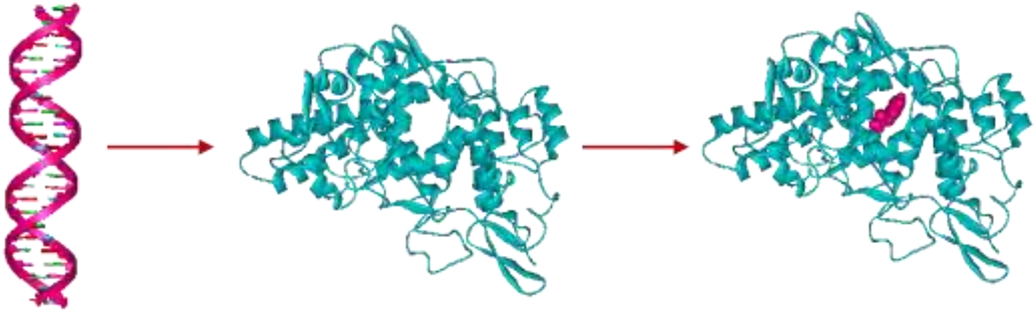
Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi
www.scfbio-iitd.res.in
A Centre of Excellence of the Department of Biotechnology, Govt. of India



Integrating Chemistry with Biology & IT:
Towards a disease-free Planet

Genomes to Hit Molecules in silico:
A country path today, A highway tomorrow

Prof. B. Jayaram

Department of Chemistry &

Kusuma School of Biological Sciences

Supercomputing Facility for Bioinformatics & Computational Biology &

Indian Institute of Technology Delhi

SCFBio: An Overview (2002 -2017)

SCFBio, IIT Delhi was created in July 2002 with funding from Department of Biotechnology, Govt. of India, under the guidance of Principal Investigator, Prof. B. Jayaram with a vision to develop novel scientific methods and new softwares for genome analysis, protein structure prediction, *in silico* drug design and for human resource training. The facility was inaugurated by Hon'ble Minister of Science and Technology and Human Resource Development Shri Murli Manohar Joshi in presence of Hon'ble Minister of State for S&T Shri Bachi Singh Rawat, IITD Director Prof. R.S. Sirohi, DBT Secretary Dr. Manju Sharma and other dignitaries. IITD adopted SCFBio as a Central Facility of National Importance in March, 2003.

In Dec 2013, SCFBio was recognized as a "**Centre of Excellence (CoE) in the area of Bioinformatics & Computational Biology**" by Dept. of Biotechnology, Govt. of India.

Upgradation to Multi-Tera Facility

- SCFBio was upgraded to a multi Teraflop facility under the Programme Support from DBT and inaugurated on 17th Sep, 2009 by Hon'ble Secretary, DBT, Dr. M.K. Bhan in presence of IITD Director, Prof. Surendra Prasad and other dignitaries.

- The aggregate compute power of the facility was over 6 Teraflops with a data storage of ~ 50 Terabytes. A modern data center was created to host the infrastructure.

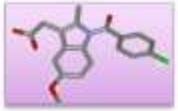
- Subsequently, the facility's capacity was upgraded to 16 Tera Flops of CPU+GPU based Clusters along with 200 Terabytes of Parallel File System Storage.

- The facility is connected via a 30 Mbps dedicated line.

The facility's compute capacity is soon (2018) to be upgraded to ~ 50 Teraflops

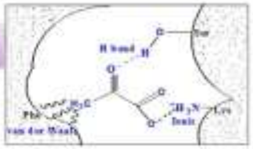
A Dream Plan

atggccctgtggatgcgcctcctgcccctgctggcgctgctggccctctggggacctgac.......
MALWMRLLPLLALLALWGPD.......



Sanjeevini App on mobile

Docking on a Supercomputer



Best Docked Structure

Send the lead molecule to xyz Pharmaceuticals for synthesis, clinical trials, approvals and recommendations on dosage etc. by a panel of doctors, all online

Medicine taken by patient

Turns him/her into a healthy person

10 AM: Disease reported on a smart phone along with genome card and other reports/ symptoms.

5 PM: Drug delivered at door step what ever the disease!

Goal: Personalized medicine:

Tools: Genomics + Proteomics + Information Technology + Chemistry

Some Achievements of SCFBio (2002-2017)

- *SCFBio is interpreting the language of Genomic DNA from a new physico-chemical perspective (Chemgenome). [Goal: One should be able to read genomes (including human) like Harry Potter novels!]*

- *SCFBio is addressing the Grand Challenge problem of protein tertiary structure prediction. SCFBio is the only Participant from India in the server category in the global Protein structure prediction Olympics called CASP. (BhageerathH+). SCFBio shares first rank globally for low resolution models and 11th rank officially for high resolution models (CASP12,2016).*

