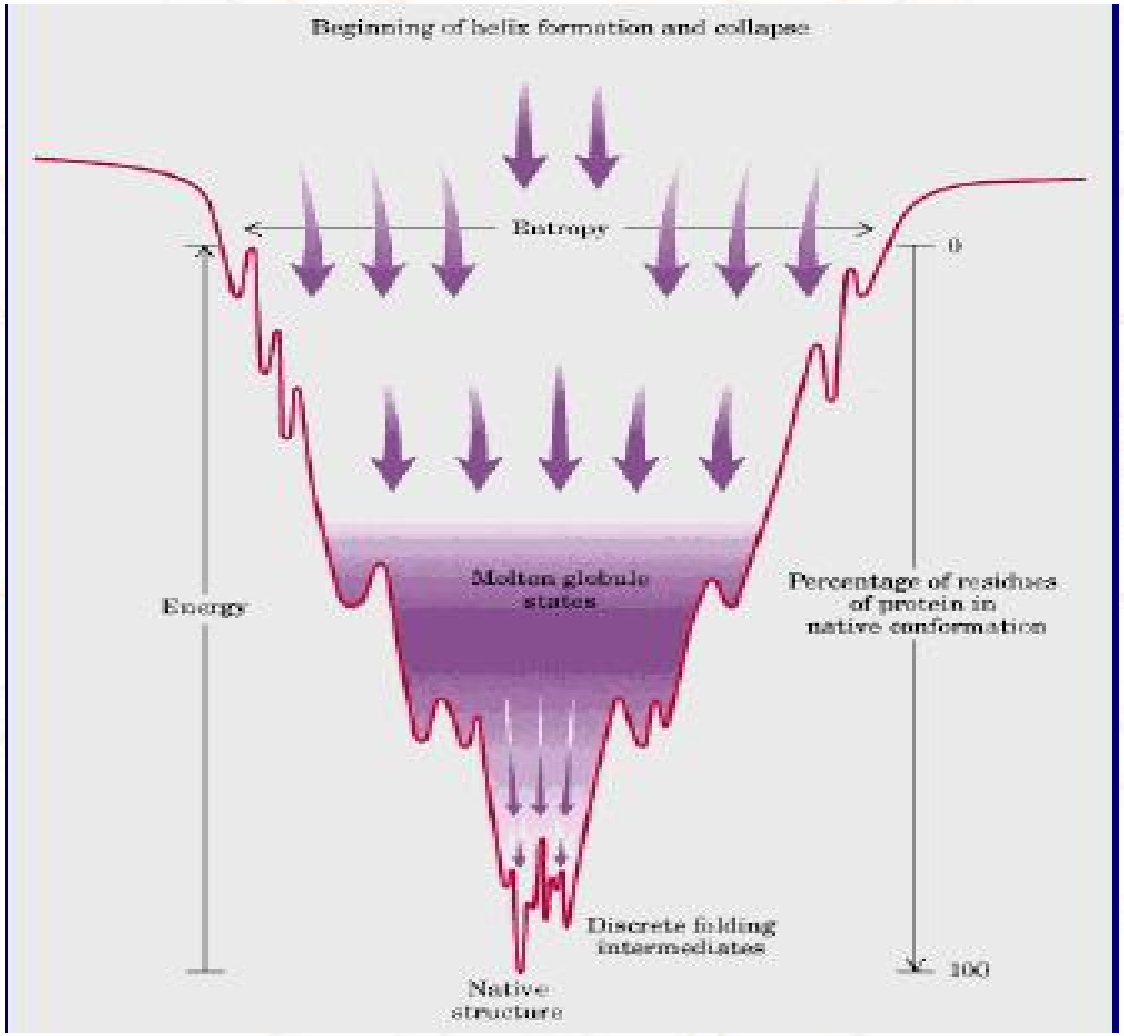
Protein Folding Problem

Recognized as a Grand Challenge / NP Complete (hard) problem



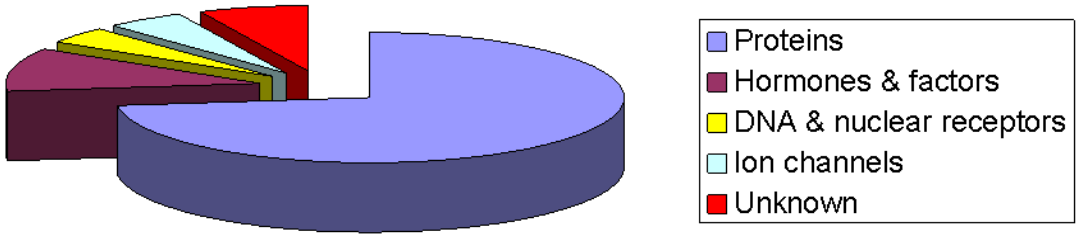
PROTEIN FOLDING LANDSCAPE

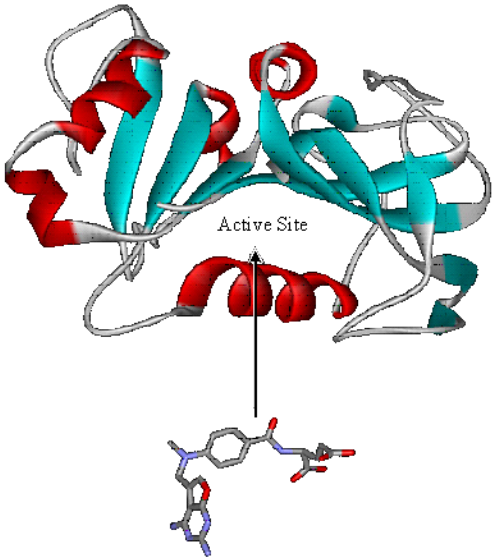
"Native structure" at the bottom of the free energy well is the folded (native) protein – Thermodynamic hypothesis of Anfinsen



WHY FOLD PROTEINS ?

One of the several compelling reasons comes from Pharmaceutical/Medical Sector





Majority of Drug Targets

are Proteins

- Structure-based drug-design
- Mapping the functions of proteins in metabolic pathways.

Experimental methods such as X-Ray & NMR provide the true structures but these are not cost and time effective and hence the need for computational models.

Comparative Modeling Approaches (knowledge-based methods) for Protein Tertiary Structure Prediction

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)

These work best when sequence matches, global or local, are found in databases (RCSB/PDB) of known structures

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 = 3×10^{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

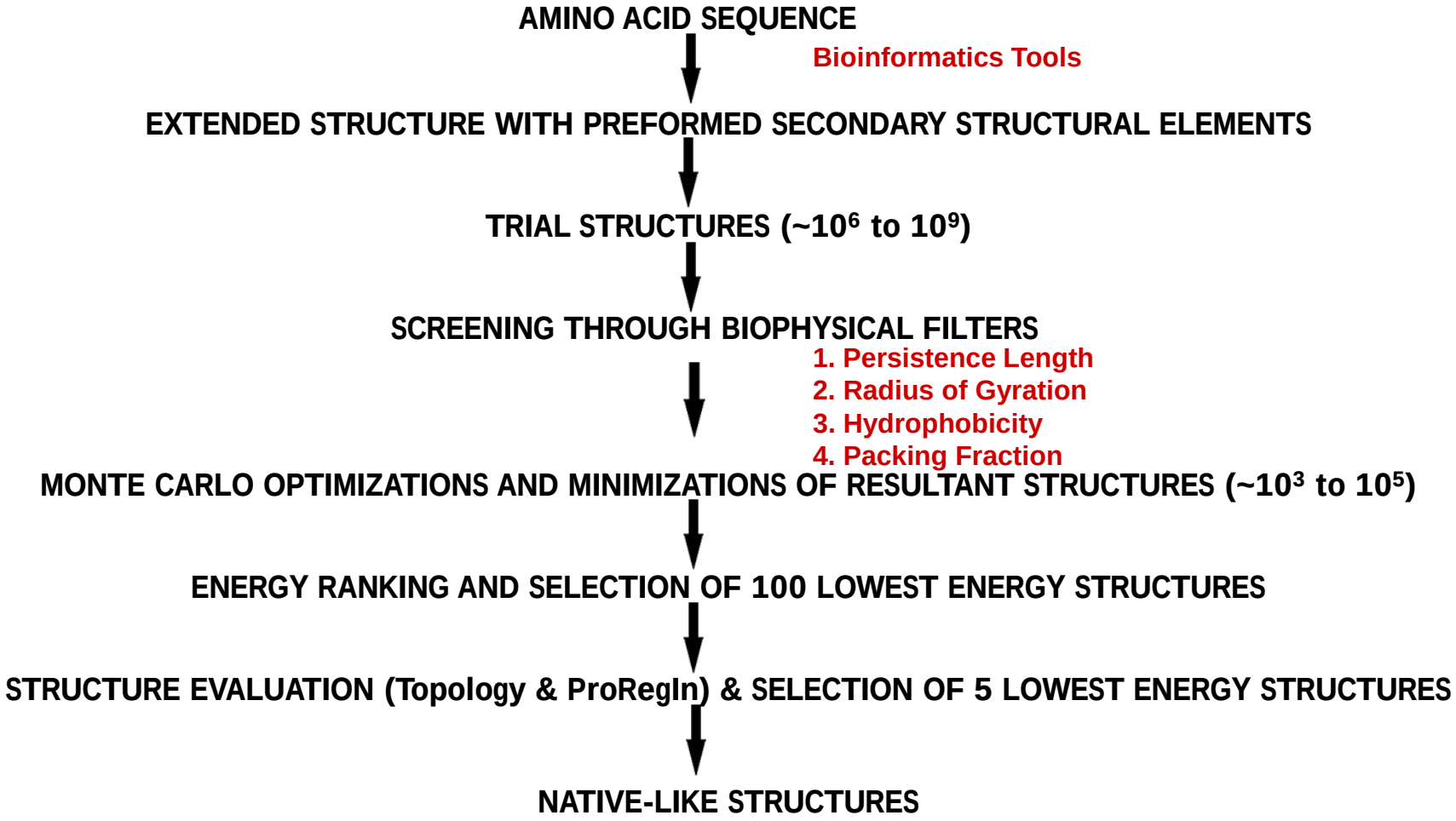
With million processors ~ 1 year

Anton machine is making 'Strategy B' viable for small proteins: David E. Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O. Dror, Michael P. Eastwood, Joseph A. Bank, John M. Jumper, John K. Salmon, Yibing Shan, and Willy Wriggers, "Atomic-Level Characterization of the Structural Dynamics of Proteins," *Science*, vol. 330, no. 6002, 2010, pp. 341–346.

Some online software tools available for protein tertiary structure prediction

Sl. No	Softwares	URLs	Description
1	CPHModels3.0	http://www.cbs.dtu.dk/services/CPHmodels/	Protein homology modeling server
2	SWISS-MODEL	http://swissmodel.expasy.org/SWISS-MODEL.html	A fully automated protein structure homology-modeling server
3	Modeller	http://salilab.org/modeller/	Program for protein structure modeling by satisfaction of spatial restraints
4	3D-JIGSAW	http://3djigsaw.com/	Server to build three-dimensional models for proteins based on homologues of known structure
5	PSIPRED	http://bioinf.cs.ucl.ac.uk/psipred/	A combination of methods such as sequence alignment with structure based scoring functions and neural network based jury system to calculate final score for the alignment
6	3D-PSSM	http://www.sbg.bio.ic.ac.uk/~3dpssm/index2.html	Threading approach using 1D and 3D profiles coupled with secondary structure and solvation potential
7	ROBETTA	http://robetta.bakerlab.org	*De novo* Automated structure prediction analysis tool used to infer protein structural information from protein sequence data
8	PROTINFO	http://protinfo.compbio.washington.edu/	*De novo* protein structure prediction web server utilizing simulated annealing for generation and different scoring functions for selection of final five conformers
9	SCRATCH	http://scratch.proteomics.ics.uci.edu/	Protein structure and structural features prediction server which utilizes recursive neural networks, evolutionary information, fragment libraries and energy
10	I-TASSER	http://zhanglab.ccmb.med.umich.edu/I-TASSER/	Predicts protein 3D structures based on threading approach
11	BHAGEERATH	http://www.scfbio-iitd.res.in/bhageerath/index.jsp	Energy based methodology for narrowing down the search space of small globular proteins
12	BHAGEERATH-H	http://www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp	A Homology *ab-initio* Hybrid Web server for Protein Tertiary Structure Prediction

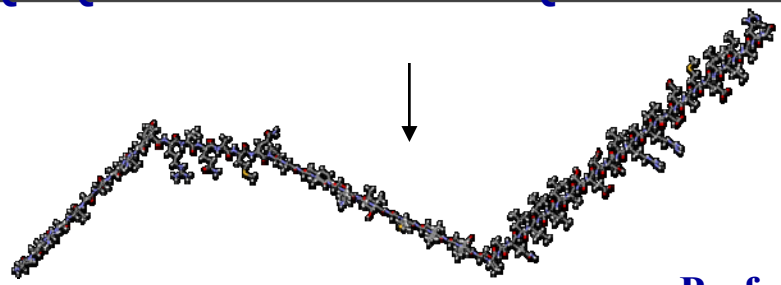
From Sequence to Structure: The *Bhageerath* Pathway



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.

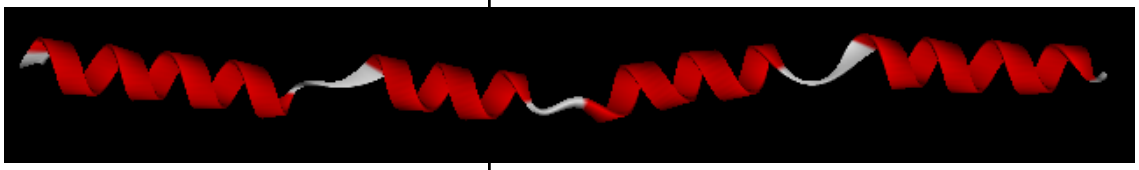
Sampling 3D Space

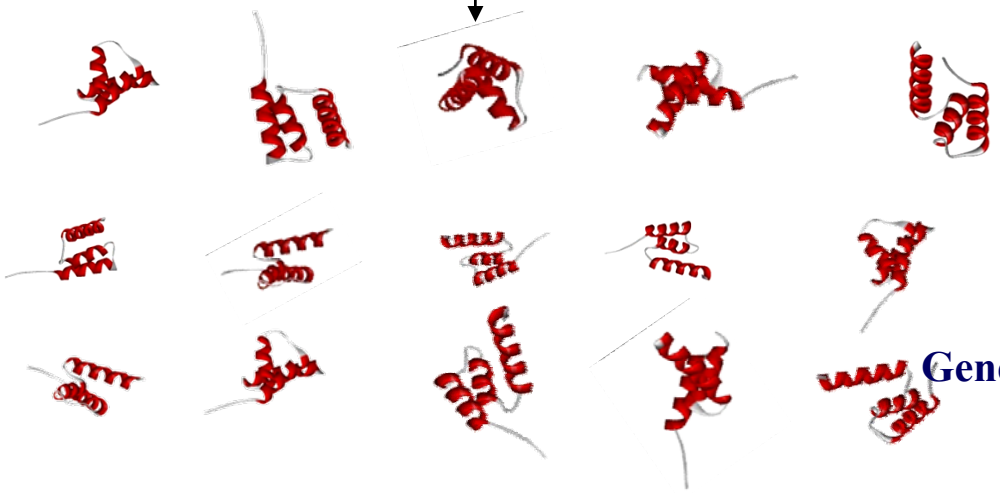
HRQALGERLYPRVQAMQPAFASKITGMLLELSPAQLLLLLASENSLRARVNEAMELIIAHG



Extended Chain

Preformed Secondary Structural Units

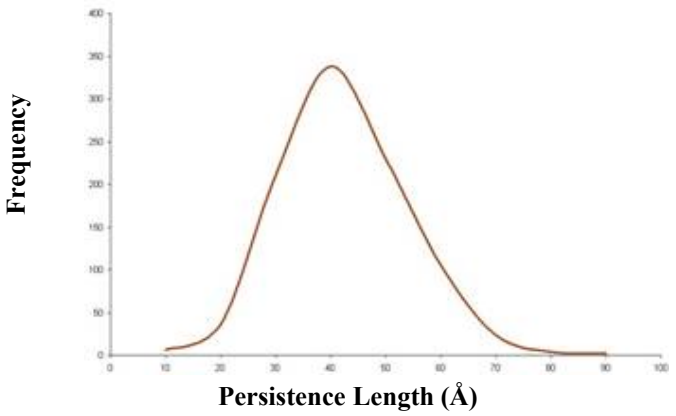




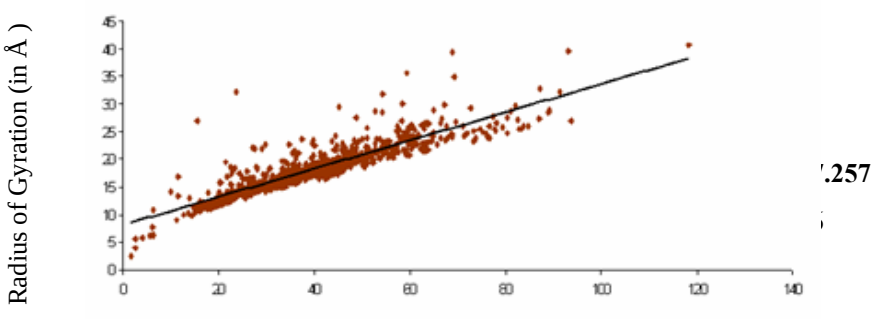
Generation of Trial Structures

Filter-Based Structure Selection

Persistence Length Analysis of 1,000 Globular Proteins



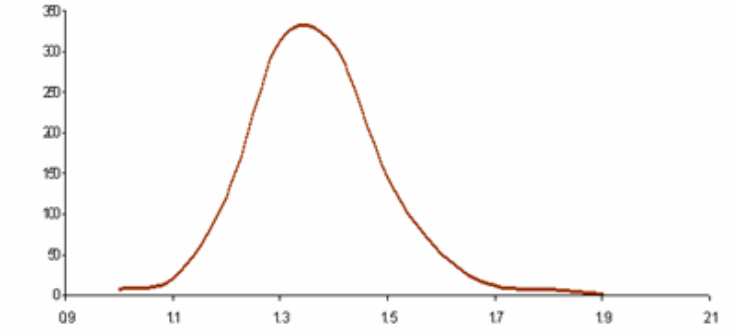
Radius of Gyration vs $N^{3/5}$ of 1,000 Globular Proteins



$N^{3/5}$ (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

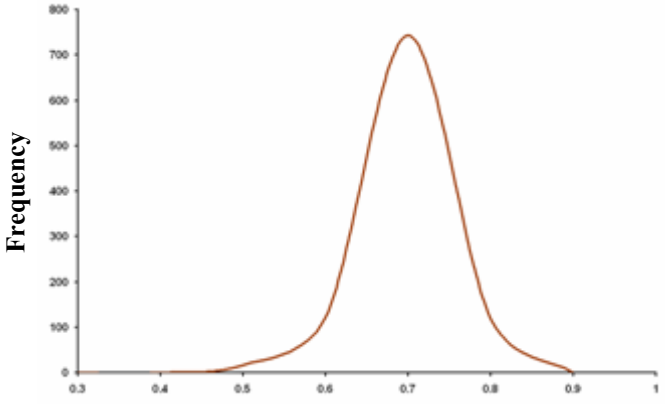


Hydrophobicity Ratio (Φ_H)

