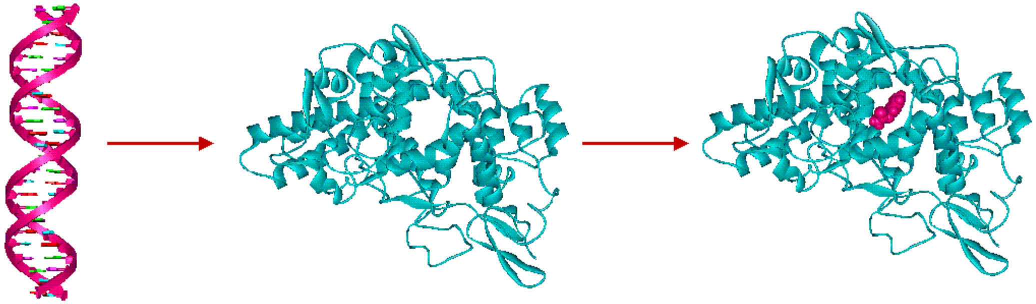


Some Challenges in Biomolecular Recognition

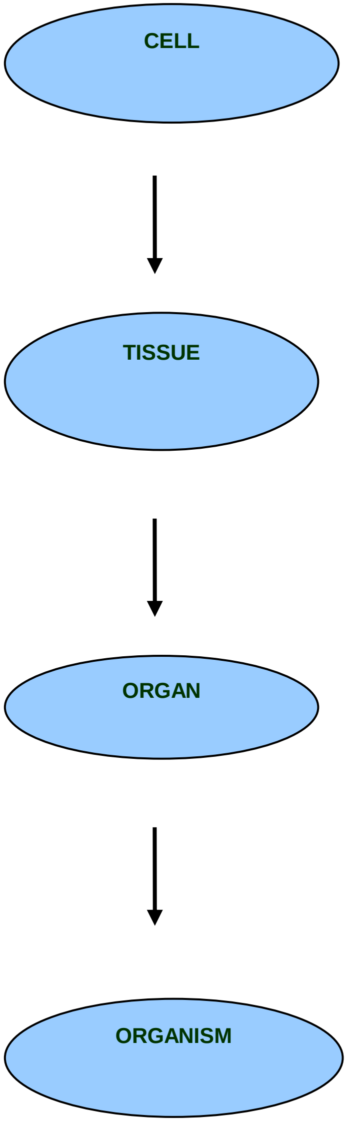
Gene to Drug in Silico:

A Molecular Bioinformatics Approach

Prof. B. Jayaram

Department of Chemistry, Supercomputing Facility for Bioinformatics & Computational Biology & School of Biological Sciences

Indian Institute of Technology Delhi

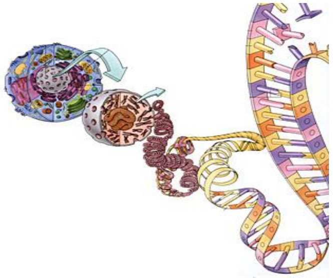


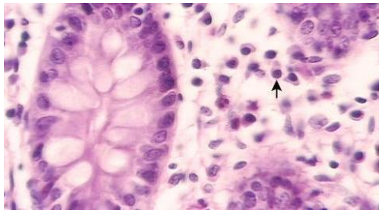
CELL

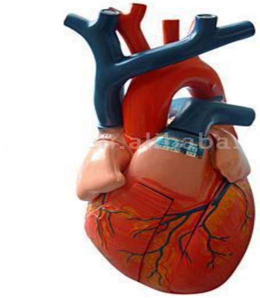
TISSUE

ORGAN

ORGANISM

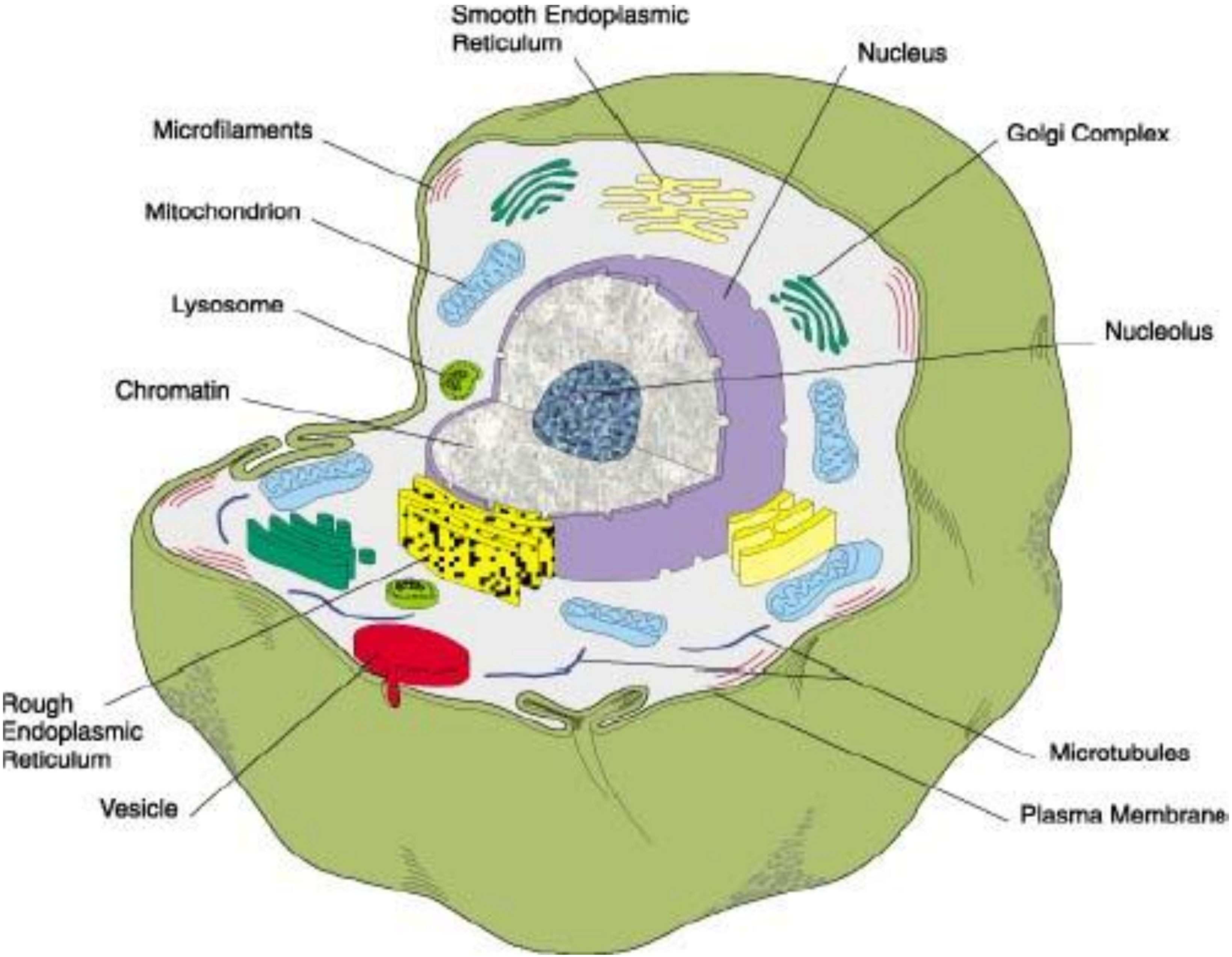


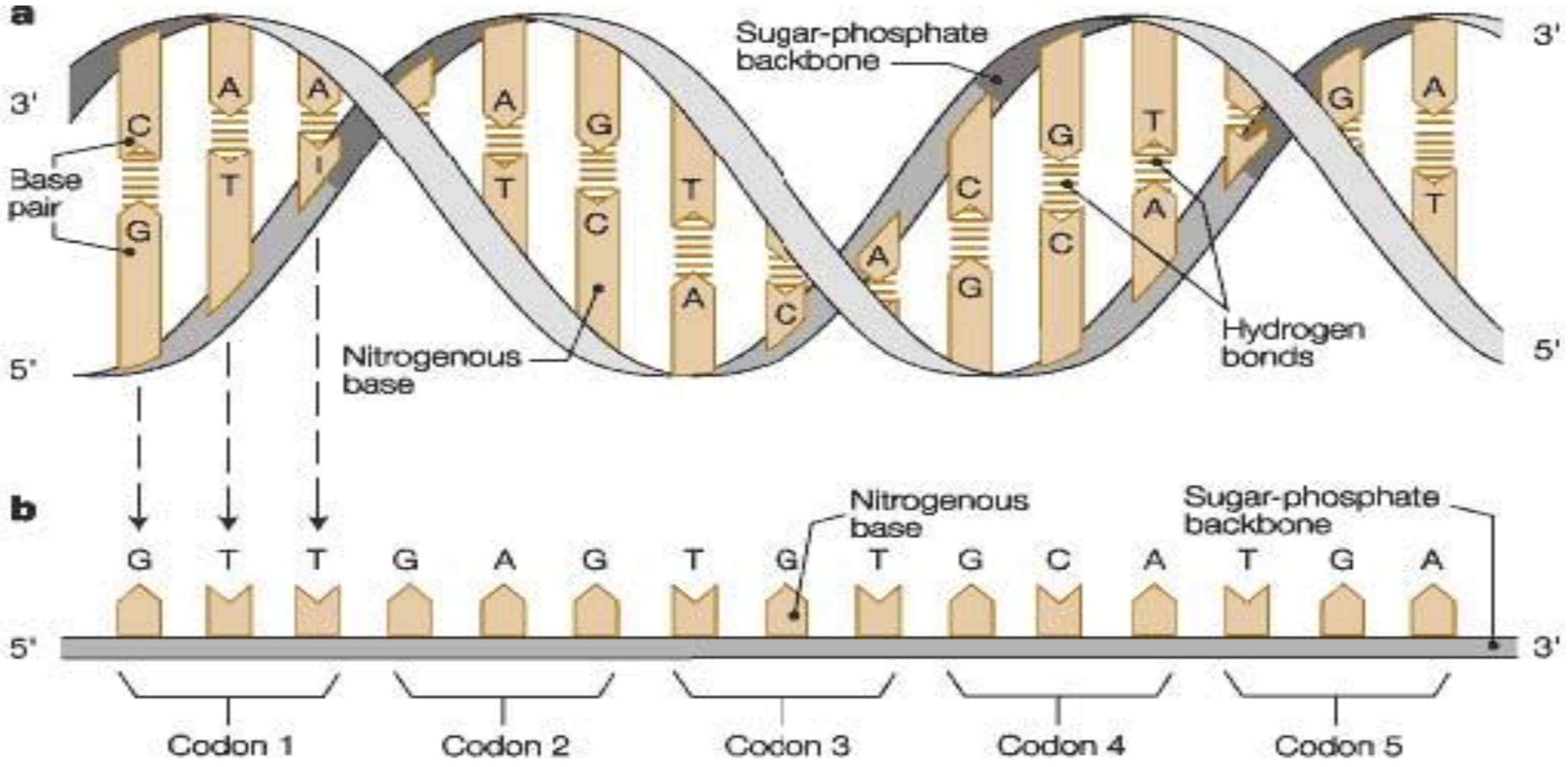


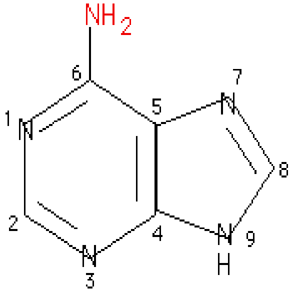


The Post Genomic Challenge

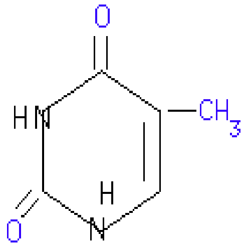
Developing A Molecular level understanding of the entire Organism


Smooth Endoplasmic Reticulum
Nucleus
Microfilaments
Golgi Complex
Mitochondrion
Lysosome
Nucleolus
Chromatin
Rough Endoplasmic Reticulum
Vesicle
Microtubules
Plasma Membrane

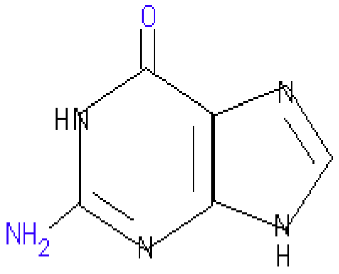




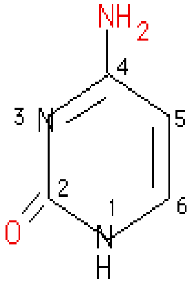
Adenine



Thymine

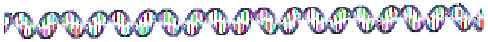


Guanine



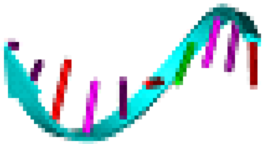
Cytosine

Central Dogma of Life…



DNA (Genome)

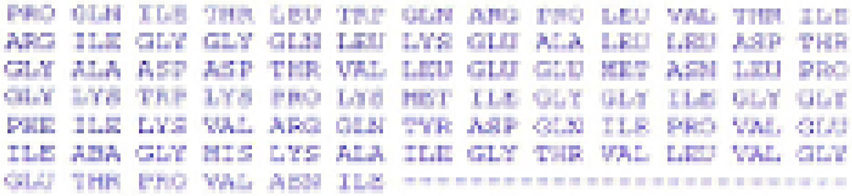
Transcription



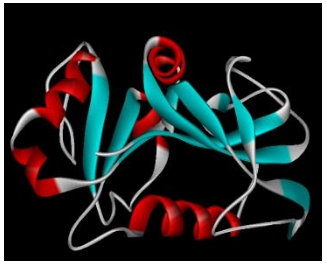
RNA

Translation

PROTEIN



```
PRO GLN ILE THR LEU TRP GLN ARG PRO LEU VAL THR ILE
ARG ILE GLY GLY GLN LEU LYS GLU ALA LEU LEU ASP THR
GLY ALA ASP ASP THR VAL LEU GLU GLU MET ASN LEU PRO
GLY LYS TRP LYS PRO LYS MET ILE GLY GLY ILE GLY GLY
PHE ILE LYS VAL ARG GLN TYR ASP GLN ILE PRO VAL GLU
ILE ABA GLY HIS LYS ALA ILE GLY THR VAL LEU VAL GLY
GLU THR PRO VAL ASN ILE ---------------------------
```



Genome sizes of some organisms

Organism	***Genome size***
	***((Mb)** (Mb=Mega base)*
Eschericia coli	4.6
Sacchromyces cerevisiae (Yeast)	15
M tuberculosis	4.4
H.Influenza	1.83
C. elegans (Nematode)	100
Drosophila melanogaster (Fruit fly)	120
Gallus gallus (Chicken)	120
Homo sapiens (humans)	3300
Mouse	3000
Rice	430
Wheat	13500

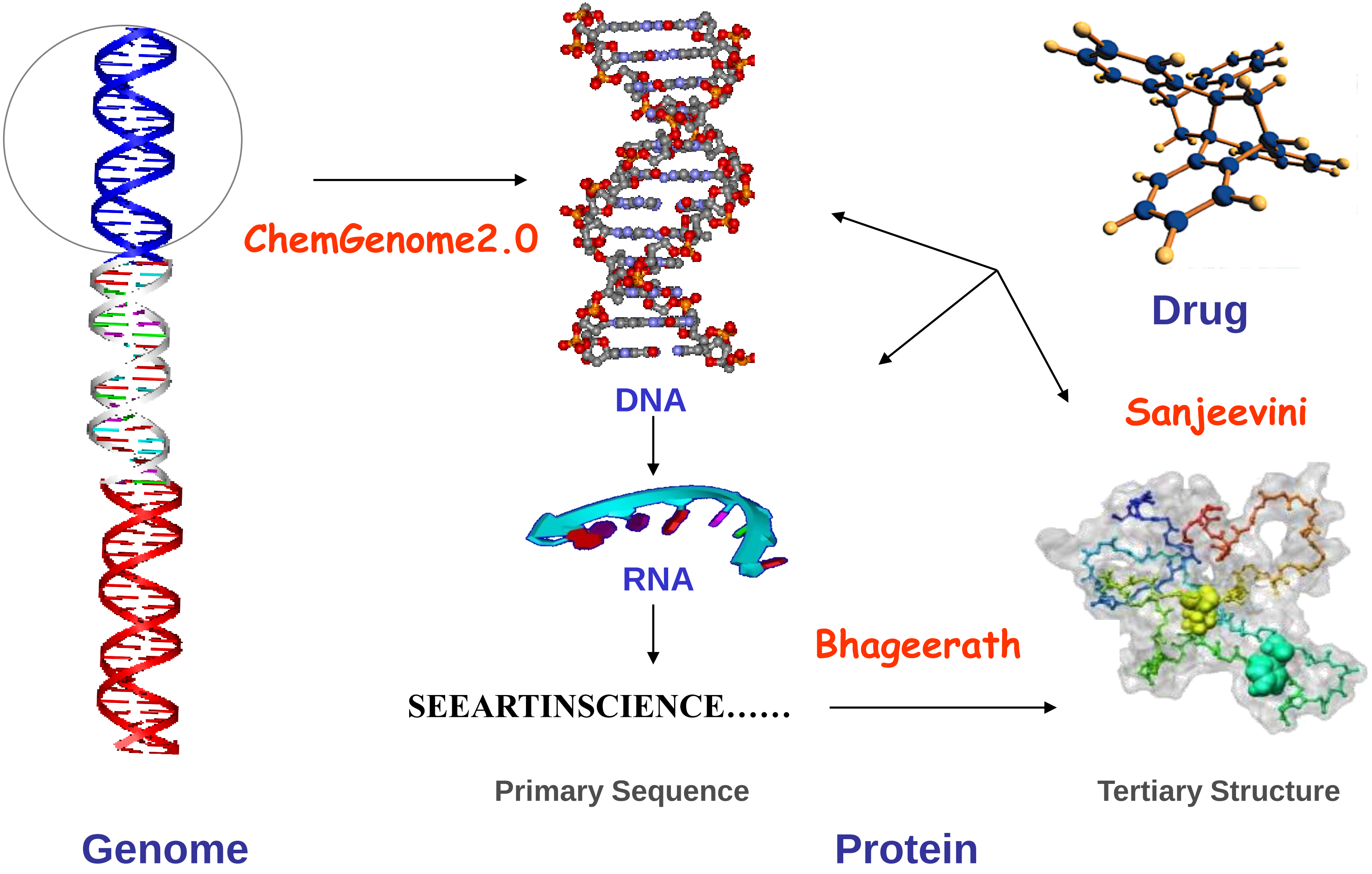
(source: www.wormlab.caltech.edu/briggsae/genomeSize.html)

Specific genetic disorders

Genetic Disorder	*Reason*
Huntington’s Disease	Excessive repeats of a three-base sequence, "CAG" on chromosome
Parkinson’s Disease	Variations in genes on chromosomes 4,6.
Sickle Cell	DiseaseMutation in hemoglobin-b gene on chromosome 11
Tay-Sachs Disease	Controlled by a pair of genes on chromosome 15
Cystic Fibrosis	Mutations in a single (CFTR) gene
Breast Cancer	Mutation on genes found on chromosomes 13 & 17
Leukemia	Exchange of genetic material between the long arms of chromosome 6 & 22
Colon cancer	Proteins MSH2, MSH6 on chromosome 2 & MLH1 on chromosome 3 are mutated.
Asthma	Disfunctioning of genes on chromosome 5, 6, 11, 14&12
Rett Syndrome	Disfunctioning of a gene on the X chromosome.
Brukitt lymphoma	Translocations on chromosome 8
Alzheimer disease	Mutations on four genes located on chromosome 1, 14, 19 & 21.
Werner Syndrome	Mutations on genes located on chromosome 8
Angelman Syndrome	Deletion of a segment on maternally derived chromosome 15.

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

From Gene to Drug : The Dream @ SCFBio



ChemGenome2.0

DNA

RNA

SEEARTINSCIENCE……

Primary Sequence

Genome

Drug

Sanjeevini

Bhageerath

Tertiary Structure

Protein

- **Genome Analysis - *ChemGenome***

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

- Protein Structure Prediction – *Bhageerath*

A *de novo* energy based protein structure prediction software

- Drug Design – *Sanjeevini*

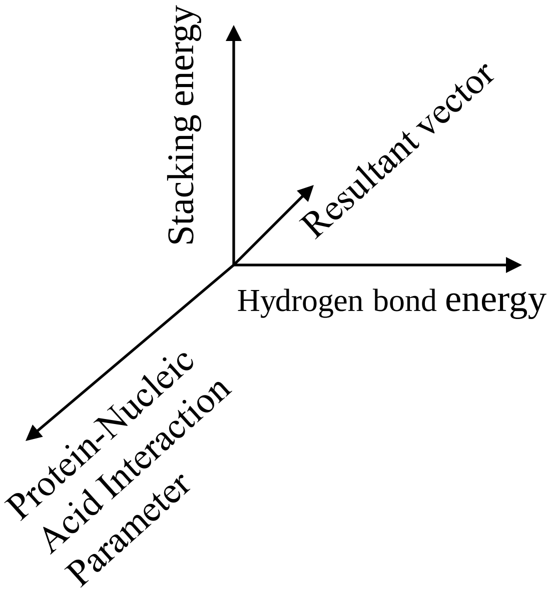
A comprehensive indigenous active site directed lead molecule design protocol

Arabidopsis Thaliana
(Thale Cress)

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

ChemGenome

A Physico-Chemical Model to Distinguish Genes from Non-Genes

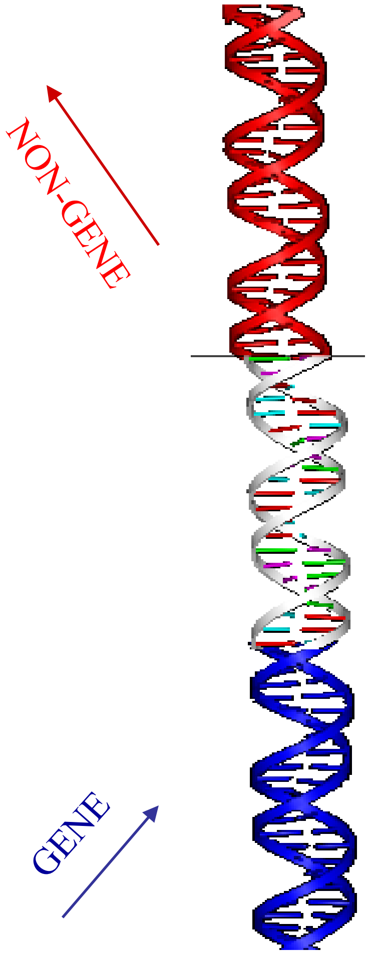


Stacking energy

Resultant vector

Hydrogen bond energy

Protein-Nucleic Acid Interaction Parameter



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

i……l
j…..m
k…..n

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

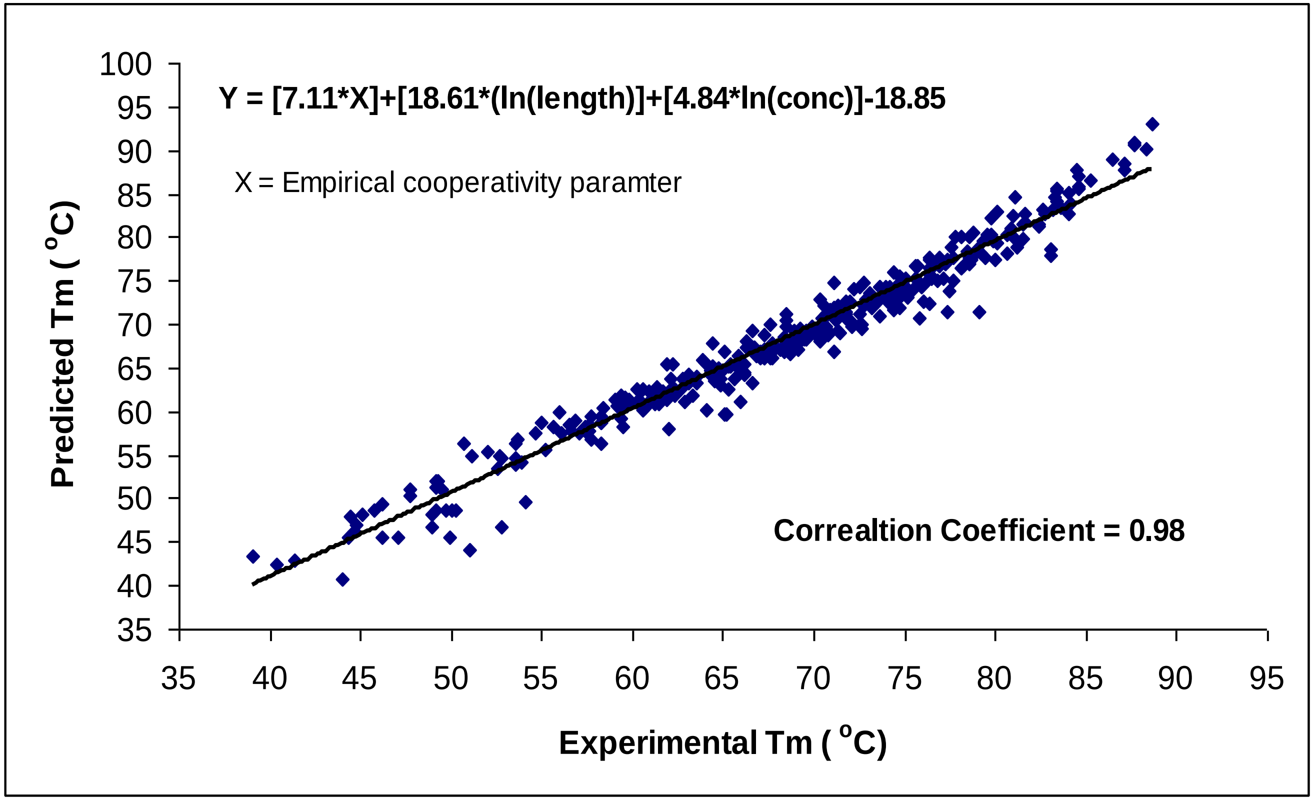
$$E_{Stack} = (E_{i\text{-}m}+E_{i\text{-}n}) + (E_{j\text{-}l}+E_{j\text{-}n}) + (E_{k\text{-}l}+E_{k\text{-}m}) +(E_{i\text{-}j}+E_{i\text{-}k}+ E_{j\text{-}k}) + (E_{l\text{-}m}+E_{l\text{-}n}+ E_{m\text{-}n})$$

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

[*]Beveridge et al. (2004). ***Biophys J*** **87, 3799-813.**

[#]Dixit et al. (2005). ***Biophys J*** **89, 3721-40.**

Prediction of Melting Temperatures of 348 Oligonucleotides (Theory vs Experiment)



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

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$_{1,2}$:

eg. A$_2$ x C$_2$ & G$_2$ x U$_2$

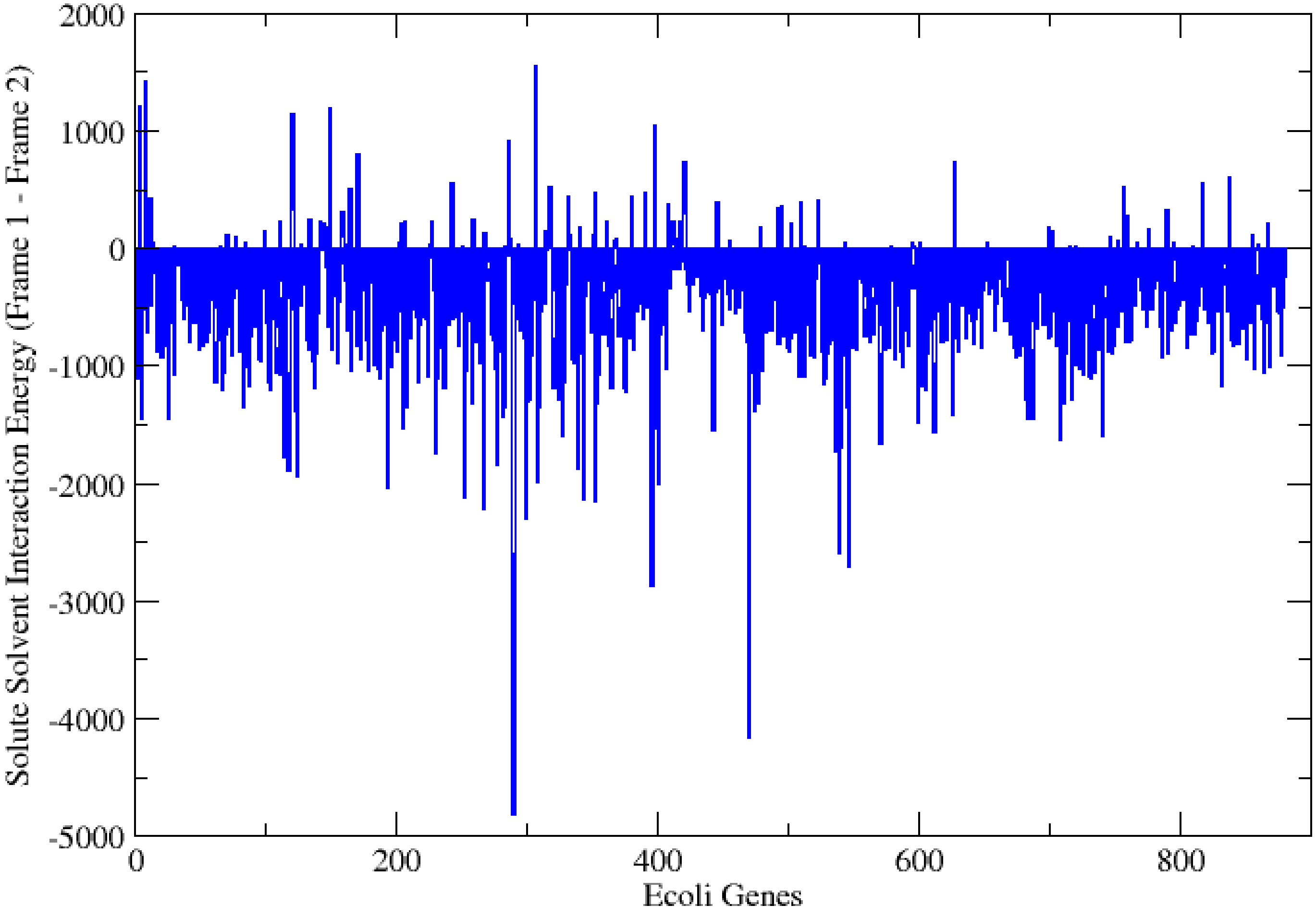
With 6 h-bonds at positions 1 and 2 between codon and anticodon, third base is inconsequential

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

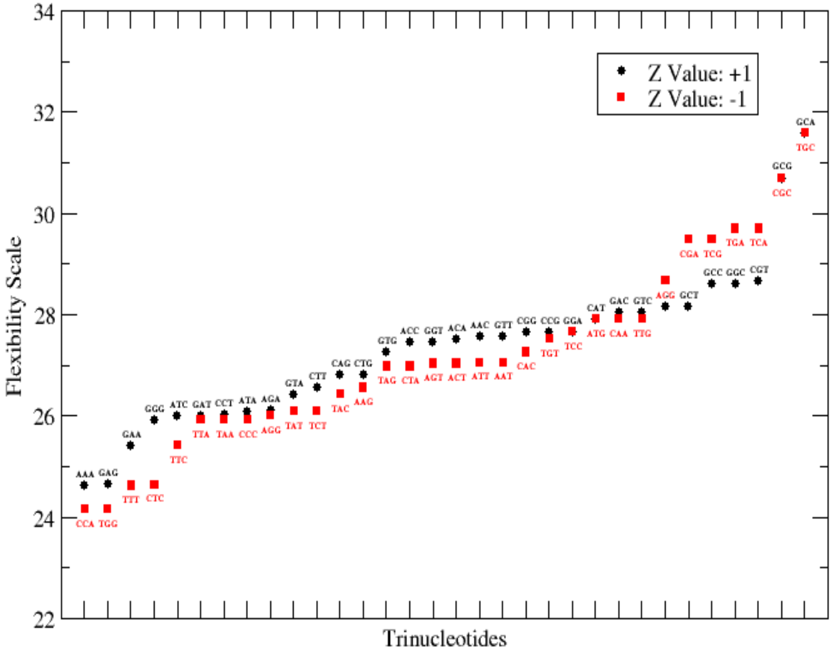
With 5 h-bonds middle pyrimidine renders third base inconsequential; middle purine requires third base.