- *SCFBio developed a complete, freely accessible, indigenous, software suite for computer aided Drug Discovery (Sanjeevini) based on physico-chemical principles. SCFBio is called upon to implement Sanjeevini on National Supercomputing Mission (NSM) platform. A few molecules against malaria, Alzheimer's, breast cancer, HAV & HBV infections have been developed & published/patented.*

- *SCFBio developed over 43 freely accessible webservers (Complete list of software developed at SCFBio is available at http://www.scfbio-iitd.res.in/bioinformatics/bioinformaticssoftware.htm)*

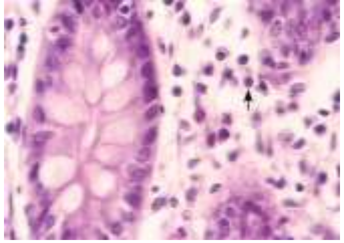
- *SCFBio published ~ 90 papers with an average impact factor of 4 + a Nature paper. (Complete list of publications is available at http://www.scfbio-iitd.res.in/publication/publication.htm)*

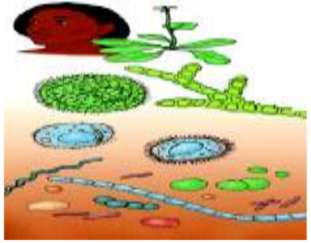
- *SCFBio organized an international conference (INCOB-2006), two Indo-Japan workshops (2010) and three national conferences (2011, 2012 & 2017).*

- *SCFBio is providing free access to its resources. (~ 20,000 hits per day from users in ~ 30 countries). (Hardware is accessible to users from India and software to users from across the world)*

- *SCFBio trained ~ 1000 students through short- and long-term training programmes & produced 18 PhDs. (For a list of trainees, please see http://scfbio-iitd.res.in/training/training.htm)*

- *Two start-up companies evolved (Leadinvent and Novo informatics) so far from SCFBio.*





Today’s challenge: (Big) Data → Information → Knowledge → Products useful to Society

Hypotheses generation & validation

What is big data in biology?

Vivien Marx, Biology: The big challenges of big data, Nature, 498, 255-260 (2013)

NCBI: http://www.ncbi.nlm.nih.gov/ : Genomic information of more than 70000 organisms

UNIPROT: http://www.uniprot.org/: Protein sequence information of more than 88 million entries

RCSB: Protein Data Bank: http://www.rcsb.org/: Structural information of ~ 134000 biomolecules

Zinc Database: http://zinc.docking.org/: Over 35 million purchasable compounds

Human genome ~ 3 GB
Genomes + Proteomes + Small molecules ~ hundreds of Petabytes
+ Gene expression data + PPI Networks + ….. ~ Exabytes

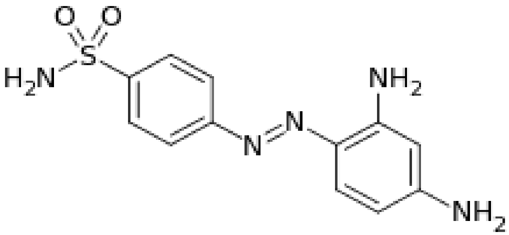
Let us use the big data, make some drugs and get rid of diseases!

Note: Information on Nobels and Nobel laureates in this presentation is collected from www.nobelprize.org

Some milestones in drug discovery

The Nobel Prize in Physiology or Medicine 1939

"for the discovery of the antibacterial effects of prontosil"



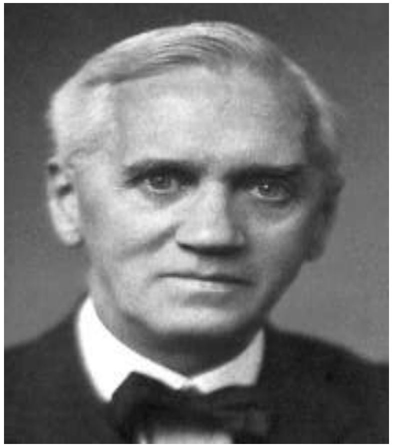
Prontosil – a synthetic antibacterial compound

Gerhard Domagk
Munster U, Germany
b. 1895 (Germany)

Prontosil is a derivative of sulfanilamide (*p*-aminobenzenesulphonamide). Some thousands of derivatives of sulphanilamide have been produced and tested for their antibacterial properties. Domagk's work has thus given to medicine, and also to surgery, a whole new series of weapons that are effective against many infectious diseases. Later, he attacked the problem of the chemotherapy of tuberculosis, developing for this the thiosemicarbazones (Conteben) and isonicotinic acid hydrazide (Neoteben). The supreme aim of chemotherapy is, in Domagk's opinion, the cure and control of carcinoma and he was convinced that this will be, in the future, achieved.