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

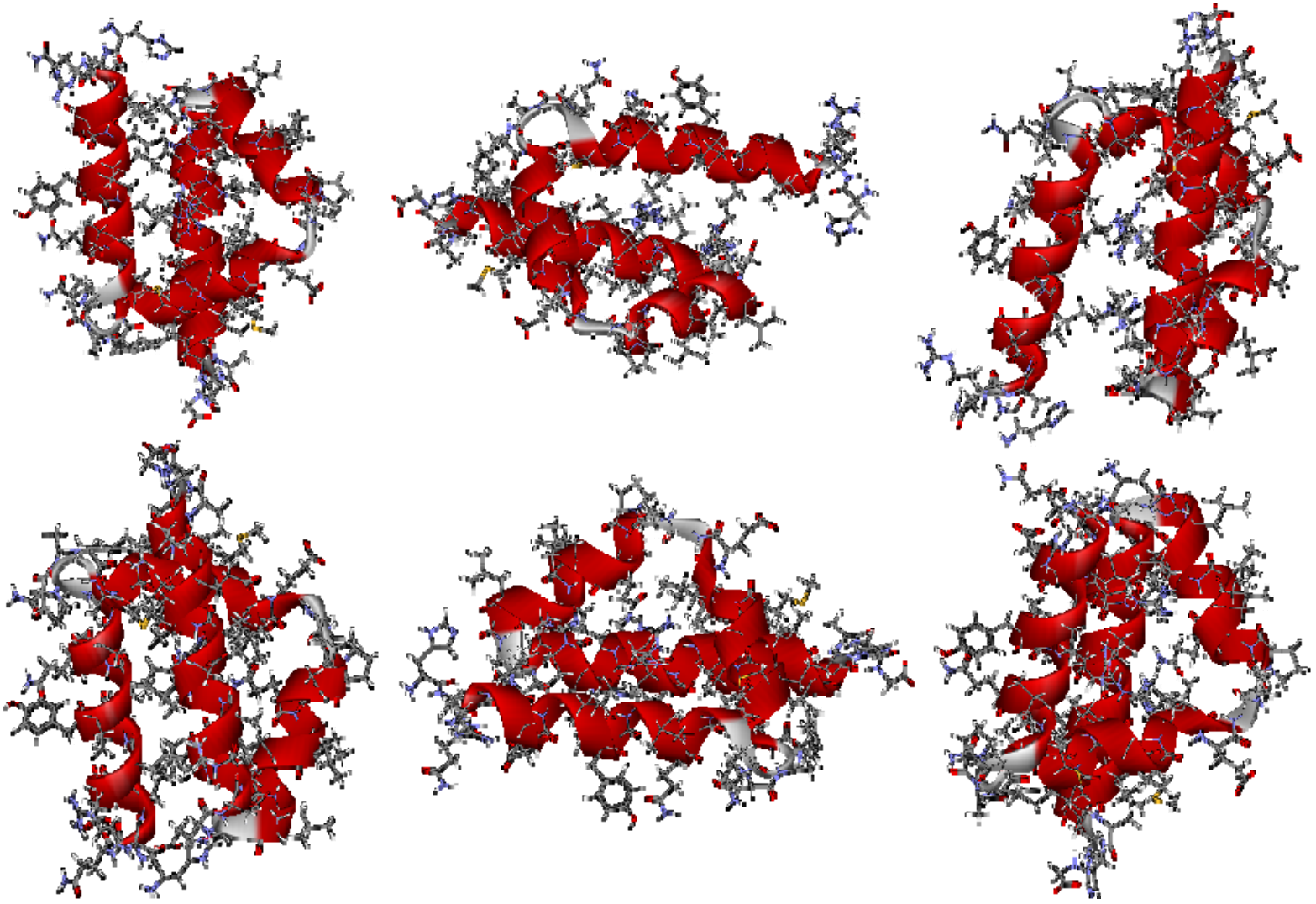
ASA : Accessible surface area

Frequency vs Packing Fraction of 1,000 Globular Proteins



Globular proteins are known to exhibit packing fractions around 0.7

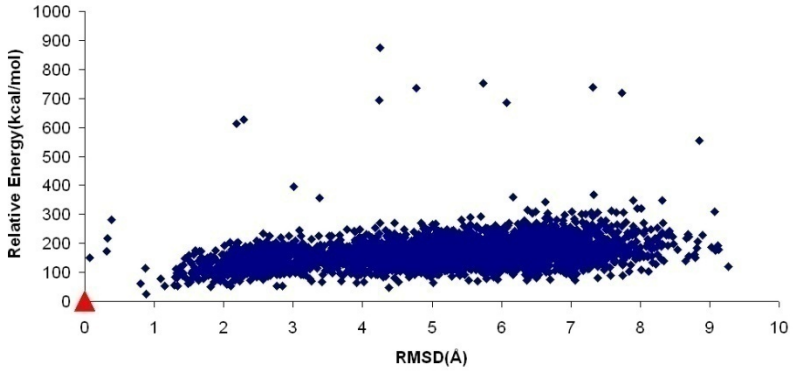
Removal of Steric Clashes in Selected Structures (Distance Based Monte Carlo)



Validation of Empirical Energy Based Scoring Function

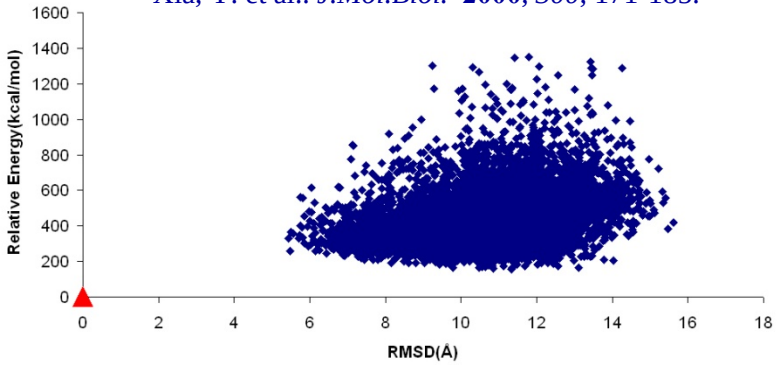
Four-state reduced decoy set

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



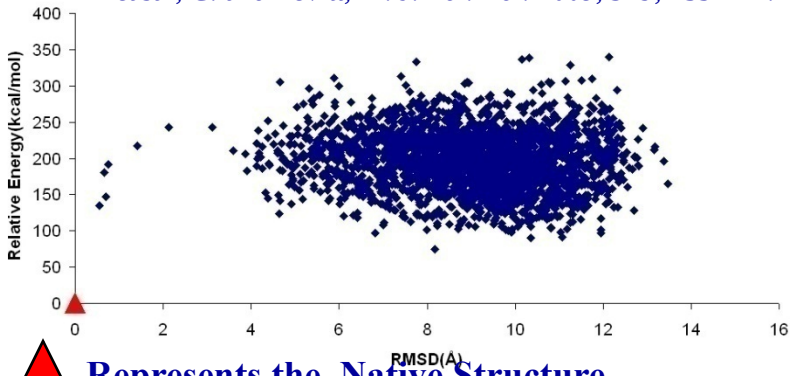
Lattice_ssfit decoy set

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



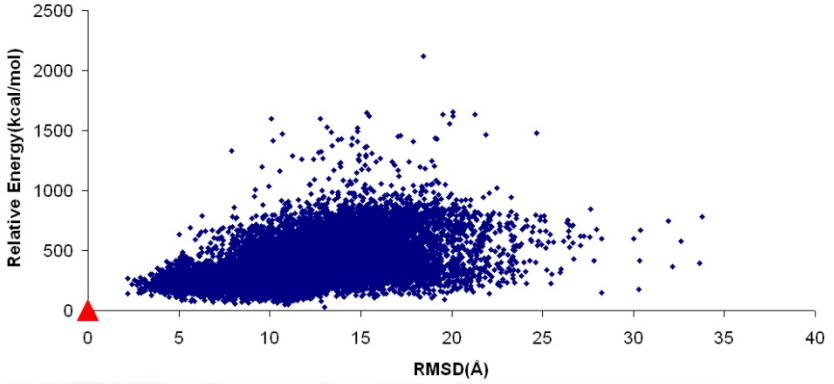
Lmds decoy set

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



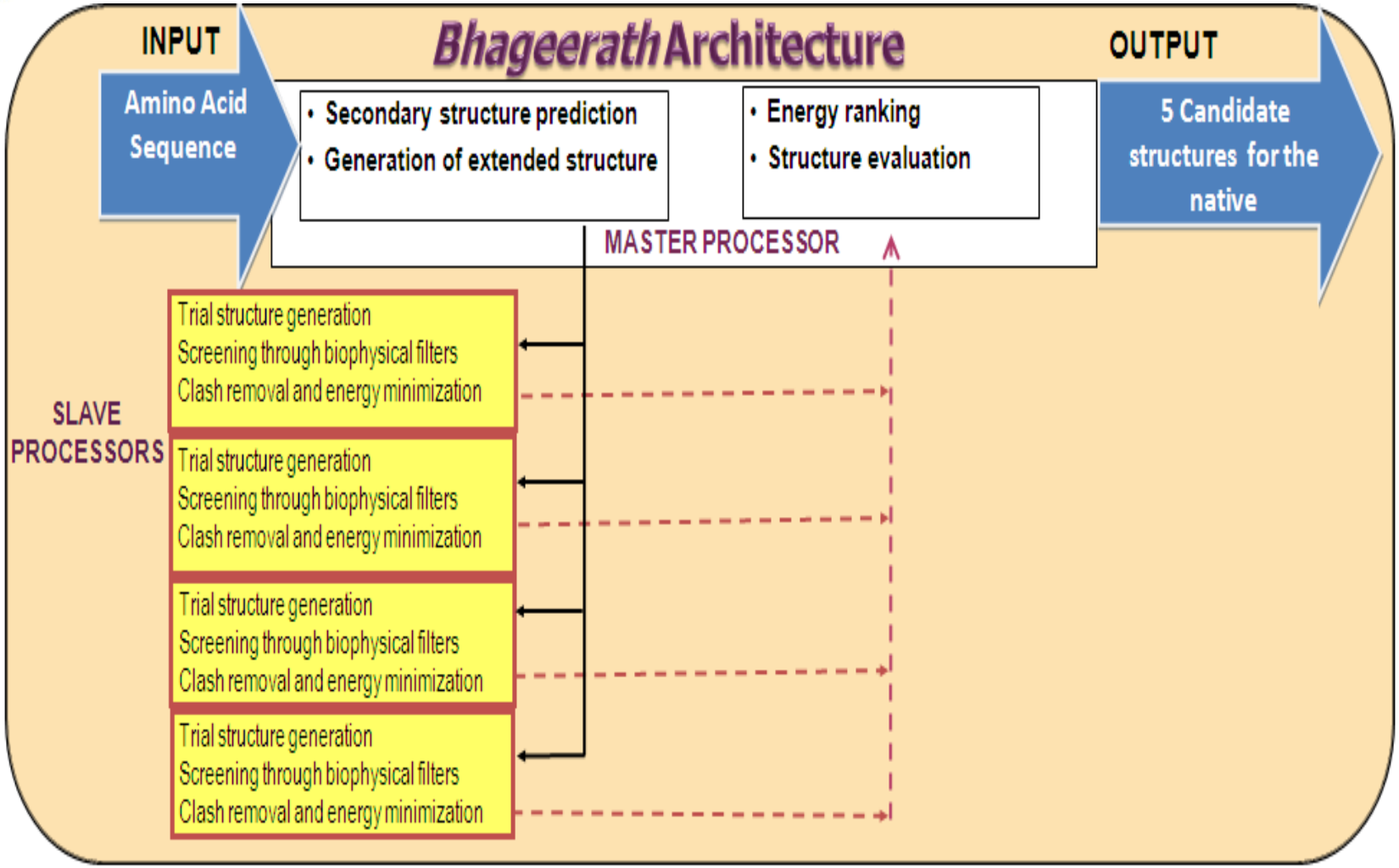
Rosetta decoy set

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 teraflop) cluster**

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

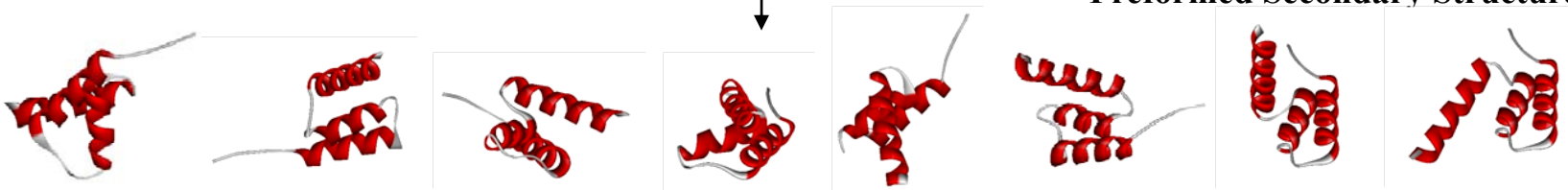
A Case Study of Mouse C-Myb

DNA Binding (52 AA)

LIKGPWTKEEDQRVIELVQKYGPKRWSVIAKHLKGRIGKQCRERWHNHLNPE

Sequence

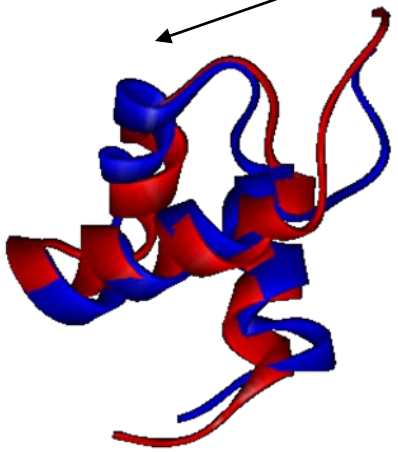
Preformed Secondary Structure



16384 Trial Structures

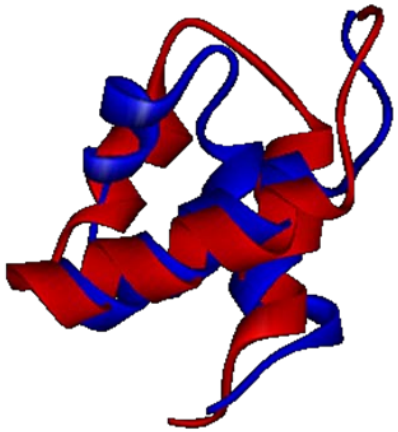
Biophysical Filters & Clash Removal

10632 Structures



RMSD=2.87, Energy Rank=1774

Energy Scans



RMSD=4.0 Ang, Energy Rank=4

Blue: Native; Red: Predicted

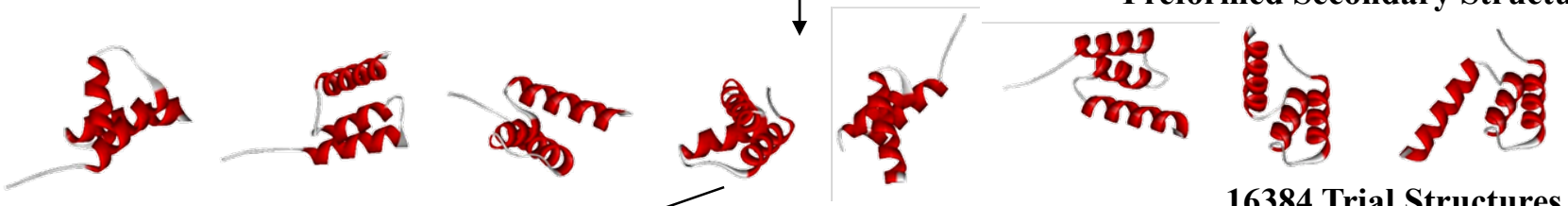
A Case Study of *S.aureus* Protein A

Immunoglobulin Binding (60 AA)

RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKS

Sequence

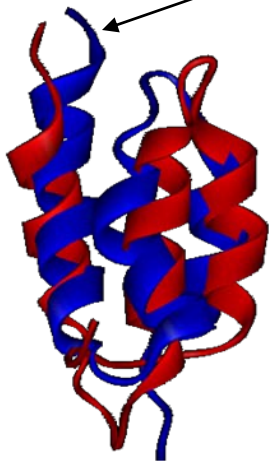
Preformed Secondary Structure



16384 Trial Structures

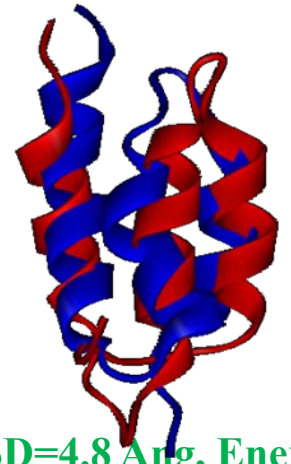
Biophysical Filters & Clash Removal

11255 Structures



RMSD=4.2, Energy Rank=44

Energy Scans



RMSD=4.8 Ang, Energy Rank=5

Blue: Native; Red: Predicted

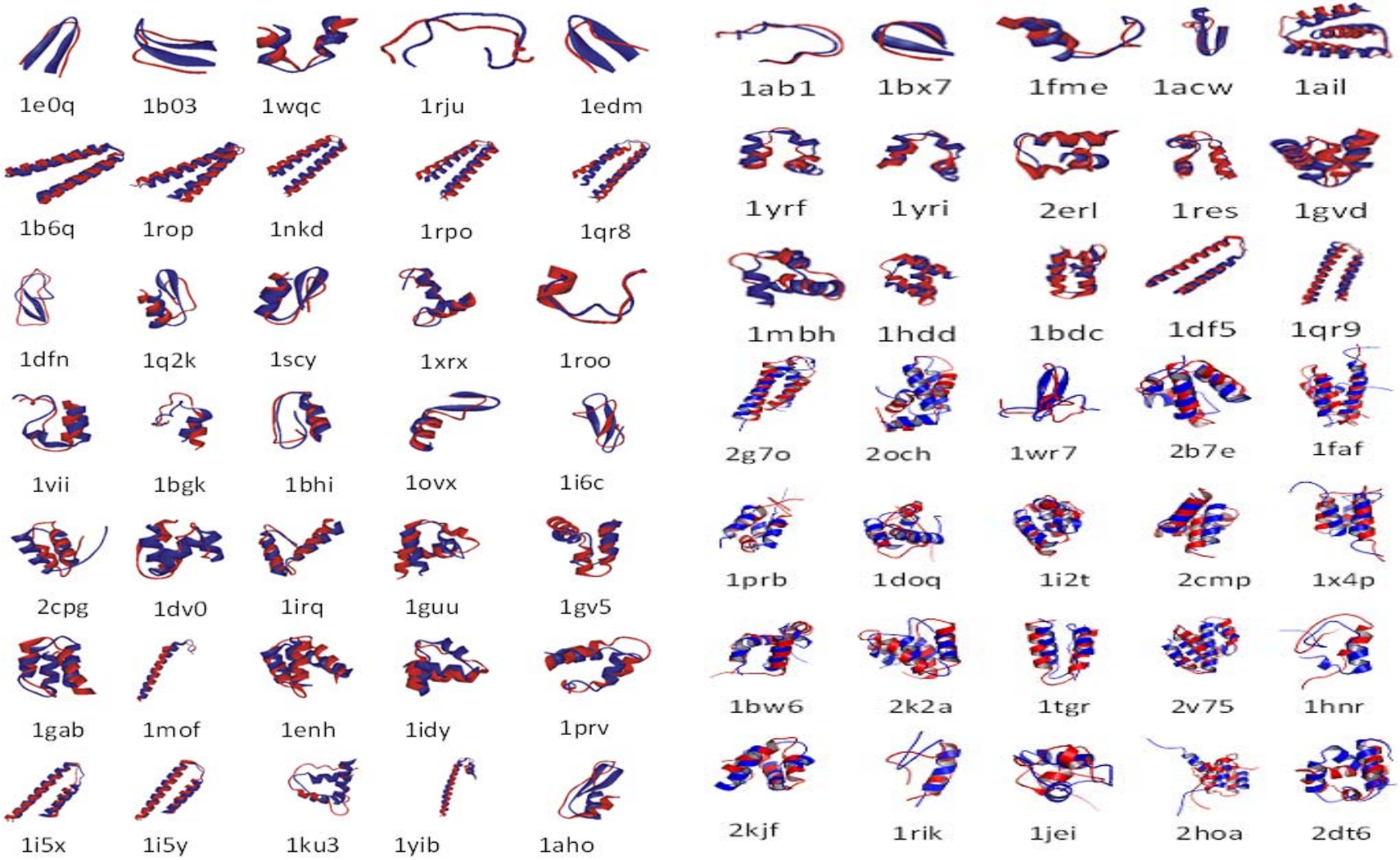
Performance of *Bhageerath* on 70 Small Globular Proteins

S.No.	PDBID	No of Amino Acids	No. of Secondary Structure elements	Lowest RMSD Å	Energy rank of lowest 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 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 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



Native structure Predicted structure

Jayaram et al., *Nucl. Acids Res.*, 2006, *34*, 6195-6204.

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 methods are simply superb where the similarities between the query sequence and a template in the protein structural database are high. Where there is no match/similarity, ab initio / de novo methods such as Bhageerath are the only option.