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

Solute-Solvent Interaction Energy for Genes/Non-genes

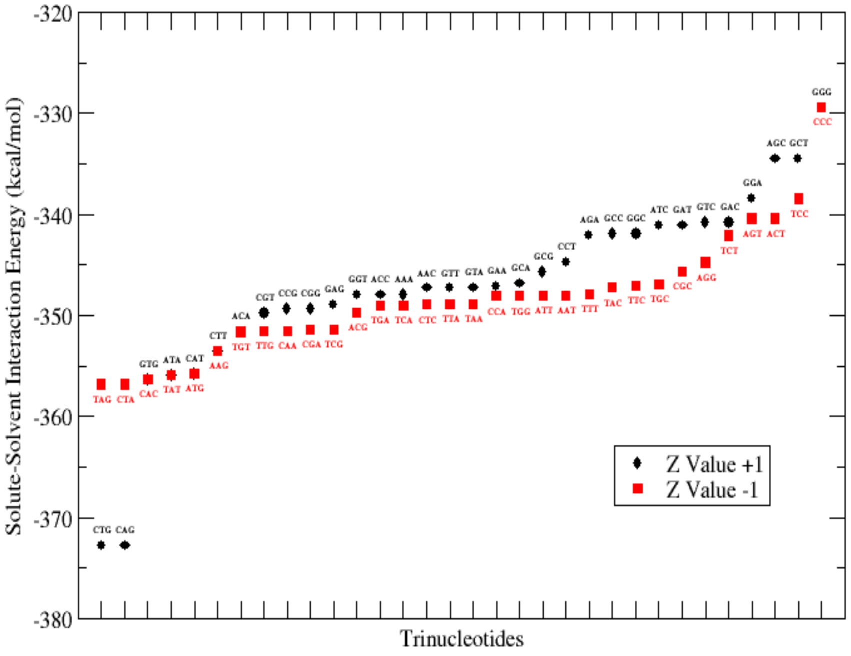


Correlation of Protein-Nucleic Acid Interaction Parameter (Z) with Physical Properties of Codons

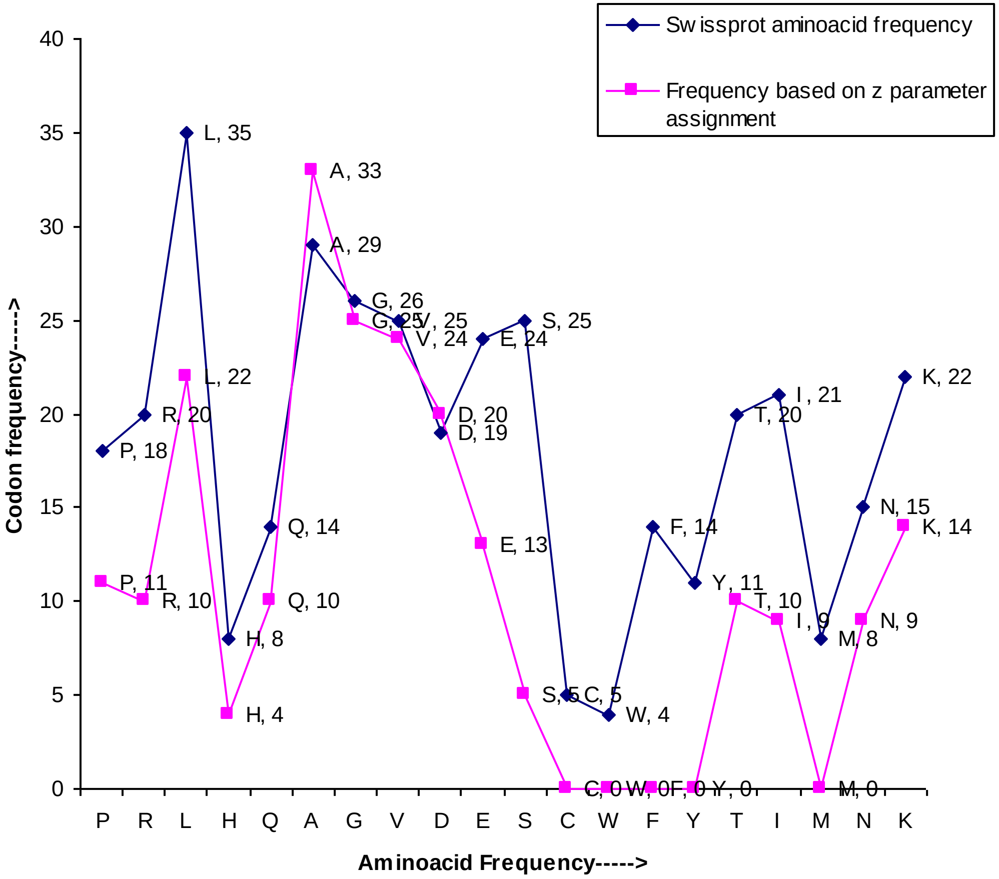


Flexibility of Trinucleotides Based on Molecular Dynamics Simulations

Solute-Solvent Interaction Energy of Trinucleotides Based on Molecular Dynamics Simulations

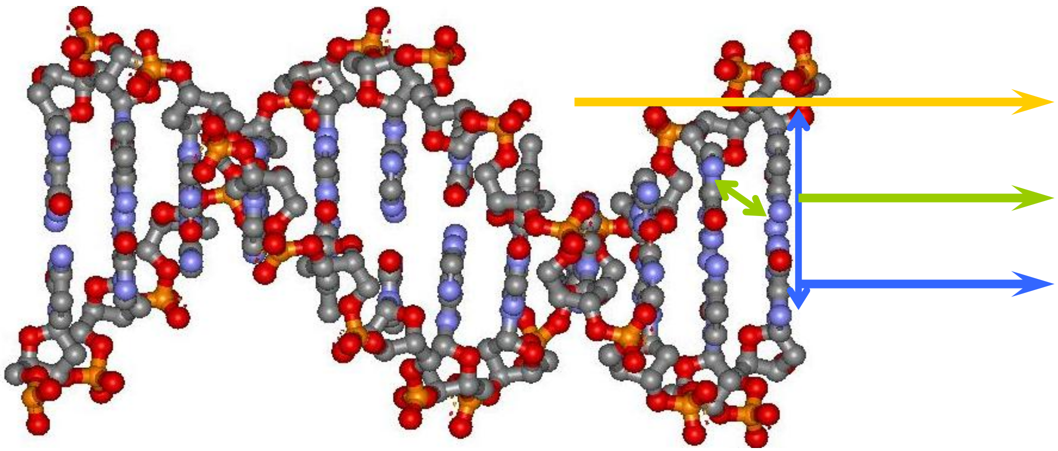


Swissprot Amino acid Frequency for 175000 Proteins vs. Codon Frequency Based on Protein-Nucleic Acid Interaction Parameter (Z) Assignment



ChemGenome

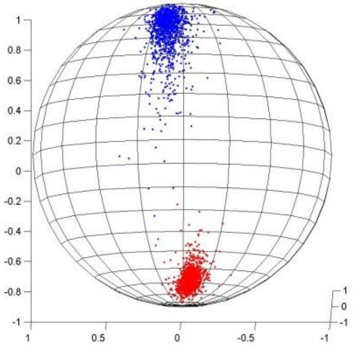
A Physico-Chemical Model to Distinguish Genes from Non-Genes

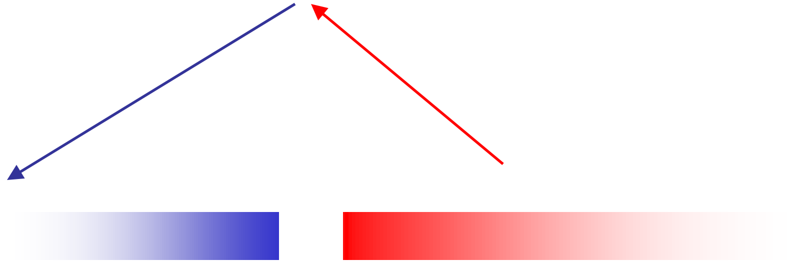


Protein-nucleic acid interaction propensity parameter

Stacking Energy

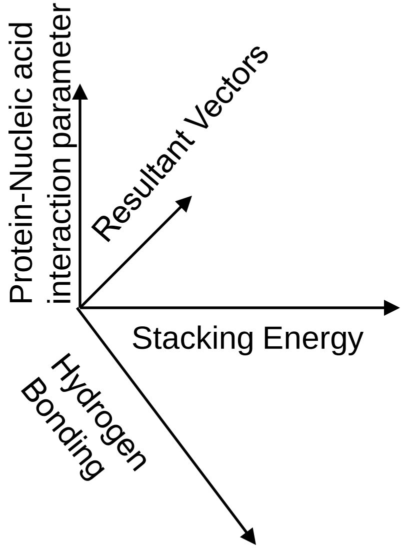
Hydrogen Bonding





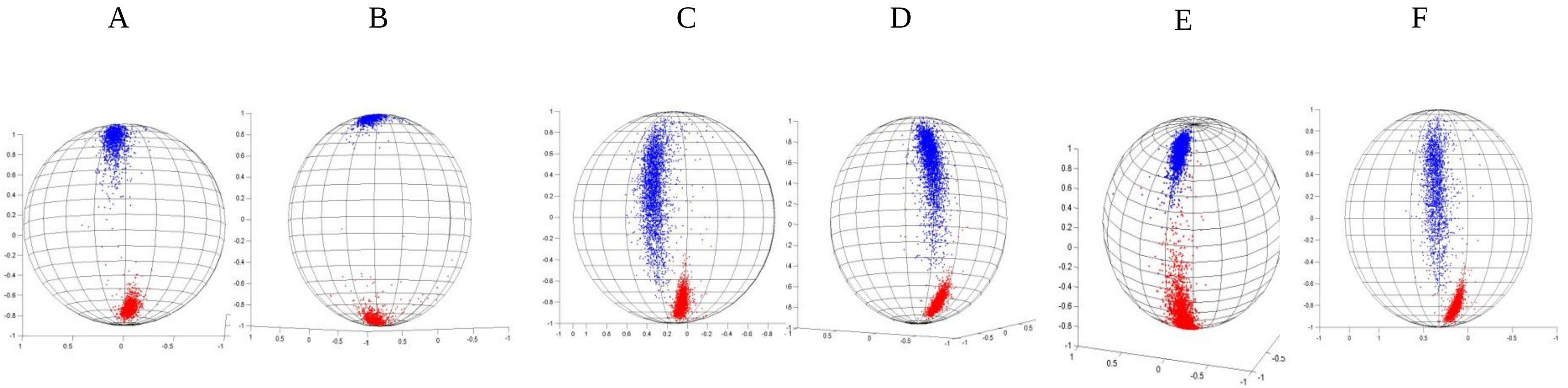
Gene

Non Gene



"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.**

Distinguishing Genes (blue) from Non-Genes (red) in 372 Prokaryotic Genomes



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

Computational Protocol Designed for Gene Prediction

Read the complete genome sequence in the FASTA format

↓

Search for all possible ORFs in all the six reading frames

↓

Calculate resultant unit vector for each of the ORFs

↓

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

↓

Genes and false positives

↓

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

↓

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

↓

Potential protein coding genes

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

Genes Predicted using *ChemGenome2.0* for Prokaryotic Genomes

S.No.	NCBI_ID	Initial Orfs	SS	SP	ChemGenome (DNA Space)	SS	SP	Chemgenome (Protein Space)	SS	SP	Chemgenome (Swissprot Space)	Annotated Genes	SS	SP
1	NC_000117	6773	99.78	13.18	4558	98.32	19.31	2135	94.97	39.81	1284	895	92.07	64.17
2	NC_000853	15104	99.46	12.24	10688	99.30	17.26	4991	96.50	35.92	3037	1858	92.47	56.57
3	NC_000854	11774	99.95	15.63	9616	98.59	18.87	5273	91.31	31.88	2282	1841	80.66	65.07
4	NC_000868	11066	99.89	17.12	6598	98.95	28.43	3524	96.62	51.99	2232	1896	90.08	76.52
5	NC_000907	11945	99.64	13.82	6582	96.68	24.34	3064	92.76	50.16	1926	1657	90.53	77.88
6	NC_000908	3334	70.66	10.26	1930	63.84	16.01	1035	48.55	22.71	602	484	43.80	35.22
7	NC_000909	7829	99.54	21.98	3786	98.67	45.06	2450	96.59	68.16	1488	1729	80.05	93.01
8	NC_000911	28534	99.91	11.09	20656	98.48	15.10	10459	95.17	28.82	5891	3167	92.86	49.92
9	NC_000912	5998	75.47	8.67	3628	67.05	12.73	1577	51.38	22.45	935	689	44.99	33.16
10	NC_000913	41399	99.54	10.36	30642	99.10	13.94	15618	97.01	26.78	8500	4311	94.22	47.79
11	NC_000915	9647	98.29	16.06	5829	96.38	26.06	3227	90.04	43.97	1807	1576	85.98	74.99
12	NC_000916	14586	99.89	12.83	10537	99.47	17.68	6315	97.17	28.82	3024	1873	91.40	56.61
13	NC_000917	17584	99.13	13.64	11988	98.64	19.91	6121	96.32	38.08	3584	2420	90.08	60.83
14	NC_000918	10140	100.00	15.08	6591	99.87	23.17	2784	97.65	53.63	1749	1529	91.24	79.76
15	NC_000919	11875	99.71	8.70	8694	98.75	11.77	4200	93.92	23.17	2165	1036	90.06	43.09
16	NC_000921	9384	98.86	15.71	5682	97.72	25.64	3155	92.49	43.71	1763	1491	89.20	75.44
17	NC_000922	7505	99.91	14.03	5040	98.01	20.50	2484	93.83	39.81	1504	1054	90.70	63.56
18	NC_000961	10026	99.95	19.50	5869	96.98	32.32	3317	93.56	55.17	2096	1956	86.04	80.30
19	NC_000962	45751	99.82	8.73	39813	99.82	10.03	21629	96.05	17.76	6342	3999	85.47	53.89
20	NC_000963	4307	100.00	19.39	2148	96.77	37.62	1271	93.05	61.13	805	835	85.87	89.07

Prediction accuracies of translation start sites using *ChemGenome2.0* on reliable datasets as test sets

Organism	Test sets	No. of genes in test set	Accurate start predictions (%)				
			Glimmer	GS-Finder	MED-Start	*ChemGenome2.0* (DNA space)	*ChemGenome2.0* (DNA+Protein space)
E.coli	Ecogene	854	63.23	91.1	92.9	96.9	94.3
	Link	195	66.67	92.3	95.4	99.5	95.9
B.subtilis	Bsub1248	1248	61.30	-	90.1	99.5	92.5
	Bsub58	58	68.96	96.6	96.6	100.0	98.3
	Bsub123	123	48.78	83.7	87.8	100.0	81.3
	Bsub72	72	48.61	90.3	93.1	100.0	84.7
	Bsub51	51	41.76	92.2	96.1	100.0	86.3

Accuracy of *ChemGenome2.0* in locating the start and stop positions without a prior knowledge of start and stop sites

S.No.	Genome version	Number of experimentally verified genes	Percentage of genes whose start site is identified to within			Percentage of genes whose stop site is identified to within		
			±10 bases	±20 bases	±30 bases	±10 bases	±20 bases	±30 bases
1.	NC_000117.1	602	66.0	78.0	83.1	60.8	79.0	85.0
2.	NC_000853.1	1084	91.0	95.0	96.3	87.2	93.1	97.1
3.	NC_002570.2	2143	82.0	90.2	93.3	82.0	89.0	93.0

ChemGenome Performance

Gene evaluation data	Accuracy
372 Prokaryotic genomes for experimentally verified genes	**96%**
21 eukaryotic genomes for experimentally verified tRNA genes.	**97%**
21 eukaryotic genomes for experimentally verified genes.	**82%**

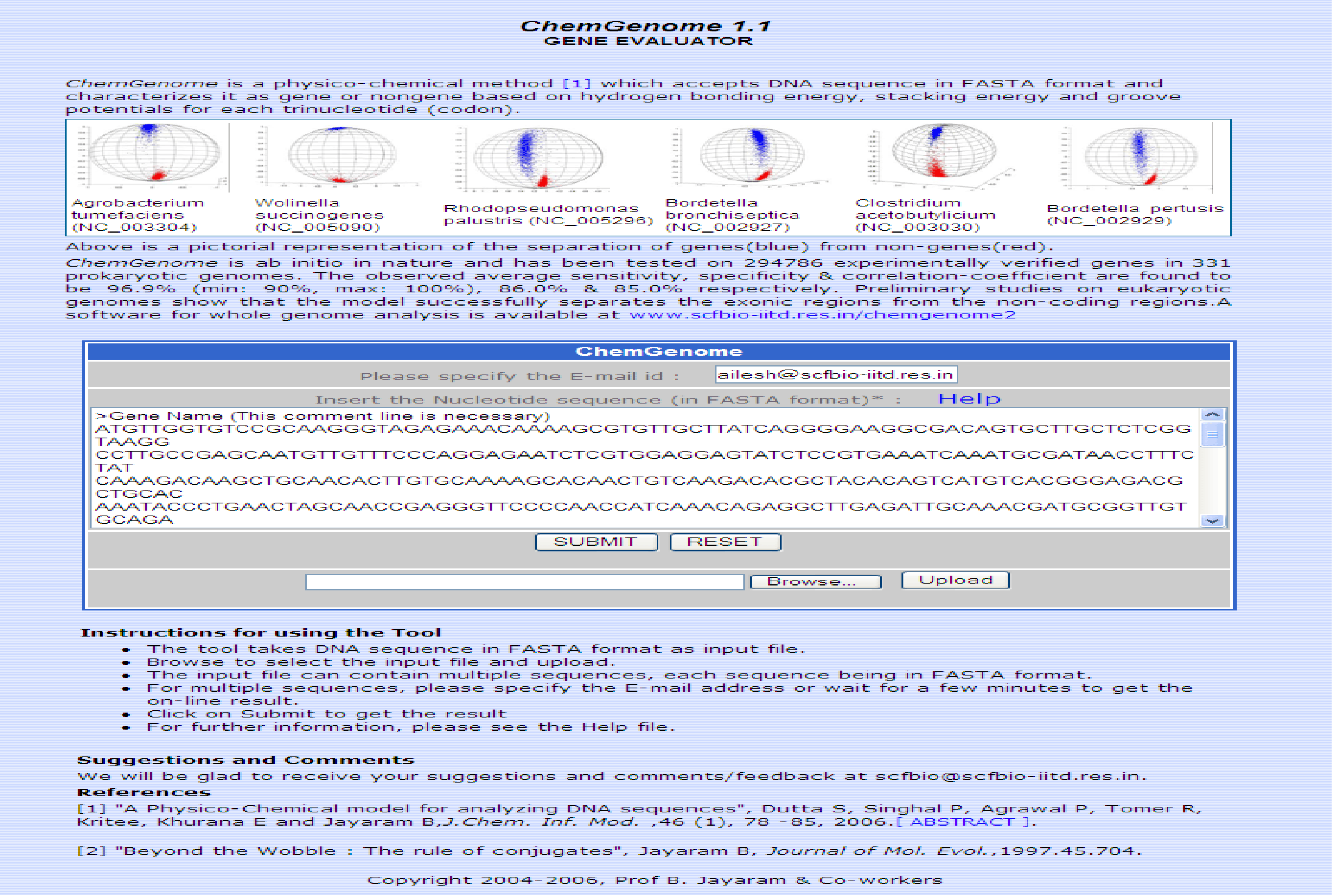
Software	Tested on Bacteria	Accuracy
ChemGenome www.scfbio-iitd.res.in/chemgenome	372 systems	96.94%
GeneMark www.ebi.ac.uk/genemark	7 systems	94.96%
Glimmer www.tigr.org/software/glimmer/	31 systems	99.36%
FgenesB www.softberry.com	1 system	98%

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

SCFBio

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).

Agrobacterium tumefaciens (NC_003304)
Wolinella succinogenes (NC_005090)
Rhodopseudomonas palustris (NC_005296)
Bordetella bronchiseptica (NC_002927)
Clostridium acetobutylicium (NC_003030)
Bordetella pertusis (NC_002929)

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)
ATGTTGGTGTCCGCAAGGGTAGAGAAACAAAAAGCGTGTTGCTTATCAGGGGAAGGCGACAGTGCTTGCTCTCGG
TAAGG
CCTTGCCGAGCAATGTTGTTTCCCAGGAGAATCTCGTGGAGGAGTATCTCCGTGAAATCAAATGCGATAACCTTTC
TAT
CAAAGACAAGCTGCAACACTTGTGCAAAAAGCACAACTGTCAAGACACGCTACACAGTCATGTCACGGGAGACG
CTGCACAC
AAATACCCTGAACTAGCAACCGAGGGTTCCCCAACCATCAAAACAGAGGCTTGAGATTGCAAACGATGCGGTTGT
GCAGA
```

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.

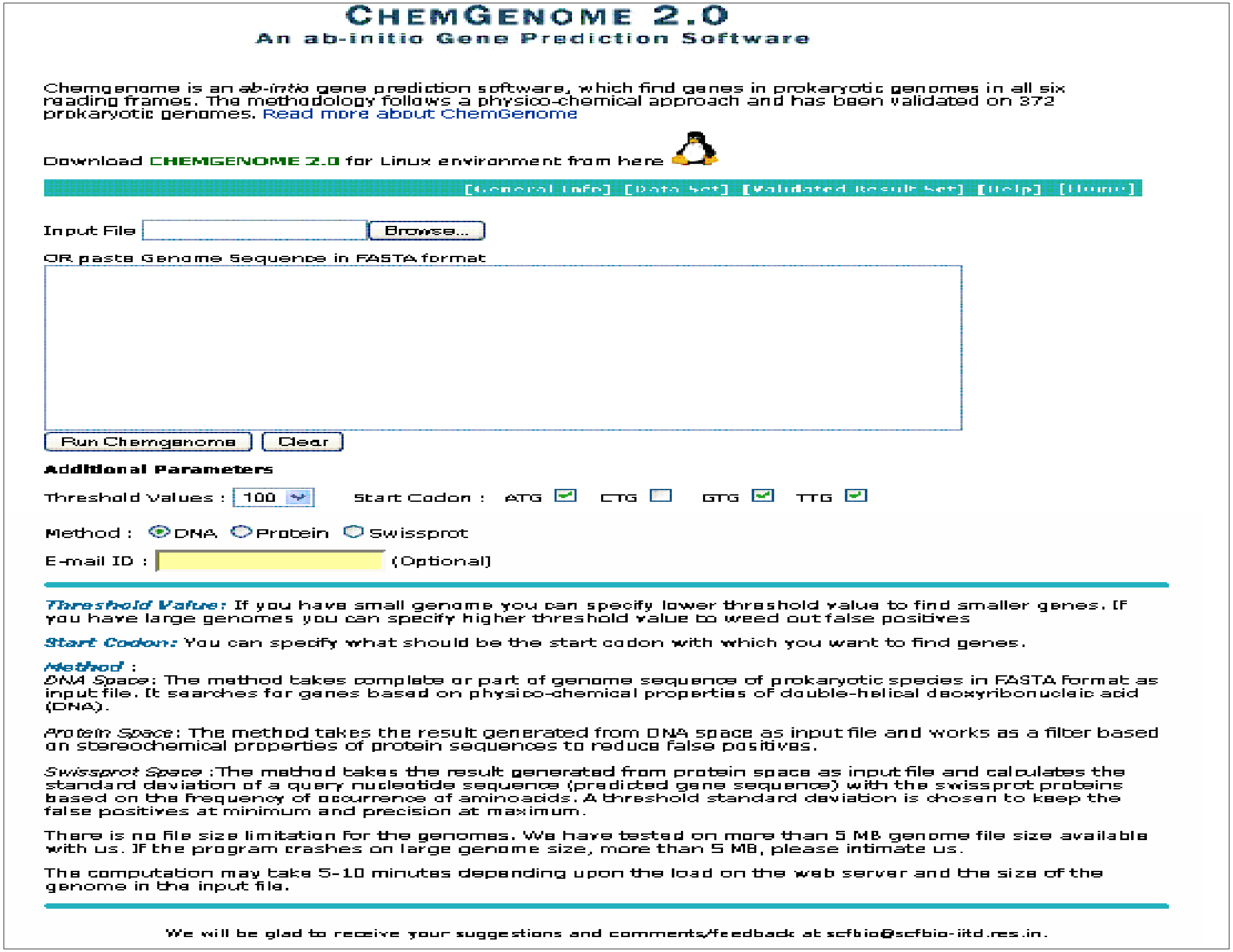
References

[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



CHEMGENOME 2.0

An ab-initio Gene Prediction Software

Chemgenome is an *ab-initio* gene prediction software, which find genes in prokaryotic genomes in all six reading frames. The methodology follows a physico-chemical approach and has been validated on 372 prokaryotic genomes. Read more about ChemGenome

Download CHEMGENOME 2.0 for Linux environment from here

[General Info] [Data Set] [Validated Result Set] [Help] [Home]

Input File [ ] Browse...

OR paste Genome Sequence in FASTA format

Additional Parameters

Threshold Values : 100

Start Codon : ATG ☑ CTG ☐ GTG ☑ TTG ☑

Method : ● DNA ○ Protein ○ Swissprot

E-mail ID : (Optional)

Threshold Value: If you have small genome you can specify lower threshold value to find smaller genes. If you have large genomes you can specify higher threshold value to weed out false positives

Start Codon: You can specify what should be the start codon with which you want to find genes.

Method :

DNA Space: The method takes complete or part of genome sequence of prokaryotic species in FASTA format as input file. It searches for genes based on physico-chemical properties of double-helical deoxyribonucleic acid (DNA).

Protein Space: The method takes the result generated from DNA space as input file and works as a filter based on stereochemical properties of protein sequences to reduce false positives.

Swissprot Space :The method takes the result generated from protein space as input file and calculates the standard deviation of a query nucleotide sequence (predicted gene sequence) with the swissprot proteins based on the frequency of occurrence of aminoacids. A threshold standard deviation is chosen to keep the false positives at minimum and precision at maximum.

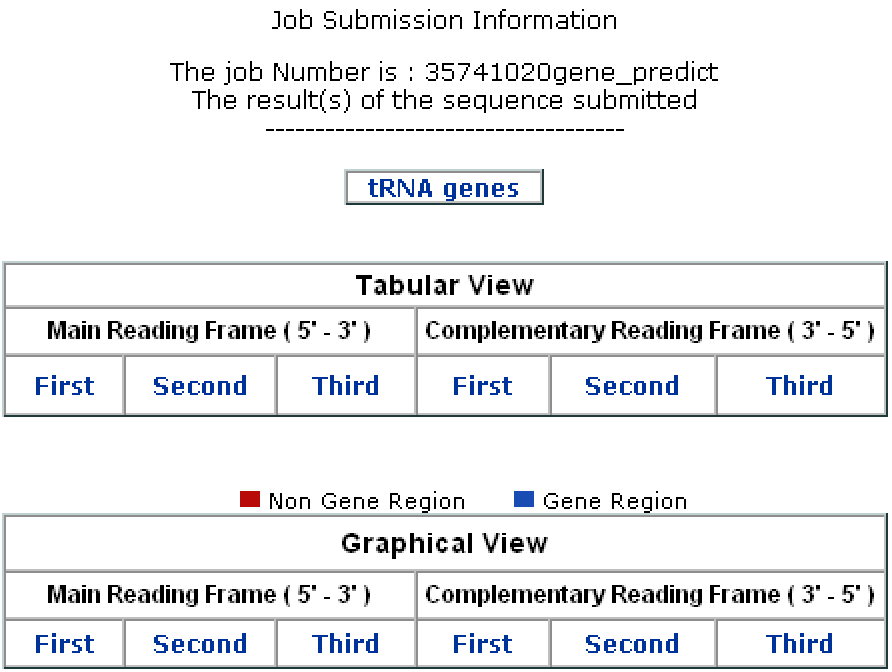
There is no file size limitation for the genomes. We have tested on more than 5 MB genome file size available with us. If the program crashes on large genome size, more than 5 MB, please intimate us.

The computation may take 5-10 minutes depending upon the load on the web server and the size of the genome in the input file.

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

Results obtained on Aeropyrum pernix (NCBI ID: NC_000854)

http://www.scfbio-itd.res.in/chemgenome/status_chemtrna.jsp?jobid=35741020gene_predict&threshold=100&method=3&email=



Job Submission Information

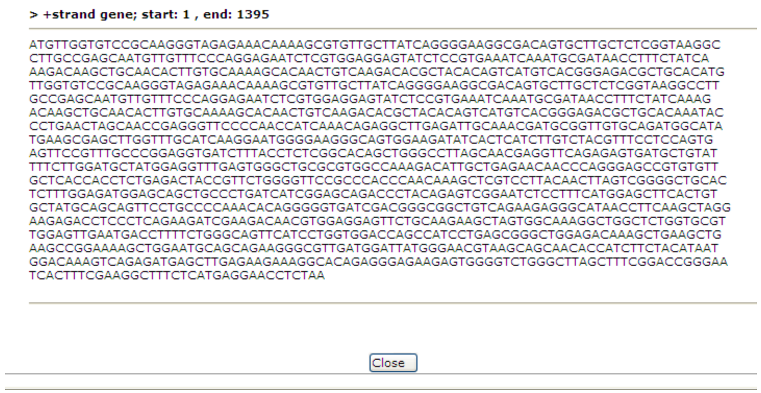
The job Number is : 35741020gene_predict
The result(s) of the sequence submitted

tRNA genes

Tabular View					
Main Reading Frame (5' - 3')			Complementary Reading Frame (3' - 5')		
First	Second	Third	First	Second	Third

🟥 Non Gene Region 🟦 Gene Region

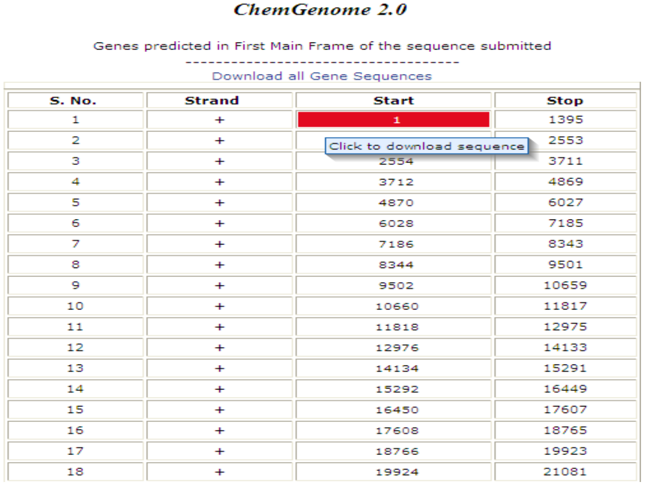
Graphical View					
Main Reading Frame (5' - 3')			Complementary Reading Frame (3' - 5')		
First	Second	Third	First	Second	Third



> +strand gene; start: 1 , end: 1395