The Nobel Prize in Physiology or Medicine 1945

"for the discovery of penicillin and its curative effect in various infectious diseases"



Sir Alexander Fleming
London U., UK, b. 1881 (Scotland)

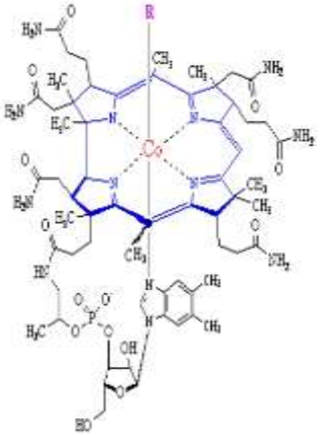
Ernst B Chain
U Oxford UK, b. 1906 (Germany)

Sir Howard Florey
U Oxford UK, b. 1898 (Australia)

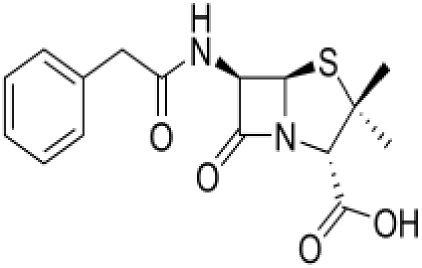
The Nobel Prize in Chemistry 1964

"for her determinations by X-ray techniques of the structures of important biochemical substances“

Chemists knew that penicillin consisted of 27 atoms: 11 hydrogen, 9 carbon, 4 oxygen, 2 nitrogen atoms and 1 sulphur atom. The trouble was that this combination of atoms could form two very different structures, and chemists couldn't decide which structure was more likely. Some chemists were convinced the structure contained two five-membered rings connected by a single bond, known as a thiazolidine-oxazolone. Others were equally sure it was a four-membered ring fused to a five-membered ring, known as a beta lactam. "The final solution of the problem of the structure of penicillin came from crystallographic X-ray studies."



Dorothy Crowfoot Hodgkin
U Oxford UK
b. 1910 (Egypt)



The enchanting β– lactam

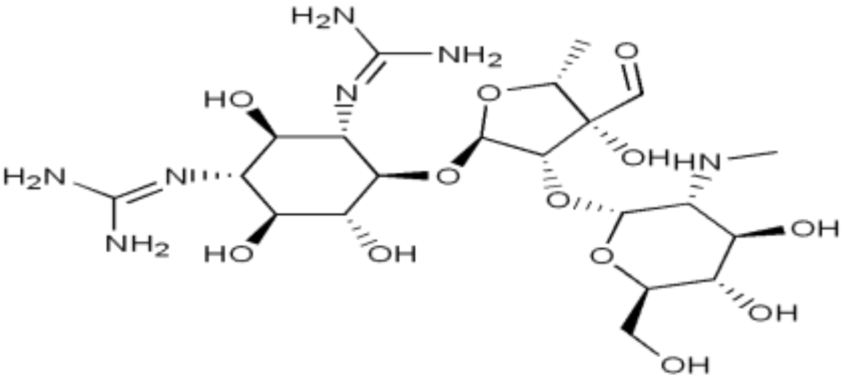
She also solved the structure of Vitamin-B12, in addition to penicillin.

Knowledge of the penicillin structure finally opened new avenues for creating and developing semi-synthetic derivatives of penicillin – such as the cephalosporines – that sparked the creation of antibiotic treatments.

Source: www.nobelprize.org

The Nobel Prize in Physiology or Medicine 1952

“for his discovery of streptomycin, the first antibiotic effective against tuberculosis“

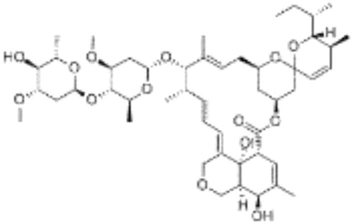


Streptomycin – an antibacterial compound – the first of a class of drugs called aminoglycosides – the first effective treatment against tuberculosis

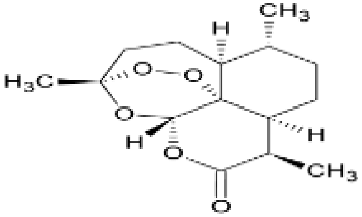
Selman A. Waksman
Rutgers U, USA
b. 1888 (Ukraine)

“He has isolated, together with his students and associates, a number of new antibiotics, including actinomycin (1940), clavacin, streptothricin (1942), streptomycin (1943), grisein (1946), neomycin (1948), fradicin, candicidin, candidin, and others. Two of these, streptomycin and neomycin, have found extensive application in the treatment of numerous infectious diseases of men, animals and plants. They have been covered by patents, that on streptomycin having been recently listed as one of the ten patents that shaped the world.”