Bhageerath vs other servers for Template free prediction in CASP9 (2010)

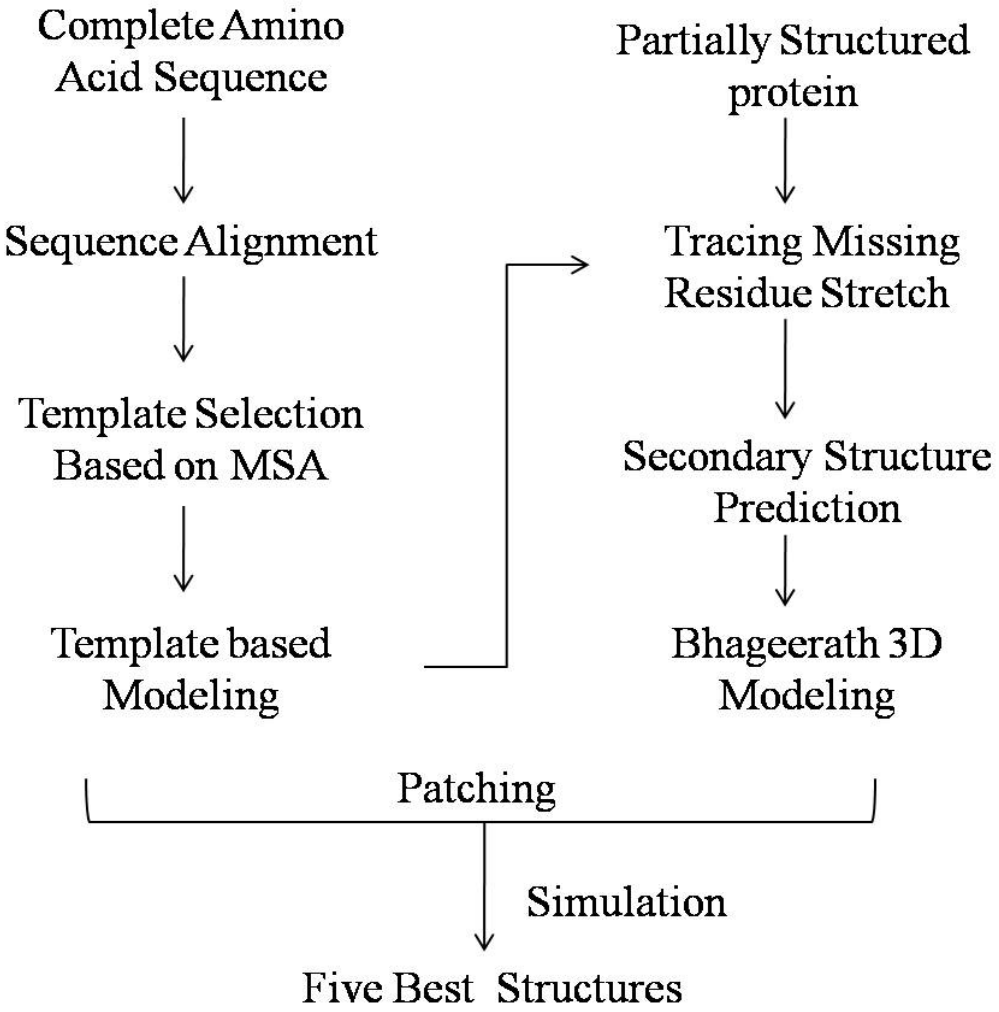
Target No.	No.of residues	PDBID	Bhageerath RMSD Å	TASSER RMSD Å	ROBETTA RMSD Å	SAM-T08 RMSD Å
T0531	65	2KJX	7.1	11.0	11.9	12.6
T0553	141	2KY4	9.6	6.0	11.5	8.6
T0581	136	3NPD	15.8	11.6	5.3	15.1
T0578	164	3NAT	19.2	11.6	15.5	19.1

While *Bhageerath* works well for small proteins (< 100 AAs), improvements are necessary to tackle larger proteins

Development of a homology / ab initio hybrid server

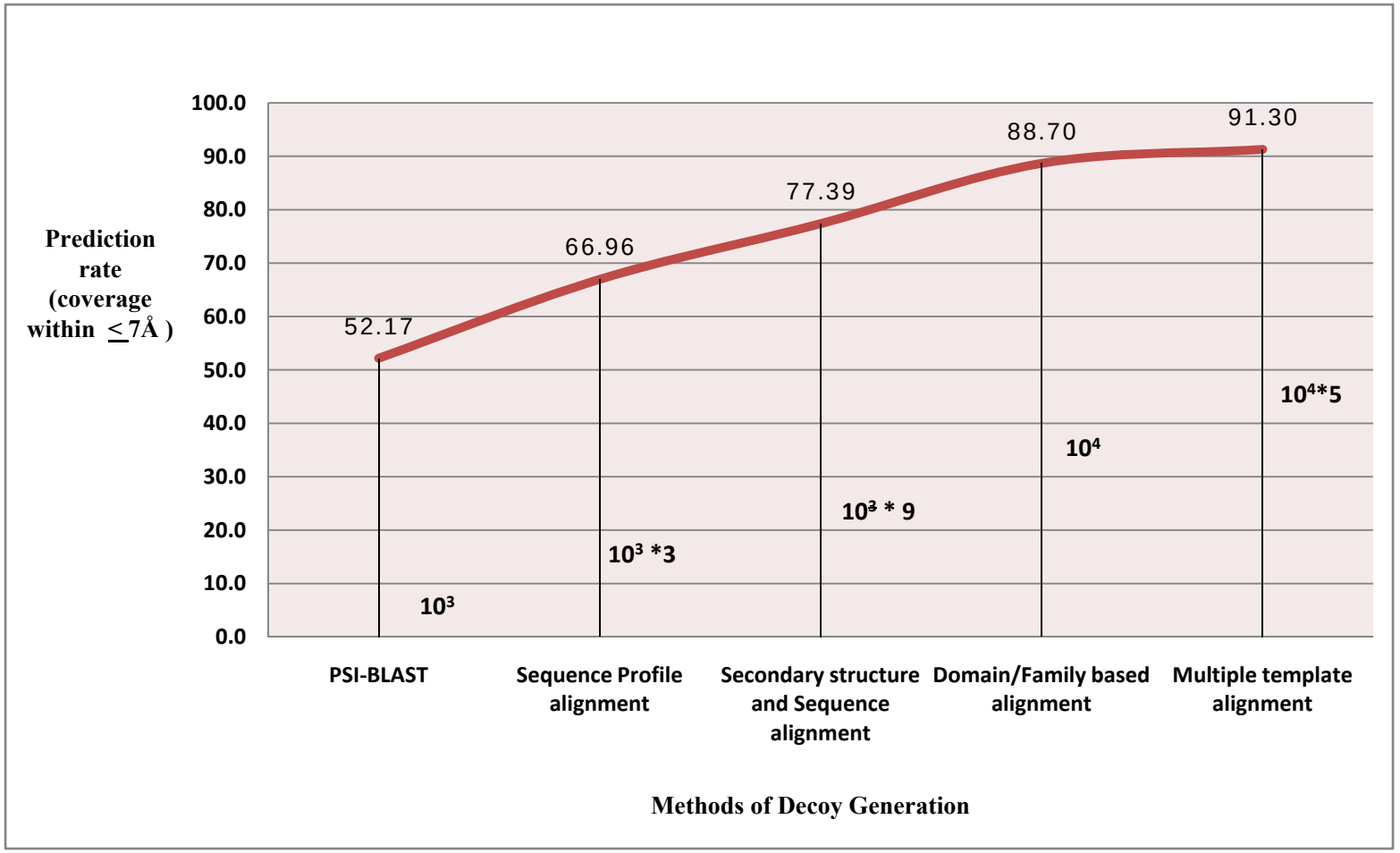
Bhageerath-H Protocol

Overall strategy:
(1) Generate several plausible candidate structures by a mix of methods &
(2) Score them to realize near-native structures



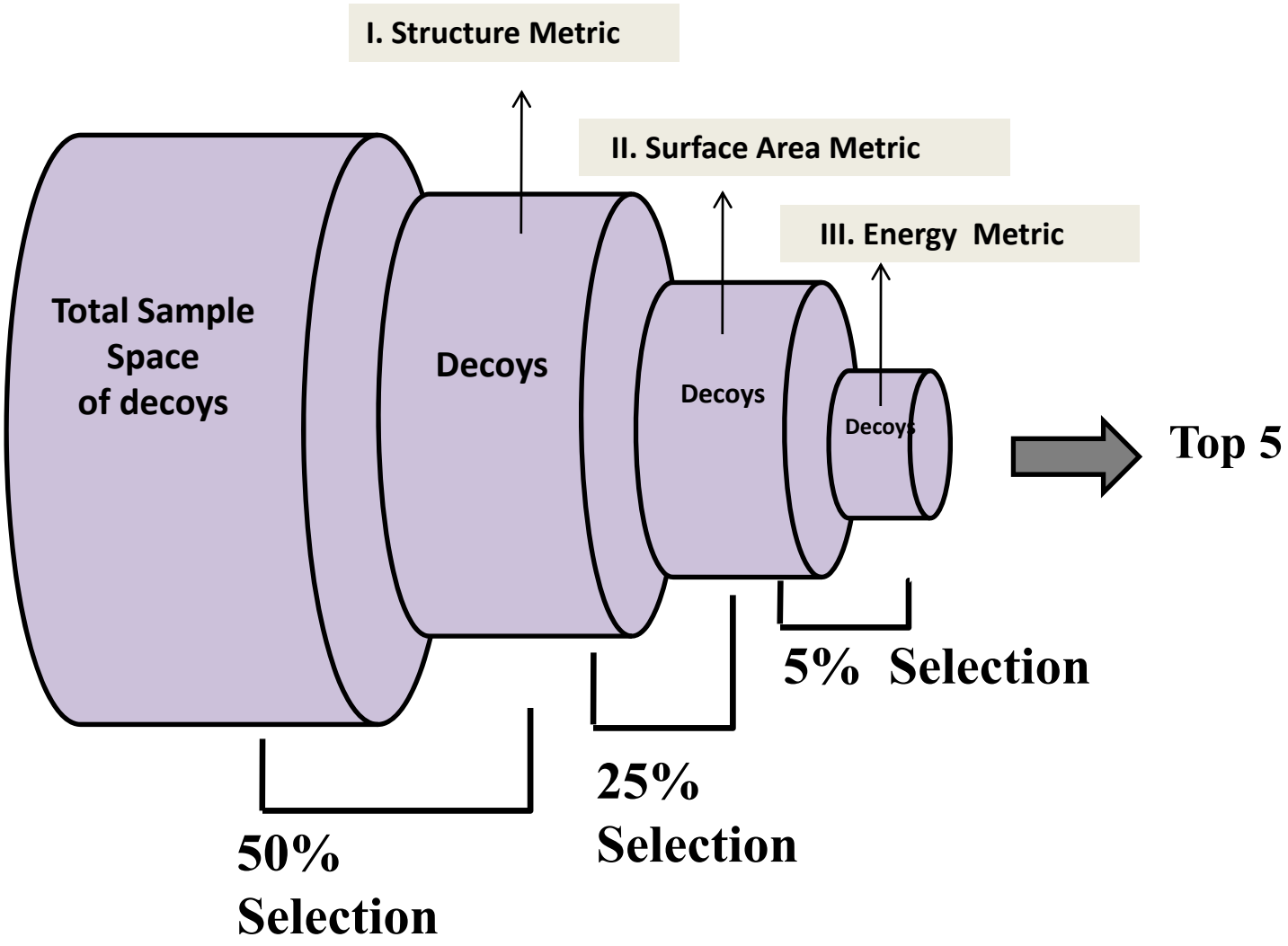
B. Jayaram, Priyanka Dhingra, Baharat Lakhani, Shashank Shekhar, "*Bhageerath*: Attempting the Near Impossible – Pushing the Frontiers of Atomic Models for Protein Tertiary Structure Prediction", *J Chemical Sciences*, 2012, *124* (1), 83-91.

Sampling near native conformations with *BHAGEERATH*-H: A hybrid software for protein tertiary structure prediction



Total number of targets fielded in CASP 9 : 115 (excluding the cancelled targets);
Number of targets with decoys within 7Å rmsd from native : 105

“Deployment” of a Structural Metric for Capturing Native



Who is the Native ?

Decoy I

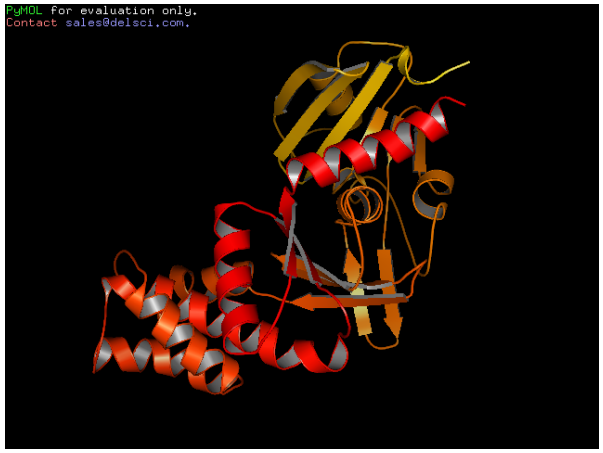
M = 4869.25

A = 0.290

E = -1.46

RMSD with Native = 0

Decoy II



M = 4875.75

A = 0.314

E = - 1.34

RMSD with Native = 1.03

Decoy III

M = 5077.55

A = 0.397

E = - 1.20

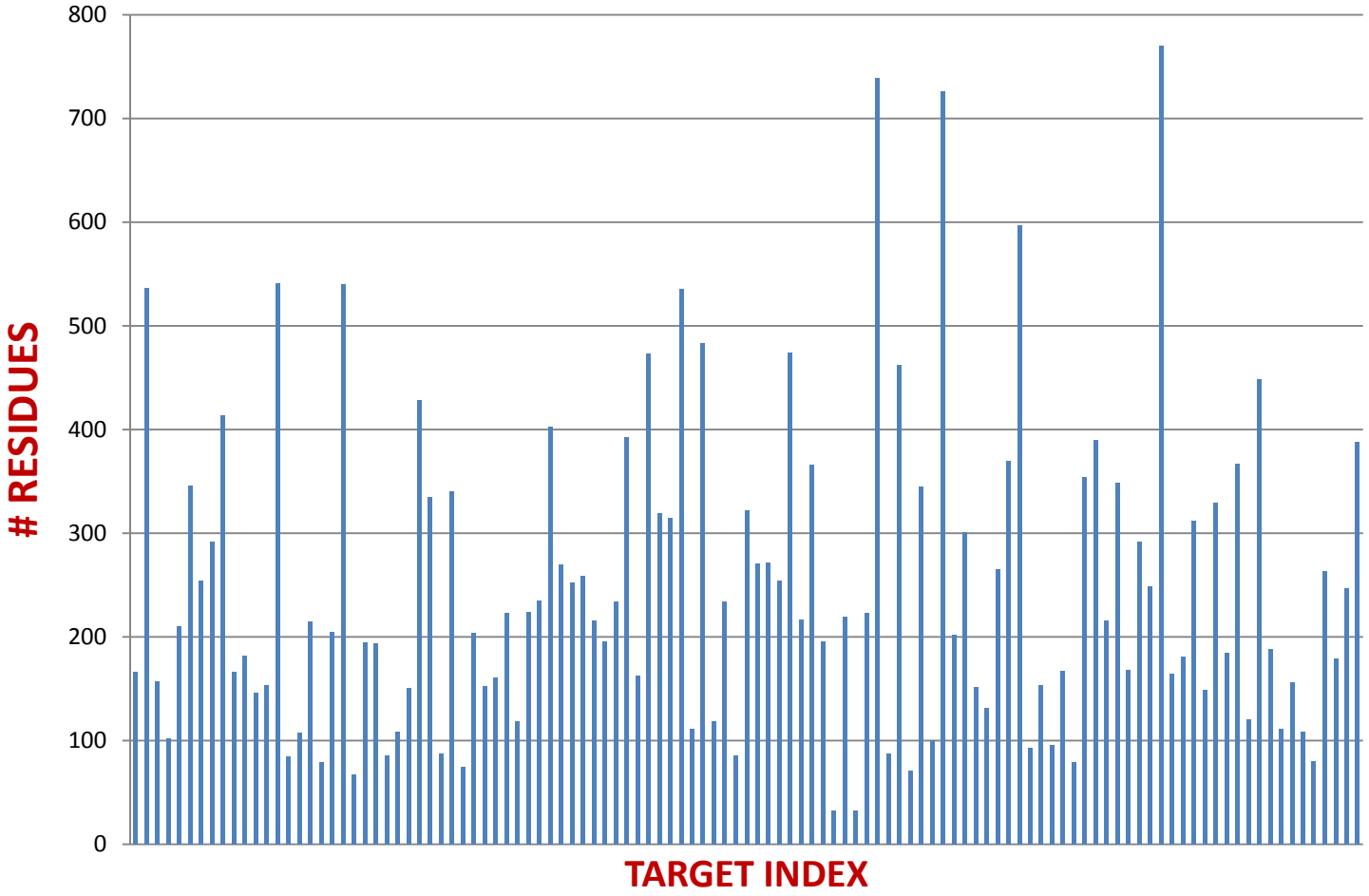
RMSD with Native = 9.14

M,A,E are the least for I. So, Decoy I is the Native

! 3NUW : 295 aa

Protein Tertiary Structure Prediction : CASP10 Experiment

(May 1st to July 17th, 2012: 113 Targets)



Minimum Target Length=33, Maximum Target Length=770

Bhageerath-H CASP10 Performance

- 58 Natives Released in PDB as of Dec., 2012 for Valid Targets
- All C-alpha RMSD comparison
- Server predicted models with lesser number of residues compared to released sequence length by CASP are discarded

Rank	Server	< 6 Å
1	QUARK	30
2	Zhang-Server	29
3	TASSER-VMT	28
4	BAKER-ROSETTASERVER	27
5	Pcons-net	26
6	Distill	26
7	PMS	25
8	PconsM	25
9	MULTICOM-REFINE	25
10	Distill_roll	25
11	chunk-TASSER	25
12	BhageerathH	24
13	RaptorX	24
14	ZHOU-SPARKS-X	24
15	STRINGS	24
16	Seok-server	24

Rank	Server	<6 Å
17	MATRIX	24
18	Jiang_Server	24
19	chuo-repack	24
20	chuo-fams-server	24
21	Bilab-ENABLE	24
22	RaptorX-ZY	24
23	slbio	23
24	Phyre2_A	23
25	MULTICOM-NOVEL	23
26	MULTICOM-CONSTRUCT	23
27	MUFOLD-Server	23
28	IntFOLD	23
29	NewSerf	22
30	MULTICOM-CLUSTER	22
31	Mufold-MD	22
32	IntFOLD2	22

Rank	Server	<6 Å
33	FALCON-TOPO	15
34	FALCON-TOPO-X	14
35	Atome2_CBS	13
36	MUFold_CRF	10
37	GSmetaserver	8
38	FFAS03	6
39	FFAS03mt	5
40	sysimm	3
41	RBO-MBS	2
42	RBO-MBS-BB	2
43	FFAS03hj	2
44	FFAS03c	2
45	3D-JIGSAW_V5-0	2
46	RBO-i-MBS	1
47	RBO-i-MBS-BB	1
48	HOMER	1

Rank	Server	<6 Å
49	HHpredA	22
50	HHpredAQ	22
51	FRESS_server	22
52	Jiang_Threader	21
53	HHpred-thread	21
54	PconsD	20
55	Jiang_Fold	20
56	hGen3D	20
57	SAM-T08-server	19
58	PROTAGORAS	19
59	AOBA-server	19
60	YASARA	18
61	samcha-server	17
62	SAM-T06-server	16
63	UGACSBL	15
64	panther	15

Rank	Server	<6Å
65	RaptorX-Roll	0
66	Pcons-net	0
67	Lenserver	0
68	FALCON-TOPO-X	0
69	confuzzGS	0
70	confuzz3d	0

Expectation: More, preferably all, predicted structures under < 3 Ang.

Homology / *ab initio* hybrid methods are getting better with every passing year.

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.

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

The user inputs the amino acid sequence & five candidate structures for the native are emailed back to the user

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.