```
ATGTTGGTGTCCGCAAGGGTAGAGAAACAAAAGCGTGTTGCTTATCAGGGGAAGGCGACAGTGCTTGCTCTCGGTAAGGC
CTTGCCGAGCAATGTTGTTTCCCAGGAGAATCTCGTGGAGGAGTATCTCCGTGAAATCAAATGCGATAACCTTTCTATCA
AAGACAAGCTGCAACACTTGTGCAAAAGCACAACTGTCAAGACACGCTACACAGTCATGTCACGGGAGACGCTGCACATG
TTGGTGTCCGCAAGGGTAGAGAAACAAAAGCGTGTTGCTTATCAGGGGAAGGCGACAGTGCTTGCTCTCGGTAAGGCCTT
GCCGAGCAATGTTGTTTCCCAGGAGAATCTCGTGGAGGAGTATCTCCGTGAAATCAAATGCGATAACCTTTCTATCAAAG
ACAAGCTGCAACACTTGTGCAAAAGCACAACTGTCAAGACACGCTACACAGTCATGTCACGGGAGACGCTGCACAAATAC
CCTGAACTAGCAACCGAGGGTTCCCCAACCATCAAACAGAGGCTTGAGATTGCAAACGATGCGGTTGTGCAGATGGCATA
TGAAGCGAGCTTGGTTTGCATCAAGGAATGGGGAAGGGCAGTGGAAGATATCACTCATCTTGTCTACGTTTCCTCCAGTG
AGTTCCGTTTGCCCGGAGGTGATCTTTACCTCTCGGCACAGCTGGGCCTTAGCAACGAGGTTCAGAGAGTGATGCTGTAT
TTTCTTGGATGCTATGGAGGTTTGAGTGGGCTGCGCGTGGCCAAAGACATTGCTGAGAACAACCCAGGGAGCCGTGTGTT
GCTCACCACCTCTGAGACTACCGTTCTGGGGTTCCGCCCACCCAACAAAGCTCGTCCTTACAACTTAGTCGGGGCTGCAC
TCTTTGGAGATGGAGCAGCTGCCCTGATCATCGGAGCAGACCCTACAGAGTCGGAATCTCCTTTCATGGAGCTTCACTGT
GCTATGCAGCAGTTCCTGCCCCAAACACAGGGGGTGATCGACGGGCGGCTGTCAGAAGAGGGCATAACCTTCAAGCTAGG
AAGAGACCTCCCTCAGAAGATCGAAGACAACGTGGAGGAGTTCTGCAAGAAGCTAGTGGCAAAGGCTGGCTCTGGTGCGT
TGGAGTTGAATGACCTTTTCTGGGCAGTTCATCCTGGTGGACCAGCCATCCTGAGCGGGCTGGAGACAAAGCTGAAGCTG
AAGCCGGAAAAGCTGGAATGCAGCAGAAGGGCGTTGATGGATTATGGGAACGTAAGCAGCAACACCATCTTCTACATAAT
GGACAAAGTCAGAGATGAGCTTGAGAAGAAAGGCACAGAGGGAGAAGAGTGGGGTCTGGGCTTAGCTTTCGGACCGGGAA
TCACTTTCGAAGGCTTTCTCATGAGGAACCTCTAA
```

Close

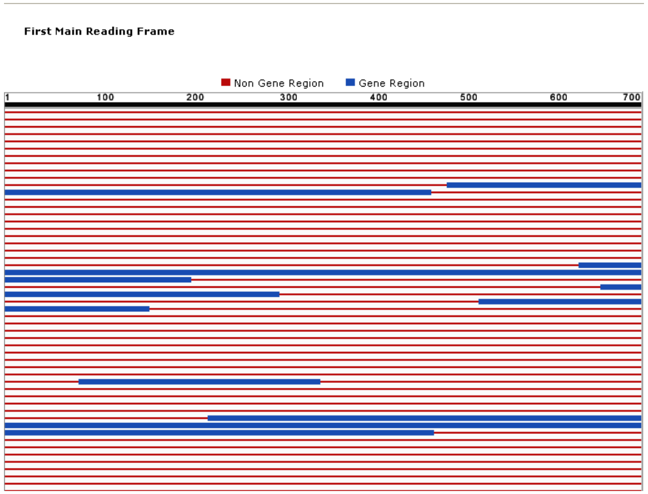


ChemGenome 2.0

Genes predicted in First Main Frame of the sequence submitted

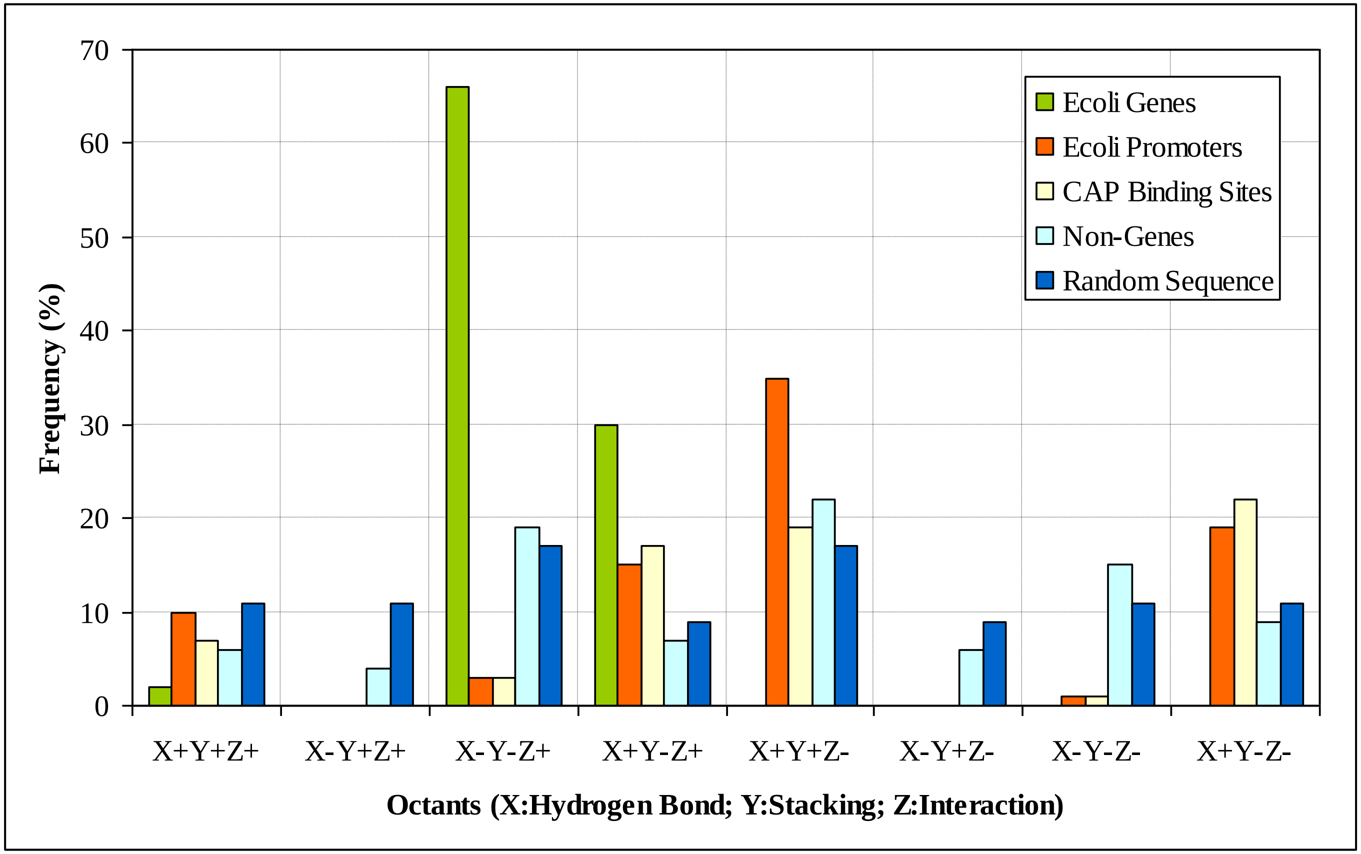
Download all Gene Sequences

S. No.	Strand	Start	Stop
1	+	1	1395
2	+	Click to download sequence	2553
3	+	2554	3711
4	+	3712	4869
5	+	4870	6027
6	+	6028	7185
7	+	7186	8343
8	+	8344	9501
9	+	9502	10659
10	+	10660	11817
11	+	11818	12975
12	+	12976	14133
13	+	14134	15291
14	+	15292	16449
15	+	16450	17607
16	+	17608	18765
17	+	18766	19923
18	+	19924	21081



Towards Designer Genomes?

An Orientational Analysis of Physico-chemical Vectors of DNA

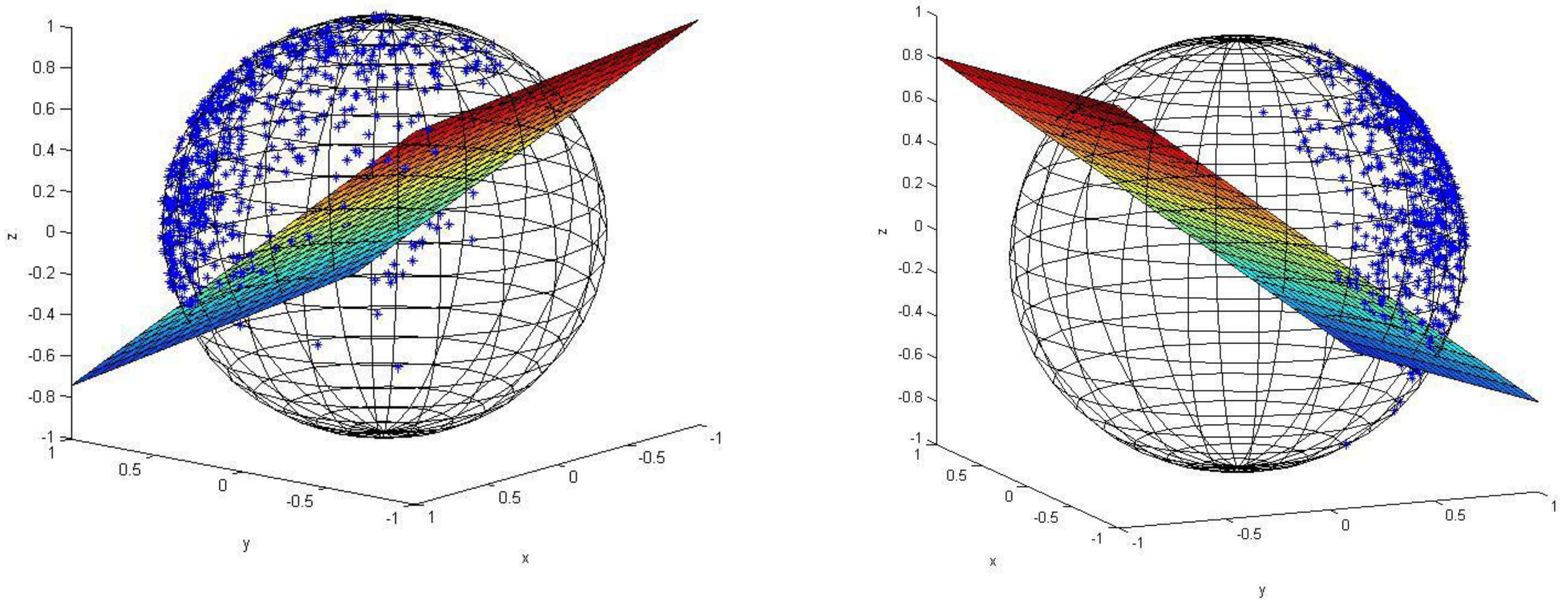


Promoter Prediction Results in E. coli

Method	Sensitivity	Specificity
Chemgenome	**0.959**	**0.734**
TLS-NNPP	**0.452**	**0.188**
NNPP	**0.443**	**0.109**
Novel method *(Manju Bansal & coworkers)*	**0.910**	**0.350**

Chemgenome on Eukaryotes

Exon data plot

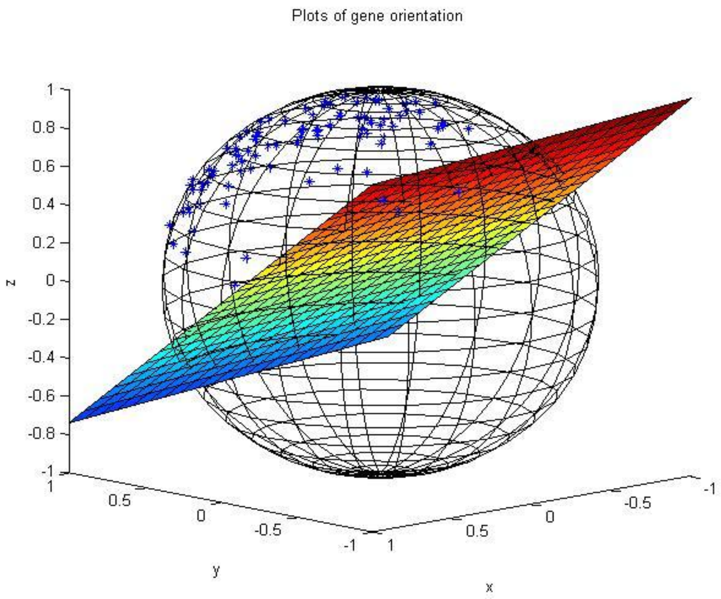


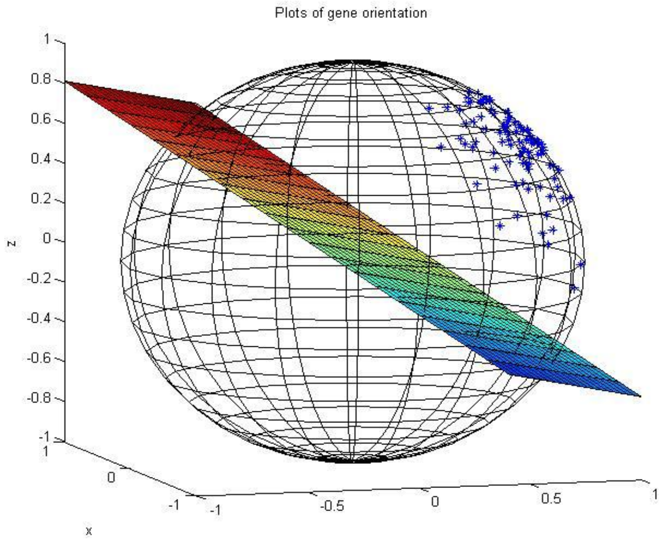
Sensitivity= (645/668)=.9655

Database :Intron Exon database University of toledo
http://hsc.utoledo.edu/depts/bioinfo/database.html

Chemgenome on Eukaryotes

Gene Vectors of Experimentally verified Proteins from SwissProt

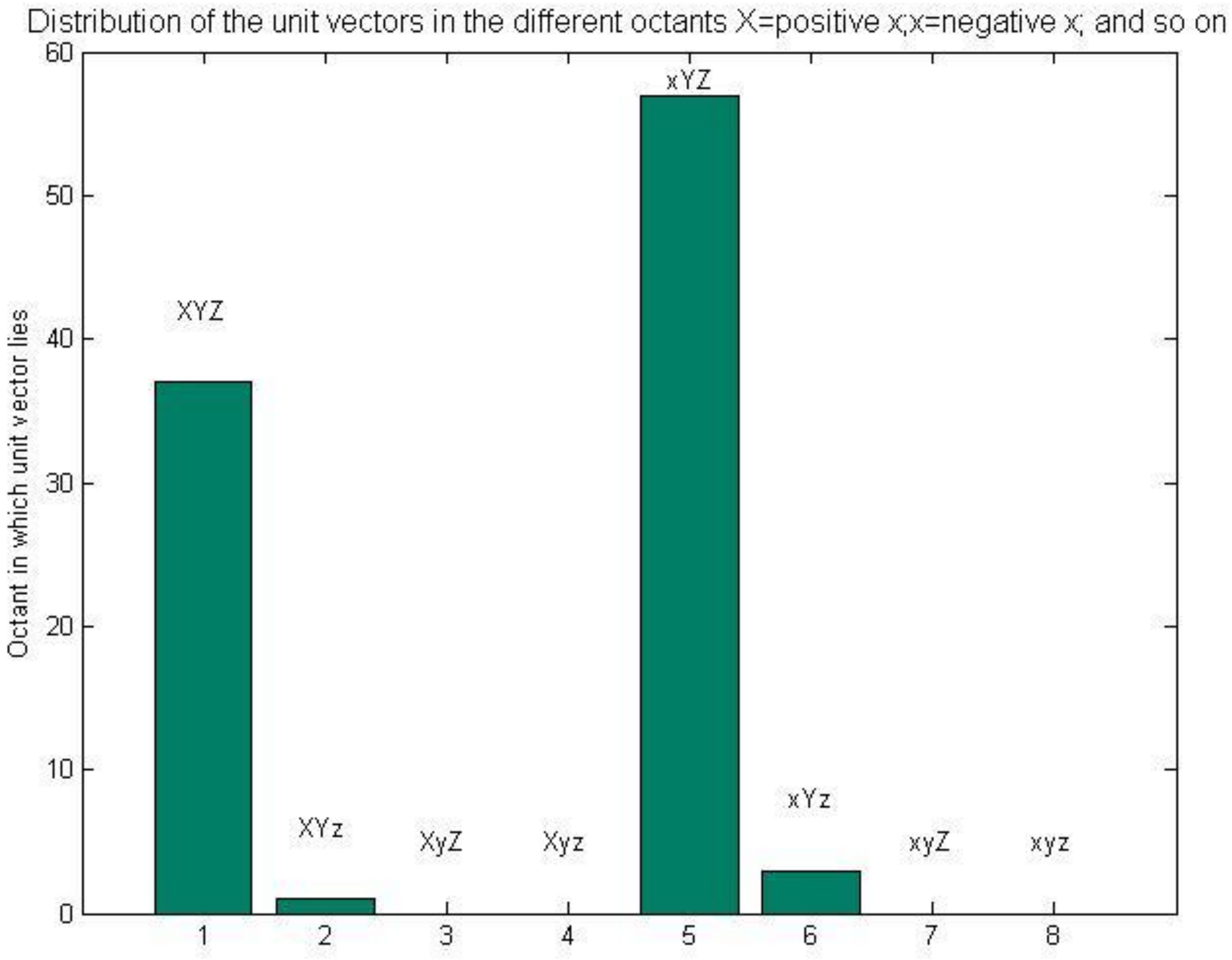


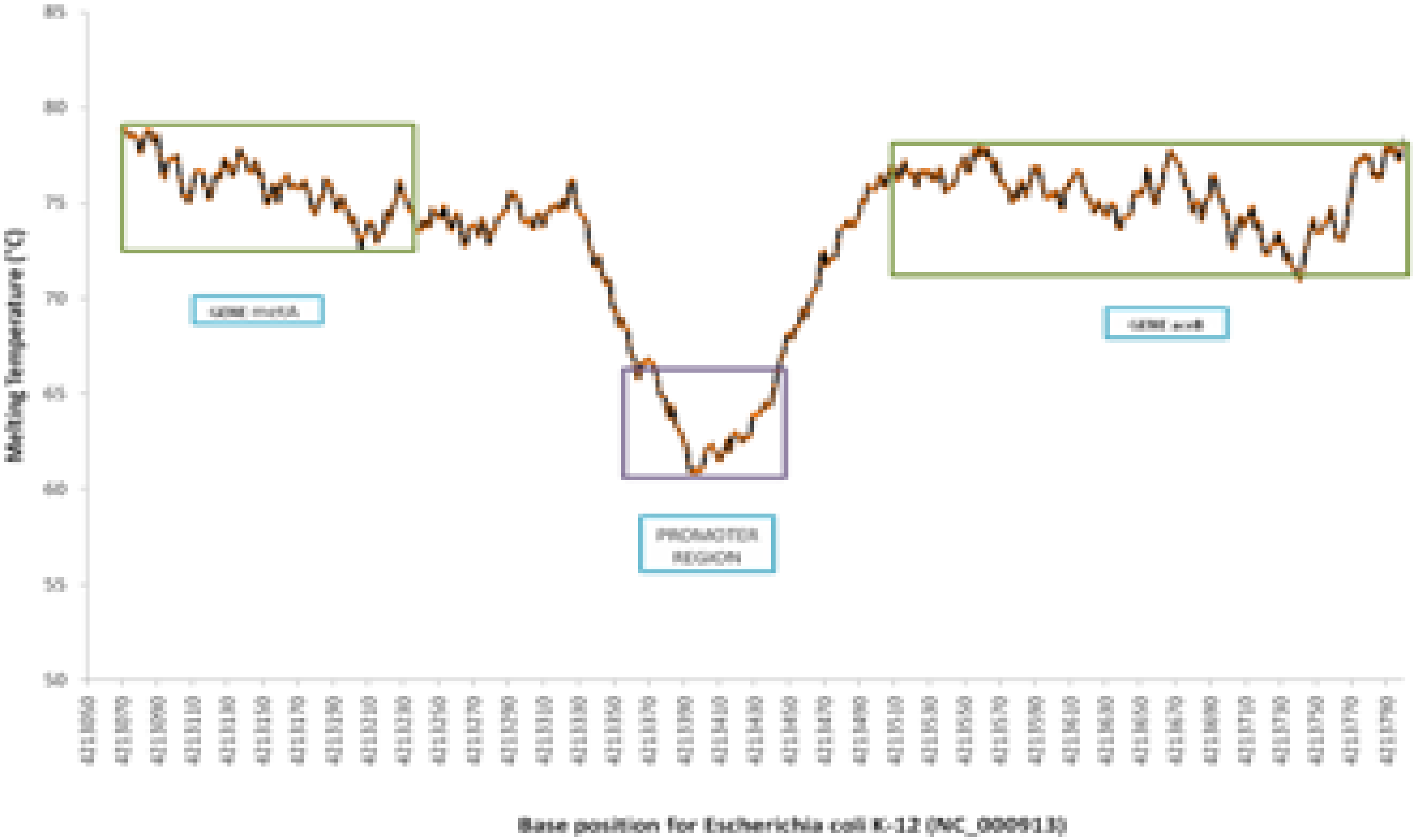


Sensitivity= (97/98)= .9897

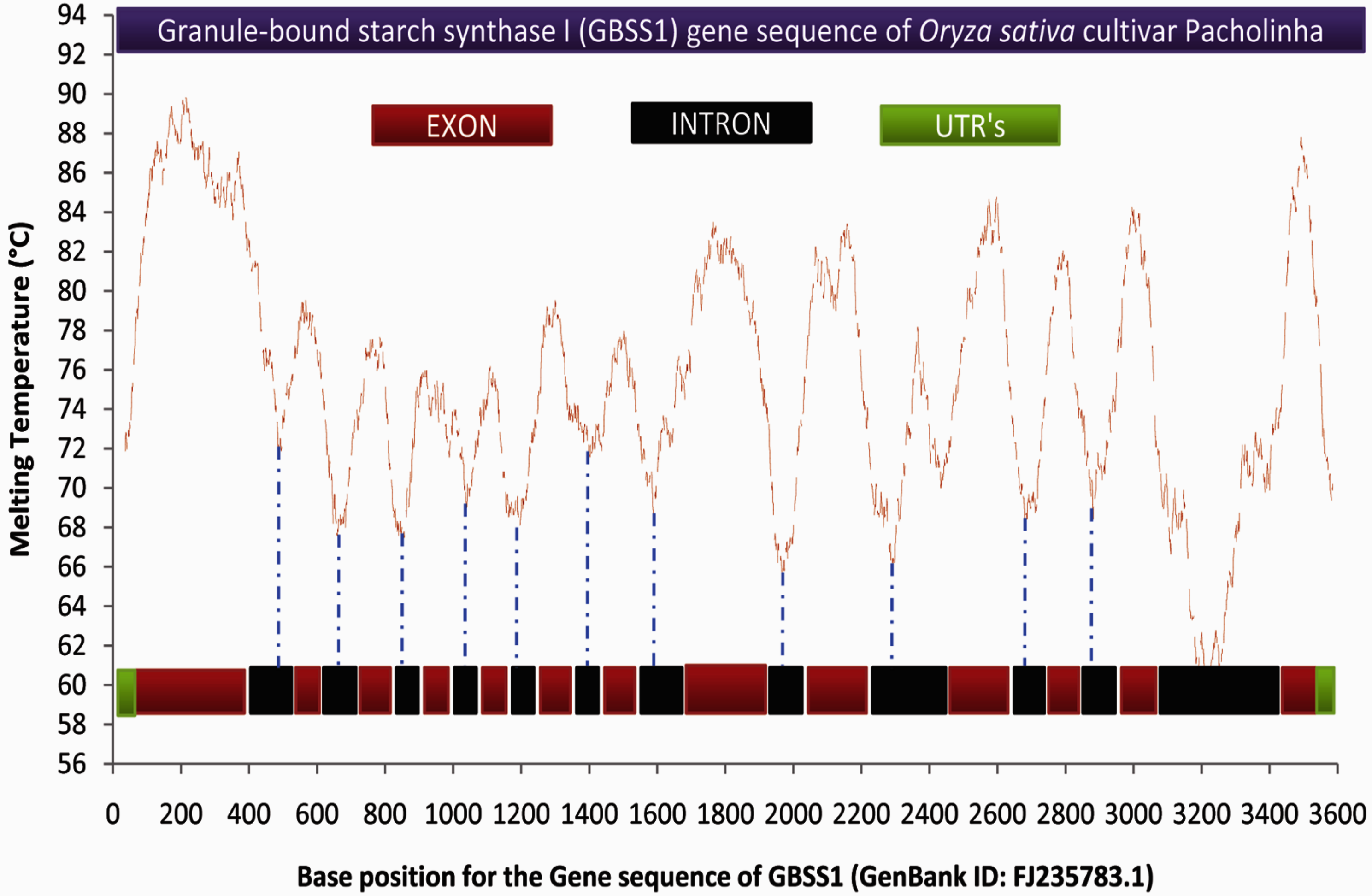
Chemgenome on Eukaryotes

Octant analysis of Experimentally verified Proteins from Swiss-Prot





Melting profile for an experimentally verified gene and its corresponding experimentally verified promoter sequence for Escherichia coli K-12 genome (NC_000913)


Granule-bound starch synthase I (GBSS1) gene sequence of *Oryza sativa* cultivar Pacholinha
EXON
INTRON
UTR's
Melting Temperature (°C)
94
92
90
88
86
84
82
80
78
76
74
72
70
68
66
64
62
60
58
56
0
200
400
600
800
1000
1200
1400
1600
1800
2000
2200
2400
2600
2800
3000
3200
3400
3600
Base position for the Gene sequence of GBSS1 (GenBank ID: FJ235783.1)

ChemGenome Summary

- An *ab-initio* physico-chemical model is proposed to characterize DNA sequences as genes and non-genes.

- Analyses of 372 bacterial genomes and 21 eukaryotic genomes present a proof of concept.

- Gene and non-gene regions separate out.
- Consequences of frame-shift mutations are correctly predicted.
- The specificities and sensitivities achieved are >90% (with reliable datasets)
- The methodology captures more than 90% genes without a prior knowledge of start and stop sites.

- Whole genome analysis software for gene prediction is available at

www.scfbio-iitd.res.in/chemgenome2

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.

- Genome Analysis - *ChemGenome*

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

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

A *de novo* energy based protein structure prediction software

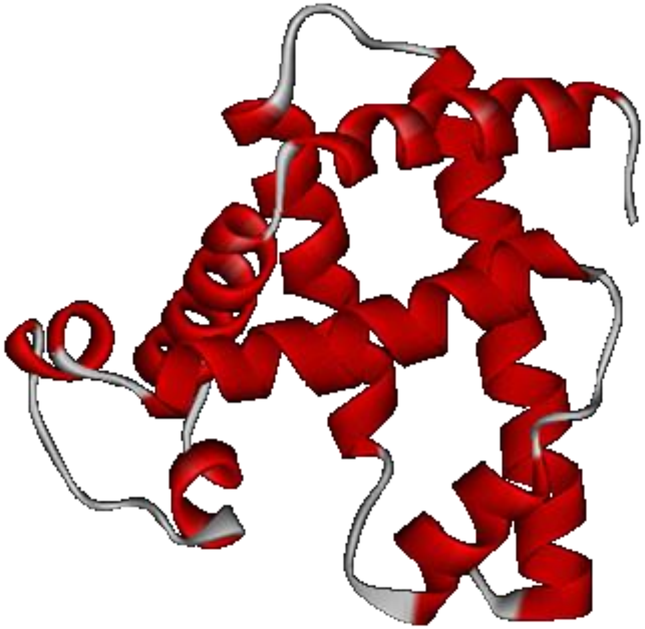
- Drug Design – *Sanjeevini*

A comprehensive indigenous active site directed lead molecule design protocol

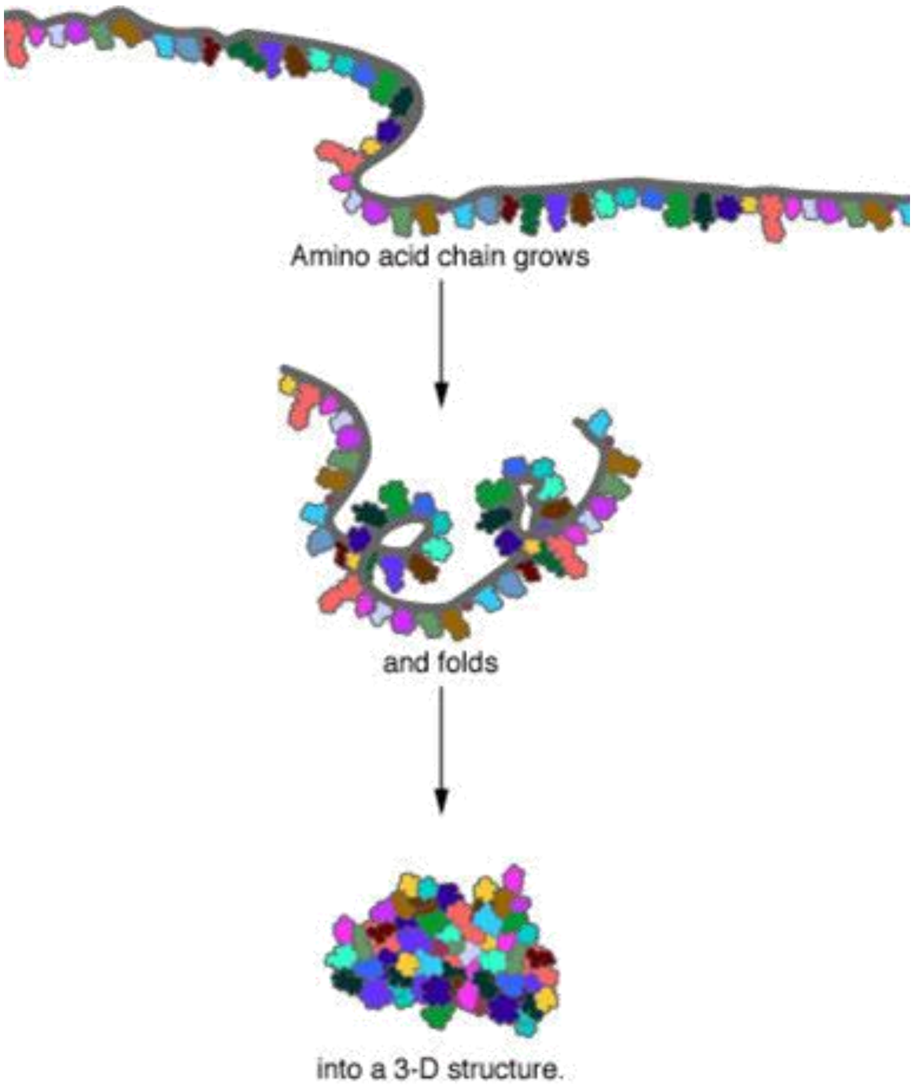
Bhageerath

Protein Tertiary Structure Prediction

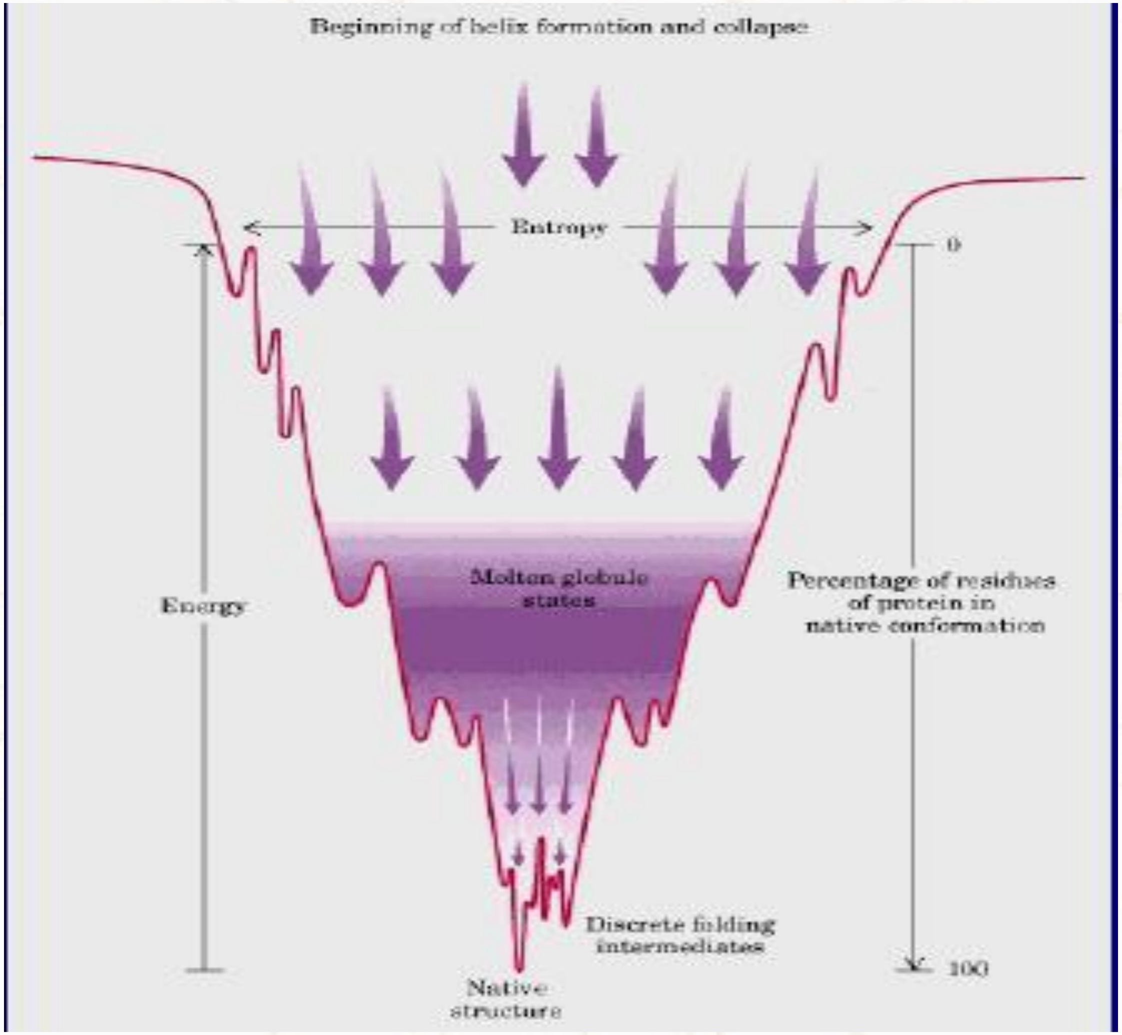
**................GLU ALA GLU MET LYS ALA SER GLU ASP LEU LYS
LYS HIS GLY VAL THR VAL LEU THR ALA LEU GLY ALA ILE LEU
LYS LYS LYS GLY HIS HIS GLU ALA GLU LEU LYS PRO LEU ALA
GLN SER HIS ALA THR LYS HIS LYS ILE PRO ILE LYS TYR LEU
GLU PHE ILE SER GLU ALA ILE ILE HIS LEU HIS.......................**



Protein Folding Problem

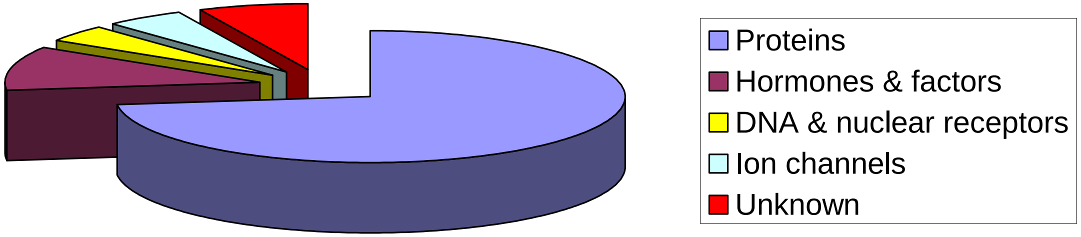


PROTEIN FOLDING LANDSCAPE

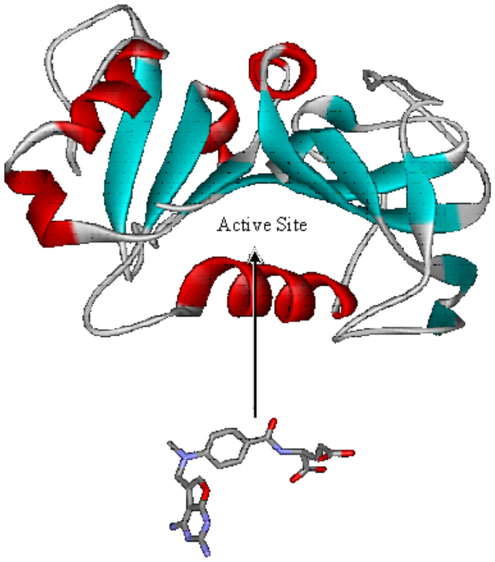

Beginning of helix formation and collapse
Entropy
0
Energy
Molten globule states
Percentage of residues of protein in native conformation
Discrete folding intermediates
100
Native structure

WHY FOLD PROTEINS ?

Pharmaceutical/Medical Sector

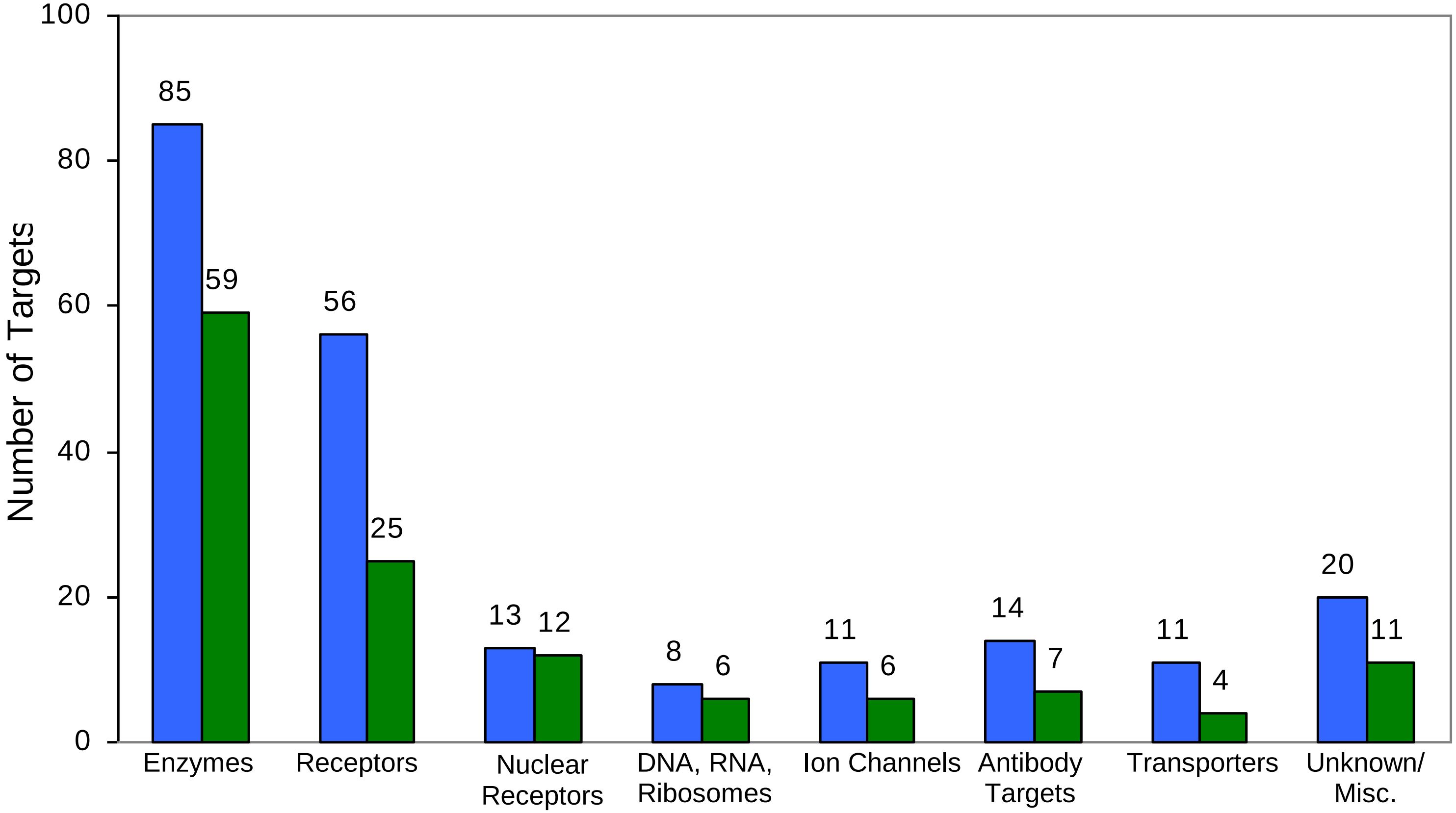


Drug Targets



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

Present Scenario of Drug Targets