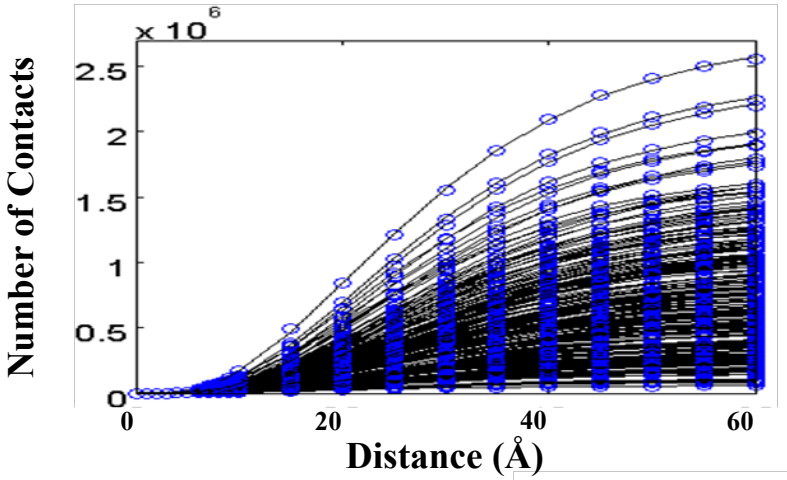
Amino Acid	Protein sequences confirmed by annotation and experiments (mean ± std, n = 131855)
A	7.2 ± 3.0
V	6.3 ± 2.1
I	5.1 ± 2.2
L	9.6 ± 2.9
Y	3.0 ± 1.5
F	3.9 ± 1.8
W	1.2 ± 0.9
P	5.4 ± 2.6
M	2.2 ± 1.3
C	1.9 ± 2.3
T	5.5 ± 1.8
S	7.9 ± 2.8
Q	4.3 ± 2.0
N	4.2 ± 1.9
D	5.2 ± 1.9
E	6.8 ± 2.8
H	2.4 ± 1.3
R	5.3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

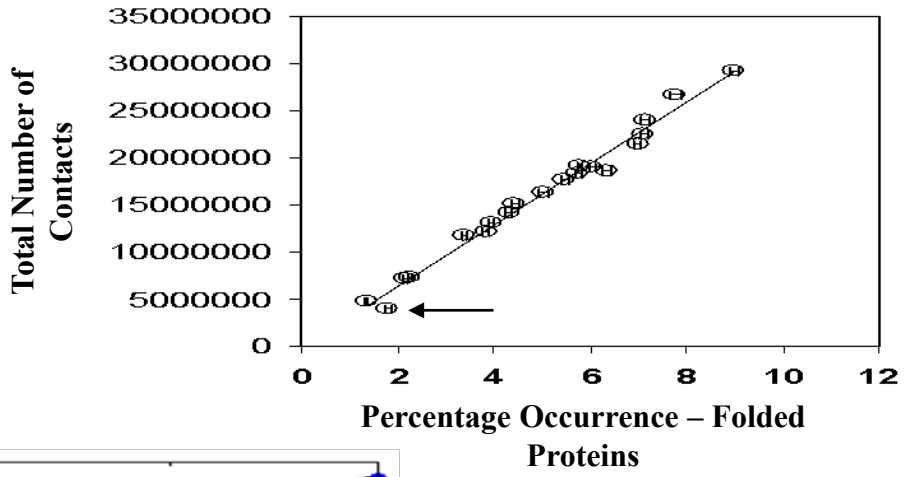
The average percentage occurrence of each amino acid, their STD as observed and as calculated from the binomial distribution.

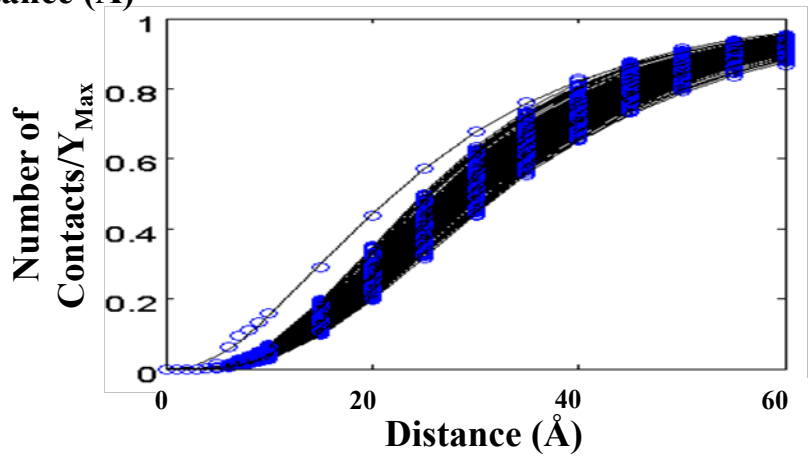
	P (%)	STD (observed)	STD (random)
A	7.8	3.4	7.2
V	7.1	2.4	6.6
I	5.8	2.4	5.5
L	9.0	2.9	8.2
Y	3.4	1.7	3.3
F	3.9	1.8	3.7
W	1.3	1.0	1.3
P	4.4	2.0	4.2
M	2.2	1.3	2.2
C	1.8	1.5	1.8
T	5.5	2.4	5.2
S	6.0	2.5	5.6
Q	3.8	2.0	3.7
N	4.3	2.2	4.1
D	5.8	2.0	5.5
E	7.0	2.7	6.5
H	2.3	1.4	2.2
R	5.0	2.3	4.8
K	6.3	2.8	5.9
G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

Mittal et al. JBSD, 2010 & 2011 & Mezei, JBSD, 2011

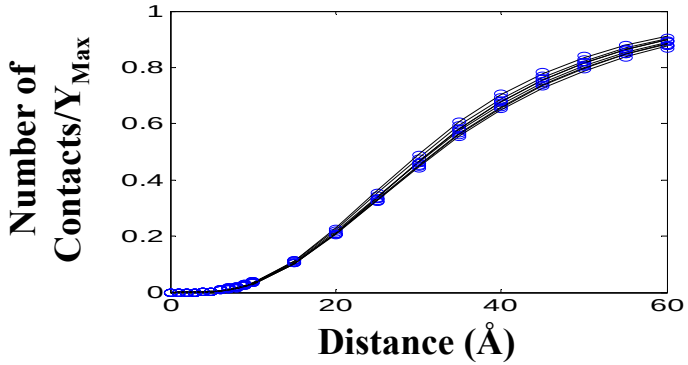
In search of rules of protein folding: Cα spatial distributions show universality

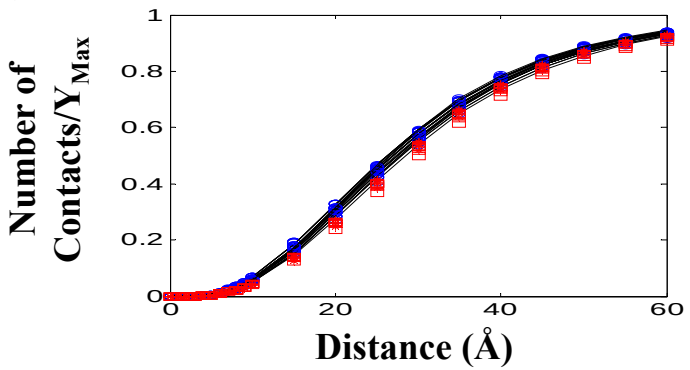






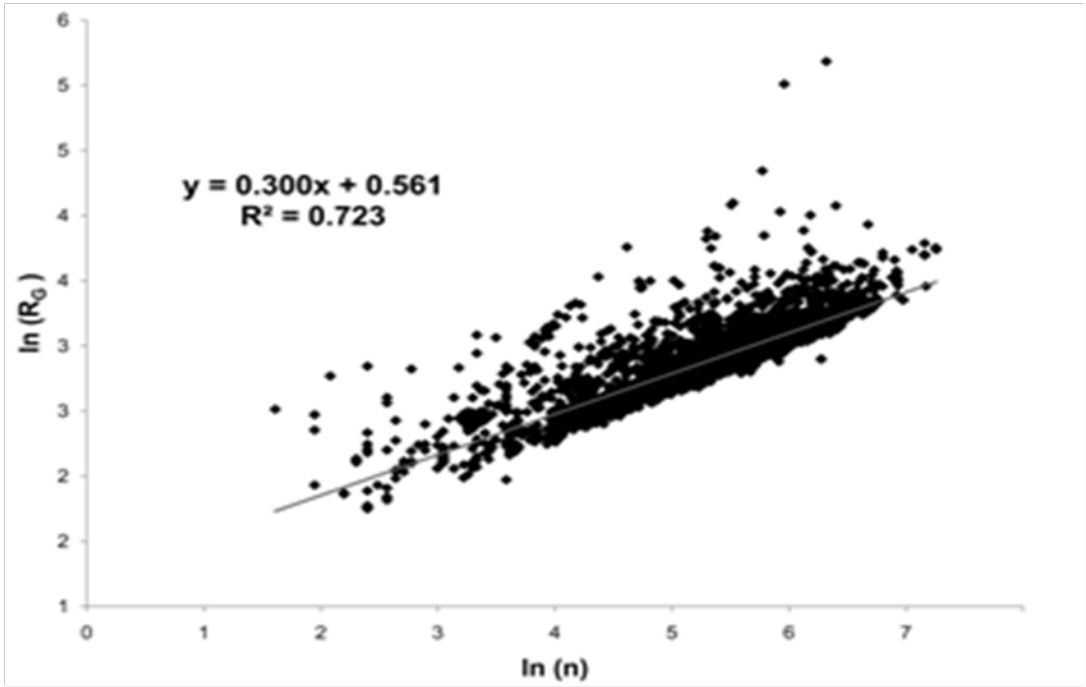
$$Y = Y_{Max}(1-e^{-kX})^n$$





A. Mittal, B. Jayaram et al. *J. Biomol. Struc. Dyn.*, 2010, *Vol. 28 (2)*, 133-142;2011, *28(4)*, 443 -454; 2011, *28(4)*, 669-674.

Size



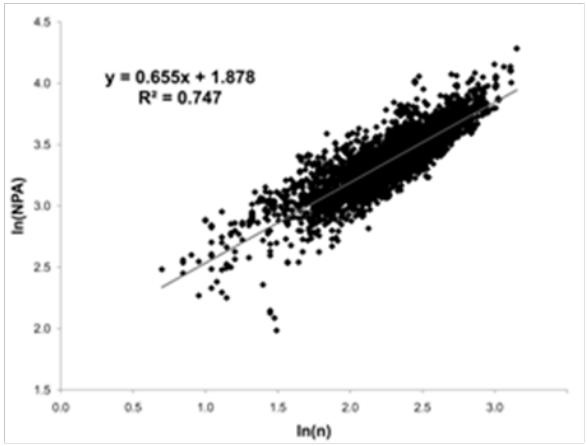
Radius of gyration plotted against number of residues as a log-log plot for ~ 6750 proteins. Proteins are seen to be extremely compact compared to random chains and synthetic polymers in good solvents. In the parlance of Flory, water is not a "good solvent" for proteins.

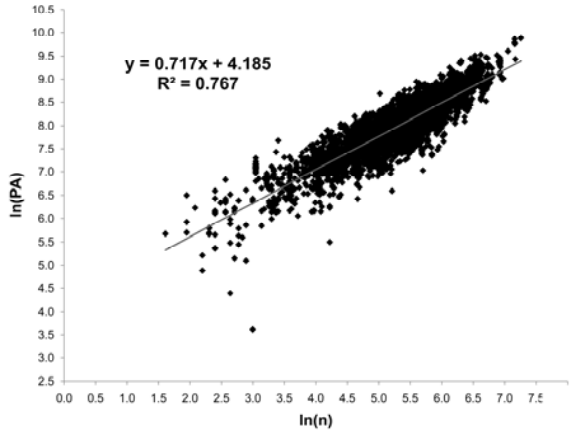
B. Jayaram, Aditya Mittal, Avinash Mishra, Chanchal Acharya, Garima Khandelwal "Universalities in Protein Tertiary Structures: Some New Concepts", in *Biomolecular Forms and Functions*, 2013, World Scientific Publishing Co. Pte. Ltd., Singapore, Eds; Manju Bansal & N. Srinivasan, pp 210-219.

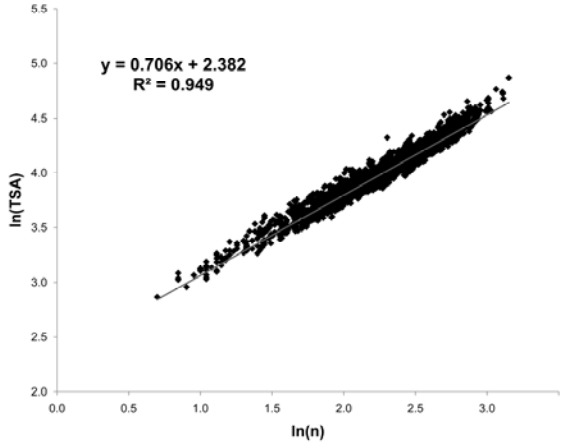
Area

Solvent accessible surface areas Nonpolar (top panel), polar (middle panel), total (bottom panel) versus number of residues (n) in ~6750 proteins shown as log-log plots.

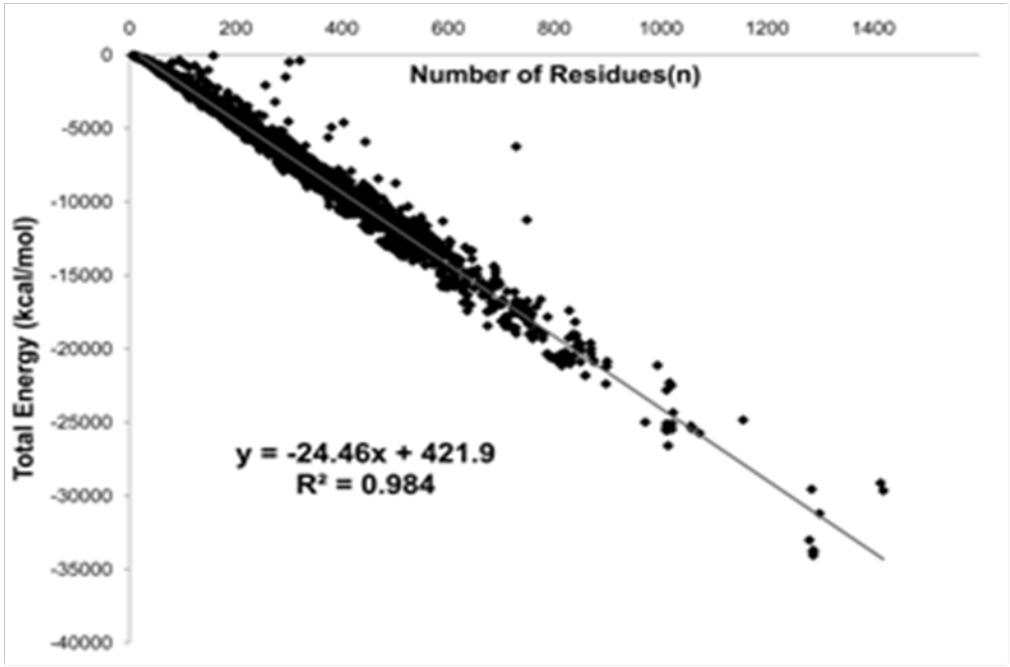
An invariant area/ residue metric appears to exist.







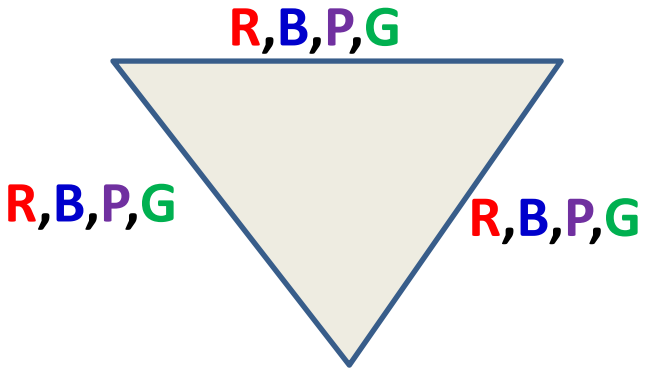
Energy



Total energy of 6750 proteins shown as a function of number of residues

An invariant energy/residue metric appears to exist.

Reexamining the language of amino acids



64 coloured triangles are possible. By virtue of the symmetries of the triangle, only 20 of these are unique.

(1) RRR	(5) BBR	(9) PPR	(13) GGR	(17) RBG
(2) RRB	(6) BBB	(10) PPB	(14) GGB	(18) RBP
(3) RRP	(7) BBP	(11) PPP	(15) GGP	(19) RPG
(4) RRG	(8) BBG	(12) PPG	(16) GGG	(20) BGP

Some observations

I. *Any color occurs in exactly 10 triangles*

R (1,2,3,4,5,9,13,17,18,19); B (2,5,6,7,8,10,14,17,18,20);
P (3,7,9,10,11,12,15,18,19,20); G (4,8,12,13,14,15,16,17,19,20)

II. *Any two distinct colors occur together in 4 triangles*

R & B (2,5,17,18); R & P (3,9,18,19); R & G (4,13,17,19)
B & P (7,10,18,20); B & G (8,14,17,20) ; P & G (12,15,19,20)

III. *Any three distinct colors occur together in only one triangle*

R, B & G (17); R, B & P (18); R, P & G (19); B, P &G (20)

IV. *All sides with same color occurs only once*

R (1); B (6); P (11); G (16)

Rule 1. **Amino acid side chains have evolved based on four chemical properties. A minimum of one and a maximum of three properties are used to specify each amino acid.**

Rule 2. **Each property occurs in exactly 10 amino acids.**

Rule 3. **Any two properties occur simultaneously in only four amino acids.**

Rule 4. **Any three properties occur simultaneously in only one amino acid.**

Rule 5. **Amino acids characterized by a single property occur only once.**

Text book classifications do not satisfy the above rules!
Either the above rules are irrelevant to amino acids or
we need to revise our understanding of the language of proteins.

Property (I): Presence of sp^3 hybridized γ carbon atom. (a) Exactly 10 amino acids {E, I, K, L, M, P, Q, R, T, V} possess this property as required by Rule 2 above.

Property (II): Hydrogen bond donor ability. (a) Exactly 10 amino acids {C, H, K, N, Q, R, S, T, W, Y} possess this property. (b) Also, only four amino acids (K, Q, R, T) exhibit both properties (I & II together) as required by Rule 3.

Property (III): Absence of δ carbon. (a) Exactly 10 amino acids {A, C, D, G, I, M, N, S, T, V} have this property. Ile is included in this set as one of the branches of its side chain is lacking in a δ carbon. (b) I and III occur simultaneously in only four amino acids (I, M, T, V) and similarly II and III occur simultaneously in only four amino acids (C, N, S, T). (c) Rule 4 requires that the above three properties (I, II and III) occur simultaneously in only one amino acid (T) and this conforms to the expectation.

The most likely candidate for property (IV): Absence of branching. Linearity of the side chains / non-occurrence of bidentate forks with terminal hydrogens in the side chains. (a) This pools together 10 amino acids in the set {A, D, E, F, H, K, M, P, S, Y}. Side chains with single rings are treated as without forks. The sulfhydryl group in Cys and its ability to form disulfide bridges requires it to be treated as forked. Accepting that this property (IV) satisfies Rule 2, (b) Rule 3 is satisfied by I and IV (E, K, M, P); by II and IV (H, K, S, Y) and by III and IV (A, D, M, S). (c) Also, Rule 4 is satisfied by I, II and IV (K), by I, III and IV (M) and by II, III and IV (S).

With all the four properties (I, II, III and IV) specified, amino acids characterized by a single property occur only once: property I (L), property II (W), property III (G) and property IV (F), consistent with Rule 5.

The 20 amino acids and some stereochemical properties of their side chains.

Amino acid	I. Presence of sp^3 hybridized γ carbon (g)	II. Presence of hydrogen bond donor group (d)	III. Absence of δ carbon (s)	IV. Absence of forks with hydrogens (l)	Assignment #
A Alanine	No	No	Yes	Yes	$g_0d_0s_2l_1$
C Cysteine	No	Yes	Yes	No	$g_0d_1s_2l_0$
D Aspartate	No	No	Yes	Yes	$g_0d_0s_1l_2$
E Glutamate	Yes	No	No	Yes	$g_1d_0s_0l_2$
F Phenylalanine	No	No	No	Yes	$g_0d_0s_0l_3$
G Glycine	No	No	Yes	No	$g_0d_0s_3l_0$
H Histidine	No	Yes	No	Yes	$g_0d_2s_0l_1$
I Isoleucine	Yes	No	Yes	No	$g_2d_0s_1l_0$
K Lysine	Yes	Yes	No	Yes	$g_1d_1s_0l_1$
L Leucine	Yes	No	No	No	$g_3d_0s_0l_0$
M Methionine	Yes	No	Yes	Yes	$g_1d_0s_1l_1$
N Asparagine	No	Yes	Yes	No	$g_0d_2s_1l_0$
P Proline	Yes	No	No	Yes	$g_2d_0s_0l_1$
Q Glutamine	Yes	Yes	No	No	$g_1d_2s_0l_0$
R Arginine	Yes	Yes	No	No	$g_2d_1s_0l_0$
S Serine	No	Yes	Yes	Yes	$g_0d_1s_1l_1$
T Threonine	Yes	Yes	Yes	No	$g_1d_1s_1l_0$
V Valine	Yes	No	Yes	No	$g_1d_0s_2l_0$
W Tryptophan	No	Yes	No	No	$g_0d_3s_0l_0$
Y Tyrosine	No	Yes	No	Yes	$g_0d_1s_0l_2$

'Yes' indicates that the property is satisfied and 'No' indicates that the property is not satisfied.
Subscript refers to the number of times each property occurs in the corresponding amino acid.

In a nut-shell

Protein tertiary structure prediction attempts for soluble proteins are progressing.

Structures of membrane bound proteins are intractable still.

Rules of protein folding continue to be elusive.

Structure & dynamics => function of proteins

Suggested reading: Aditya K. Padhi, B. Jayaram, James Gomes, "Prediction of Functional Loss of Human Angiogenin Mutants Associated with ALS by Molecular Dynamics Simulations", 2013, Scientific Reports (NPG), 3:1225, DOI: 10.1038/srep01225.

www.scfbio-iitd.res.in

•Genome Analysis - *ChemGenome*

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

•Protein Structure Prediction – *Bhageerath*

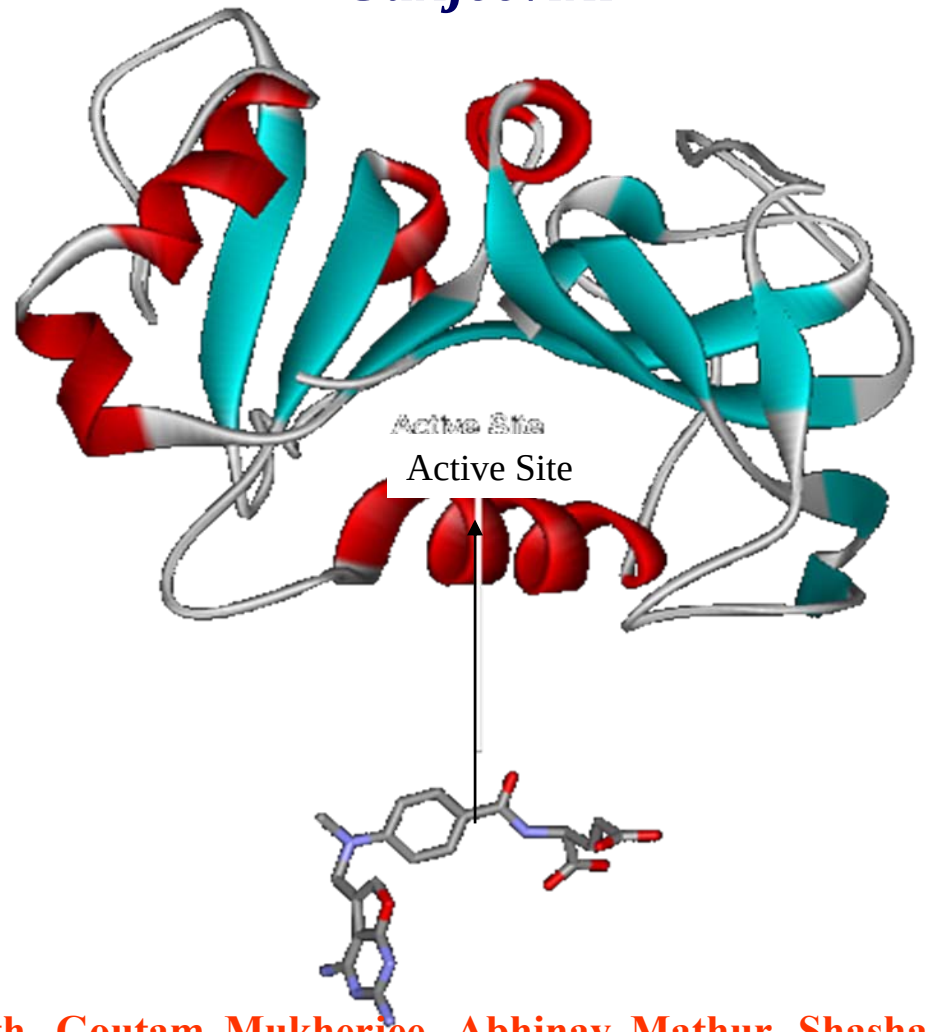
A *de novo* energy based protein structure prediction software

•**Drug Design – *Sanjeevini***

A comprehensive target directed lead molecule design protocol

Target Directed Lead Molecule Design

Sanjeevini



B. Jayaram, Tanya Singh, Goutam Mukherjee, Abhinav Mathur, Shashank Shekhar, and Vandana Shekhar, "*Sanjeevini:* A Freely Accessible Web-Server for Target Directed Lead Molecule Discovery", 2012, *BMC Bioinformatics* 2012, 13(Suppl 17):S7 doi:10.1186/1471-2105-13-S17-S7.

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

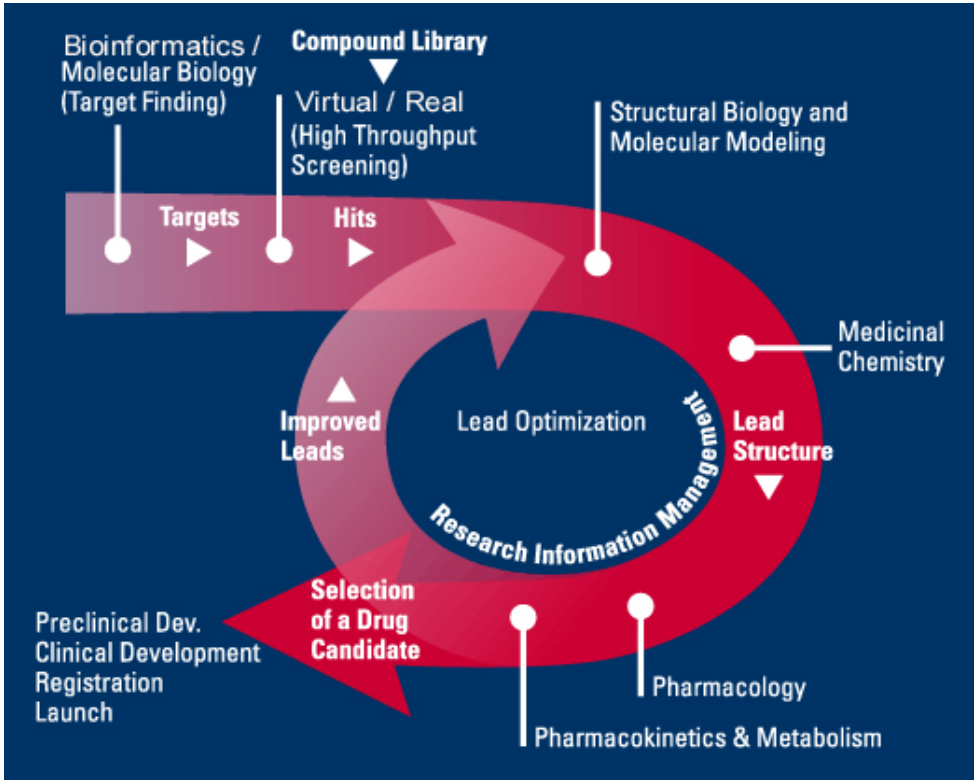
6.0yrs — *68%*

FDA Review & Approval

1.5yrs — *3%*

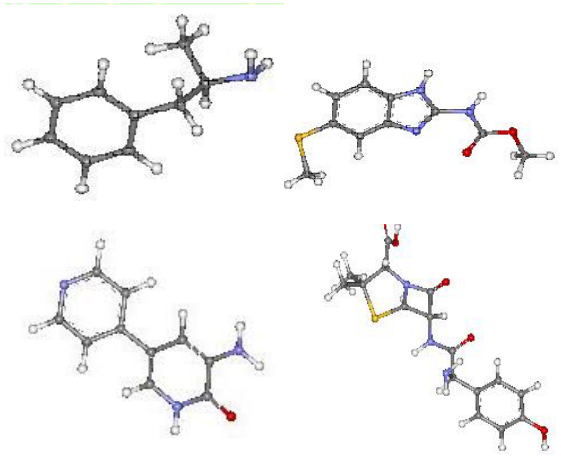
Drug to the Market

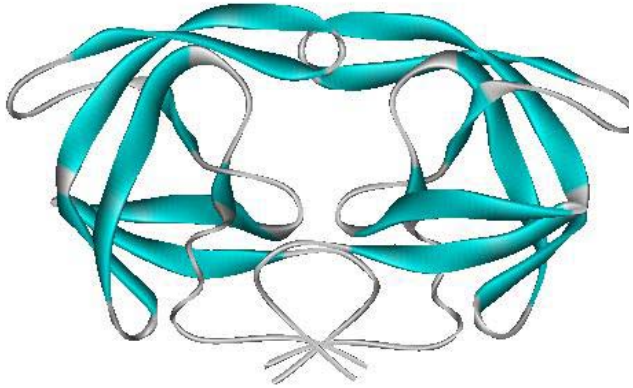
14 yrs **$1.4 billion**



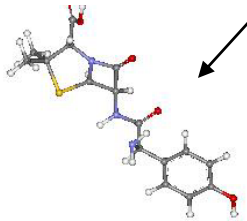
Source: PAREXEL's Pharmaceutical R&D Statistical Sourcebook, 2001, p96.; Hileman, Chemical Engg. News, 2006, 84, 50-1.

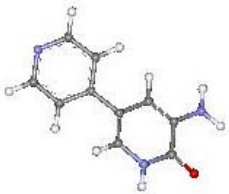
Active Site Directed Lead Molecule Design





Computer Aided Drug Design

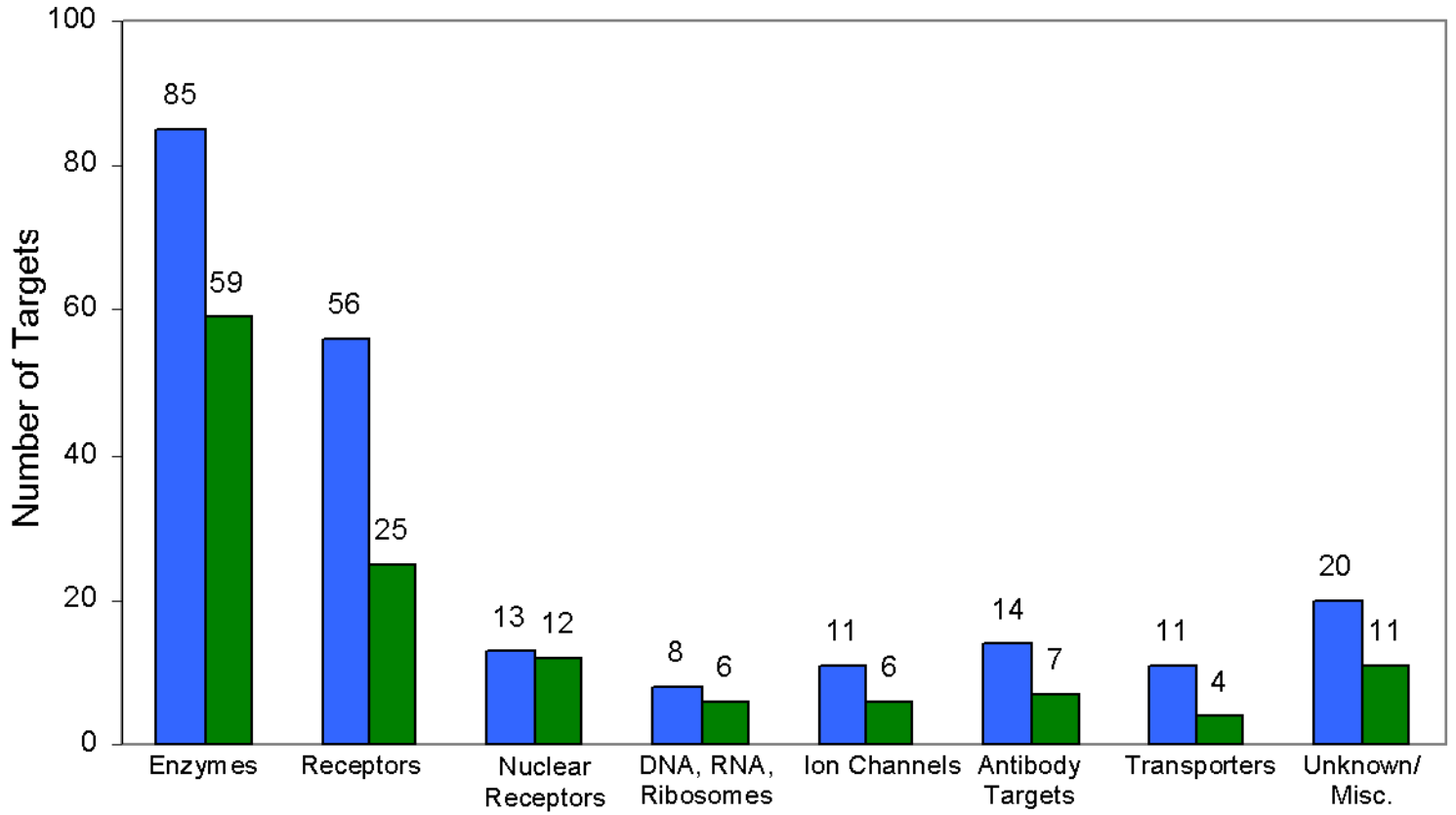




DRUG

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)

S. A. Shaikh, T. Jain, G. Sandhu, N. Latha, B. Jayaram, "From drug target to leads- sketching, A physicochemical pathway for lead molecule design in silico", *Current Pharmaceutical Design,* 2007, *13,* 3454-3470.

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
- ADMET (Acceptable Absorption, Distribution, Metabolism, Excretion & Toxicity Profiles)

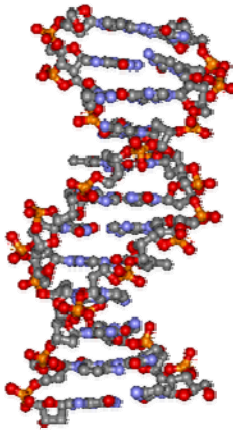
A list of some popular softwares for drug design

Sl. No.	Softwares	URL	Description
1	Discovery studio	http://accelrys.com/products/discovery-studio/structure-based-design.html	Molecular modeling and *de novo* drug design
2	Sybyl	http://www.tripos.com/	Computational software for drug discovery
3	Bio-Suite	http://www.staff.ncl.ac.uk/p.dean/Biosuite/body_biosuite.html	Tool for Drug Design, structural analysis and simulations
4	Molecular Operating Environment (MOE)	http://www.chemcomp.com/	Structure-based drug design, molecular modeling and simulations
5	Glide	https://www.schrodinger.com/products/14/5	Ligand-receptor docking
6	Autodock	http://autodock.scripps.edu/	Protein-ligand docking
7	DOCK	http://dock.compbio.ucsf.edu/	Protein-ligand docking
8	*Sanjeevini*	http://www.scfbio-iitd.res.in/sanjeevini/sanjeevini.jsp	A complete software suite for structure-based drug design
9	ArgusLab	http://www.arguslab.com/arguslab.com/ArgusLab.html	Ligand-receptor docking
10	eHITS	http://www.simbiosys.ca/ehits/index.html	Ligand-receptor docking
11	FlexX	http://www.biosolveit.de/FlexX/	Ligand-receptor docking
12	FLIPDock	http://flipdock.scripps.edu/	Ligand-receptor docking
13	FRED	http://www.eyesopen.com/oedocking	Ligand-receptor docking
14	GOLD	http://www.ccdc.cam.ac.uk/products/life_sciences/gold/	Protein-ligand docking
15	ICM-Docking	http://www.molsoft.com/docking.html	Protein-ligand docking
16	PLANTS	http://www.tcd.uni-konstanz.de/research/plants.php	Protein-ligand docking
17	Surflex	http://www.biopharmics.com/	Protein-ligand docking

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

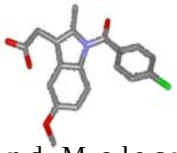
Jayaram, B., Latha, N.,Jain, T., Sharma, P., Gandhimathi, A., Pandey, V.S., Indian J. Chemistry-A. 2006, 45A, 1834-1837.

Sanjeevini Pathway

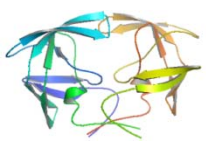
NRDBM/Million Molecule Library/Natural Products and Their Derivatives

Molecular Database

or,



Ligand Molecule



or

Target Protein/DNA

Upload

Bioavalibality Check (Lipinski Compliance)

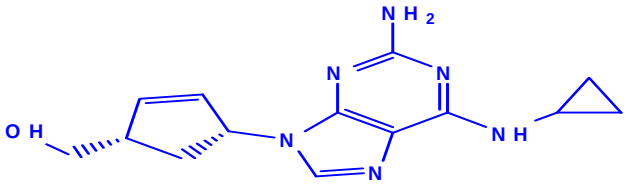
Binding EnergyEstimation by RASPD protocol

Geometry Optimization TPACM4/Quantum Mechanical Derivation of Charges

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

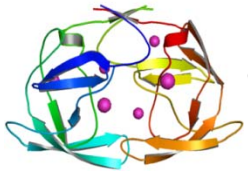
Assignment of Force Field Parameters

Ligand Molecule ready for Docking

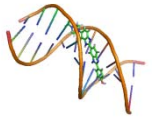


+

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

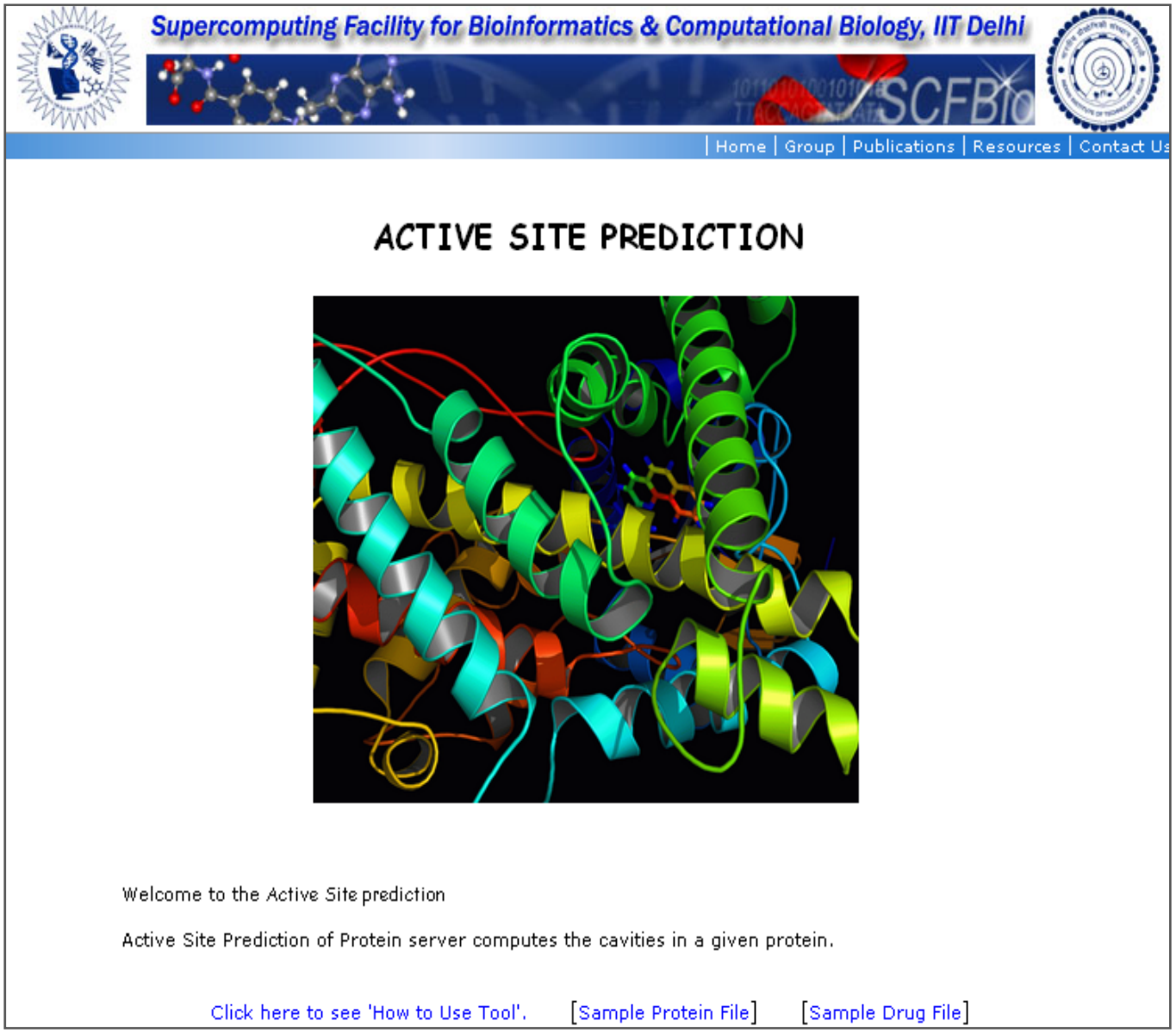
Rank of the cavity points vs. cumulative percentage prediction

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/dock/ActiveSite_new.jsp



Tanya Singh, D. Biswas, B. Jayaram, 2011, *J. Chem. Inf. Modeling*,51 (10), 2515-2527.