BLUE: Number of targets in each class. (Imming P, Sinning C, Meyer A. *Nature Rev Drug Discov* **2006**;5: 821) **(Total 218 targets & 8 classes)**

GREEN: Number of 3D structures available in each class (Total: 130) (Protein Data Bank)

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

Comparative Modeling Approaches

Homology

Similar sequences adopt similar fold is the basis.

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

Threading

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

Computational Requirements for *ab initio* Protein Folding

Strategy A

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

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

Strategy B

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

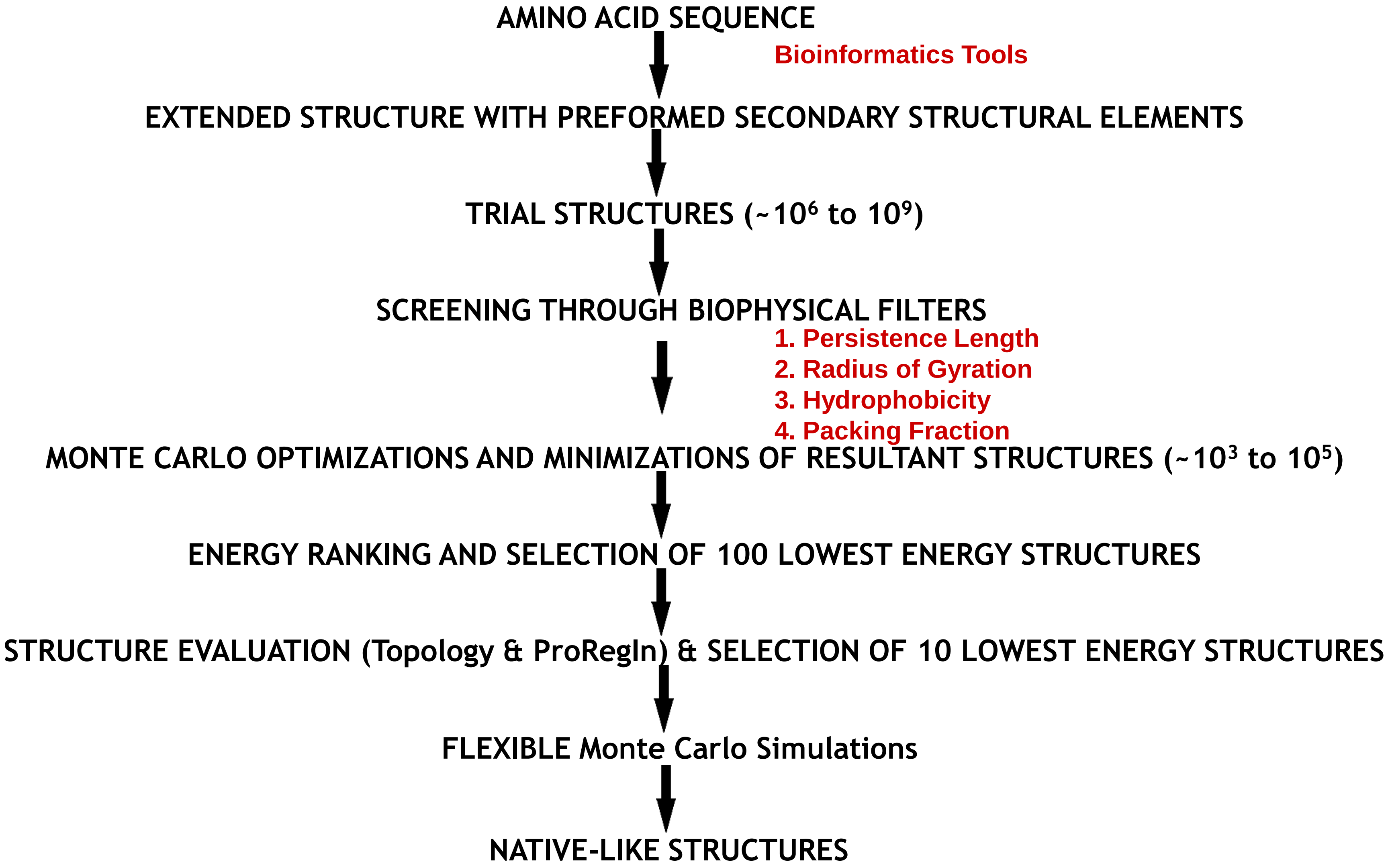
~ 10^{-10} sec / day / processor

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

~ 10^{6} years

With 10^{6} processors ~ 1 Year

From Sequence to Structure: The IITD Pathway



AMINO ACID SEQUENCE

Bioinformatics Tools

EXTENDED STRUCTURE WITH PREFORMED SECONDARY STRUCTURAL ELEMENTS

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

SCREENING THROUGH BIOPHYSICAL FILTERS

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

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

ENERGY RANKING AND SELECTION OF 100 LOWEST ENERGY STRUCTURES

STRUCTURE EVALUATION (Topology & ProRegIn) & SELECTION OF 10 LOWEST ENERGY STRUCTURES

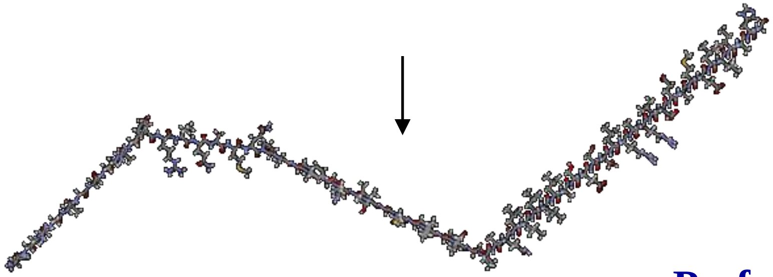
FLEXIBLE Monte Carlo Simulations

NATIVE-LIKE STRUCTURES

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

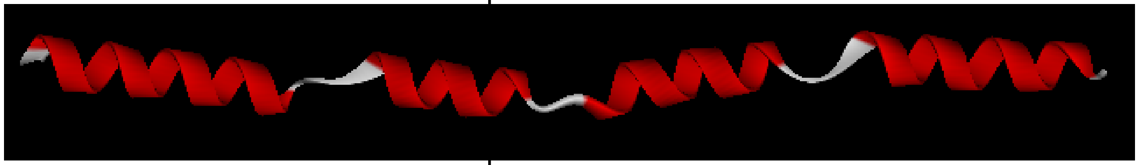
Sampling 3D Space

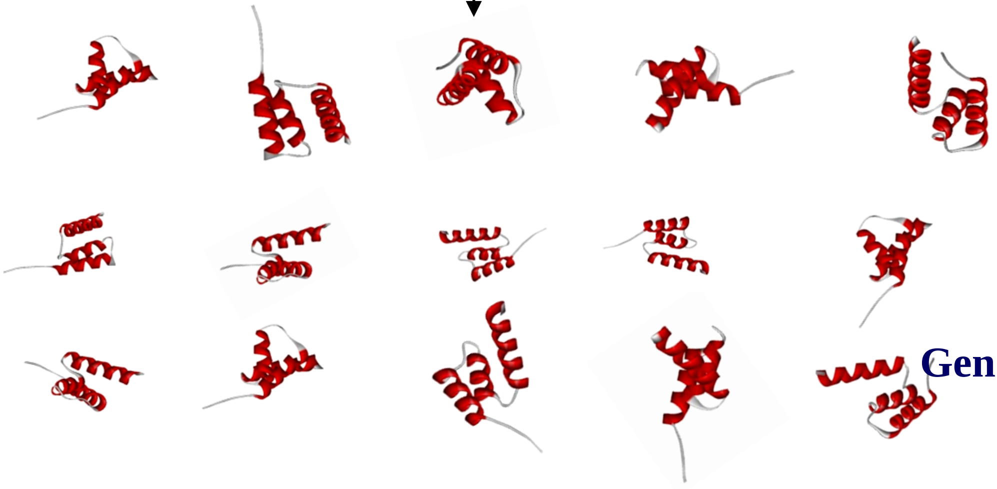
HRQALGERLYPRVQAMQPAFASKITGMLLELSPAQLLLLLASENSLRARVNEAMELIIAHG



Extended Chain

Preformed Secondary Structural Units

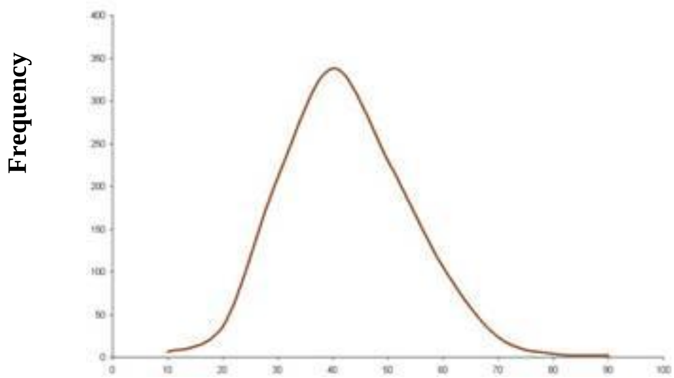




Generation of Trial Structures

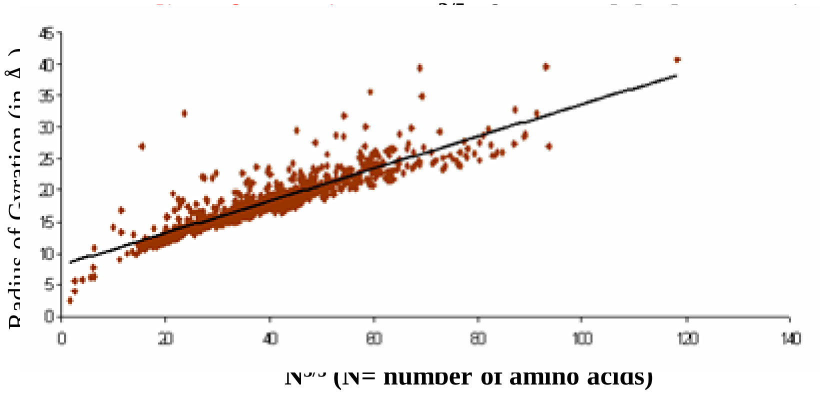
Filter-Based Structure Selection

Persistence Length Analysis of 1,000 Globular Proteins



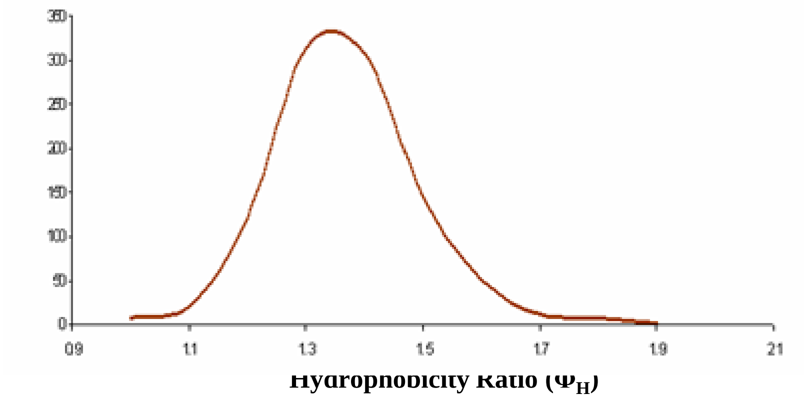
s

.257



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

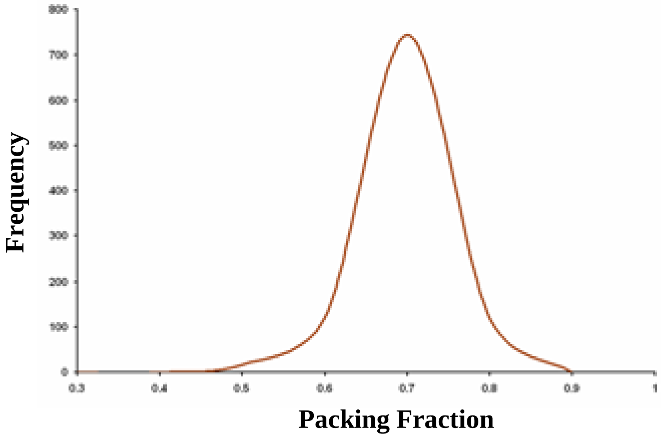
Frequency vs Hydrophobicity Ratio of 1,000 Globular Proteins



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

ASA : Accessible surface area

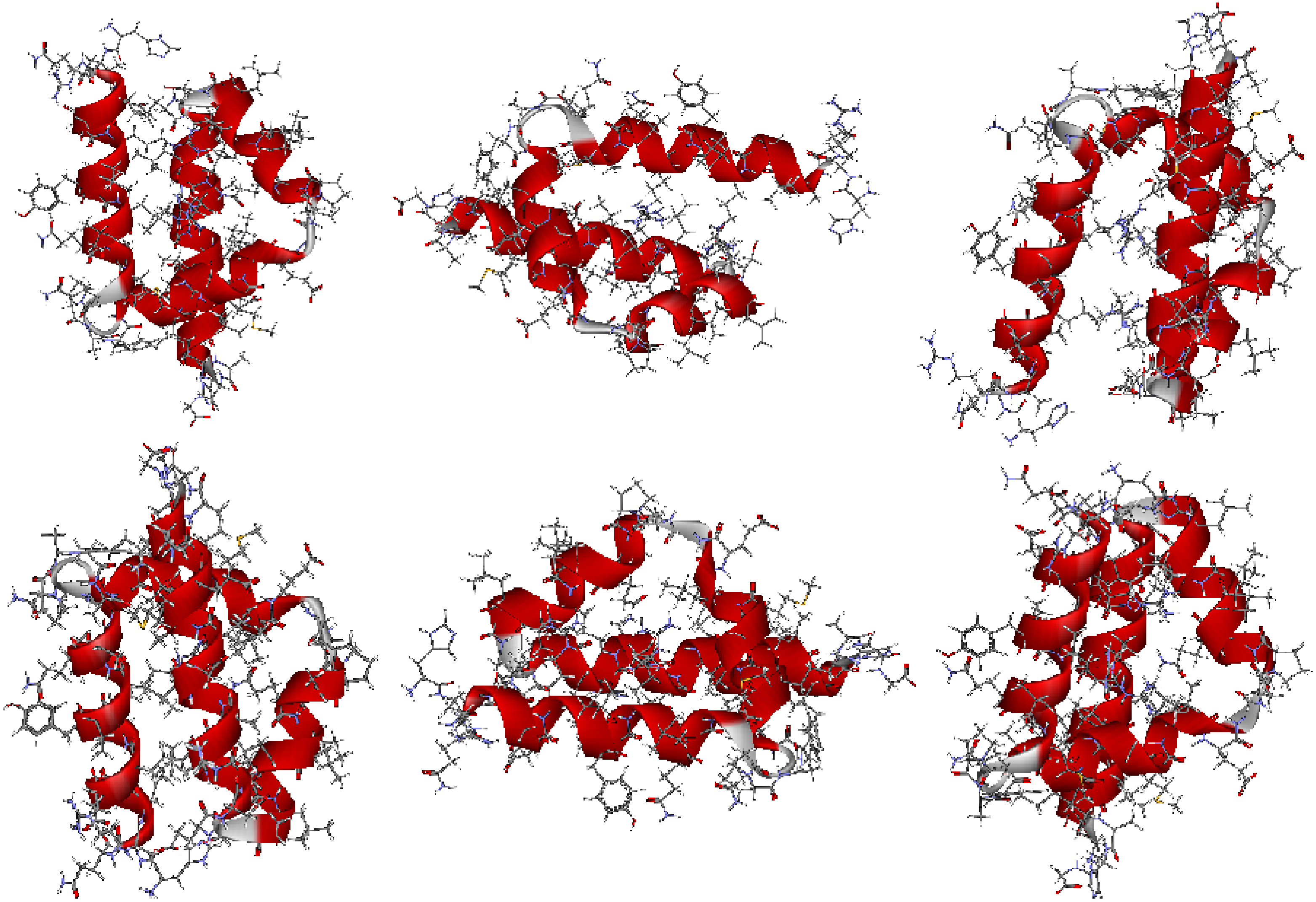
Frequency vs Packing Fraction of 1,000 Globular Proteins



Globular proteins are known to exhibit packing fractions around 0.7

Removal of Steric Clashes in Selected Structures

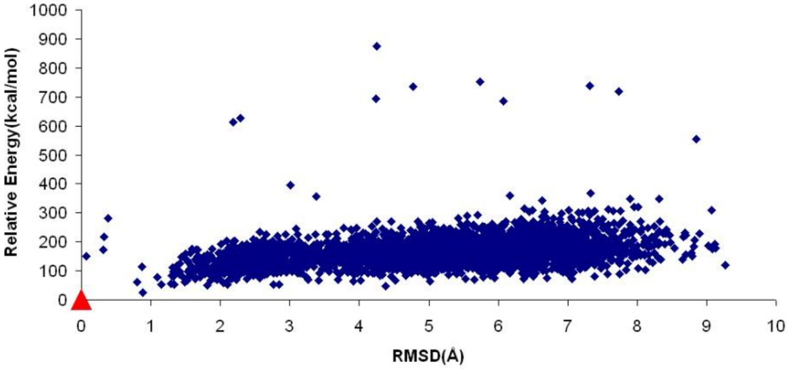
(Distance Based Monte Carlo)



Validation of Empirical Energy Based Scoring Function

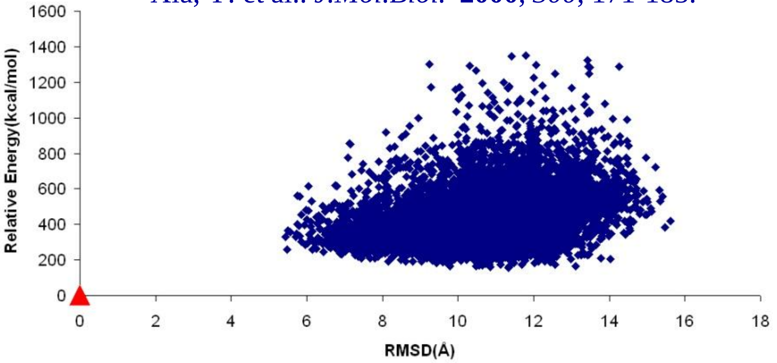
Four-state reduced decoy set

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



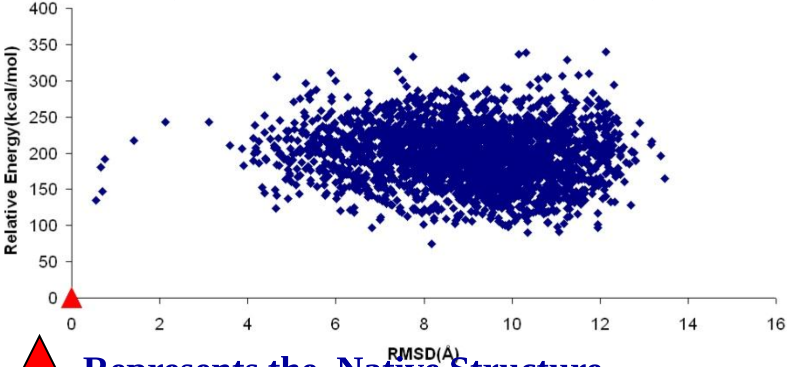
Lattice_ssfit decoy set

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



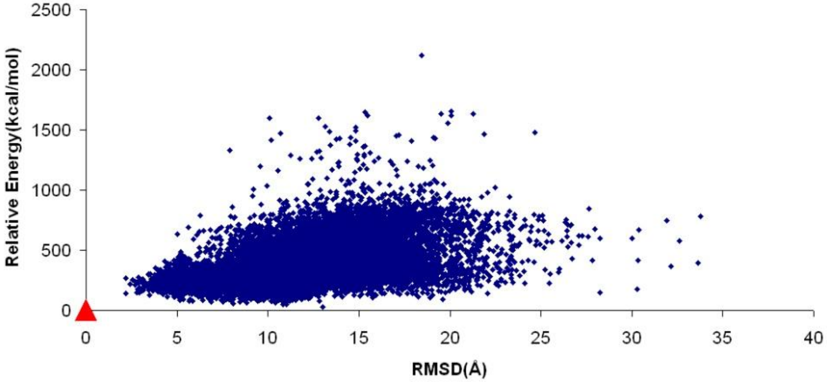
Lmds decoy set

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;

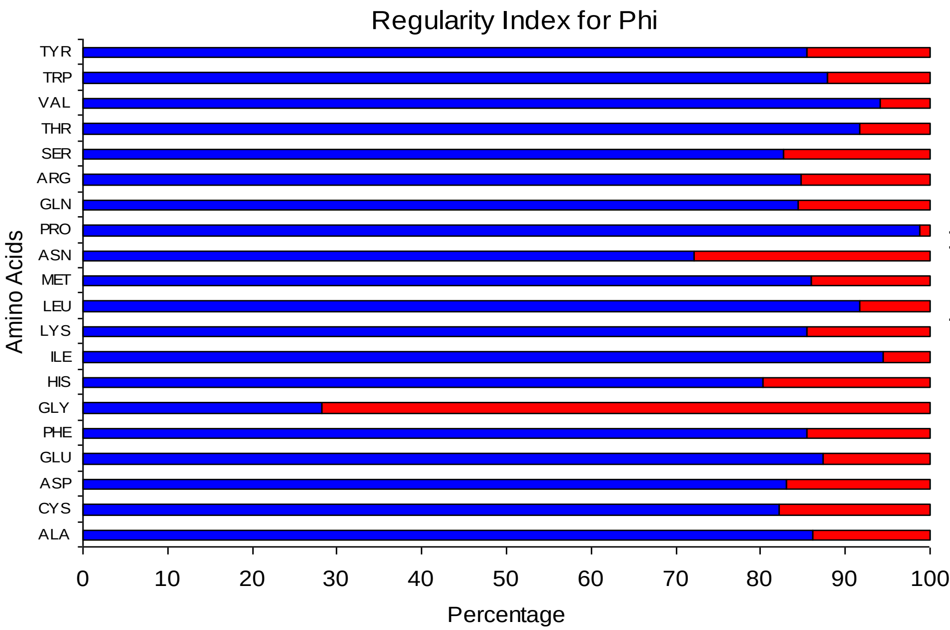
Arora N.; Jayaram B.; *J. Phys. Chem. B.* **1998**, *102*, 6139-6144;

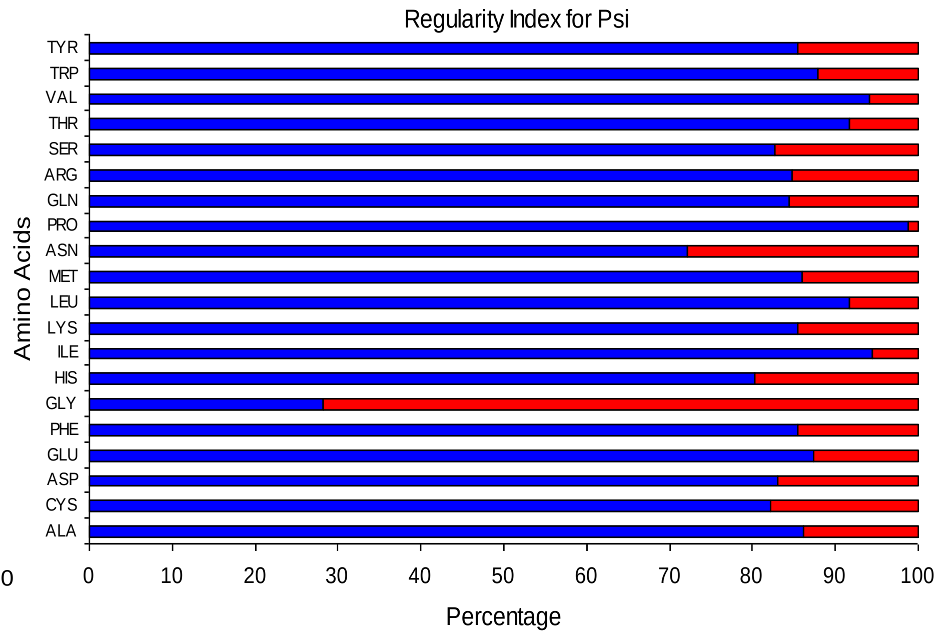
Arora N, Jayaram B, *J. Comput. Chem.*, .**1997**, *18*, 1245-1252.

ProRegIn

Protein Regularity Index for selection of native-like structures of proteins

A web-enabled tool developed based on the regularity in the φ, ψ dihedral angles of the amino acids that constitute loop regions.



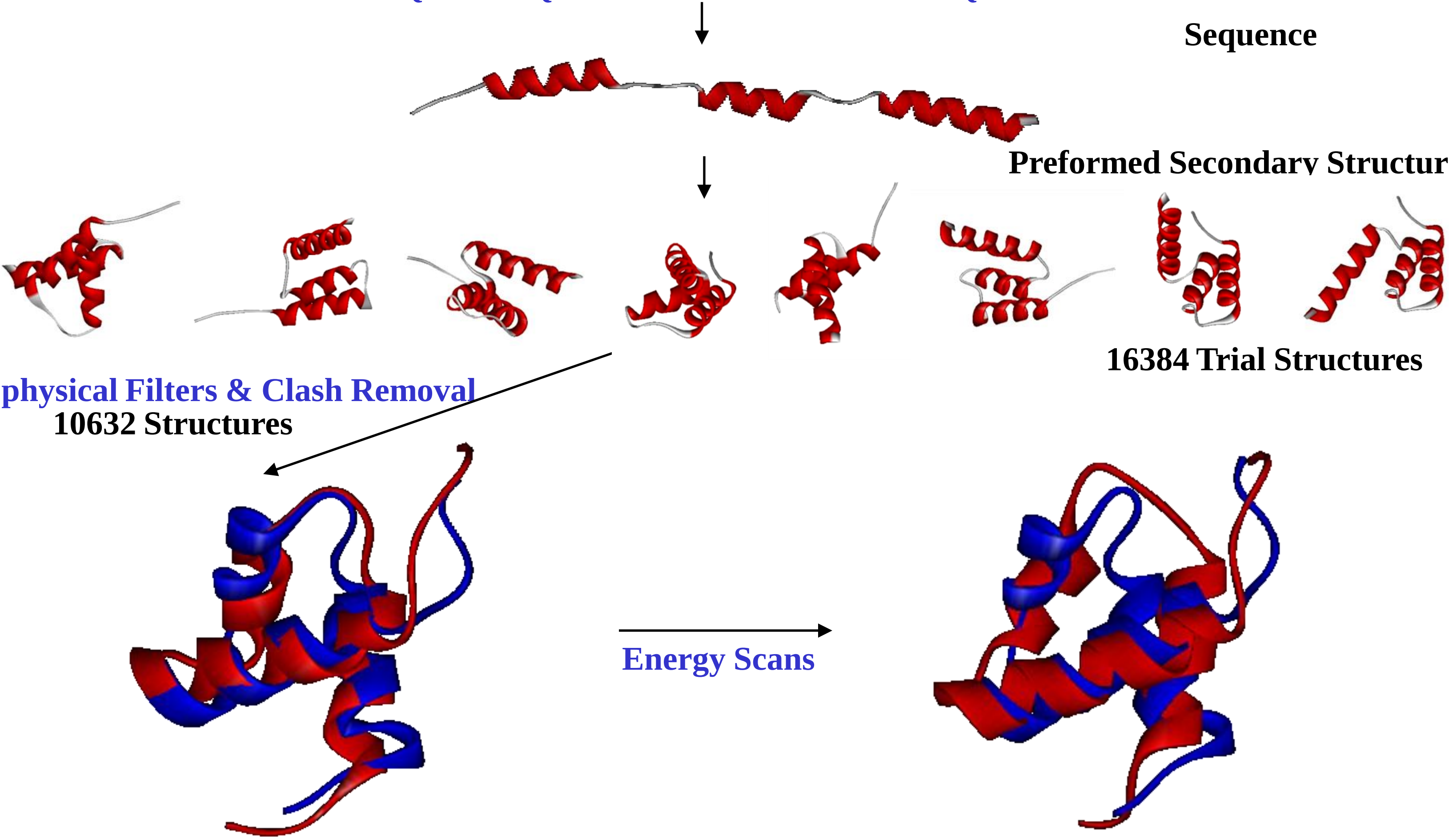


Thukral L, Shenoy S R, Bhushan K and Jayaram B. *ProRegIn* : A Regularity Index for the Selection of Native-like Tertiary Structures of Proteins. *J. Biosci.* **2007**, 32, 71-81.

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

Energy Scans

RMSD=2.87, Energy Rank=1774

RMSD=4.0, Energy Rank=4

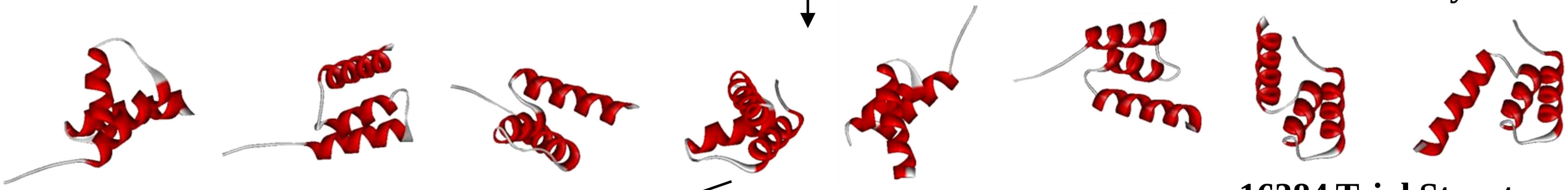
A Case Study of *S.aureus* Protein A

Immunoglobulin Binding (60 AA)

RPRTAFSSEQLARLKREFNENRYLTERRRQQLSSELGLNEAQIKIWFQNKRAKIKKS

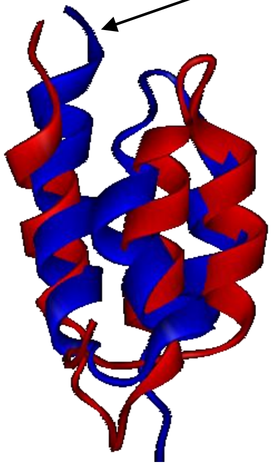
Sequence

Preformed Secondary Structure

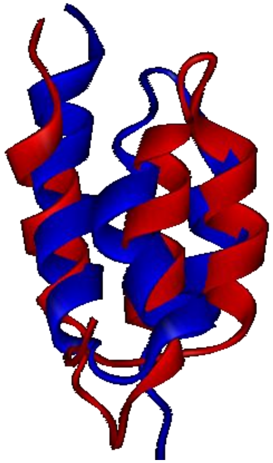


16384 Trial Structures

Biophysical Filters & Clash Removal
11255 Structures



Energy Scans



RMSD=4.2, Energy Rank=44

RMSD=4.8, Energy Rank=5

Performance of *Bhageerath* on 50 Small Globular Proteins

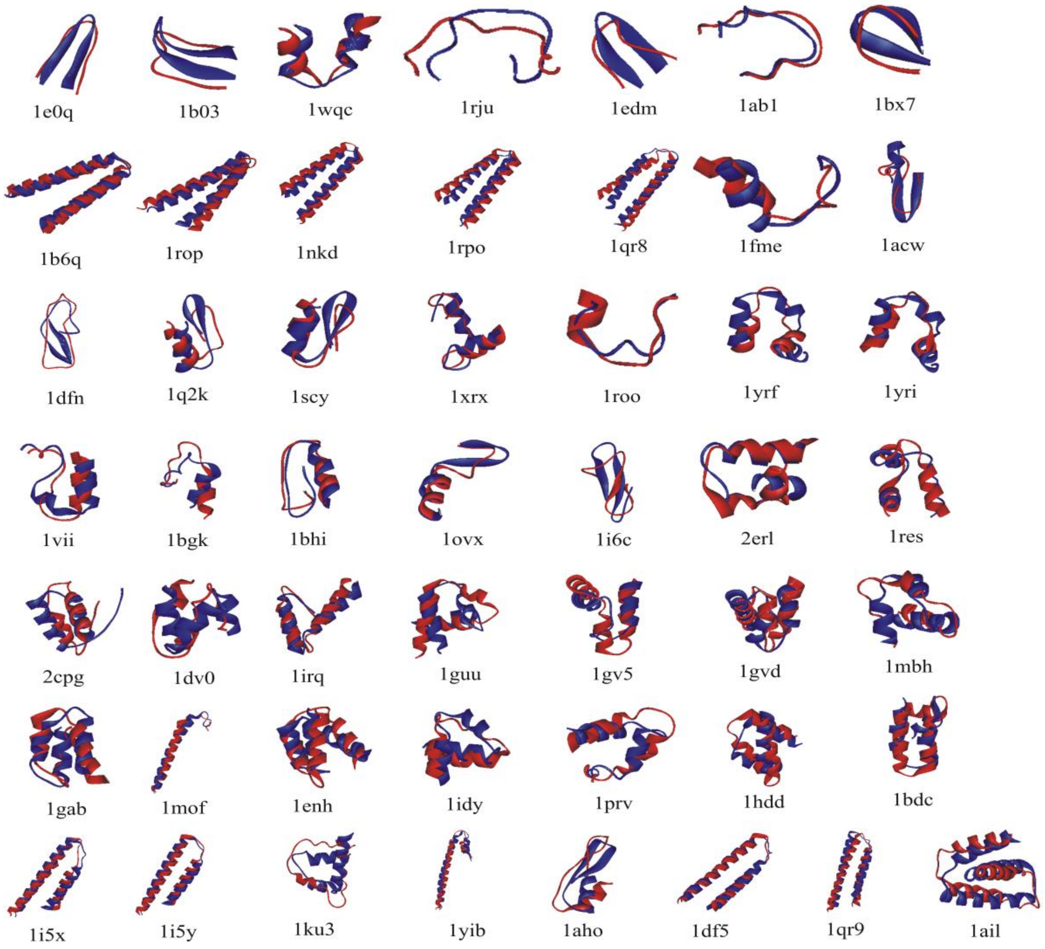
Sl. No.	PDB ID (i)	Number of amino acids (ii)	Number of secondary structure elements (iii)	Number of structures accepted after Persistence length and Radius of gyration filters (iv)	Lowest RMSD in the final 100 structures (Å) (v)	Energy Rank of the lowest RMSD structure in 100 structures (vi)	After ProRegIn Filter			After Topology and Accessible Surface Area Filter		
							Number of structures selected (Number of structures < 6 Å) (vii)	Lowest RMSD (Å) (viii)	Energy Rank of the lowest RMSD structure in 100 structures (ix)	Number of structures selected (Number of structures < 6 Å) (x)	Lowest RMSD (Å) (xi)	Energy Rank of the lowest RMSD structure in 10 structures (xii)
1.	1E0Q	17	2E	128	2.5	2	100 (29)	2.5	2	10 (10)	**2.5**	**2**
2.	1B03	18	2E	64	4.4	2	64 (5)	4.4	2	10 (5)	**4.4**	**2**
3.	1WQC	26	2H	128	2.5	6	100 (53)	2.5	6	10 (10)	**2.5**	**3**
4.	1RJU	36	2H	64	4.6	48	64 (3)	4.6	48	10 (2)	**5.9**	**6**
5.	1EDM	39	2E	128	2.9	100	100 (59)	2.9	100	10 (10)	**3.5**	**2**
6.	1AB1	46	2H	128	2.4	10	100 (82)	2.4	10	10 (10)	**2.9**	**6**
7.	1BX7	51	2E	128	2.2	71	100 (85)	2.2	71	10 (10)	**3.1**	**8**
8.	1B6Q	56	2H	128	3.1	27	100 (8)	3.1	27	10 (5)	**3.1**	**10**
9.	1ROP	56	2H	128	4.3	2	100 (6)	4.3	2	10 (2)	**4.3**	**2**
10.	1NKD	59	2H	128	3.8	8	100 (4)	3.8	8	10 (4)	**3.8**	**6**
11.	1RPO	61	2H	128	3.8	2	100 (6)	3.8	2	10 (4)	**3.8**	**2**
12.	1QR8	68	2H	128	4.4	80	100 (3)	4.4	80	10 (2)	**4.4**	**10**
13.	1FME	28	1H,2E	15592	2.9	52	100 (90)	2.9	52	10 (8)	**3.7**	**5**
14.	1ACW	29	1H,2E	15726	3.9	97	100 (45)	3.9	97	10 (5)	**5.1**	**8**
15.	1DFN	30	3E	13174	4.4	77	98 (11)	4.4	77	10 (4)	**5.0**	**1**
16.	1Q2K	31	1H,2E	16020	4.2	46	100 (20)	4.2	46	10 (4)	**4.2**	**9**

17.	1SCY	31	1H,2E	15423	3.1	10	100 (40)	3.1	10	10 (4)	**3.1**	**5**
18.	1XRX	34	1E,2H	14630	3.9	28	100 (19)	3.9	28	10 (1)	**5.6**	**1**
19.	1ROO	35	3H	1071	2.5	14	100(100)	2.5	14	10 (10)	**2.8**	**5**
20.	1YRF	35	3H	15180	3.8	16	100 (62)	3.8	16	10 (9)	**4.8**	**4**
21.	1YRI	35	3H	15180	2.8	81	100 (70)	2.8	81	10 (8)	**3.8**	**6**
22.	1VII	36	3H	16380	3.7	7	100 (50)	3.7	7	10 (6)	**3.7**	**2**
23.	1BGK	37	3H	14139	3.8	33	100 (56)	3.8	33	10 (8)	**4.1**	**3**
24.	1BHI	38	1H,2E	14923	5.3	2	100 (5)	5.3	2	10 (2)	**5.3**	**2**
25.	1OVX	38	1H,2E	12074	3.2	8	100 (76)	3.2	8	10 (5)	**4.0**	**1**
26.	1I6C	39	3E	2927	4.1	31	100 (32)	4.1	31	10 (3)	**5.1**	**2**
27.	2ERL	40	3H	16268	3.1	18	100 (32)	3.1	18	10 (2)	**3.2**	**6**
28.	1RES	43	3H	16135	4.0	30	100 (40)	4.0	30	10 (7)	**4.2**	**2**
29.	2CPG	43	1E,2H	10905	3.6	20	100 (18)	3.6	20	10 (1)	**5.3**	**2**
30.	1DV0	45	3H	14488	4.0	20	100 (21)	4.0	20	10 (1)	**5.1**	**4**
31.	1IRQ	48	1E,2H	11592	3.5	74	100 (18)	3.5	74	10 (1)	**5.3**	**9**
32.	1GUU	50	3H	13410	4.5	74	100 (42)	4.5	74	10 (7)	**4.6**	**6**
33.	1GV5	52	3H	11109	3.5	33	99 (24)	3.5	33	10 (5)	**4.1**	**2**
34.	1GVD	52	3H	10626	3.8	18	100 (35)	3.8	18	10 (6)	**4.9**	**9**
35.	1MBH	52	3H	10632	3.8	48	100 (24)	3.8	48	10 (5)	**4.0**	**4**
36.	1GAB	53	3H	14495	3.6	16	100 (12)	3.6	16	10 (3)	**3.6**	**6**
37.	1MOF	53	3H	16384	2.4	57	100 (96)	2.4	57	10 (10)	**2.9**	**5**
38.	1ENH	54	3H	13622	3.2	12	100 (23)	3.2	12	10 (3)	**4.6**	**3**
39.	1IDY	54	3H	11133	3.3	84	100 (52)	3.3	84	10 (8)	**3.5**	**6**
40.	1PRV	56	3H	5468	4.4	55	99 (25)	4.4	55	10 (7)	**4.9**	**9**

41.	1HDD	57	3H	12849	3.2	74	100 (22)	3.2	74	10 (2)	**4.8**	**8**
42.	1BDC	60	3H	11255	4.2	44	100 (19)	4.2	44	10 (2)	**4.8**	**5**
43.	1I5X	61	3H	16384	2.6	29	99 (54)	2.6	29	10 (10)	**2.6**	**6**
44.	1I5Y	61	3H	16384	2.6	20	100 (48)	2.6	20	10 (10)	**2.6**	**7**
45.	1KU3	61	3H	5701	4.9	68	100 (14)	4.9	68	10 (3)	**5.5**	**4**
46.	1YIB	61	3H	16384	2.9	7	100 (75)	2.9	7	10 (9)	**3.5**	**5**
47.	1AHO	64	1H,2E	2429	4.7	58	100 (15)	4.7	58	10 (1)	**6.0**	**6**
48.	1DF5	68	3H	16384	3.1	10	100 (41)	3.1	10	10 (6)	**3.1**	**8**
49.	1QR9	68	3H	16384	2.9	49	100 (33)	2.9	49	10 (9)	**3.8**	**2**
50.	1AIL	70	3H	16384	4.2	42	100 (5)	4.2	42	10 (3)	**4.2**	**7**

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

Predicted Structures for 50 Globular Proteins with *Bhageerath*



🟦 Native structure

🟥 Predicted structure

Bhageerath versus Homology modeling

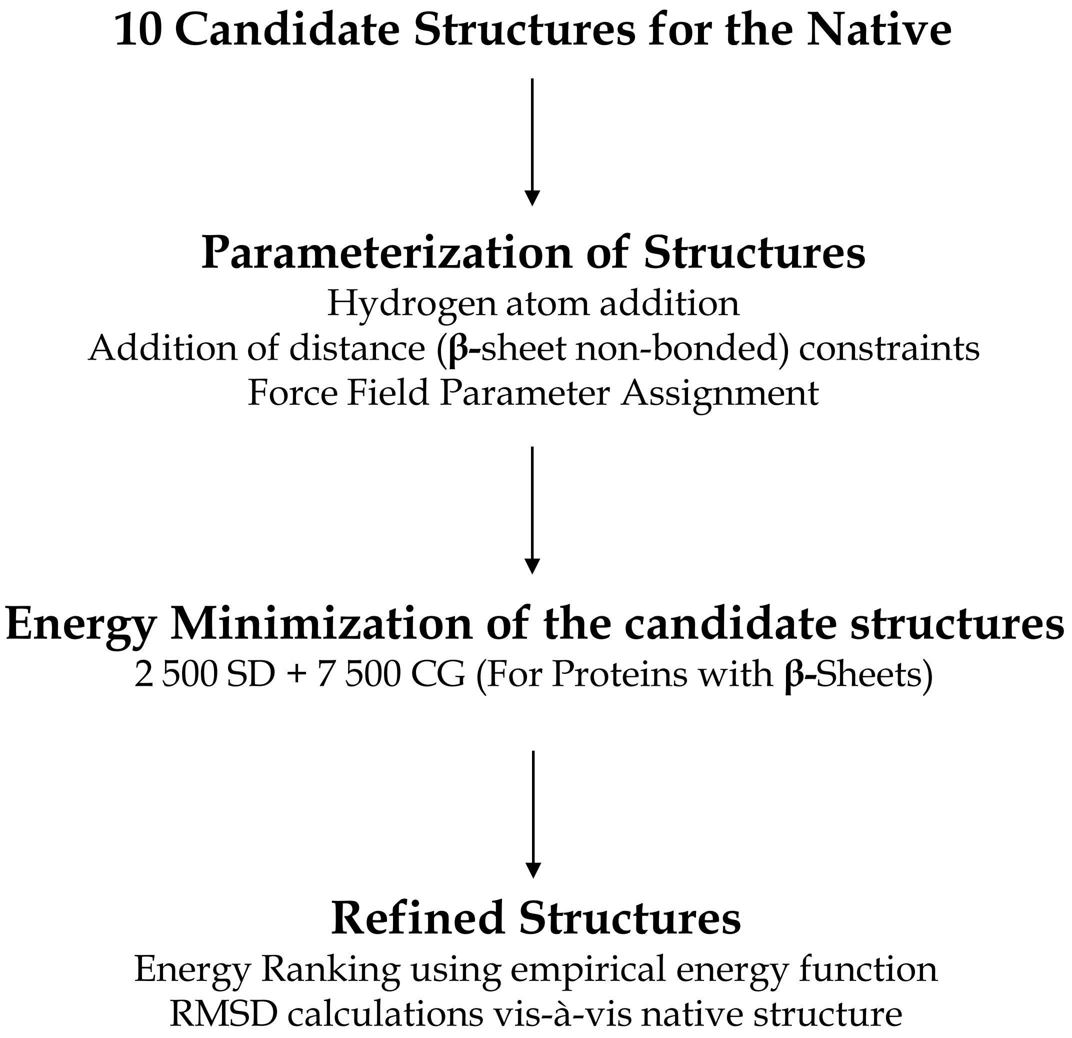
No	Protein PDB ID	CPHmodels RMSD(Å)	ESyPred3D RMSD(Å)	Swiss-model RMSD(Å)	3D-PSSM RMSD(Å)	Bhageerath# RMSD(Å)
1.	1IDY (1-54)*	3.96 (2-54)*	3.79 (2-51)*	5.73 (1-51)*	3.66 (1-51)*	3.36
2.	1PRV (1-56)*	5.66 (2-56)*	5.56 (3-56)*	6.67 (3-56)*	5.94 (1-56)*	3.87

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

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

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

Flowchart for constraint minimization of proteins with β-sheets



10 Candidate Structures for the Native

Parameterization of Structures

Hydrogen atom addition

Addition of distance (**β**-sheet non-bonded) constraints

Force Field Parameter Assignment

Energy Minimization of the candidate structures

2 500 SD + 7 500 CG (For Proteins with **β**-Sheets)

Refined Structures

Energy Ranking using empirical energy function

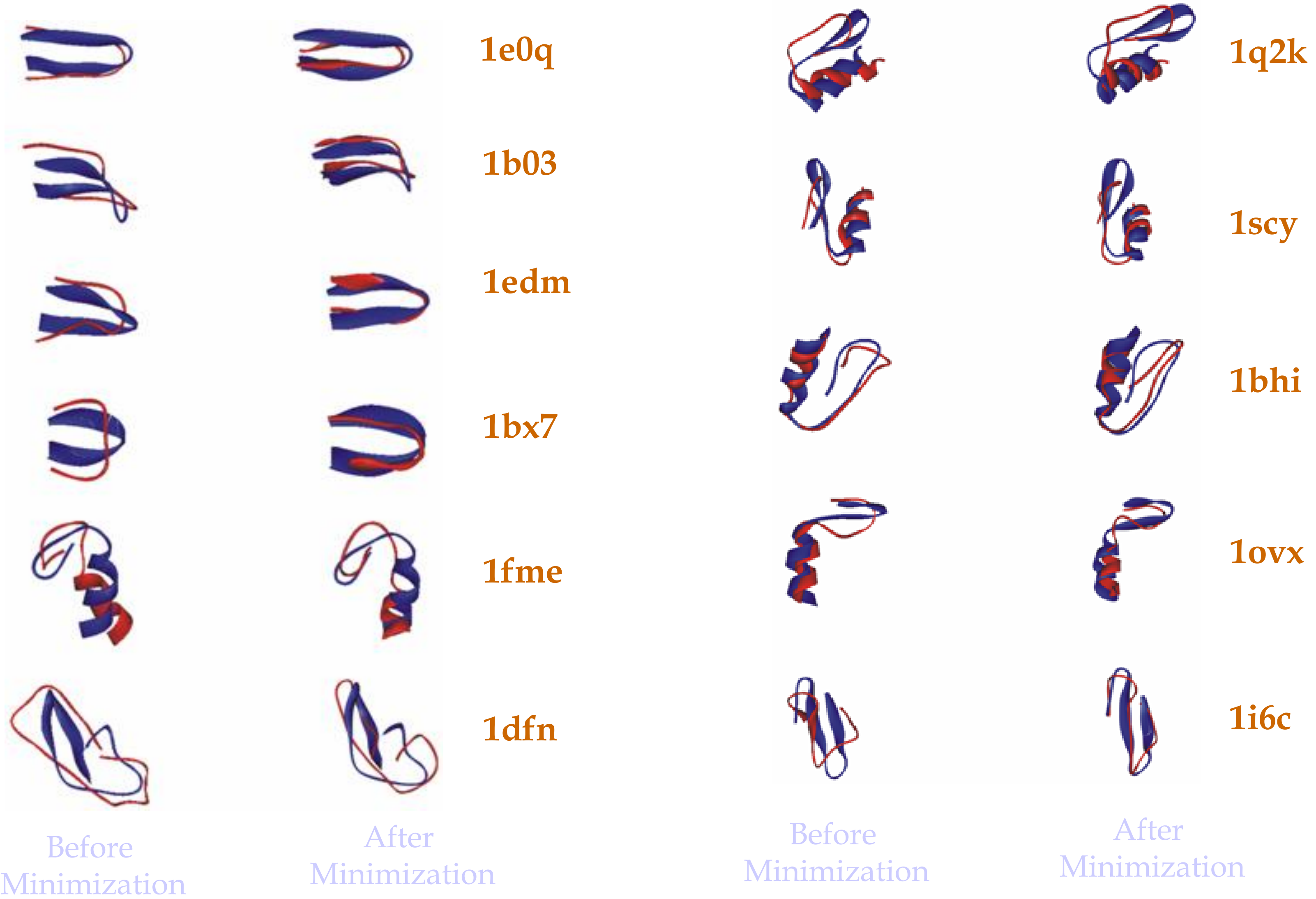
RMSD calculations vis-à-vis native structure

SD: Steepest Descent; CG: Conjugate Gradient

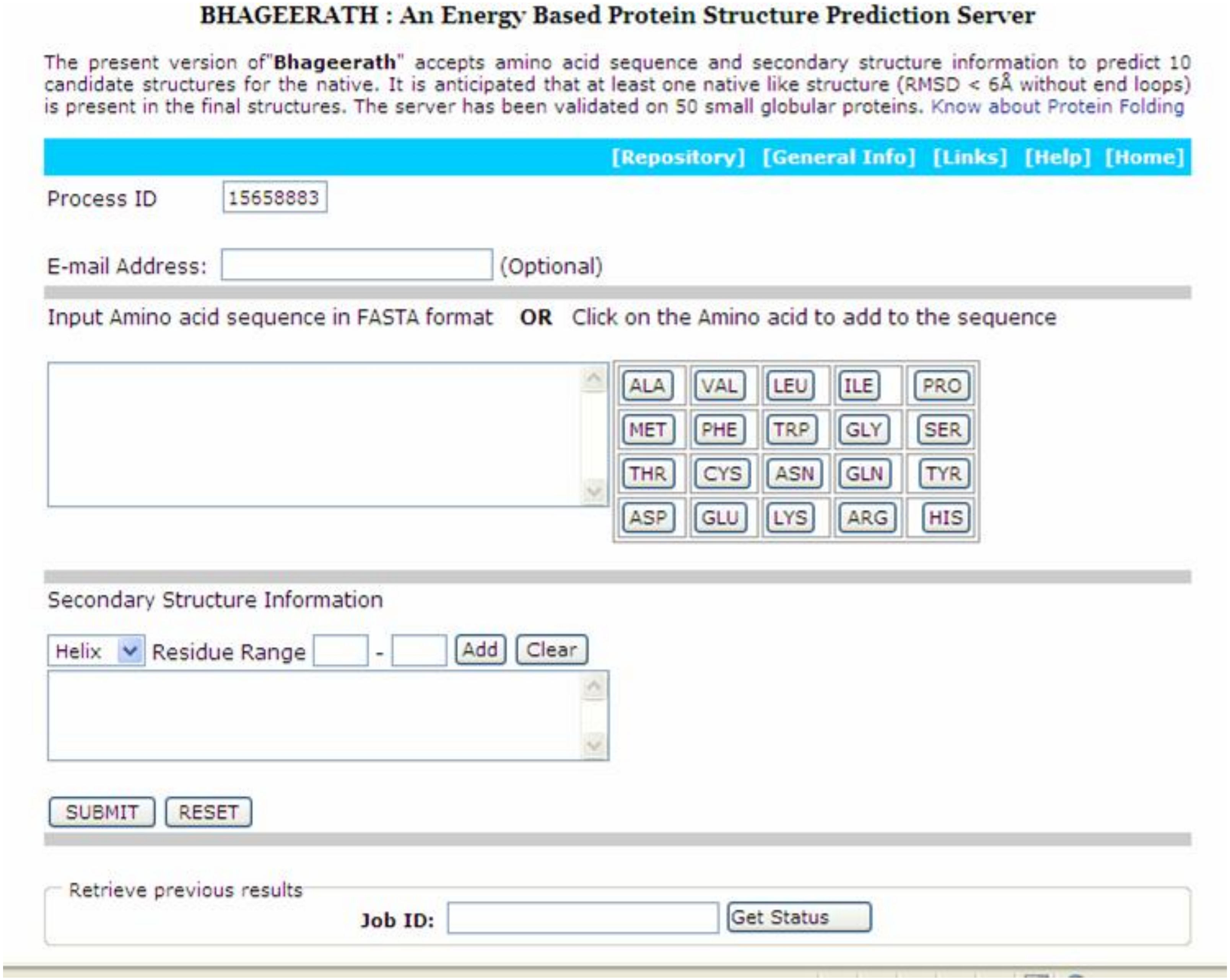
Results on β-sheet proteins after constraint minimization

Sl. No	PDB ID	Number of Amino Acids	Number of Secondary Structural Elements	After ProRegIn and Topology Filters		After distance (β sheet) constraints	
				Lowest RMSD in 10 final structures (Å)	Energy Rank	Lowest RMSD after constraint minimization (Å)	Energy Rank
1.	1E0Q	17	2E	2.5	2	**2.2**	**4**
2.	1B03	18	2E	4.4	3	**1.9**	**4**
3.	1EDM	39	2E	3.5	2	**1.5**	**9**
4.	1BX7	51	2E	3.1	8	**2.2**	**5**
5.	1FME	28	1H,2E	3.7	5	**4.1**	**8**
6.	1DFN	30	3E	5.0	1	**4.3**	**7**
7.	1Q2K	31	1H,2E	4.2	9	**3.9**	**8**
8.	1SCY	31	1H,2E	3.1	5	**3.1**	**5**
9.	1BHI	38	1H,2E	5.3	2	**3.4**	**10**
10.	1OVX	38	1H,2E	4.0	1	**3.4**	**5**
11.	1I6C	39	3E	5.1	2	**2.6**	**4**

Superimposed structures before and after constraint minimization



🟦 Native structure 🟥 Predicted structure



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

Process ID 15658883

E-mail Address: (Optional)

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

Secondary Structure Information

Helix Residue Range -

Retrieve previous results

Job ID:

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.

A stereochemical analysis of genomic (ncbi) and protein (Swissprot) sequences