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

Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi

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

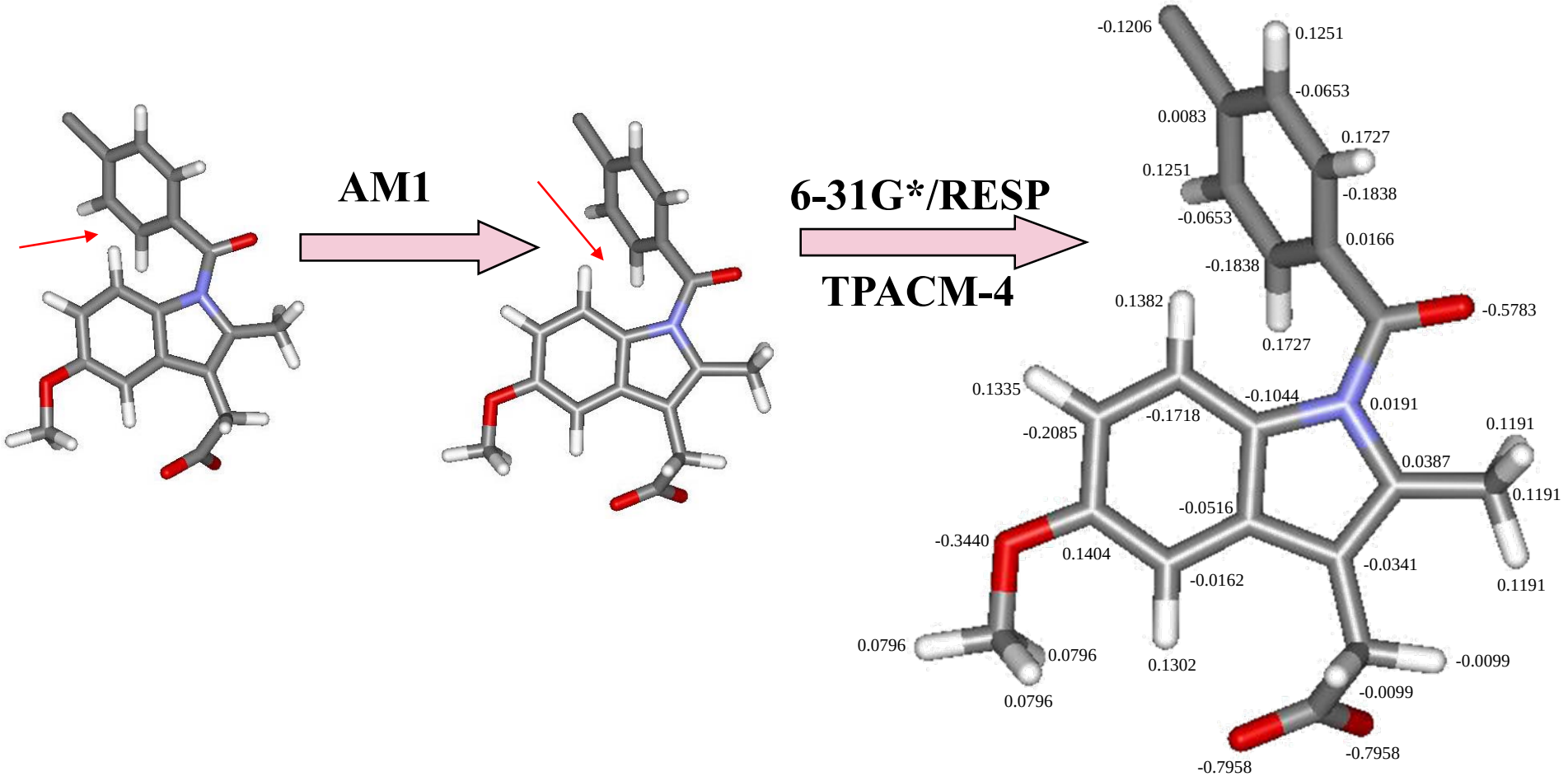
Browse...

Enter Drug Id: DRG

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

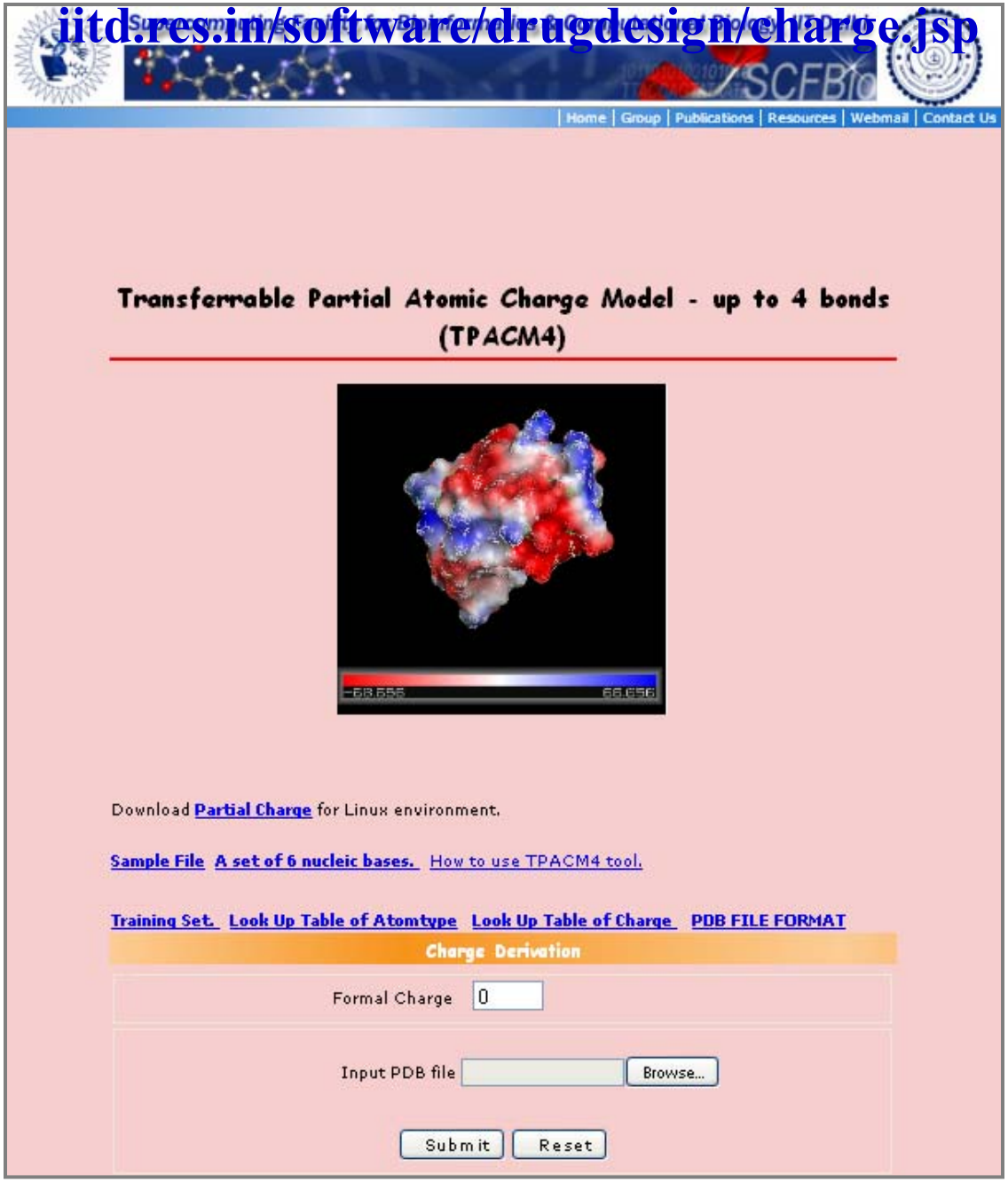
Submit

Quantum Chemistry on Candidate drugs for Assignment of Force Field Parameters



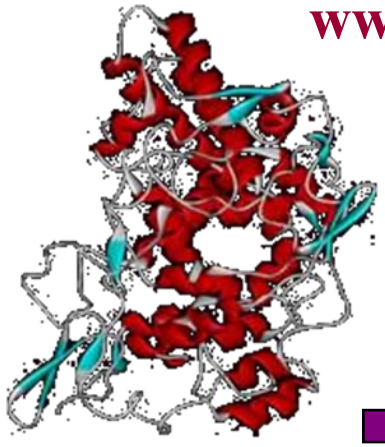
G. Mukherjee, N. Patra, P. Barua and B. Jayaram, *J. Computational Chemistry*, 32, 893-907 (2011).

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

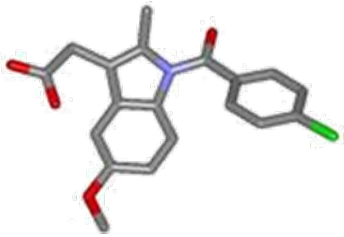


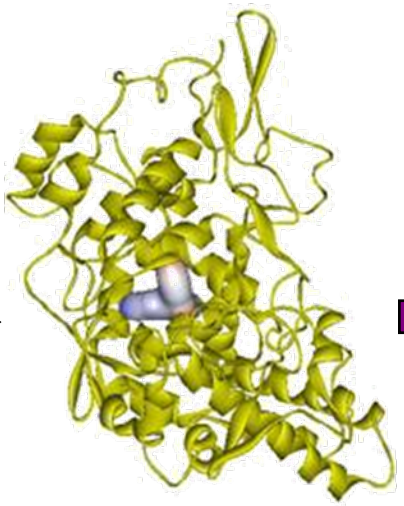
MONTE CARLO DOCKING OF THE CANDIDATE DRUG IN THE ACTIVE - SITE OF THE TARGET

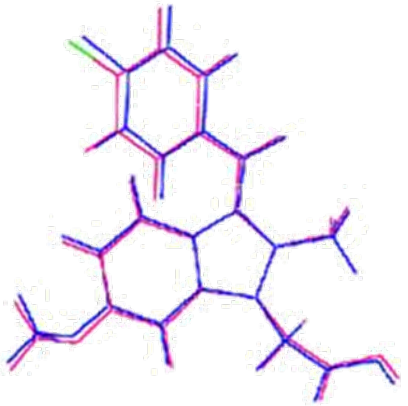
www.scfbio-iitd.res.in/dock/pardock.jsp



+





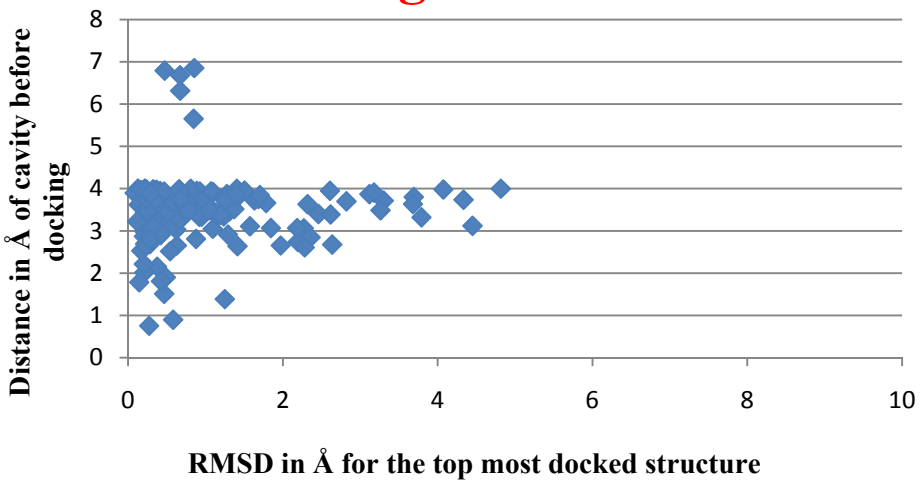


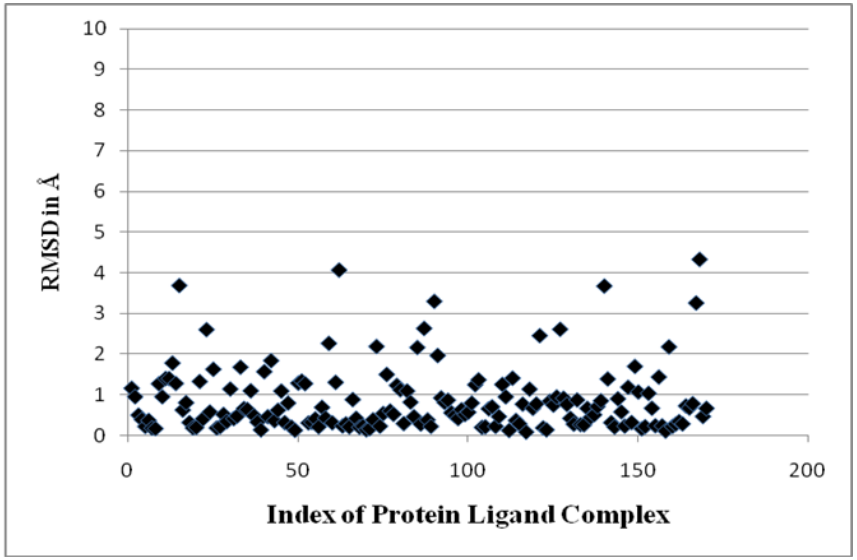
RMSD between the docked & the crystal structure is 0.2Å

ENERGY MINIMIZATION

5 STRUCTURES WITH LOWEST ENERGY SELECTED

Docking Accuracies

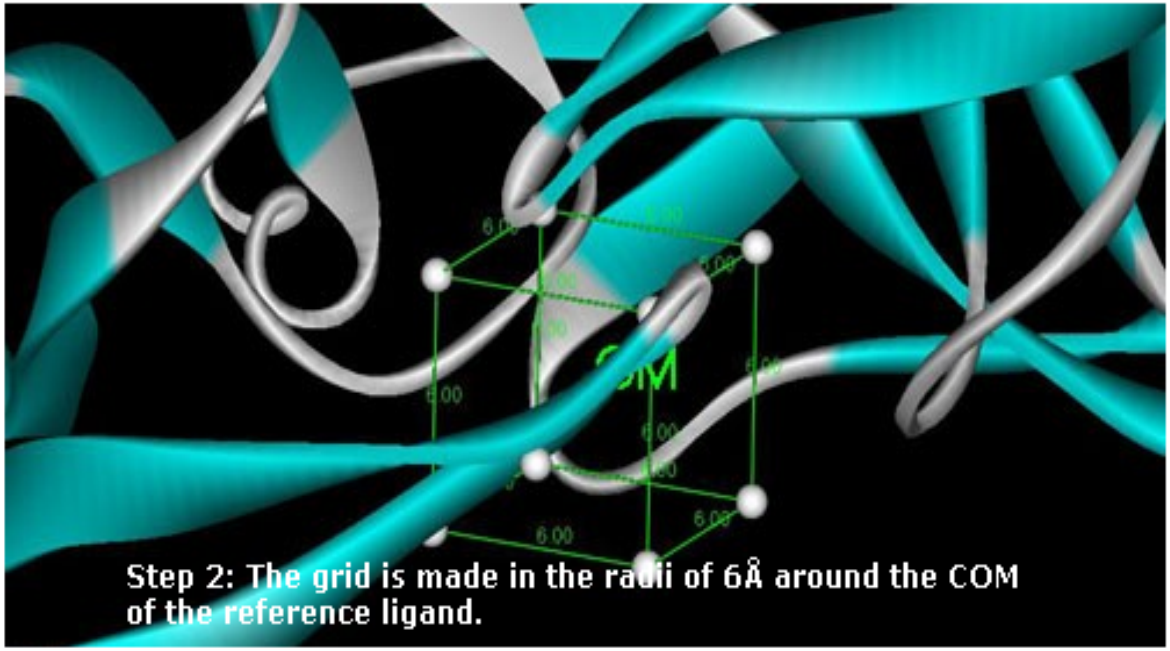




RMSD between the crystal structure and one of the top five docked structures

ParDOCK

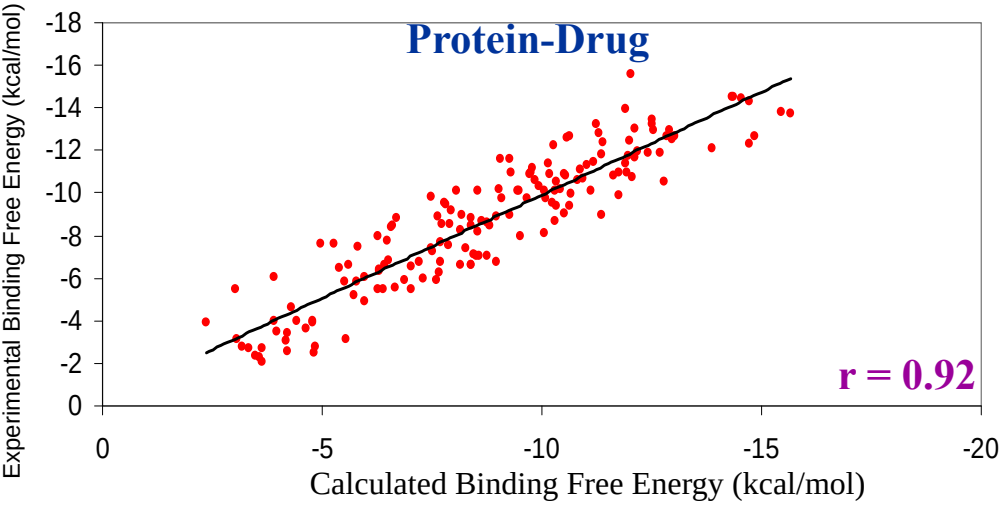
Automated Server for Protein Ligand Docking



Step 2: The grid is made in the radii of 6Å around the COM of the reference ligand.