	Swissprot# Sequences	Gene *	Intergenic (Nongene) Sequences *	Random sequences
Total Number considered	**157210**	**239418**	**204047**	**10000**
Number of proteins identified	**141784**	**227033**	**14699**	**806**

Software available at www.scfbio-iitd.res.in/software/proteomics/progenie.jsp
*Prediction Sensitivity = **0.95; Specificity = 0.94; Correlation coefficient = 0.88**
#Prediction Sensitivity = **0.90**

Jayaram, B.. Decoding the Design Principles of Amino Acids and the Chemical Logic of Protein Sequences. Available from ***Nature Precedings***. http://hdl.handle.net/10101/npre.2008.2135.1 **2008**

Conclusions and Future Perspectives

* Structures with native-like topology are bracketed within the 10 lowest energy structures. “Needle in a haystack problem” is thus reduced to finding the best 10 energy structures at least for small proteins.

* Further improvements to the methodology include introduction of Flexible MC / Explicit solvent MD so as to aid better side-chain packing, as well as usage of hydrophobicity and packing fraction filters to reduce the number of candidate structures for the native.

* The suite of programs christened “*Bhageerath*” is made accessible at www.scfbio-iitd.res.in/bhageerath

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

www.scfbio-iitd.res.in

•Genome Analysis - *ChemGenome*

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

•Protein Structure Prediction – *Bhageerath*

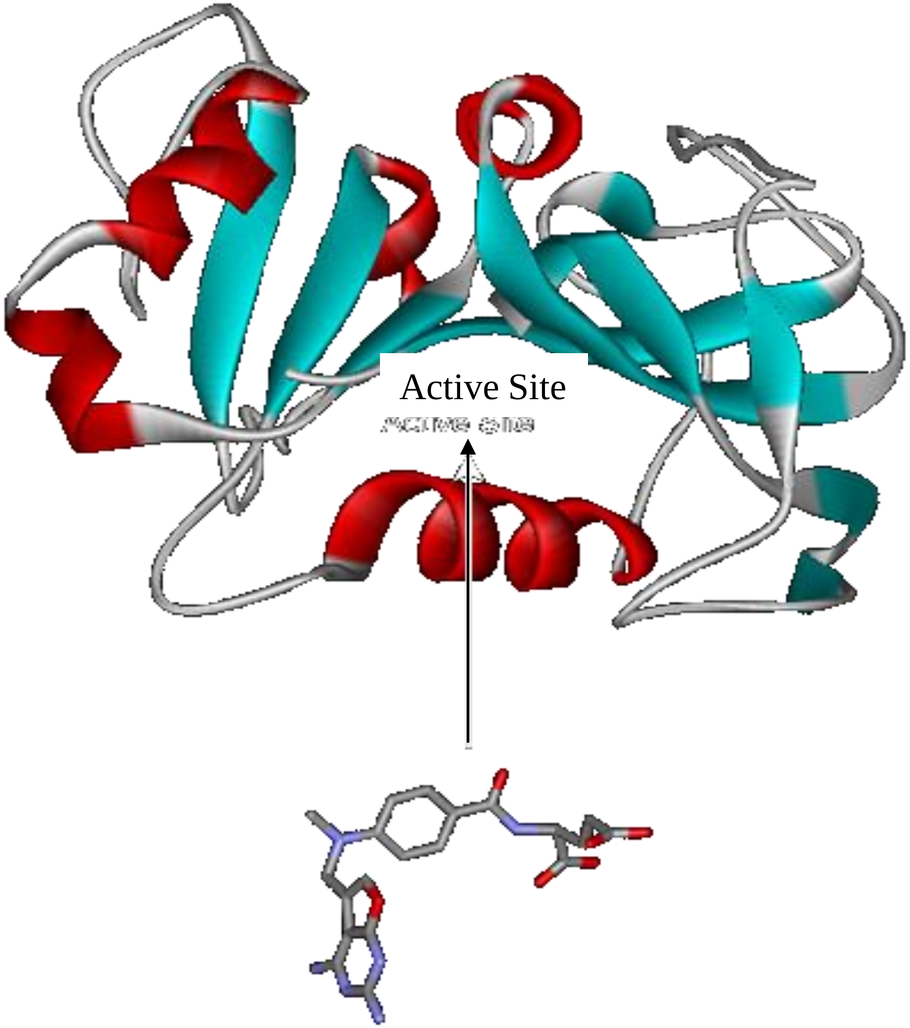
A *de novo* energy based protein structure prediction software

•Drug Design – *Sanjeevini*

A comprehensive indigenous active site directed lead molecule design protocol

Target -Site Directed Lead Design

Sanjeevini



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

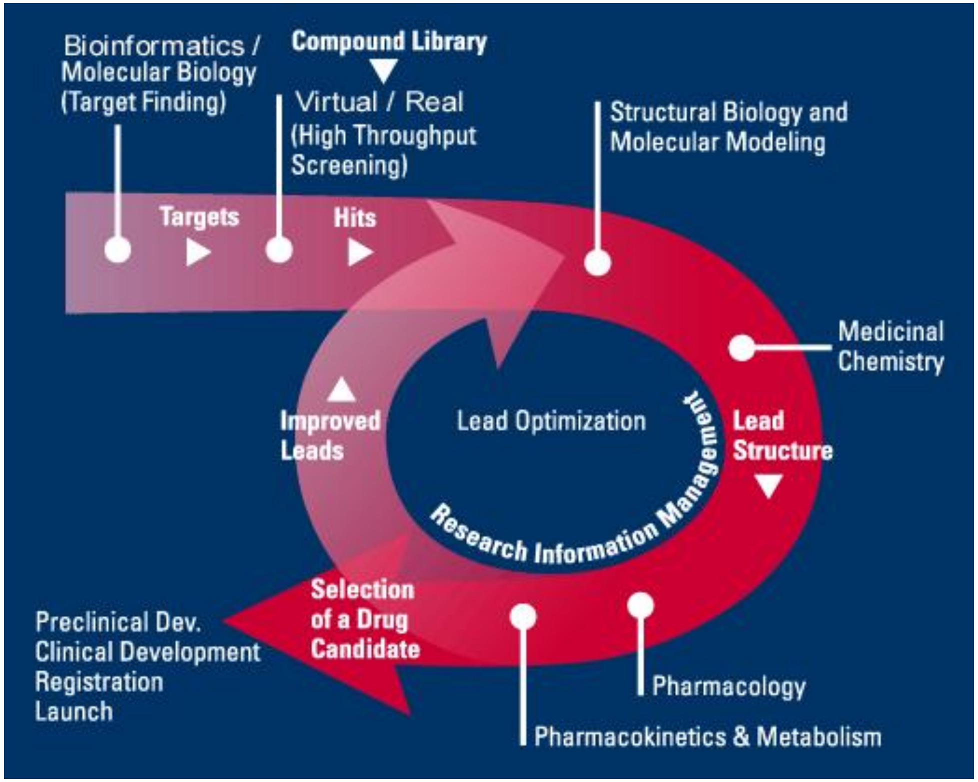
6.0yrs ↓ *68%*

FDA Review & Approval

1.5yrs ↓ *3%*

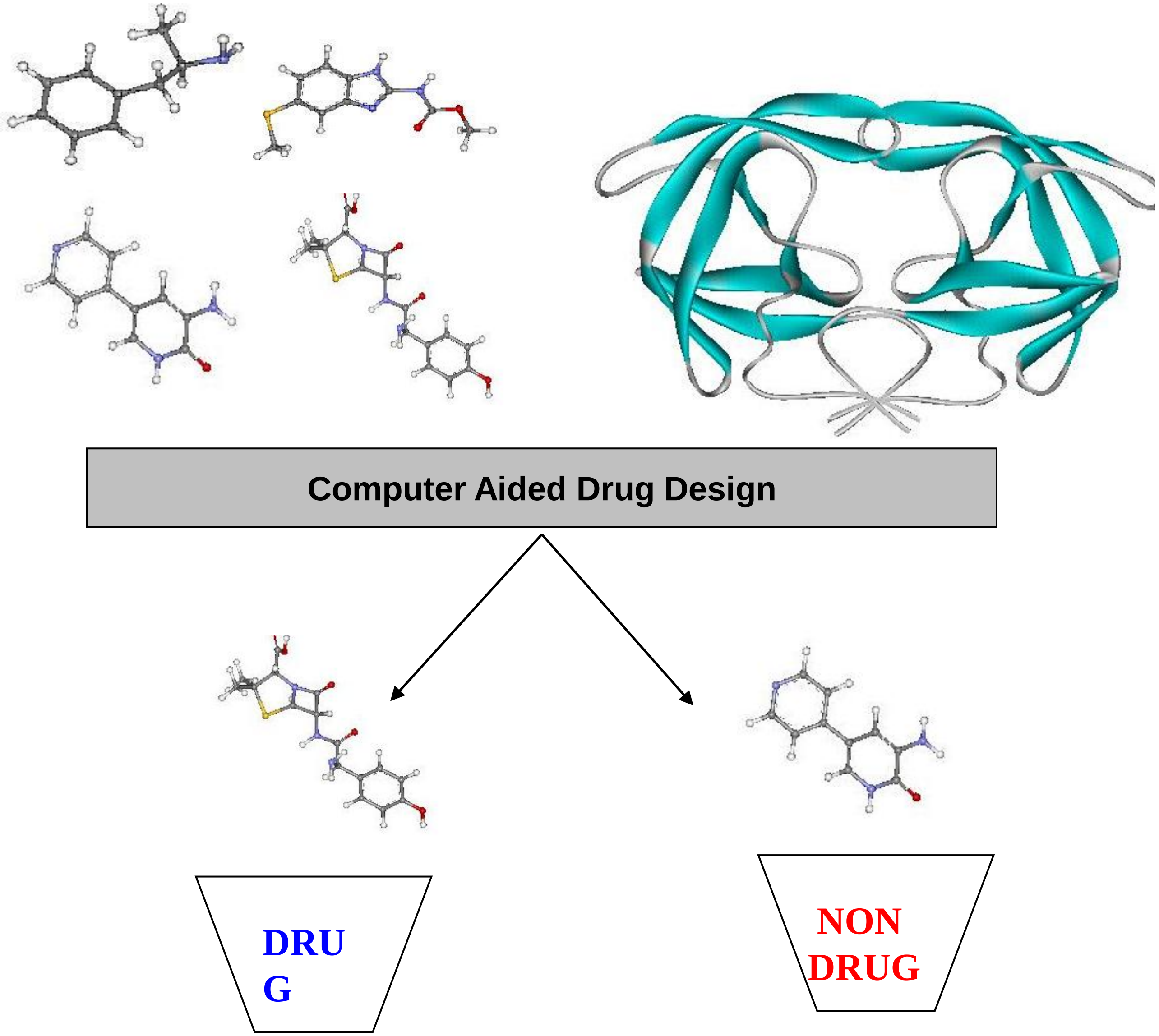
Drug to the Market

14 yrs **$1.4 billion**


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

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

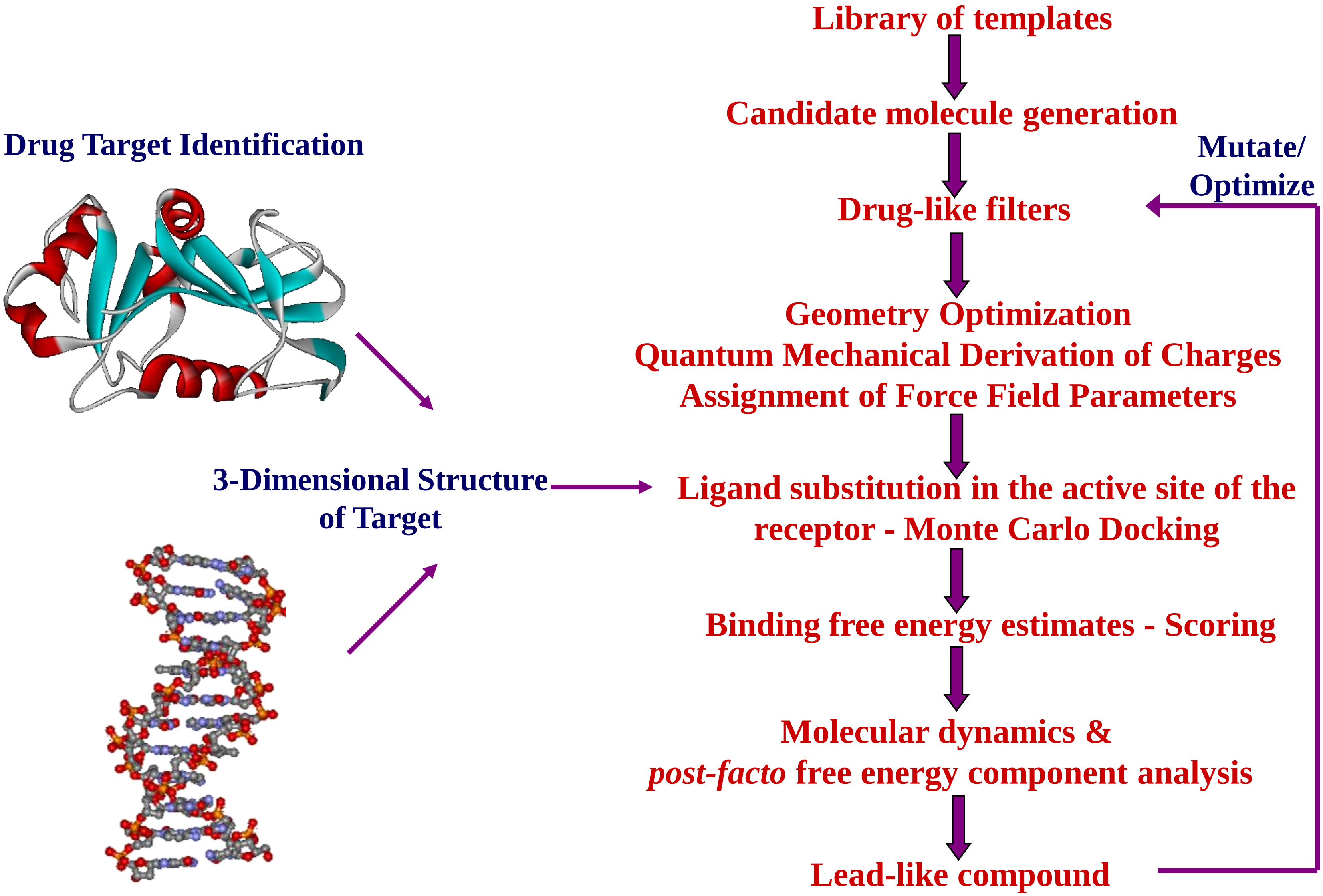
DRU
G

NON
DRUG

Some Concerns in Lead Design *In Silico*

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

De novo LEAD-LIKE MOLECULE DESIGN: THE IITD PATHWAY



Drug Target Identification

3-Dimensional Structure of Target

Library of templates

Candidate molecule generation

Drug-like filters

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

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

Binding free energy estimates - Scoring

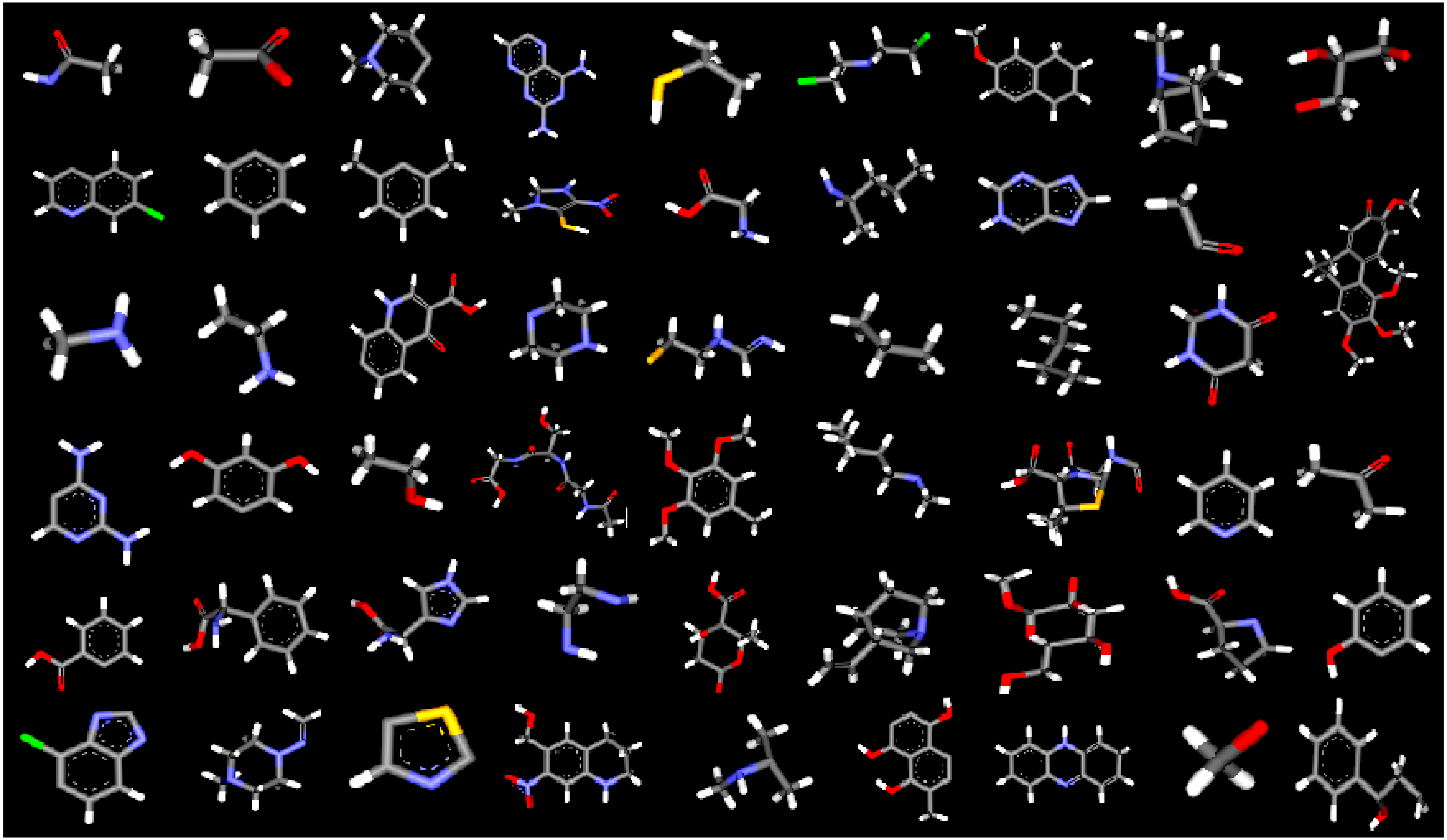
Molecular dynamics &
post-facto free energy component analysis

Lead-like compound

Mutate/ Optimize

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

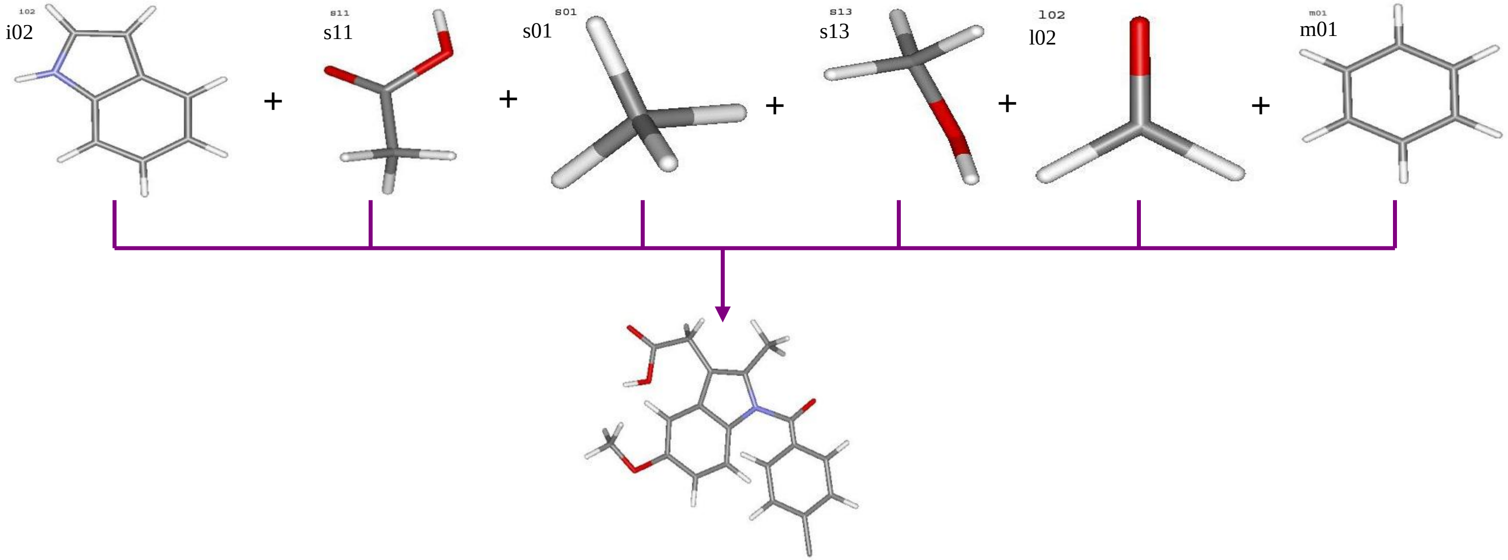
TEMPLATE LIBRARY



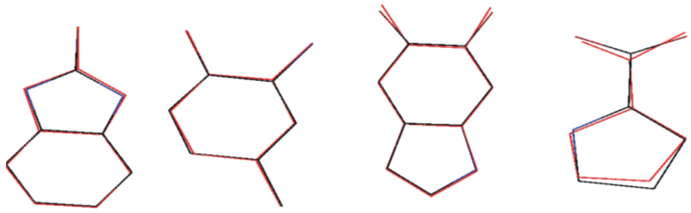
The substructure-based template library has ~ 160 chemical moieties consisting of unique rings, side chains and linkers

CANDIDATE MOLECULE GENERATION *in silico* & STRUCTURE VALIDATION

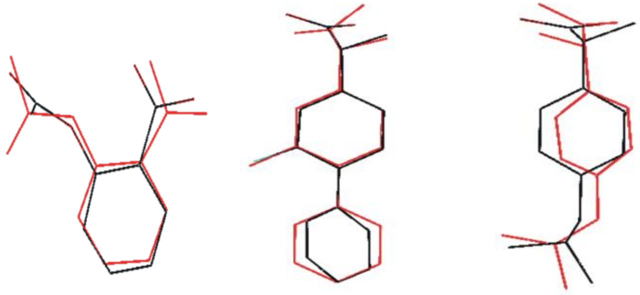
Candidate Generation



Structure Validation



Average RMSD



Average RMSD

0.07Å 0.05Å 0.08Å 0.38Å 0.15Å

0.91Å 0.54Å 0.94Å 0.80Å

Molecular Descriptors / Drug-like Filters

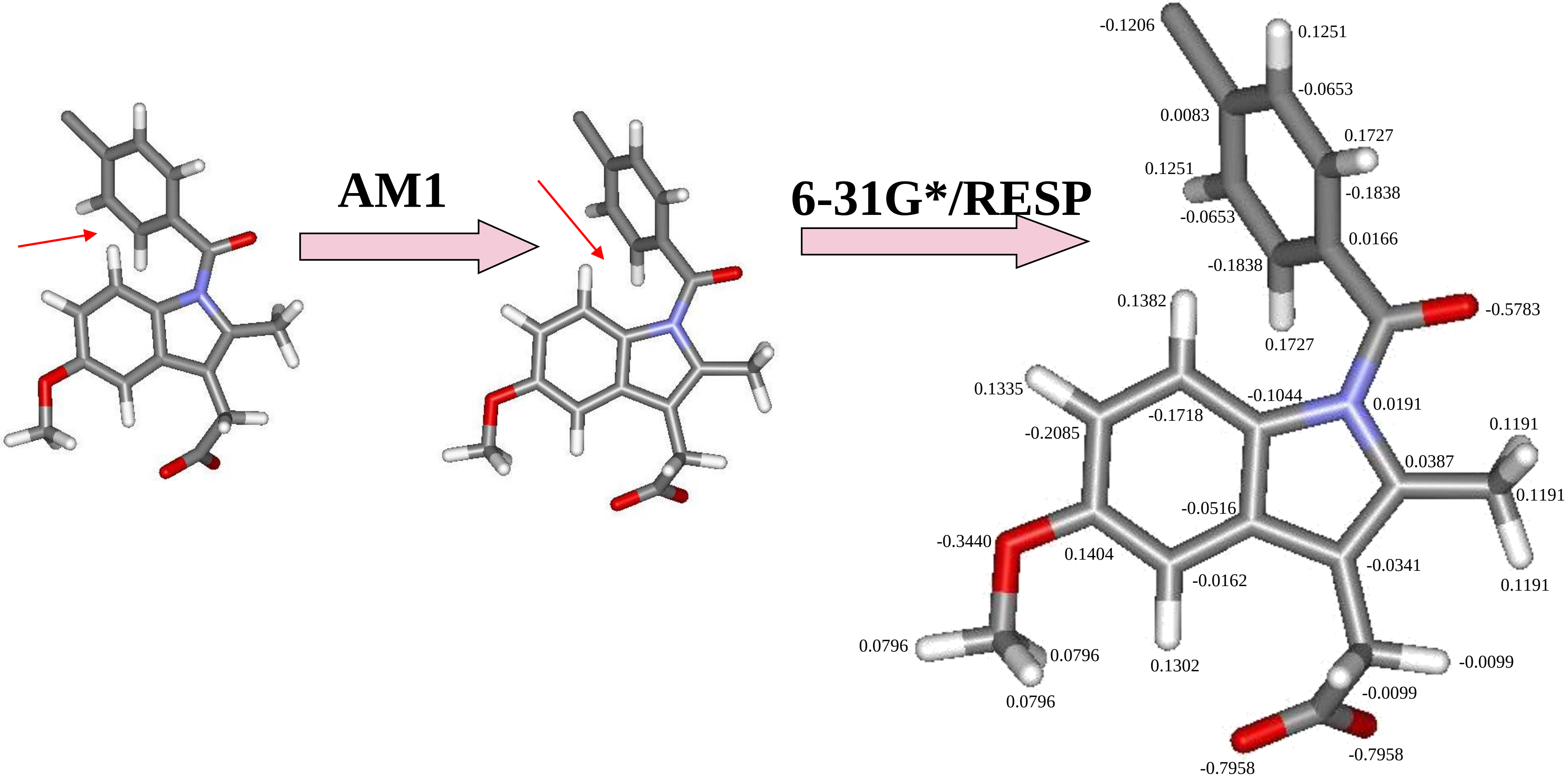
Lipinski's rule of five

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

Quantum Chemistry on Candidate drugs for Assignment of Force Field Parameters

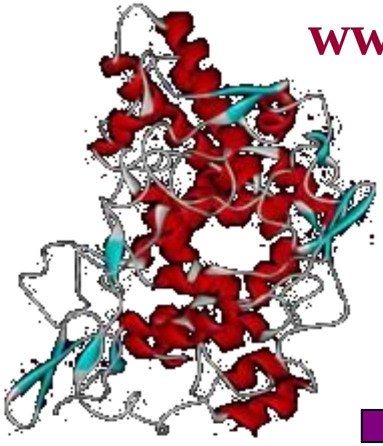


AM1

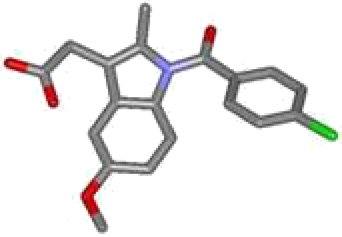
6-31G*/RESP

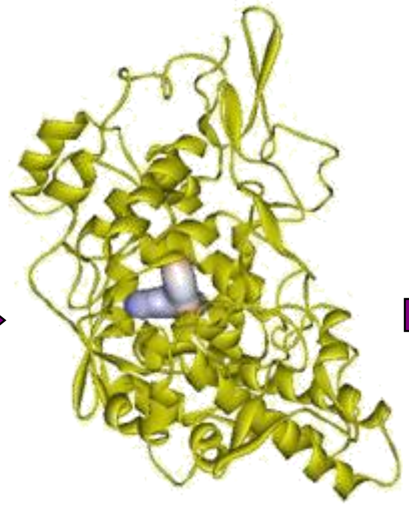
MONTE CARLO DOCKING OF THE CANDIDATE DRUG IN THE ACTIVE - SITE OF THE TARGET

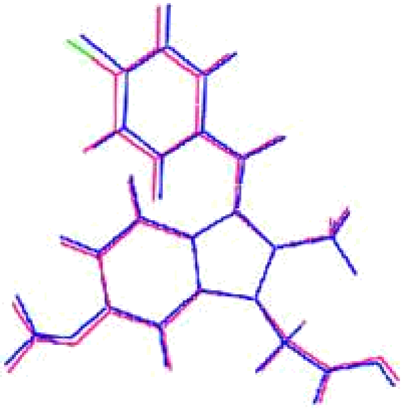
www.scfbio-iitd.res.in/dock/pardock.jsp



+



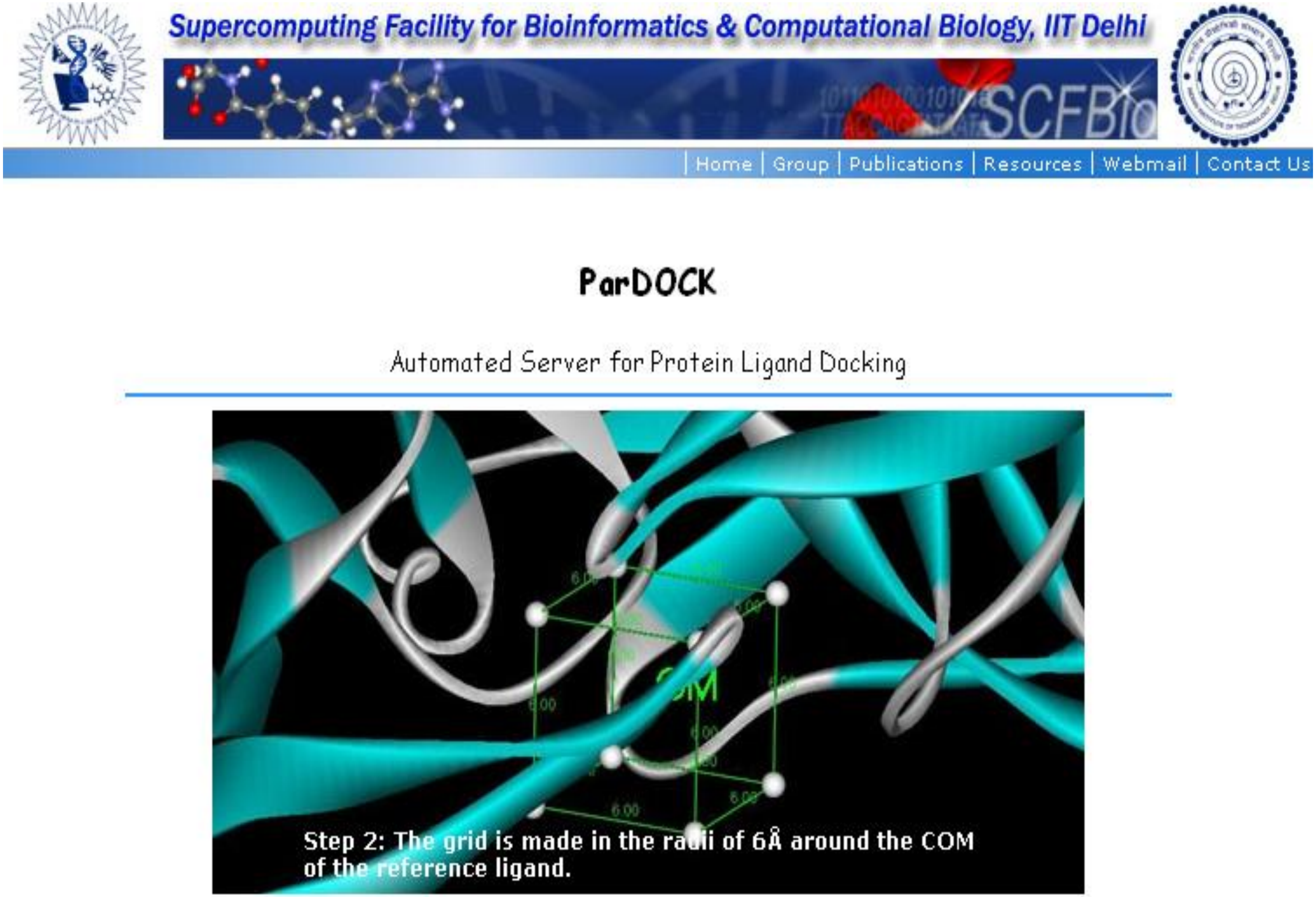




RMSD between the docked & the crystal structure is 0.2Å

ENERGY MINIMIZATION

STRUCTURES **WITH** LOWEST ENERGY SELECTED



Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi

| Home | Group | Publications | Resources | Webmail | Contact Us

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 Analysis of the Receptor (Target) -Candidate (Drug) Complex

Database for Experimental Binding Free Energy of Protein-Ligand Complexes

Protein Ligand Database *PLD* (http://www-mitchell.ch.cam.ac.uk/pld)
Ligand Protein Database *LPDB* (http://lpdb.scripps.edu/)
Protein Drug Binding Database *PDBbind* (http://www.pdbbind.org/)
Crystal Structure RCSB (http://www.rcsb.org/pdb/)

↓

Parameterization of Ligand

AM1 Geometry Optimization
HF/6-31G*/RESP Charge Derivation
Force Field Parameter Assignment

↓

Parameterization of Protein

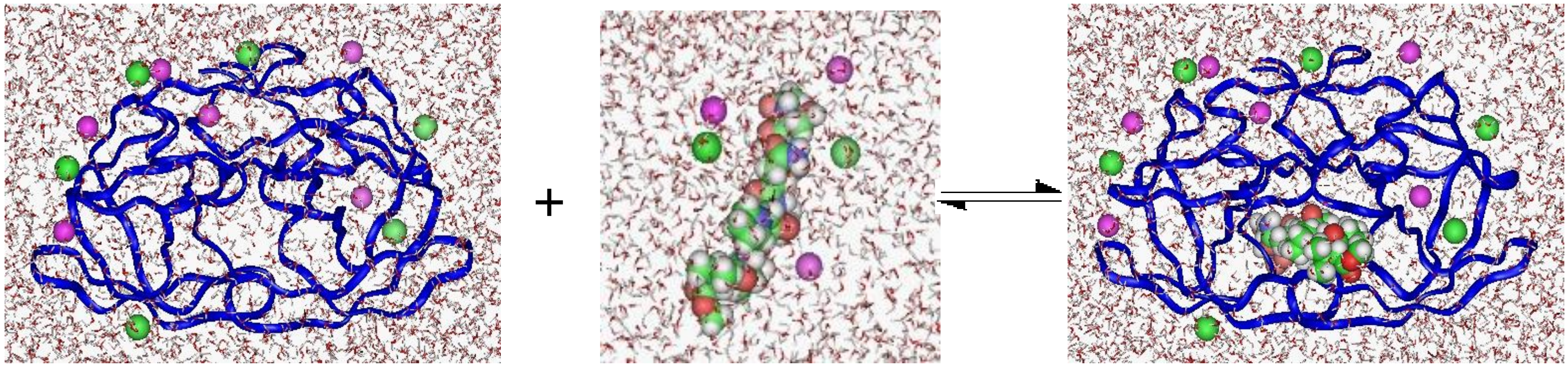
↓

Energy Optimization of the Complex

↓

Interaction Energy Calculation & Residue Wise Analysis

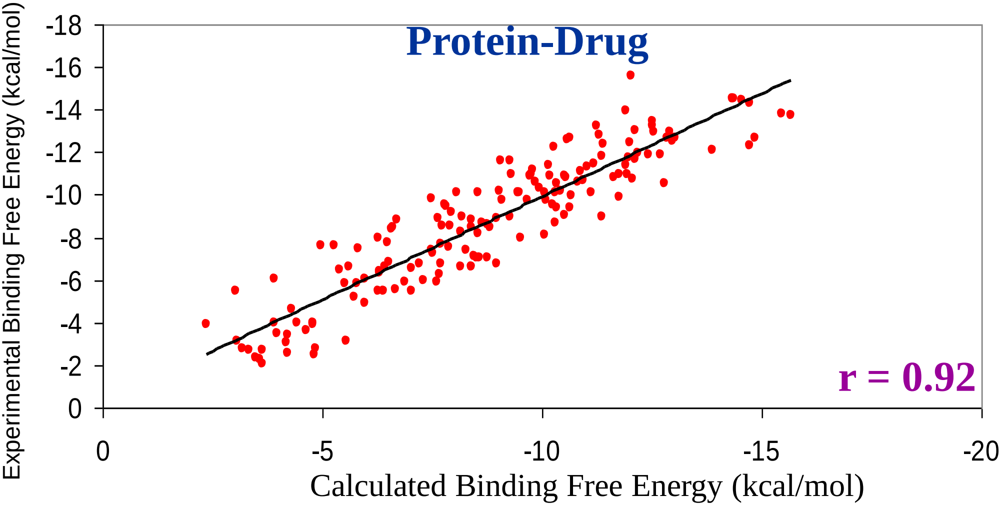
Binding Affinity Analysis



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

ENERGY BASED SCORING FUNCTION

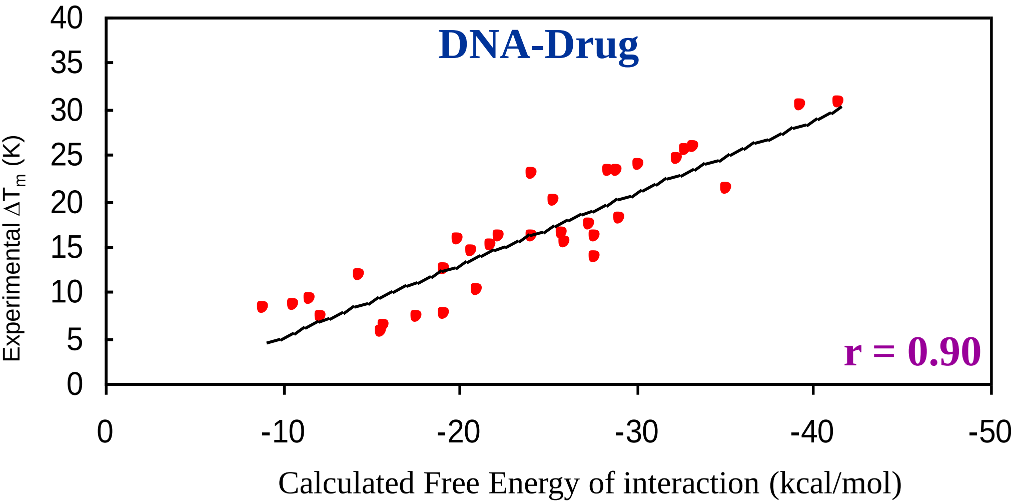
$$\Delta G_{bind} = \Delta H_{el} + \Delta H_{vdw} - T\Delta S_{rtvc} + \Delta G_{hpb}$$



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 $\Delta T_m$and calculated free energy of interaction for DNA-Drug Complexes

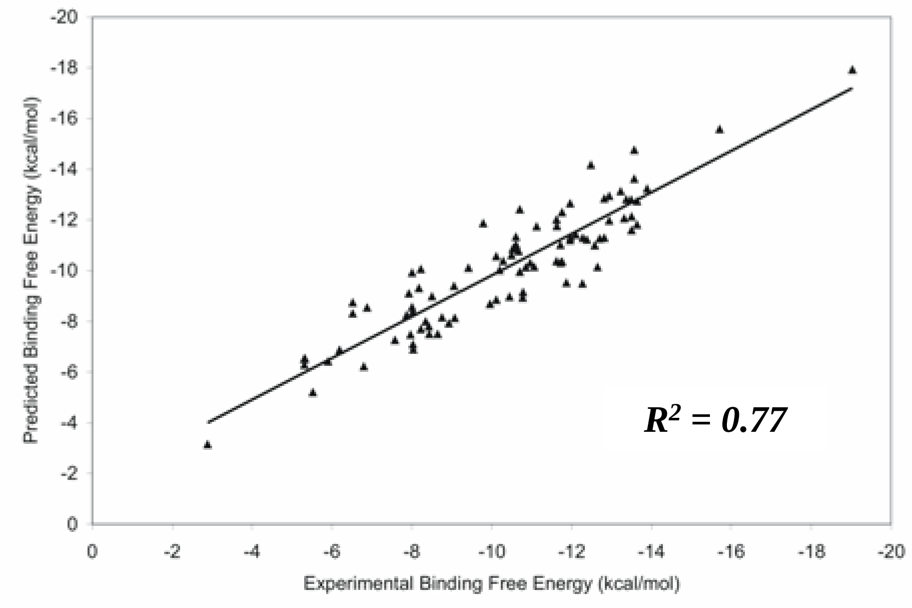
S.A Shaikh and B.Jayaram, *J. Med.Chem.* , **2007**, 50, 2240-2244

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

Comparative Evaluation of Scoring Functions