ENERGY BASED SCORING FUNCTION

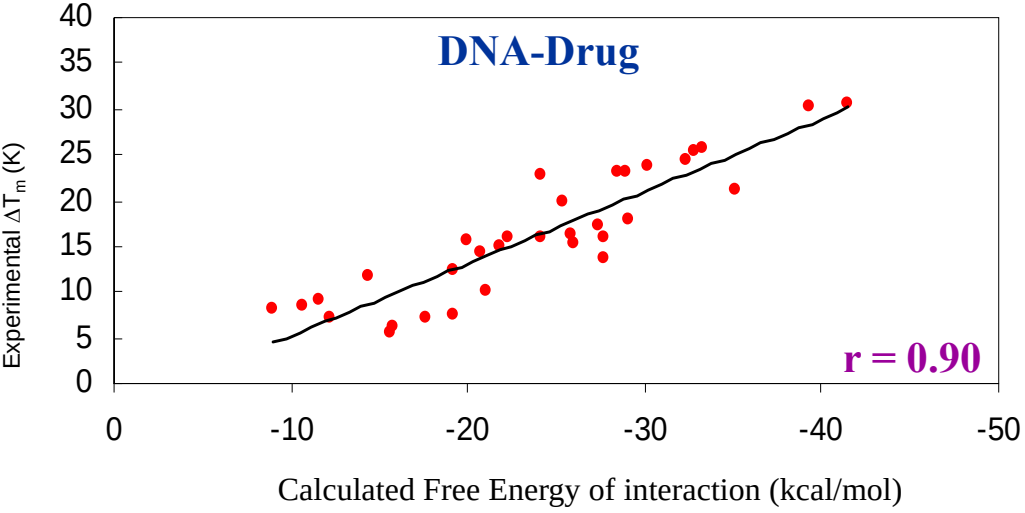
$$\Delta G^\circ_{bind} = \Delta H^\circ_{el} + \Delta H^\circ_{vdw} - T\Delta S^\circ_{rtvc} + \Delta G^\circ_{hpb}$$



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

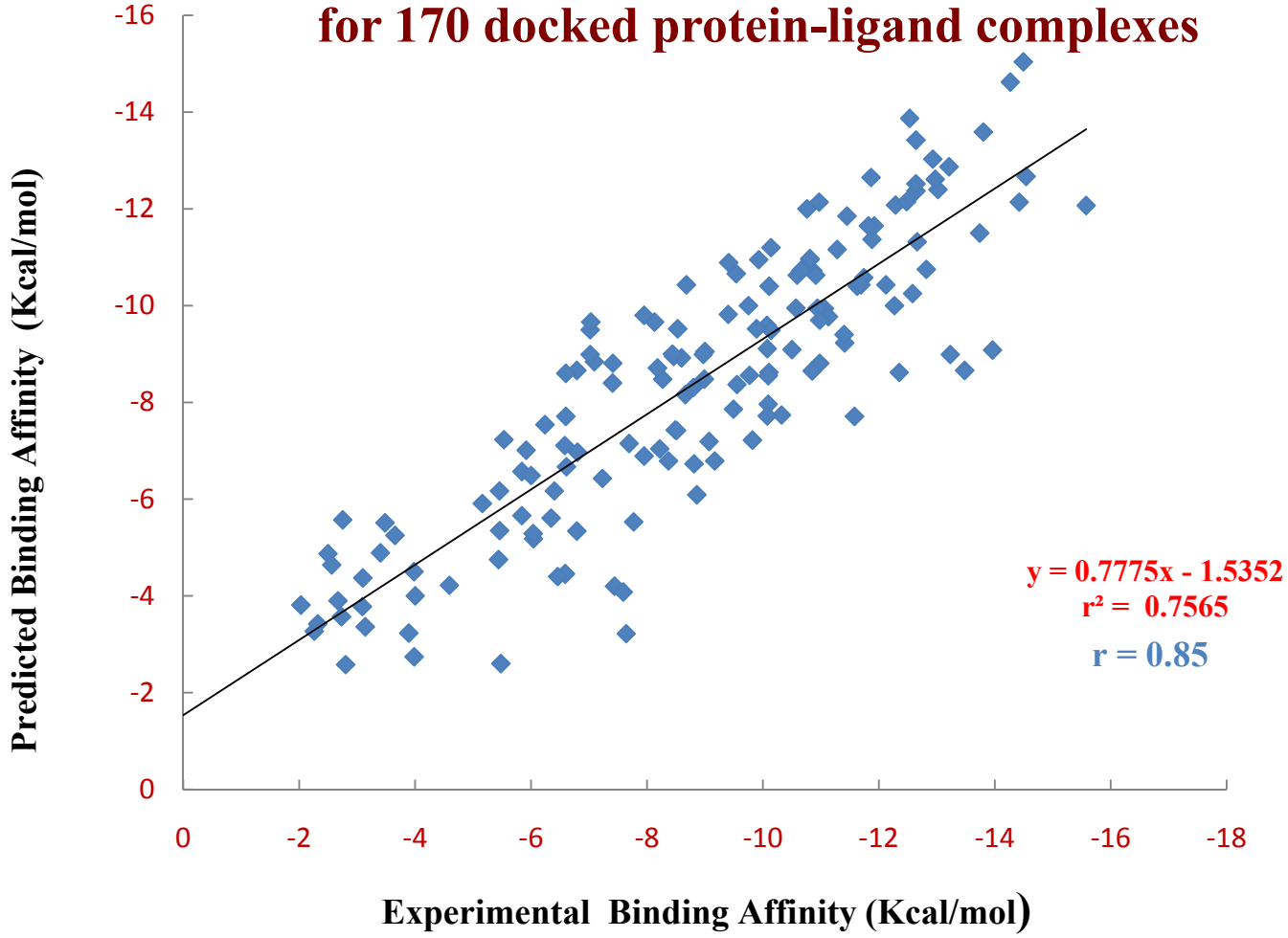


Correlation between experimental ΔT_m and calculated free energy of interaction for DNA-Drug Complexes

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

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

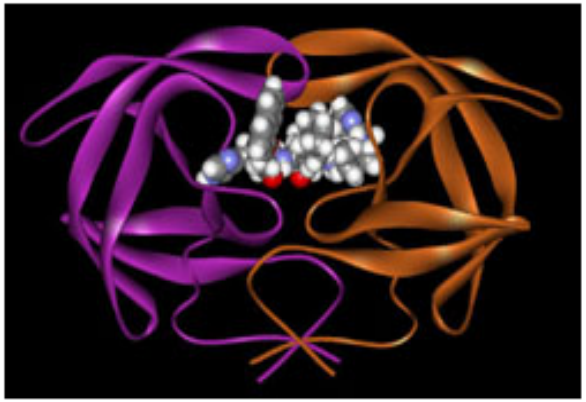
Correlation between Experimental and Predicted Binding free energies for 170 docked protein-ligand complexes



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

BAPPL server

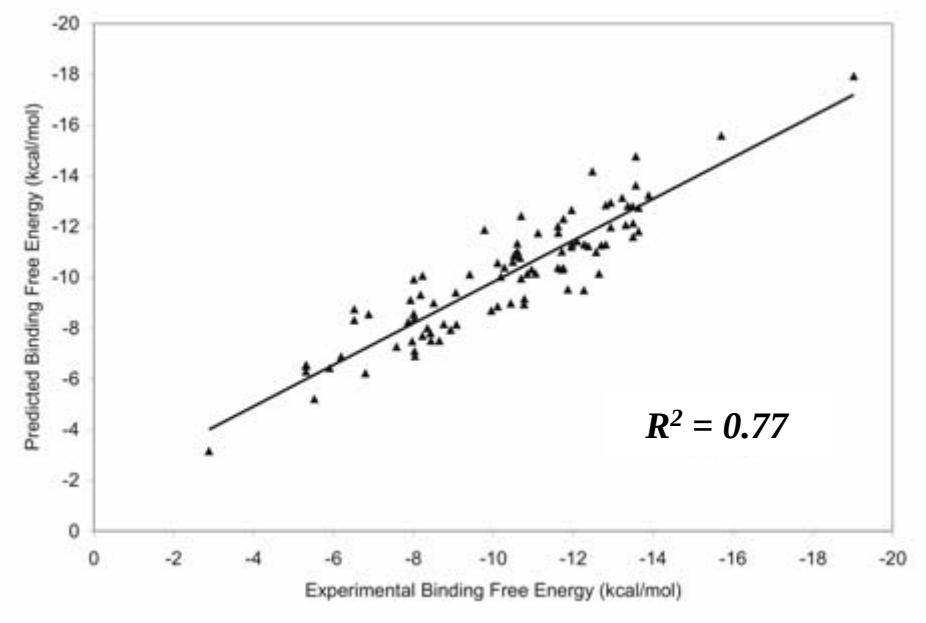


HIV-I Protease complexed with U75875 (1hiv.pdb)

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].

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

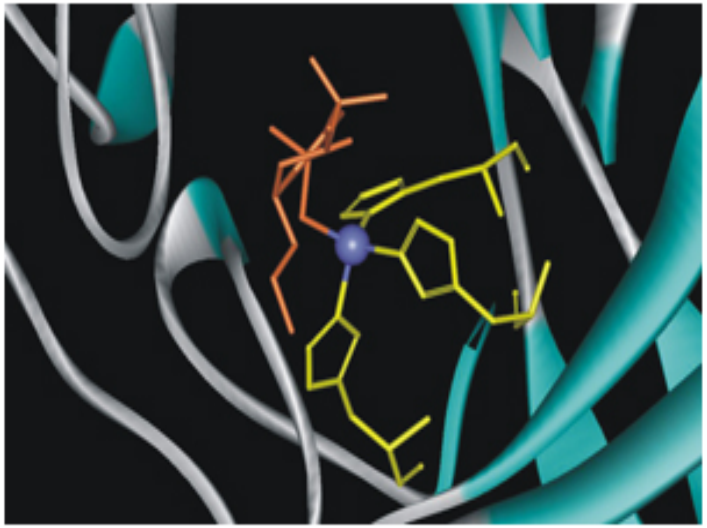
T. Jain & B. Jayaram, *Proteins: Struct. Funct. Bioinfo.* 2007, 67, 1167-1178.

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

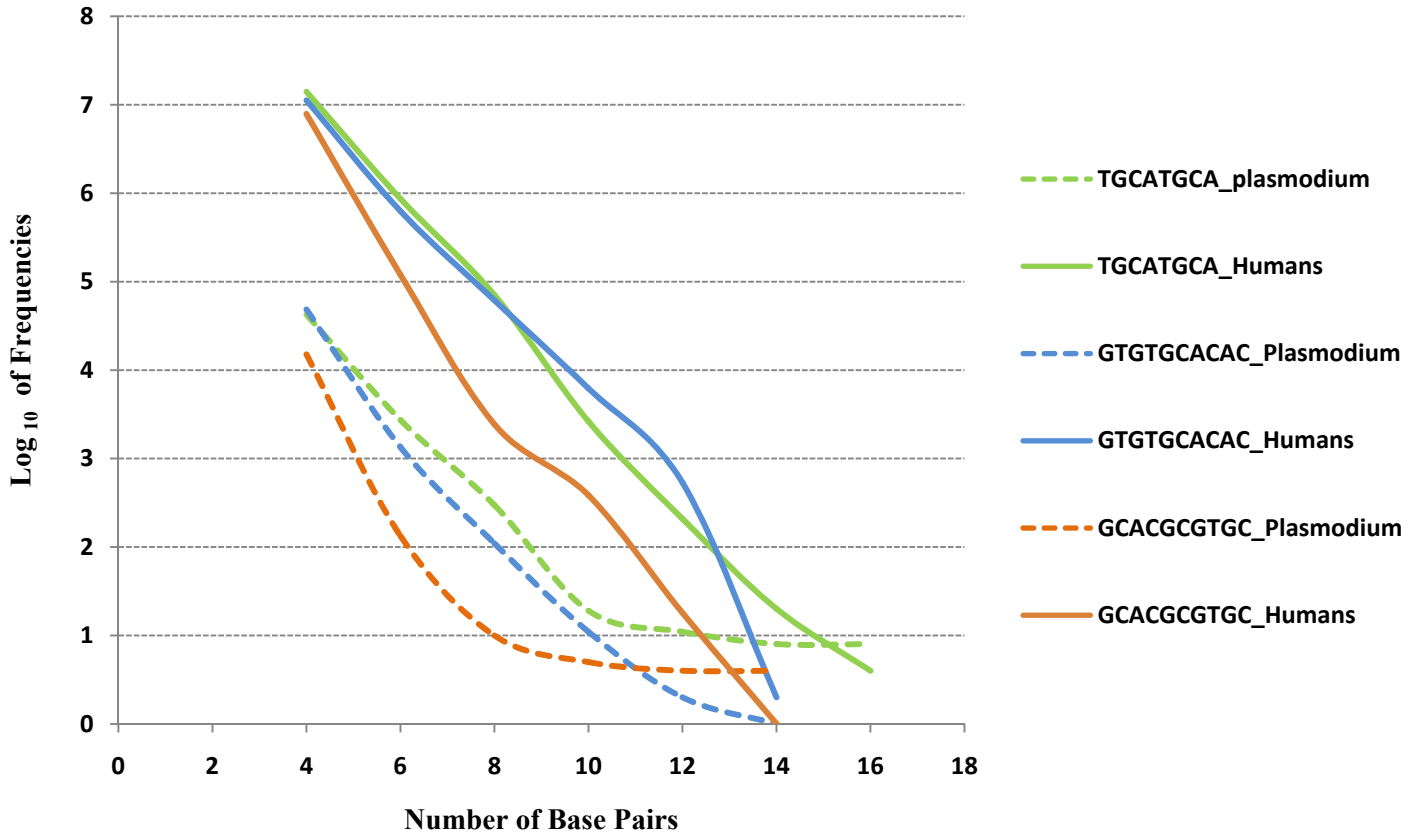
Comparative evaluation of some methodologies reported for estimating binding affinities of zinc containing metalloprotein-ligand complexes

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-Z server



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



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 => Base sequence to constitute a unique target (occurs only once) must be 18 to 20 bp long.

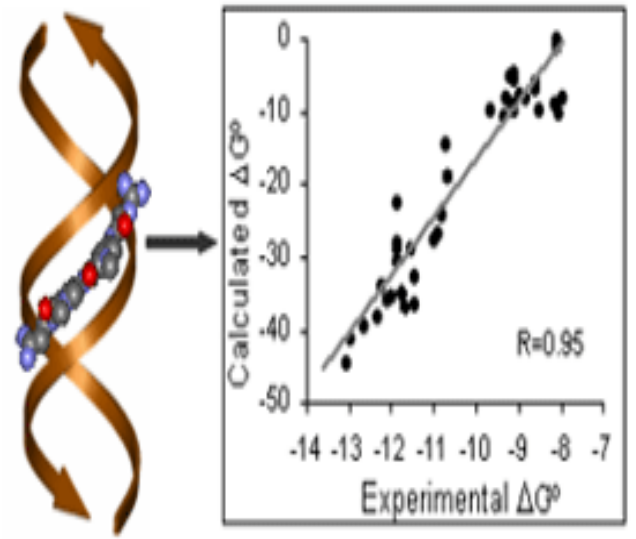
PreDDICTA

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

About Preddicta

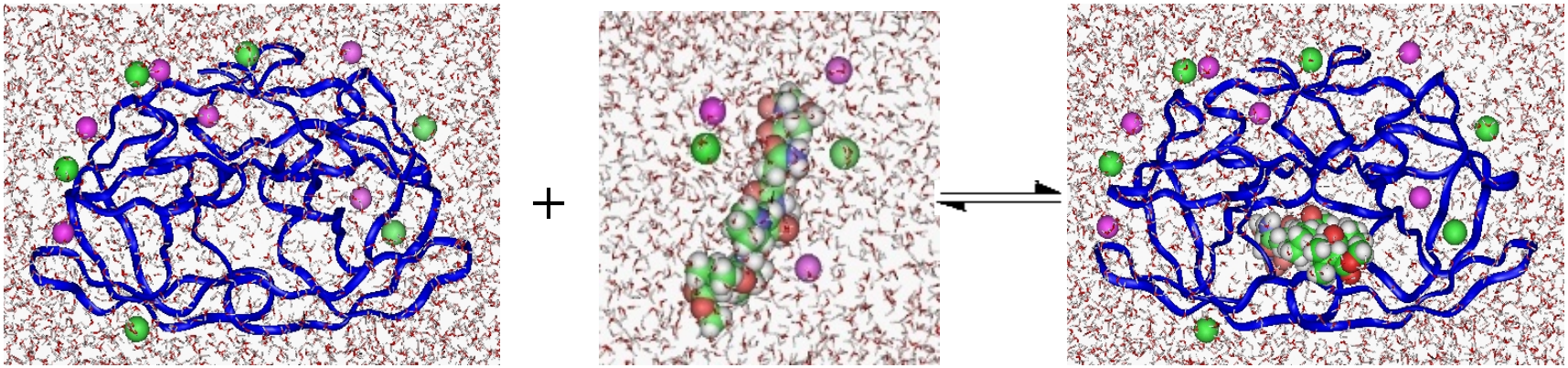
DNA Drug Interaction

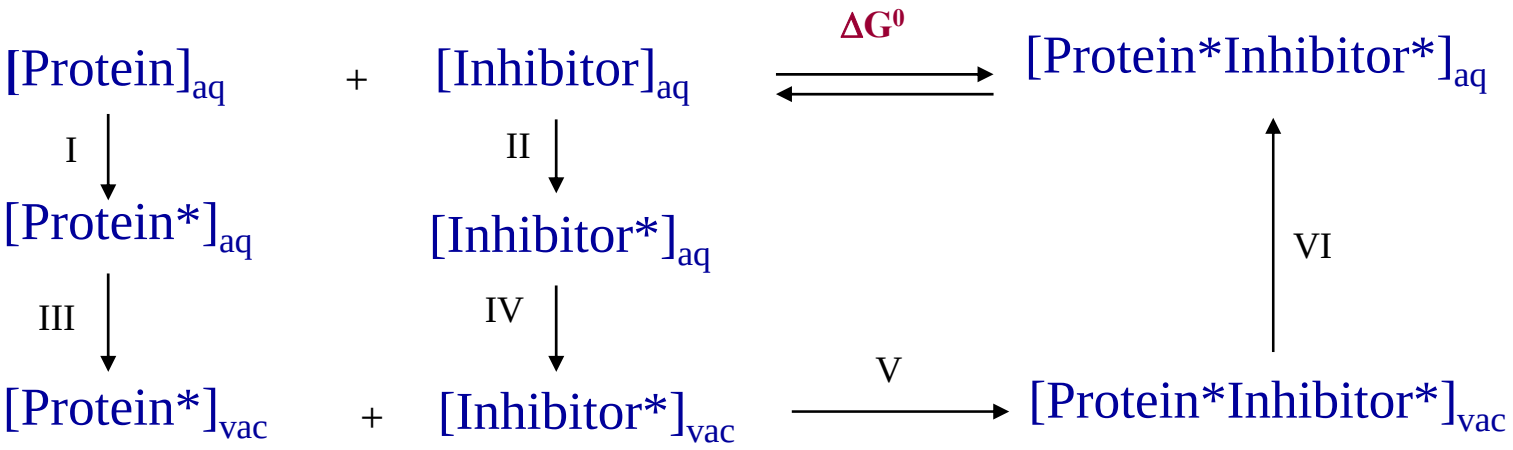
DNA Drug Complex Data Set



Binding Affinity Analysis

After obtaining candidate molecules from docking and scoring, molecular dynamics simulations followed by free energy analyses (MMPBSA/MMGBSA) are recommended.





Parul Kalra, Vasisht Reddy, B. Jayaram, "A Free Energy Component Analysis of HIV-I Protease-Inhibitor Binding", *J. Med.Chem.*, 2001, *44*, 4325-4338.

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

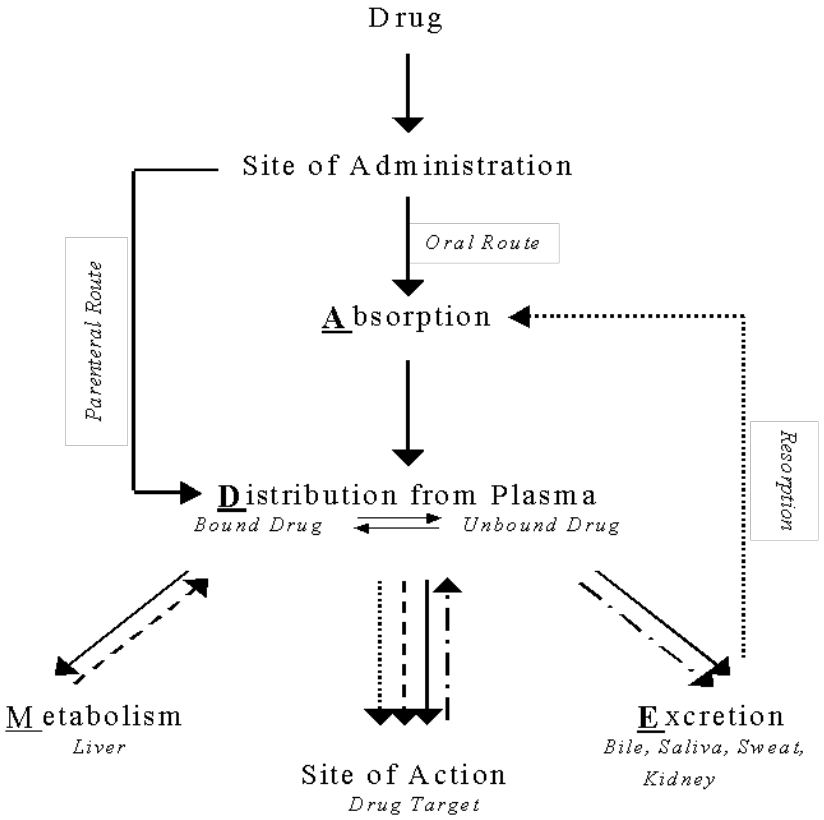
Shaikh, S., Jain. T., Sandhu, G., Latha, N., Jayaram., B., *A physico-chemical pathway from targets to leads*, 2007, *Current Pharmaceutical Design*, 13, 3454-3470.