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

Binding Affinity Analysis on Zinc Containing Metalloprotein-Ligand Complexes



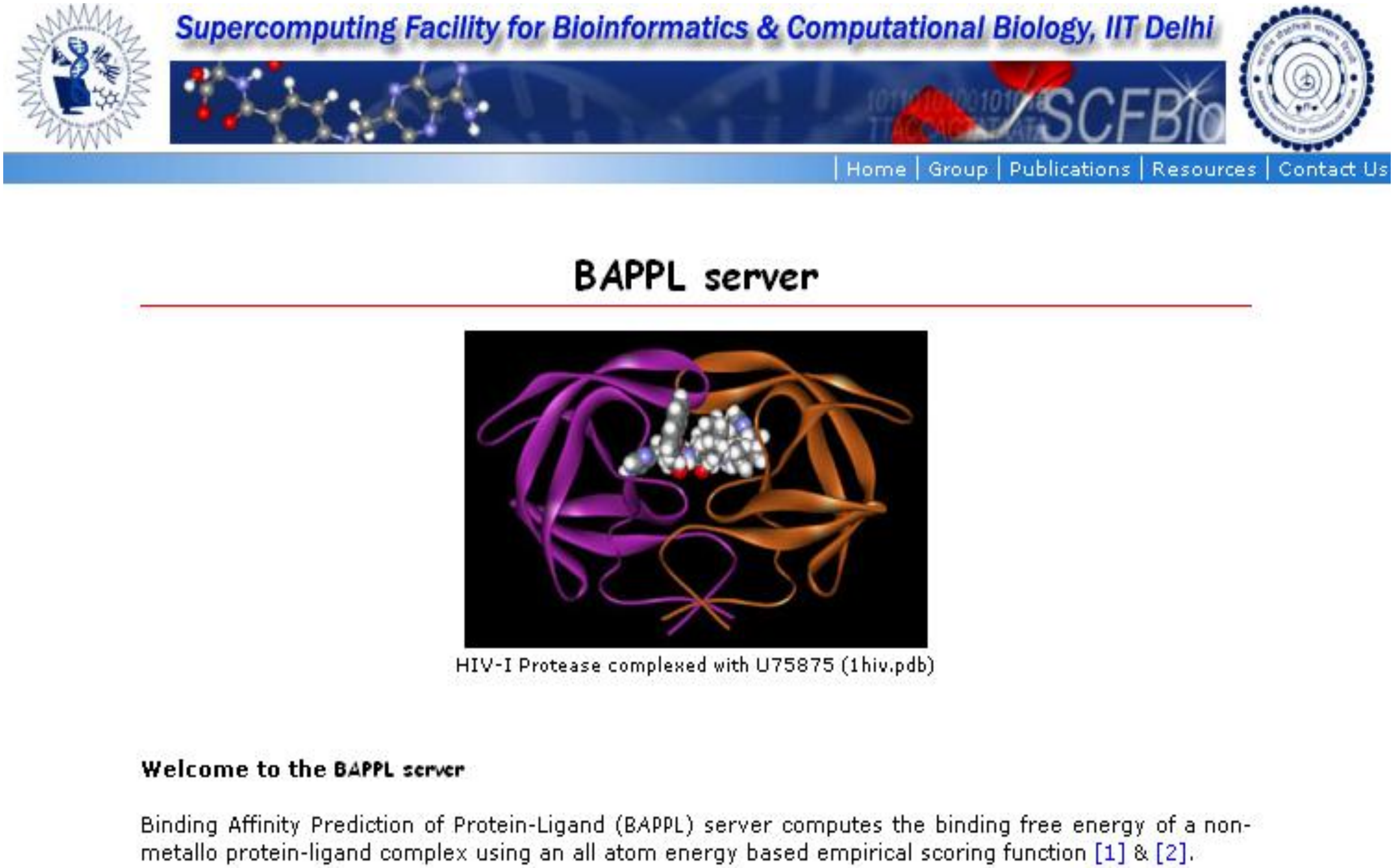
Correlation between the predicted and experimental binding free energies for 90 zinc containing metalloprotein-ligand complexes comprising 5 unique targets

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

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

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

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



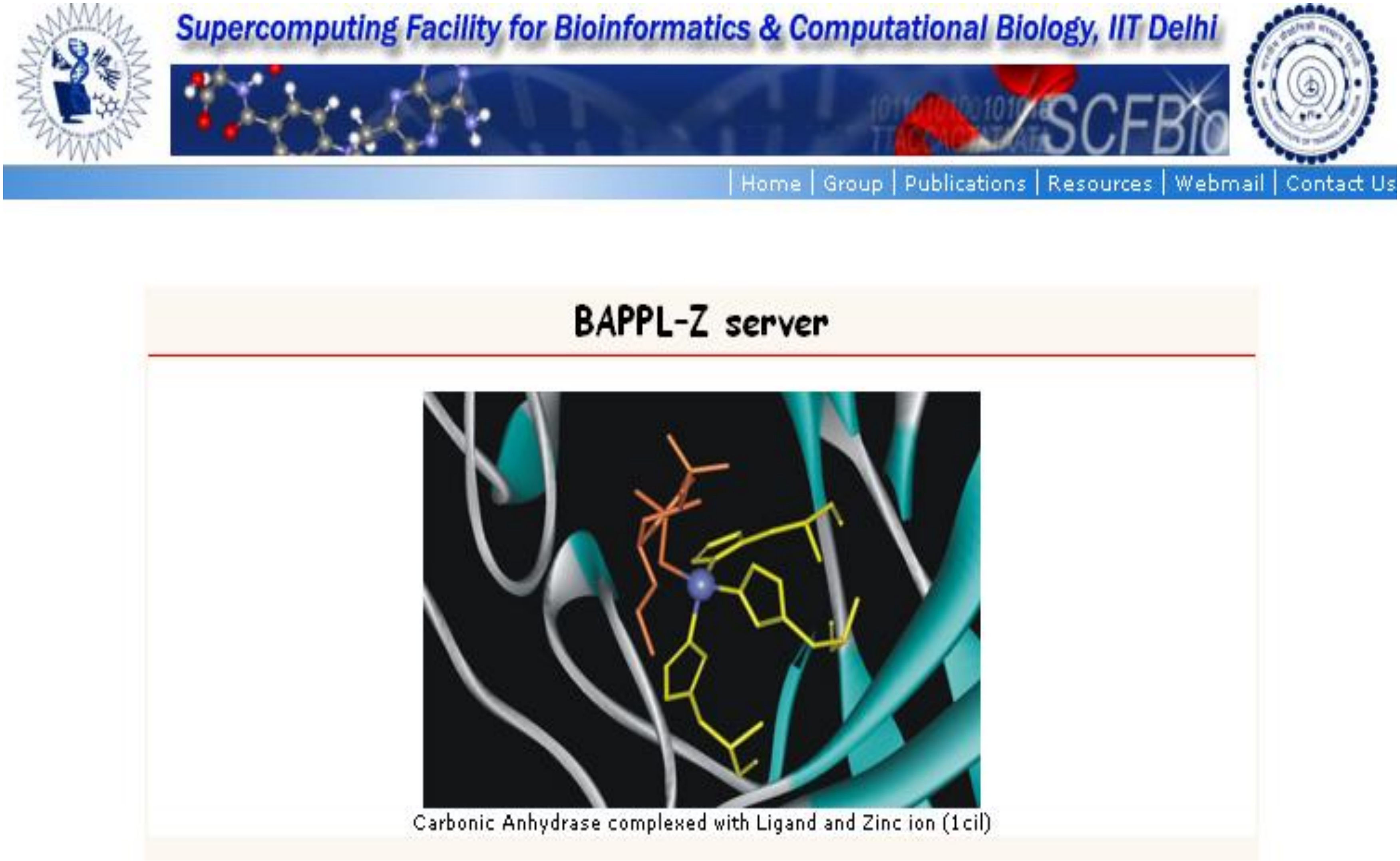
Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi
SCFBio
Home
Group
Publications
Resources
Contact Us

BAPPL server

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

Welcome to the BAPPL server

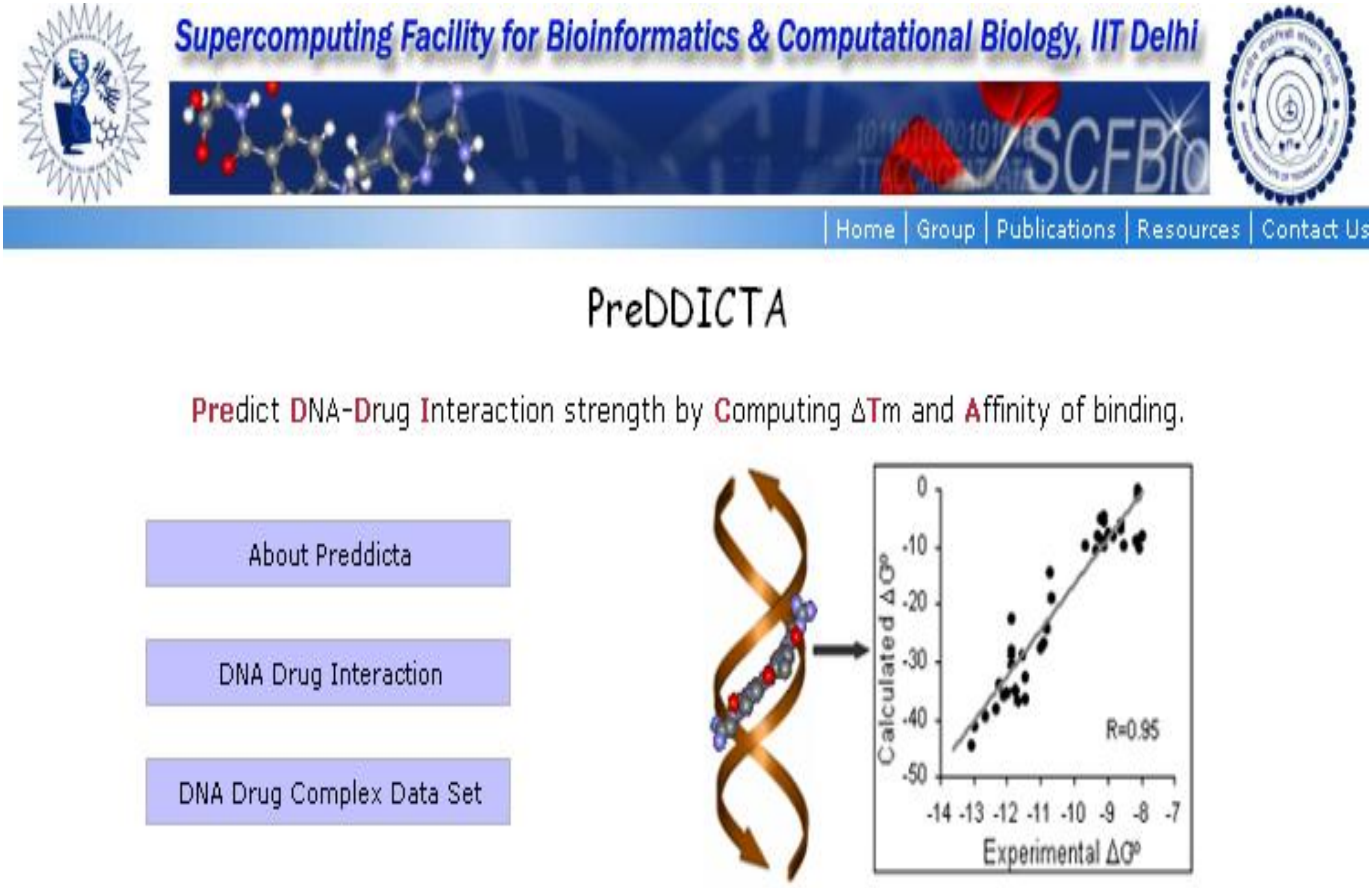
Binding Affinity Prediction of Protein-Ligand (BAPPL) server computes the binding free energy of a non-metallo protein-ligand complex using an all atom energy based empirical scoring function [1] & [2].



Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi

BAPPL-Z server

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



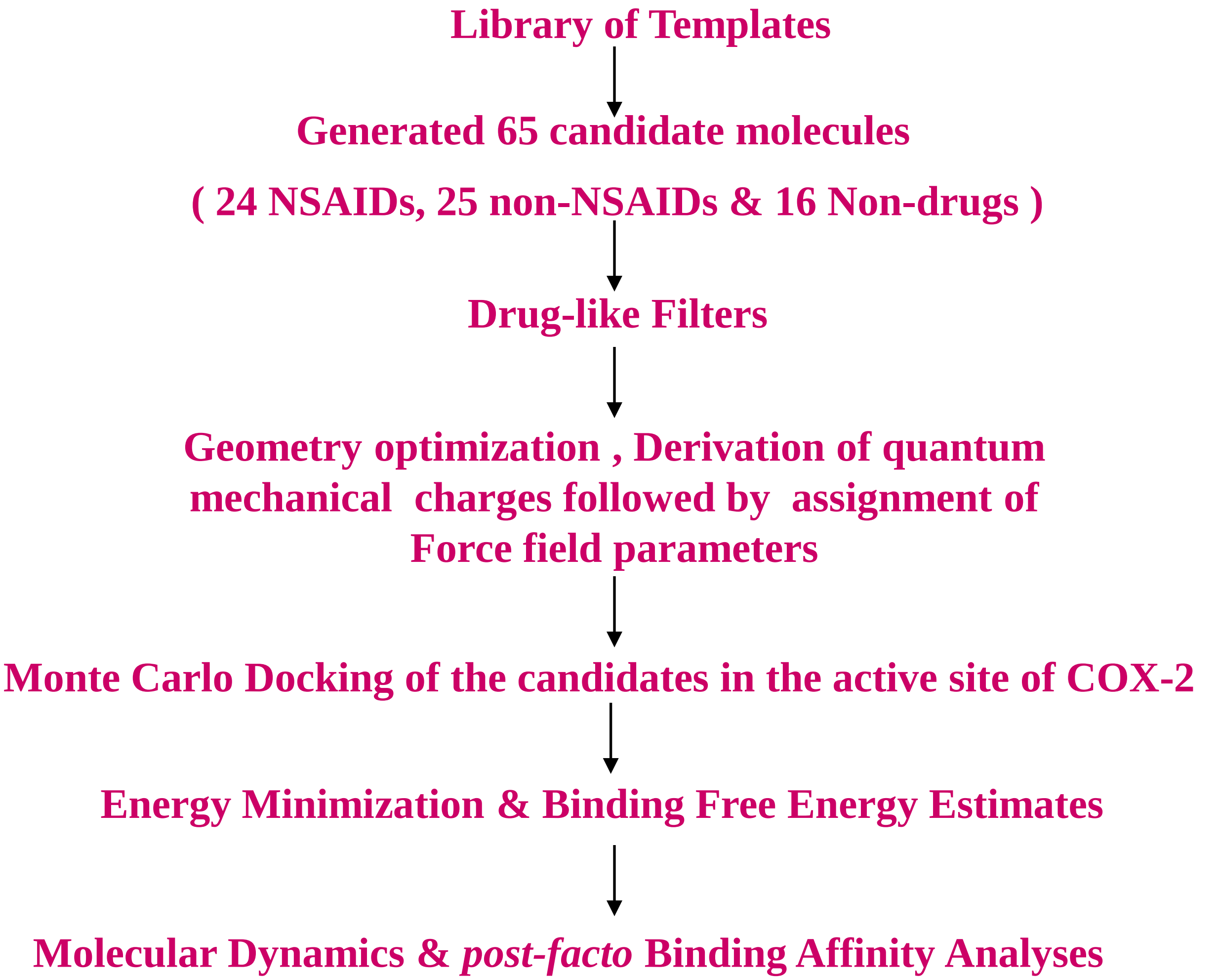
Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi

| Home | Group | Publications | Resources | Contact Us

PreDDICTA

Predict DNA-Drug Interaction strength by Computing ΔTm and Affinity of binding.

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



Library of Templates

↓

Generated 65 candidate molecules

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

↓

Drug-like Filters

↓

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

↓

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

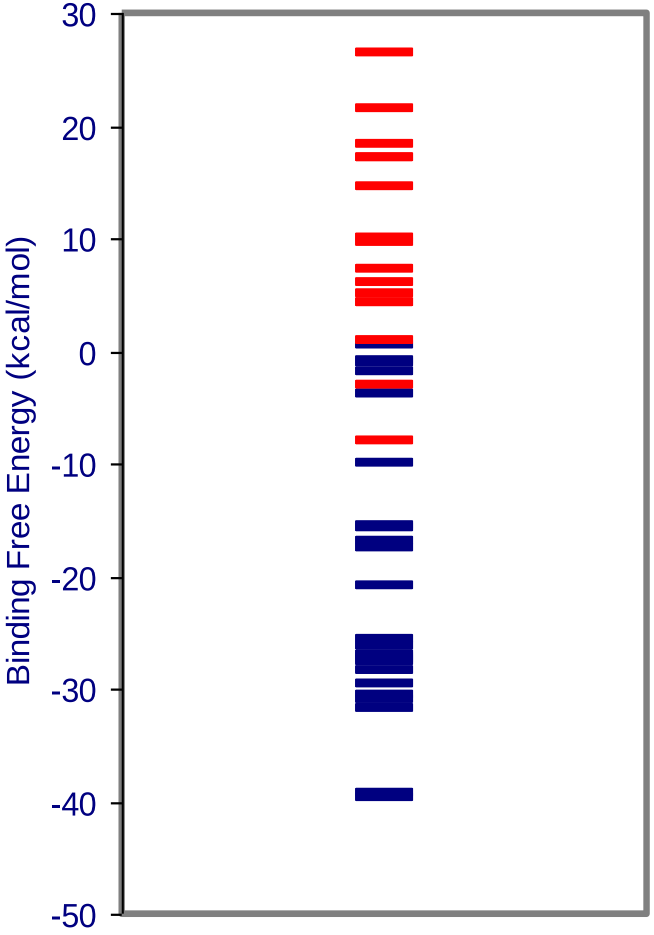
↓

Energy Minimization & Binding Free Energy Estimates

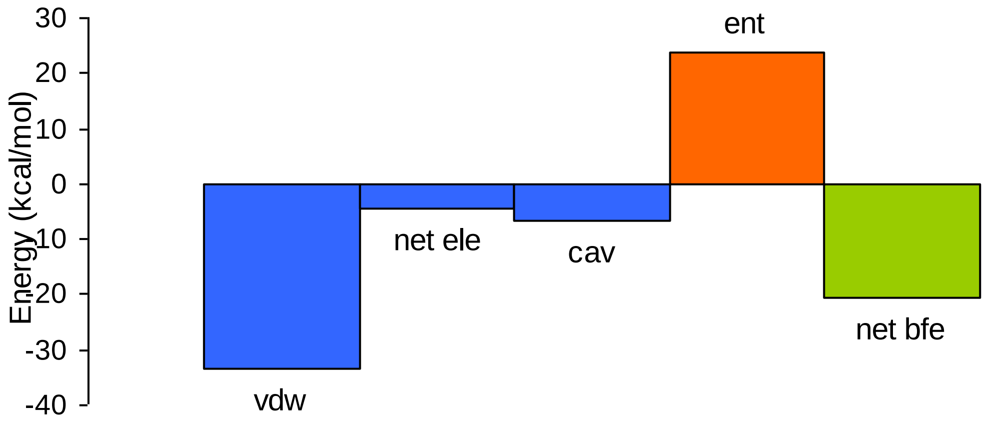
↓

Molecular Dynamics & *post-facto* Binding Affinity Analyses

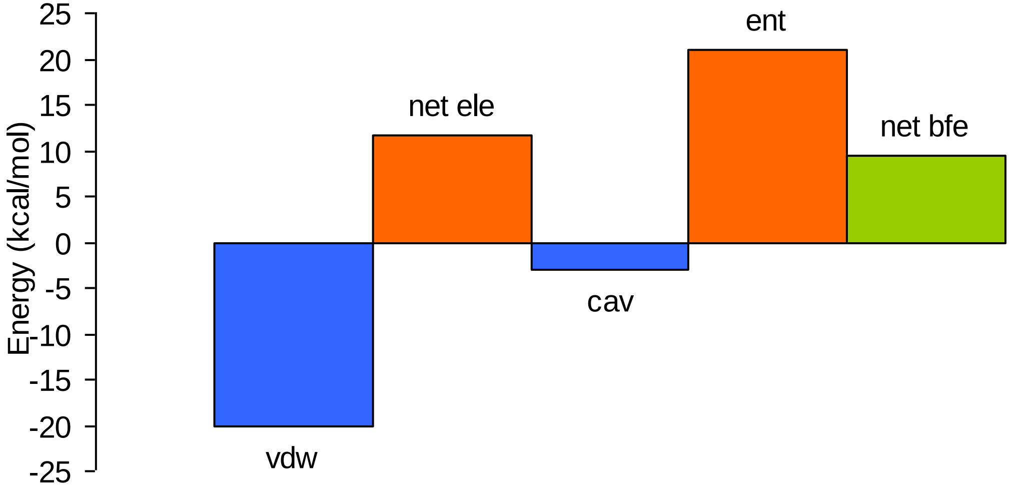
Sanjeevini1.0 distinguishes Drugs (NSAIDS, blue) from Non-Drugs (red) for Cyclooxygenase-2 Target



FREE ENERGY COMPONENT ANALYSIS



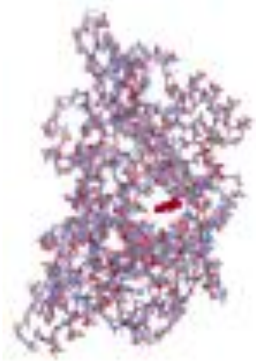
Drugs

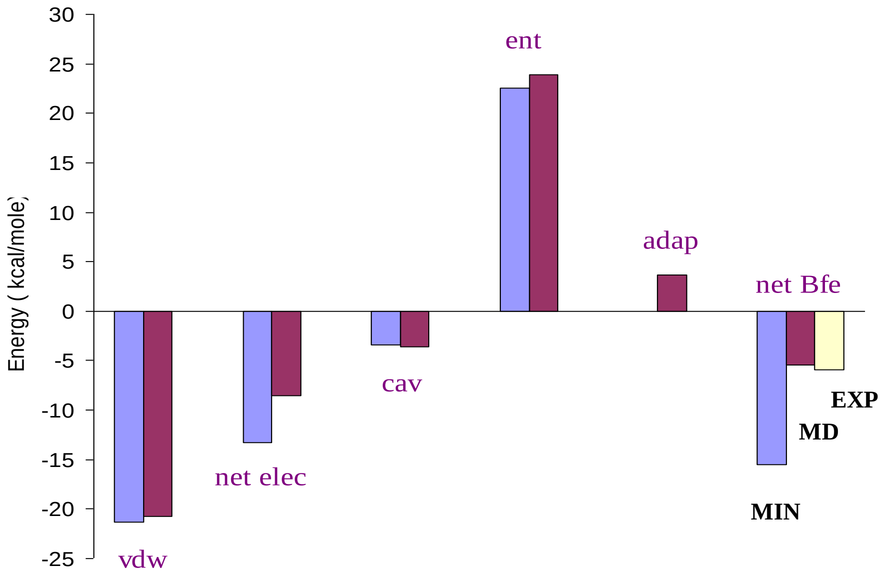


Non-Drugs

Free energy component analysis indicates how Drugs are energetically favoured over Non-drugs and facilitates further optimization of leads

Molecular Dynamics Simulations



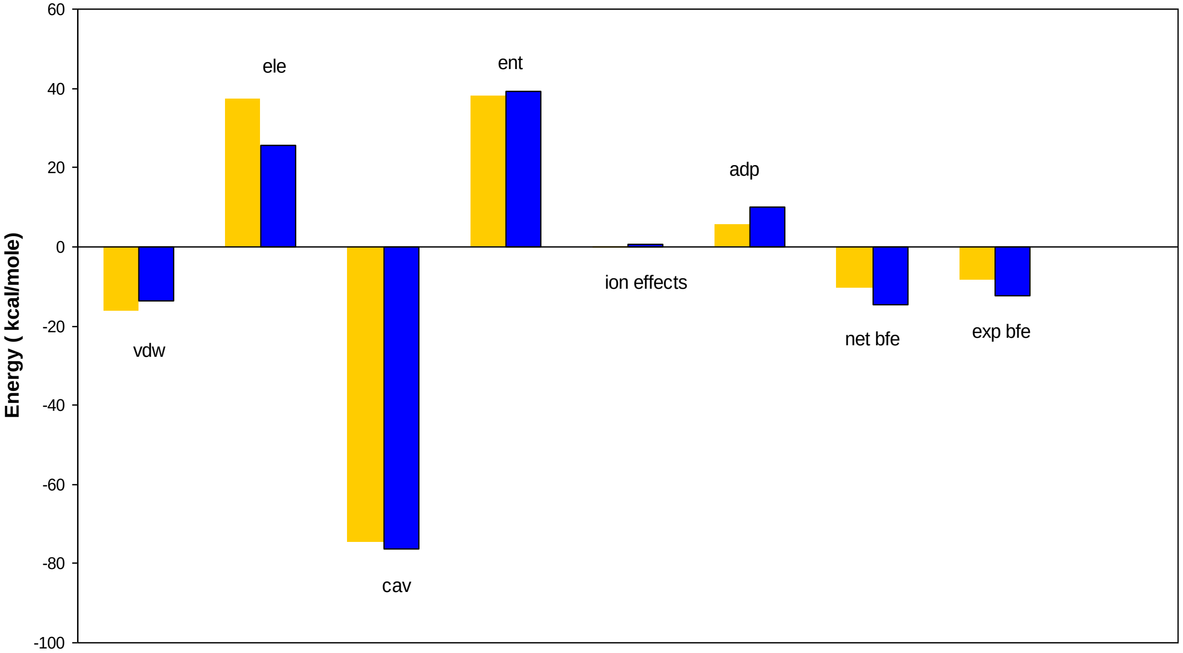


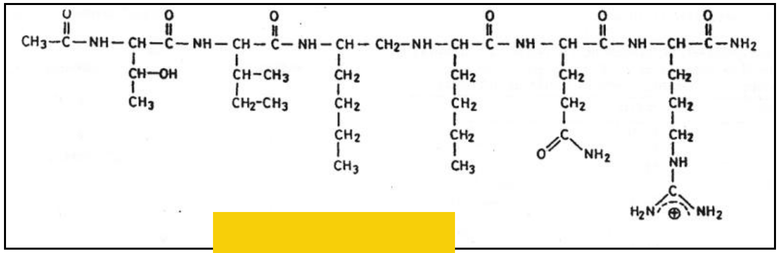
Energy components	*After minimization (kcal/mol)*	*Molecular dynamics (2 nanoseconds) (kcal/mol)*
van der Waals	- 21.3	-20.8
Net electrostatics	-13.3	-8.6
Cavitation	-3.4	-3.6
Entropy	22.5	23.9
Adaptation	0	3.7
Net binding free energy*	-15.5	- 5.4
Experimental binding free energy	-5.9	

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

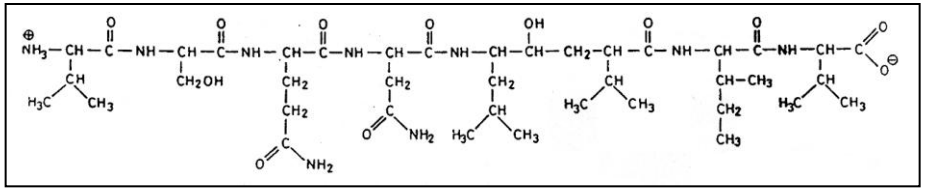
CONFIGURATIONAL AVERAGING ENHANCES THE QUALITY OF BINDING AFFINITY ESTIMATES

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





4hvp



8hvp

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

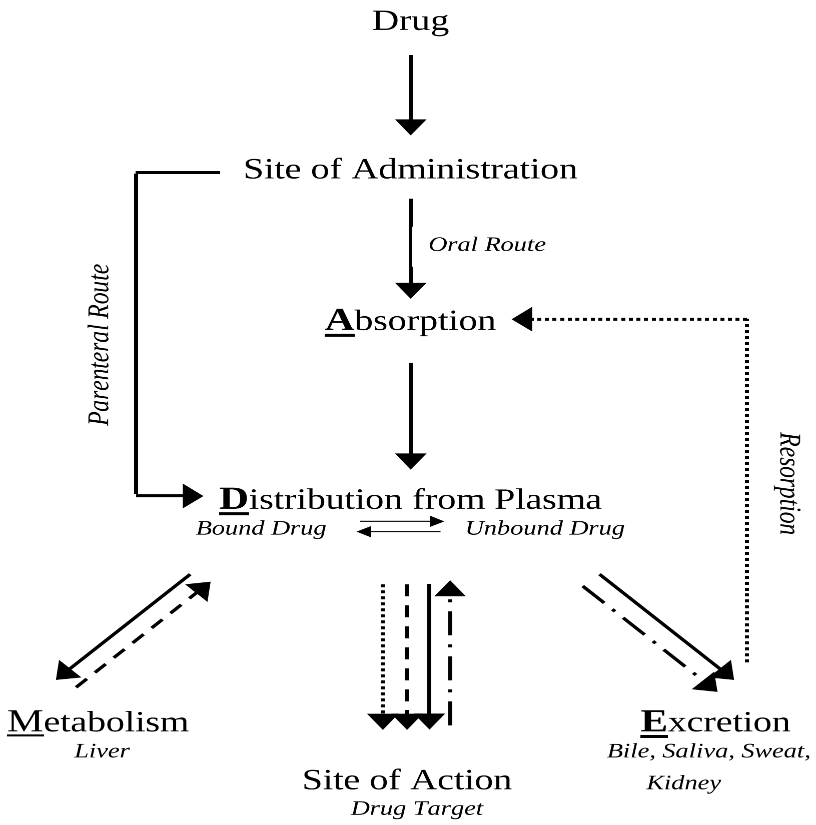
Shaikh, S., Jain. T., Sandhu, G., Latha, N., Jayaram., B., *A physico-chemical pathway from targets to leads*, 2007, *Current Pharmaceutical Design*, 13, 3454-3470.

	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



Drug

Site of Administration

Parenteral Route

Oral Route

Absorption

Resorption

Distribution from Plasma

Bound Drug ⇄ *Unbound Drug*

Metabolism

Liver

Site of Action

Drug Target

Excretion

Bile, Saliva, Sweat, Kidney

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.

SUMMARY

❖ ***Sanjeevini*** sorts out drugs from non-drugs for COX-2.

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

❖ A Scoring function has been developed for rapid assay of candidates to protein/DNA targets.

❖ Methodology has been configured in a high performance computing environment (70 UltraSparc III 900 MHz processor cluster with a compute power of over 100 Gigaflops).

❖ Work on other systems eg. nuclear receptor, DHFR, DNA targets is in progress.

❖ Development of a Lead-like molecular database with well defined force-field parameters.

❖ A number of tools for drug design are web-enabled for free access at

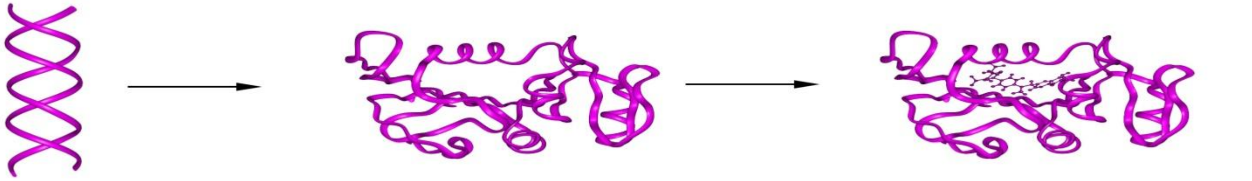
www.scfbio-iitd.res.in

Some Web Enabled Softwares at www.scfbio-iitd.res.in

Utility	Description	URL
ChemGenome 1.1	Gene Evaluator	www.scfbio-iitd.res.in/chemgenome/index.jsp
ChemGenome 2.0	A Physico-Chemical method for Whole Genome Analysis	www.scfbio-iitd.res.in/chemgenome/chemgenomenew.jsp
Bhageerath	An Energy Based Protein Structure Prediction Server	www.scfbio-iitd.res.in/bhageerath/index.jsp
ProSEE	Scoring Function for Protein Structure Evaluation	www.scfbio-iitd.res.in/utility/proteomics/energy.jsp
ProRegIn	Protein Regularity Index	www.scfbio-iitd.res.in/software/proregin/proregin.jsp
pardock	Protein-Ligand Docking	www.scfbio-iitd.res.in/dock/pardock.jsp
BAPPL	Binding Affinity Prediction of Protein-Ligand	www.scfbio-iitd.res.in/software/drugdesign/bappl.jsp
BAPPL-Z	Binding Affinity Prediction of Protein-Ligand Complexes Containing Zinc	www.scfbio-iitd.res.in/software/drugdesign/bapplz.jsp
PreDDICTA	Predict DNA-Drug Interaction strength by Computing ΔTm and Affinity of binding	www.scfbio-iitd.res.in/software/drugdesign/preddicta.jsp

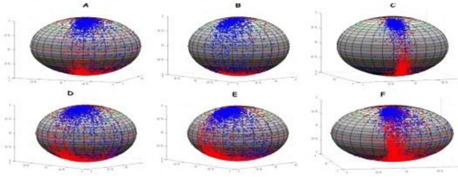
Gene to Drug

Bioinformatics suite developed at SCFBio, IIT Delhi



A Chemical Model for Genome Analysis

ChemGene 1.0

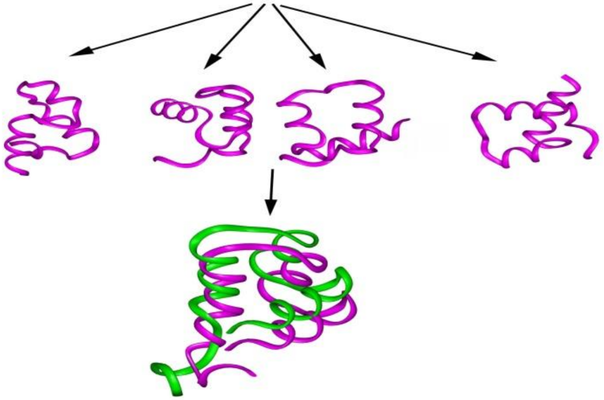


Gene (Blue) & Non Gene (Red) 120 Procaryotic genomes were evaluated & ~ 90 % sensitivity & specificity was observed

Protein Structure Prediction

Bhageerath 1.0

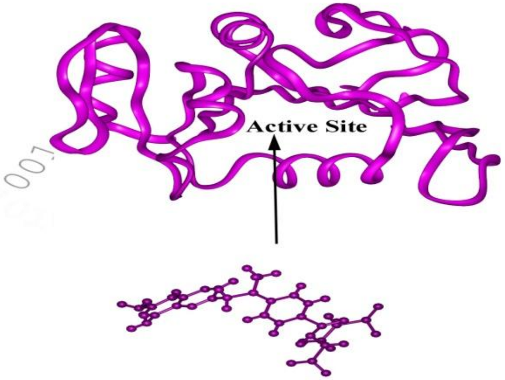
…………….GLU ALA GLU MET LYS ALA SER GLU ASP LEU LYS LYS HIS GLY VAL THR VAL LEU THR ALA LEU GLY ALA ILE LEU LYS LYS LYS GLY………………...



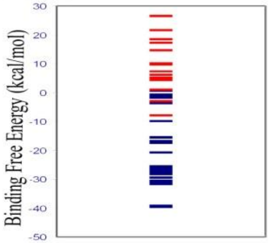
Bhageerath brackets the native like topology in the hundred best energy structure for small alpha helical proteins (green - native)

Active Site Directed Lead Design

Sanjeevini 1.0



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



BioGrid India

Vision

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

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

A Few Key References

(a) Dutta,S., Singhal,P., Agrawal,P., Tomer,R., Kritee, Khurana,E. and Jayaram.B. *A Physico-Chemical Model for Analyzing DNA sequences*, **2006**, ***Journal of Chemical Information & Modelling***, 46(1), 78-85. (b) 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.

(a), Narang,P, Bhushan,K., Bose,S. and Jayaram,B. *A computational pathway for bracketing native-like structures for small alpha helical globular proteins*. **2005**, ***Phys. Chem. Chem. Phys***., 7, 2364.; (b) Narang,P, Bhushan,K., Bose, S., Jayaram,B. *Protein structure evaluation using an all atom energy based empirical scoring function*, **2006**, ***J. Biomol. Struct. Dyn***., *23, 385-4006.* (c) Jayaram et al., Bhageerath, **2006**, ***Nucleic Acid Res***., 34, 6195-6204; (d) Jayaram, B.. Decoding the Design Principles of Amino Acids and the Chemical Logic of Protein Sequences. Available from ***Nature Precedings***. http://hdl.handle.net/10101/npre.2008.2135.1 **2008**

(a) Jain, T and Jayaram, B. *An all atom energy based computational protocol for predicting binding affinities of protein-ligand complexes*. **2005**, ***FEBS Letters***, 579, 6659; (b) Jain, T and Jayaram, B. *A computational protocol for predicting the binding affinities of zinc containing metalloprotein-ligand complexes*. **2007**, ***Proteins: Structure, Function & Bioinformatics***, 67, 1167-1178; (c) Shaikh, S., Jayaram. B., *A swift all atom energy based computational protocol to predict DNA-Drug binding affinity and ΔTm*, **2007**, ***J. Med. Chem***., 50, 2240-2244; (d) 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.

SCFBio Team

16 processor Linux Cluster

70 processor Sun Cluster; 26 dual core dual node AMD

Storage Area Network

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

Present

Dr. Sandhya Shenoy

Tanya Singh

Vandana

Pallavi Mohanty

Sanjeev Kumar

Shashank Shekhar

Priyanka Dhingra

Bharat Lakhani

Nagarajan

Garima Khandelwal

Goutam Mukherjee

Avinash Mishra

Preeti Bisht

Former

Dr. Achintya Das

Dr. Tarun Jain

Dr. Kumkum Bhushan

Dr. Nidhi Arora

Pankaj Sharma

A.Gandhimathi

Neelam Singh

Dr. N. Latha

Dr. Saher Shaikh

Dr. Poonam Singhal

Dr. E. Rajasekaran

Praveen Agrawal

Gurvisha Sandhu

Shailesh Tripathi

Dr. Pooja Narang

Dr. Parul Kalra

Dr. Surjit Dixit

Surojit Bose

Vidhu Pandey

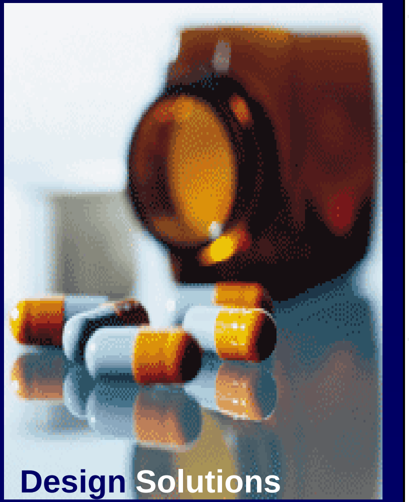
Anuj Gupta

Dhrubajyoti Biswas

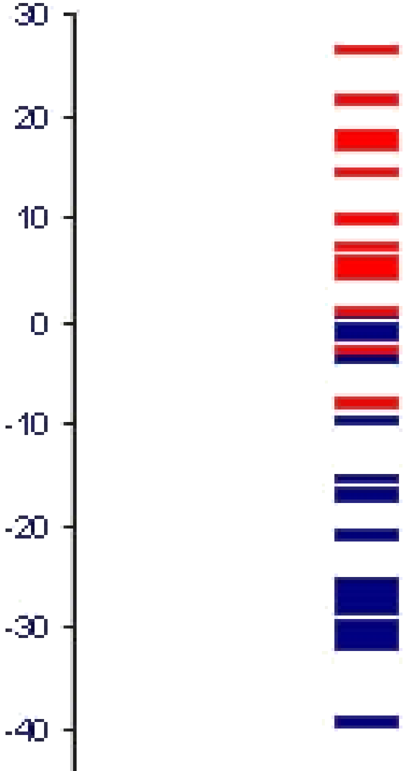
LeadInvent

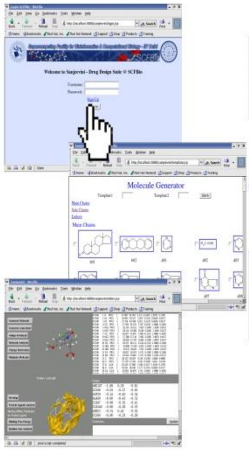
Technologies

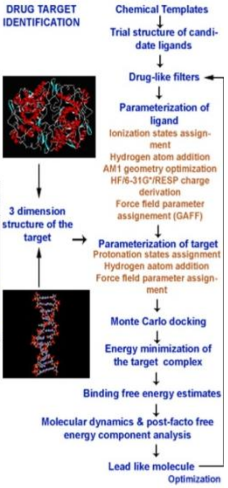
Novel Drug Discovery



Drug Design Solutions







DRUG TARGET IDENTIFICATION

3 dimension structure of the target →

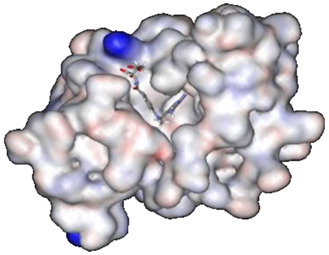
Chemical Templates ↓ Trial structure of candidate ligands ↓ Drug-like filters ↓ Parameterization of ligand

Ionization states assignment
Hydrogen atom addition
AM1 geometry optimization
HF/6-31G*/RESP charge derivation
Force field parameter assignement (GAFF)

↓ Parameterization of target

Protonation states assignment
Hydrogen aatom addition
Force field parameter assignment

↓ Monte Carlo docking ↓ Energy minimization of the target complex ↓ Binding free energy estimates ↓ Molecular dynamics & post-facto free energy component analysis ↓ Lead like molecule Optimization



An IIT Delhi Incubation

www.leadinvent.com

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)

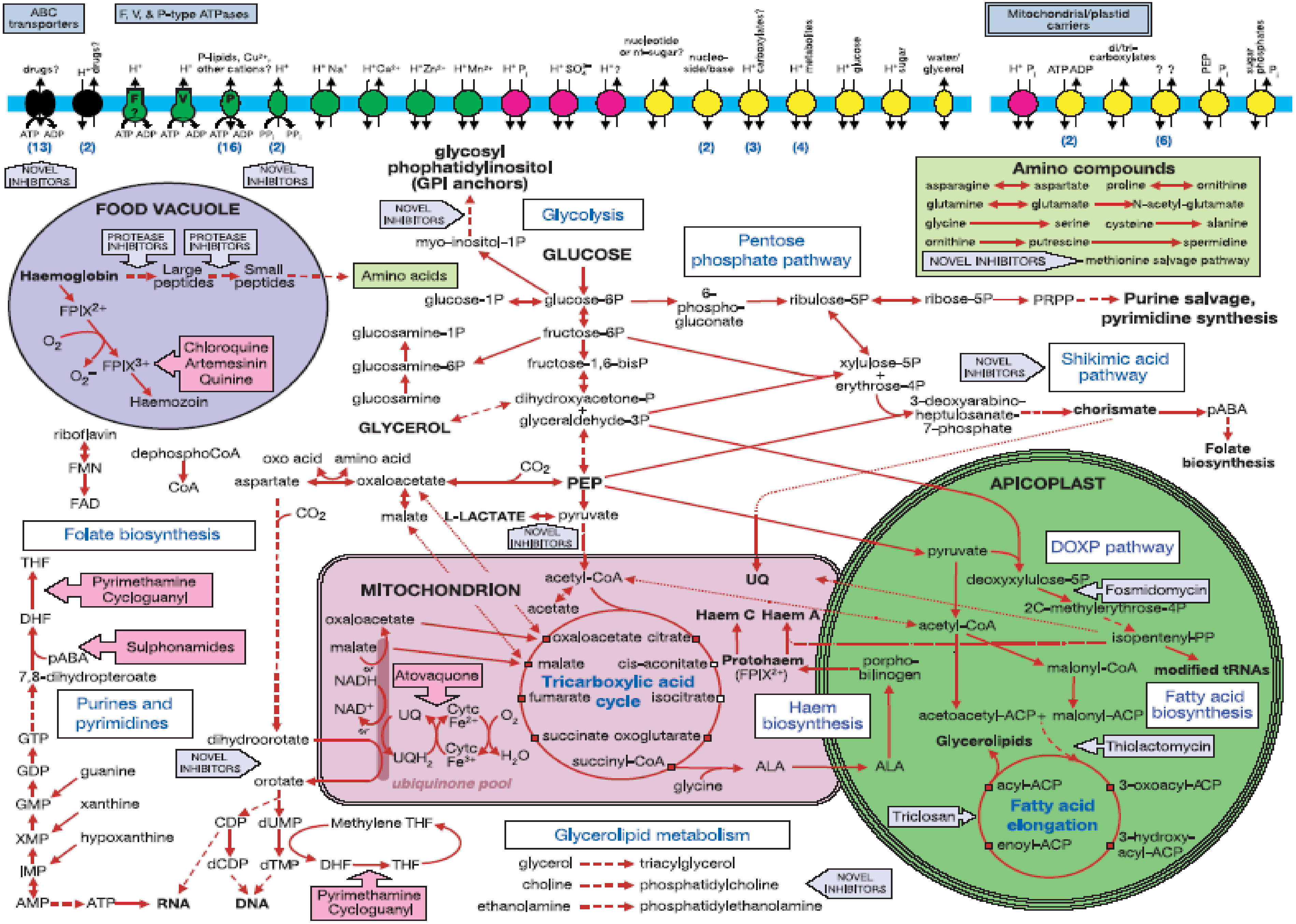
HCL Life Science Technologies

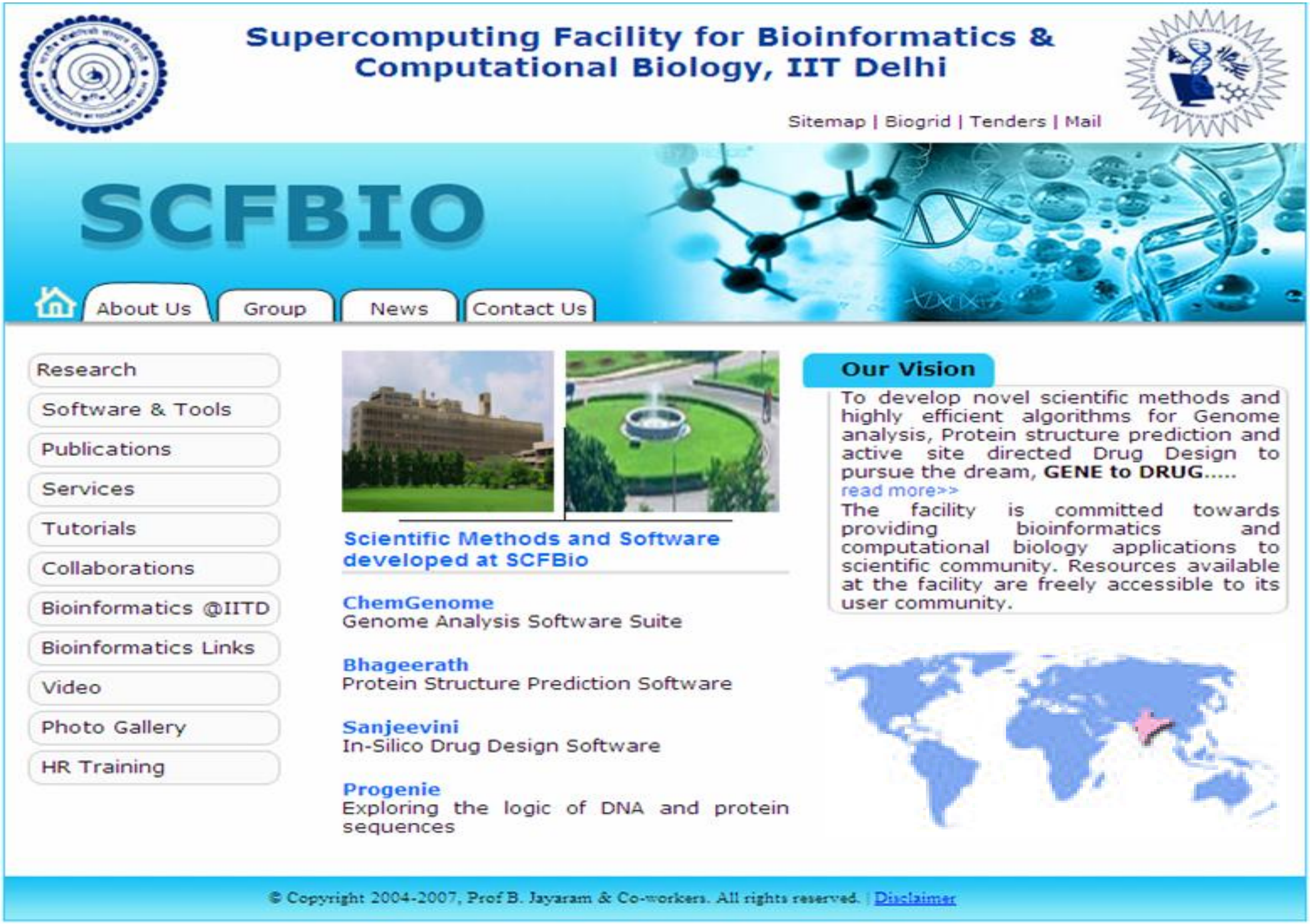
Dabur Research Foundation

Indian Institute of Technology, Delhi

Prof. D. L. Beveridge

OVERVIEW OF METABOLISM AND TRANSPORT IN *P.Falciparum*





Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi

Sitemap | Biogrid | Tenders | Mail

SCFBIO

Scientific Methods and Software developed at SCFBio

ChemGenome
Genome Analysis Software Suite

Bhageerath
Protein Structure Prediction Software

Sanjeevini
In-Silico Drug Design Software

Progenie
Exploring the logic of DNA and protein sequences

Our Vision

To develop novel scientific methods and highly efficient algorithms for Genome analysis, Protein structure prediction and active site directed Drug Design to pursue the dream, **GENE to DRUG.....**
read more>>

The facility is committed towards providing bioinformatics and computational biology applications to scientific community. Resources available at the facility are freely accessible to its user community.

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

Thank You