	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**

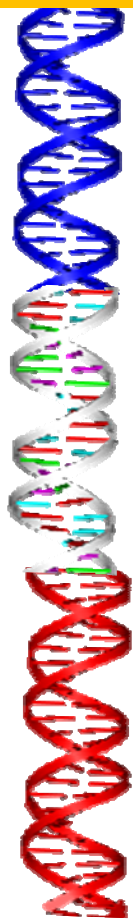
Diagonal elements represent drug-target binding affinity and off-diagonal elements show drug-non target binding affinity. Drug 1 is specific to Target 1, Drug 2 to Target 2 and so on. Target 1 is lymphocyte function-associated antigen LFA-1 (CD11A) (1CQP; Immune system adhesion receptor) and Drug 1 is 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



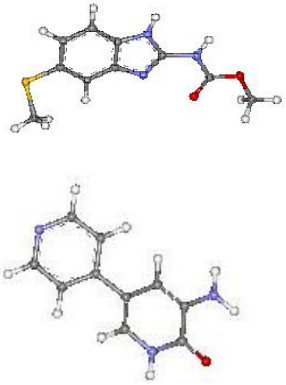
The distribution path of an orally administered drug molecule inside the body is depicted. Black solid arrows: Complete path of drug starting from absorption at site of administration to distribution to the various compartments in the body, like sites of metabolism, drug action and excretion. Dashed 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. Affinity/specificity are under control but toxicity is yet to be conquered.

From Genome to Hits



Genome

X Teraflops
Chemgenome
Bhageerath
Sanjeevini



Hits

Chikungunya Virus

Chikungunya is one of the most important re-emerging viral borne disease spreading globally with sporadic intervals. It is categorized as a BSL3 pathogen and under 'C' grade by National Institute of Allergy and Infectious Diseases (NIAID), in 2008. But, yet no approved drug/vaccine is available currently in the public domain for its treatment/prevention.

Anjali Soni, Khushhali Menaria, Pratima Ray and B. Jayaram. "Genomes to Hits *in Silico*: A Country Path Today, A Highway Tomorrow: A case study of chikungunya", Current Pharmaceutical Design, 2013, in press.

Some available information on CHIKV proteins but no structures

Protein Type	Proteins	Functions
NonStructural Proteins	nsP1	❖ Methyl transferase domain (acts as cytoplasmic capping enzyme)
	nsP2	❖ Viral RNA helicase domain (part of the RNA polyemerase complex) ❖ Peptidase C9 domain (cleaves four mature proteins from non structural polyprotein)
	nsP3	❖ Appr. 1-processng domain (minus strand and subgenomic 26S mRNA synthesis)
	nsP4	❖ Viral RNA dependent RNA polymerase domain (Replicates genomic and antigenomic RNA and also transcribes 26S subgenomic RNA which encodes for structural proteins)
Structural proteins	C	❖ Peptidase_S3 domain (autocatalytic cleavage)
	E3	❖ Alpha virus E3 spike glycoprotein domain
	E2	❖ Alpha virus E2 glycoprotein domain (viral attachment to host)
	6K	❖ Alpha virus E1 glycoprotein domain (viral glycoprotein processing and membrane permeabilization) ❖ Signal peptide domain
	E1	❖ Alpha virus E1 glycoprotein domain (class II viral fusion protein) ❖ Glycoprotein E dimerization domain (forms E1-E2 heterodimers in inactive state and E1 trimers in active state)

Flow diagram illustrating the steps involved in achieving hit molecules from genomic information

Whole genome sequence of ***Chikungunya virus*** is retrieved from NCBI: NC_004162.2
http://www.ncbi.nlm.nih.gov/nuccore/27754751?report=fasta

↓

Genes are predicted which are then translated to the protein sequences using *Chemgenome 3.0*
http://www.scfbio-iitd.res.in/chemgenome/chemgenome3.jsp

↓

These polyprotein sequences are spliced w.r.t literature and results are processed for 3-D structure prediction by *Bhageerath-H*
http://www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp

↓

Modeled structures are studied for identification of potential active sites by active site finder ***AADS***
http://www.scfbio-iitd.res.in/dock/ActiveSite.jsp

↓

A million compound library of small molecules is screened against the predicted binding sites using ***RASPD***
http://www.scfbio-iitd.res.in/software/drugdesign/raspd.jsp

↓

The screened molecules are docked, scored and optimized iteratively using ***SANJEEVINI***
http://www.scfbio-iitd.res.in/sanjeevini/sanjeevini.jsp

↓

Hits ready to be synthesized and tested in

Input the CHIKV Genome sequences to *ChemGenome 3.0*:

Chikungunya virus (strain S27-African prototype), complete genome
NCBI Ref_Sequence: NC_004162.2

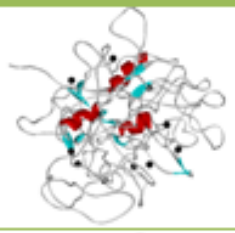
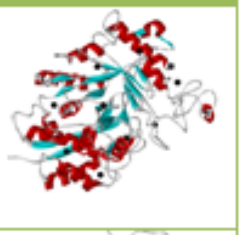
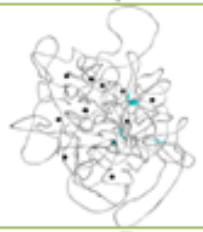
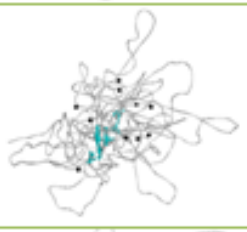
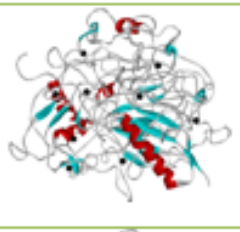
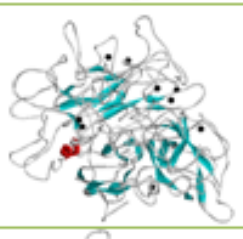
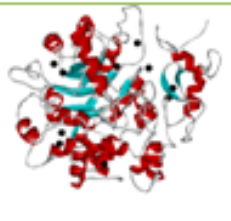
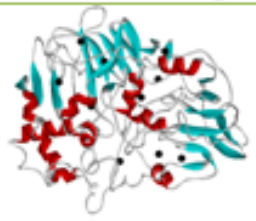
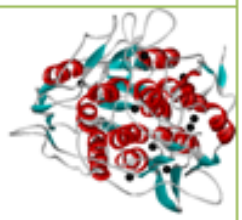
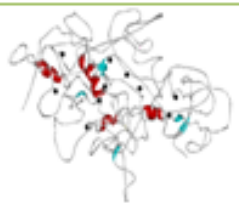
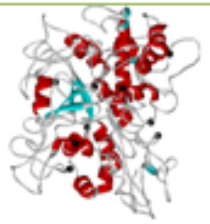
ChemGenome 3.0 output

Two protein coding regions are identified. These proteins are the polyproteins.

Genes	Start	End	Type
1	77	7501	Nonstructural Polyproteins
2	7567	11313	Structural Polyproteins

The nonstructural polyproteins are cleaved into 4 protein sequences w.r.t literature. These sequences serve as input to *Bhageerath*-H server.

Bhageerath-H output

Protein	Model-1	Model-2	Model-3	Model-4	Model-5
nsp-1					
nsp-2					
nsp-3					
nsp-4					

Input Protein Structures to an Automated version of Active site finder (*AADS/Sanjeevini*)

10 potential binding sites are identified against each model of the proteins (shown as black dots in the figure)

Scanning against a million compound library

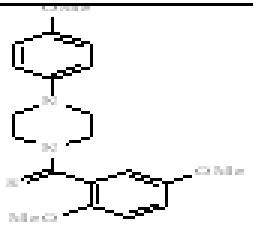
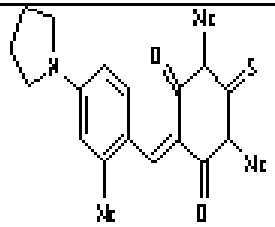
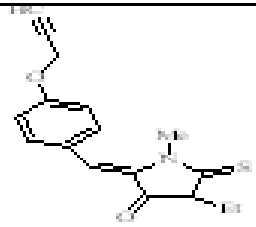
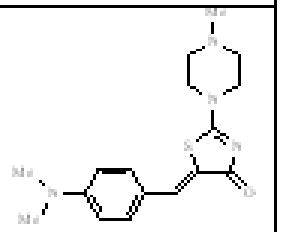
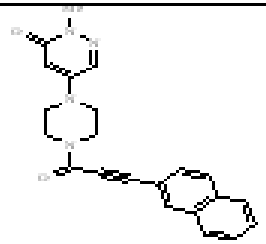
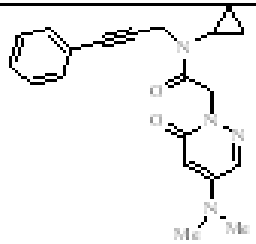
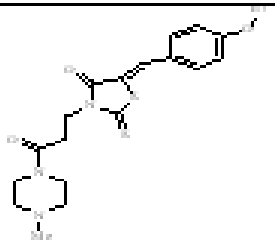
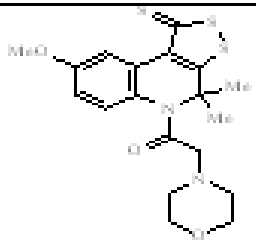
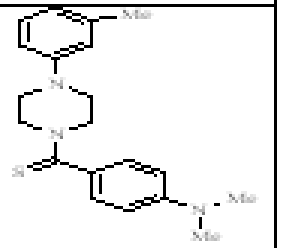
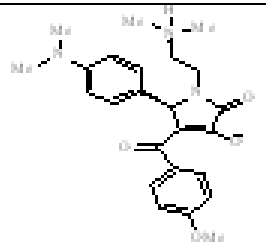
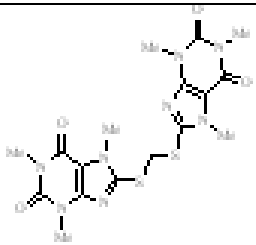
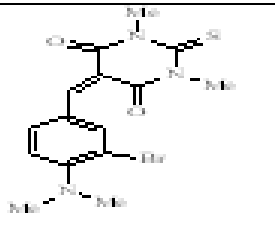
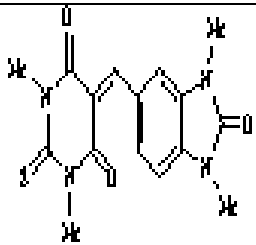
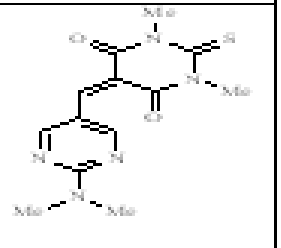
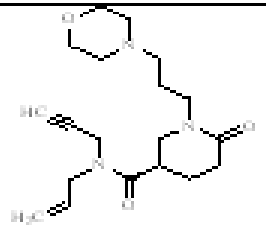
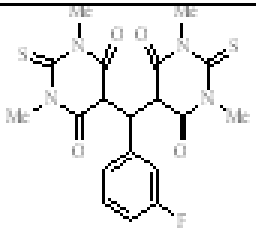
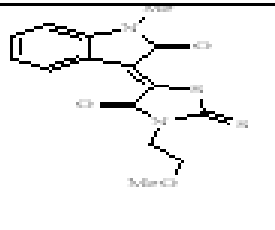
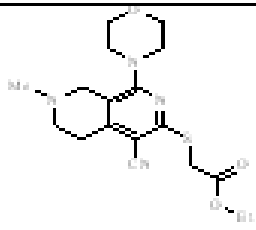
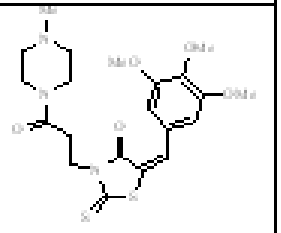
RASPD/*Sanjeevini* calculations were carried in search of the potential therapeutics with an average cut-off binding affinity to limit the number of candidates. (RASPD uses an empirical scoring function which builds in Lipinski's rules and Wiener index).

RASPD output

Top 100 molecules were screened with the cutoff binding energy to be -8.00 kcal/mol. Out of these 100, one molecule for each model is selected with good binding energy from one million molecule database corresponding to the top 5 predicted binding sites. The molecules were choosen for atomic level binding energy calculations using ParDOCK/*Sanjeevini*.

These molecules could be tested in the Laboratory.

In silico suggestions of candidate molecules against CHIKV

Protein	Model-1	Model-2	Model-3	Model-4	Model-5
nsP1	 -8.61	 -8.95	 -8.91	 -17.13	 -7.98
nsP2	 -12.22	 -11.07	 -9.21	 -8.86	 -8.43
nsP3	 -11.04	 -9.65	 -9.26	 -10.08	 -9.01
nsP4	 -15.07	 -9.25	 -8.75	 -9.01	 -9.00

SCFBio Team

16 processor Linux Cluster

Storage Area Network

~ 6 teraflops of computing; 20 terabytes of storage

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

Present

Shashank Shekhar
Tanya Singh
Avinash Mishra
Anjali Soni
Mousumi Bhattacharya
Rahul Kaushik
R. Nagarajan

Garima Khandelwal
Priyanka Dhingra
Ashutosh Shandilya
Varsha Singh
M. Hassan
Ali Khosravi
Preeti Bisht

Goutam Mukherjee
Vandana Shekhar
Abhilash Jayaraj
Ankita Singh
Prashant Rana
Kritika Karri
Sanjeev Kumar

Former

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

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

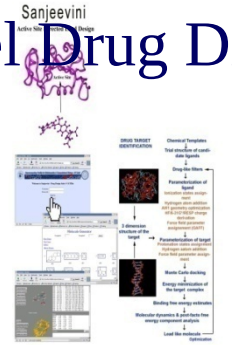
Dr. Pooja Narang
Dr. Parul Kalra
Dr. Surjit Dixit
Surojit Bose
Vidhu Pandey
Anuj Gupta
Dhrubajyoti Biswas
Bharat Lakhani
Pooja Khurana

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

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

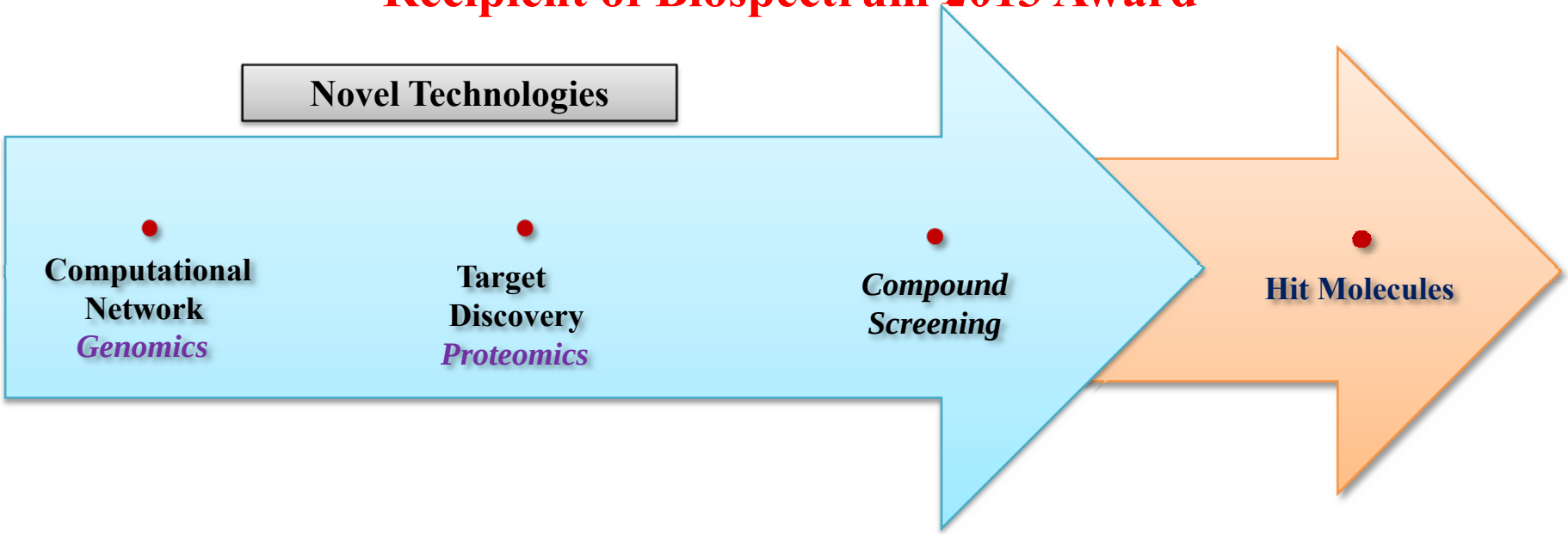
Incubated at IIT Delhi (2007-2010)

DSIR Certified (2011)

www.leadinvent.com

Novo Informatics
Smart Minds. Smart Technologies.

Under Incubation at IITD (since April, 2011)
Recipient of TATA NEN 2012 Award
Recipient of Biospectrum 2013 Award



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)

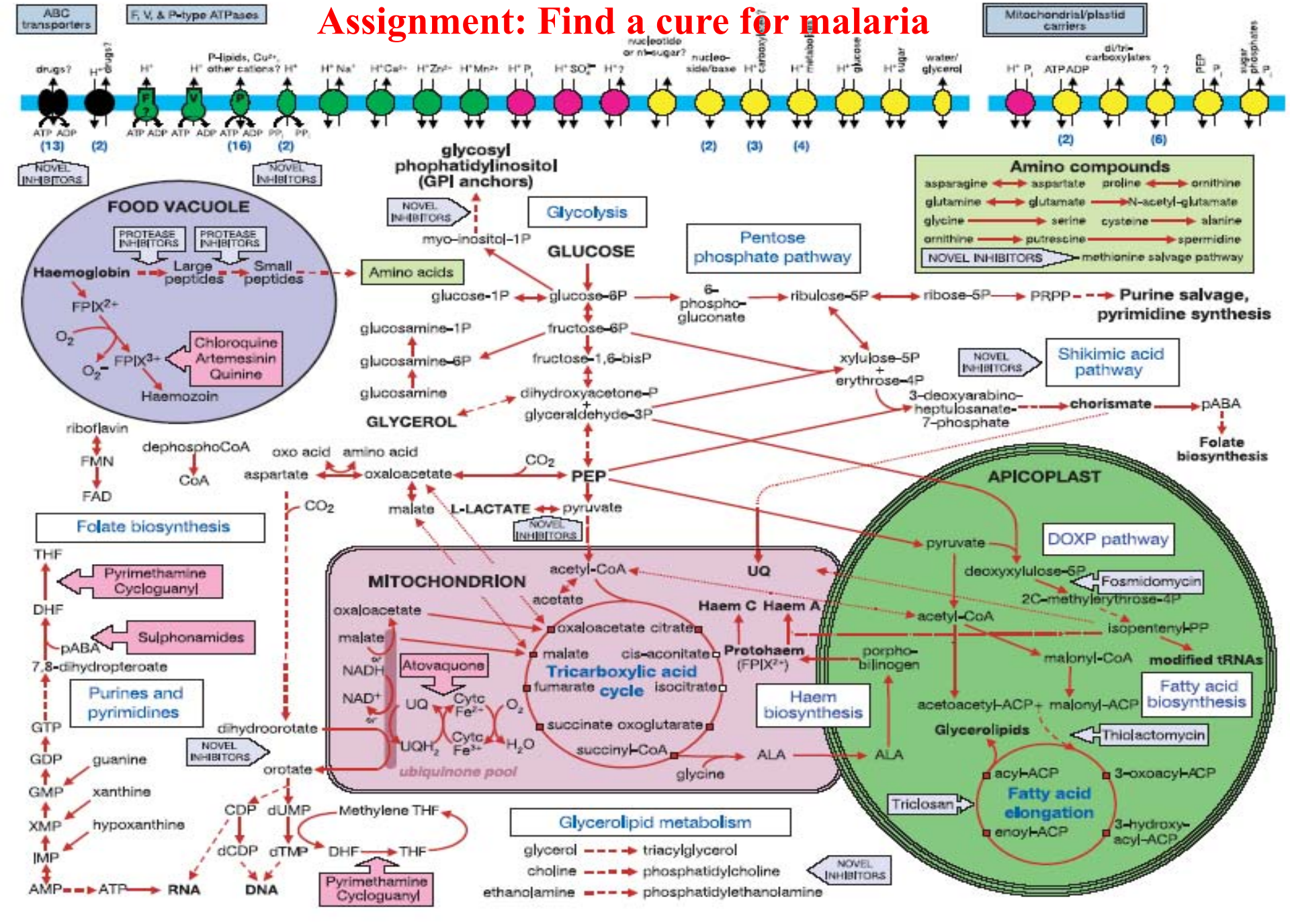
HCL Life Science Technologies

Dabur Research Foundation

Indian Institute of Technology, Delhi

OVERVIEW OF METABOLISM AND TRANSPORT IN *P.falciparum*

Assignment: Find a cure for malaria



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

Thank You