Source: www.nobelprize.org



The Nobel Prize in Physiology or Medicine 2015



for their discoveries concerning a novel therapy against infections caused by roundworm parasites and malaria

William C. Campbell
Drew U., NJ, USA
b. 1930, (Ireland)

Satoshi Omura
Kitasato U, Japan
b. 1935 (Japan)

Youyou Tu
China Academy, China
b. 1930 (China)

Avermectin for river blindness and lymphatic filariasis & Artemisinin for malaria

Source: www.nobelprize.org

How to design drugs?

The Nobel Prize in Physiology or Medicine 1988

"for their discoveries of important principles for drug treatment"

While drug development had earlier mainly been built on chemical modifications of natural products, the laureates introduced a more rational approach based on the understanding of basic biochemical and physiological processes

James W Black
London U., UK
b. 1924 (Scotland)

Gertrude B Elion
WRL USA
b. 1918 (USA)

George H Hitchings
WRL, USA
b. 1905 (USA)

JB: Pharmacotherapeutic potential of receptor blocking drugs: betablocking drug-propanolol, characterized histamine receptors, H2 receptor antagonist-cimetidine

GE & GH: Thiogaunanine, 6-mercaptopurine for leukemia, pyrimethamine for malaria, azathioprine for preventing organ rejection, allopurinol for gout etc…

www.nobelprize.org

The Nobel Prize in Chemistry 2009

"for studies of the structure and function of the ribosome”

The ribosomes (30S & 50S) in bacteria are different from their counterparts in animals / humans (40S & 60S) and hence constitute a good target for new antibiotics

Venkatraman
Ramakrishnan
Cambridge, UK
b. 1952 (India)

Thomas A. Steitz
Yale University , USA
b. 1940 (USA)

Ada E. Yonath
Weizmann Institute,
Israel
b. 1939 (Israel)

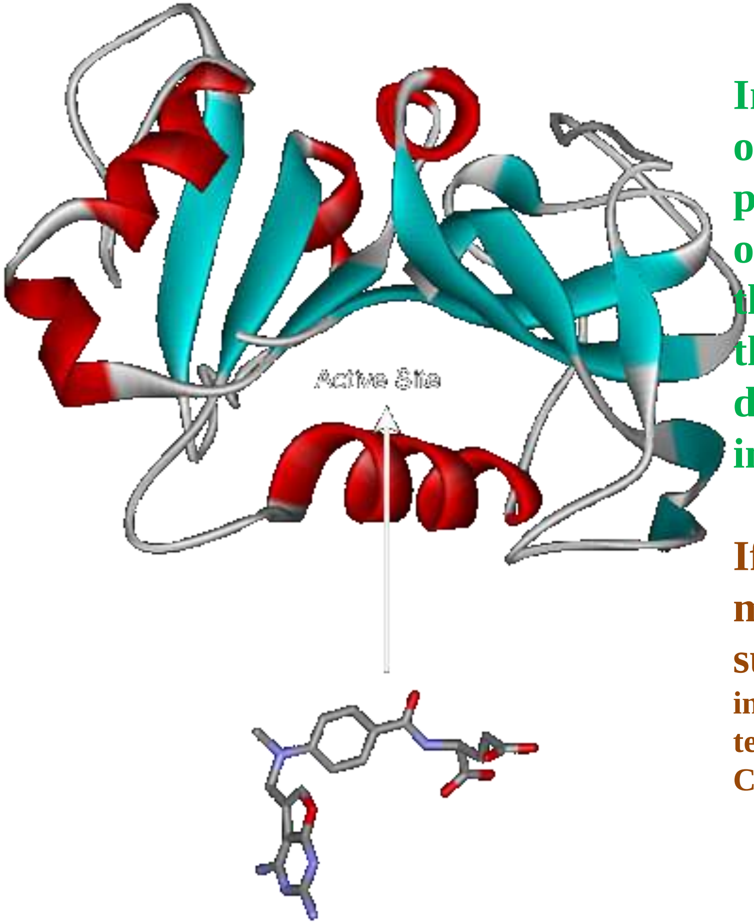
www.nobelprize.org

Rational Drug Design

Structure Based Drug Discovery

Drug molecule is like a duplicate key to jam (inhibitor) the lock (a biomolecular target) or open (activator) the lock.

Thus structure of the biomolecular target – the shape of the lock and the key hole – become important in designing the keys – the drugs. These are molecules. They are dynamic and they are surrounded by solvent, salt and other biomolecules in a cellular milieu…



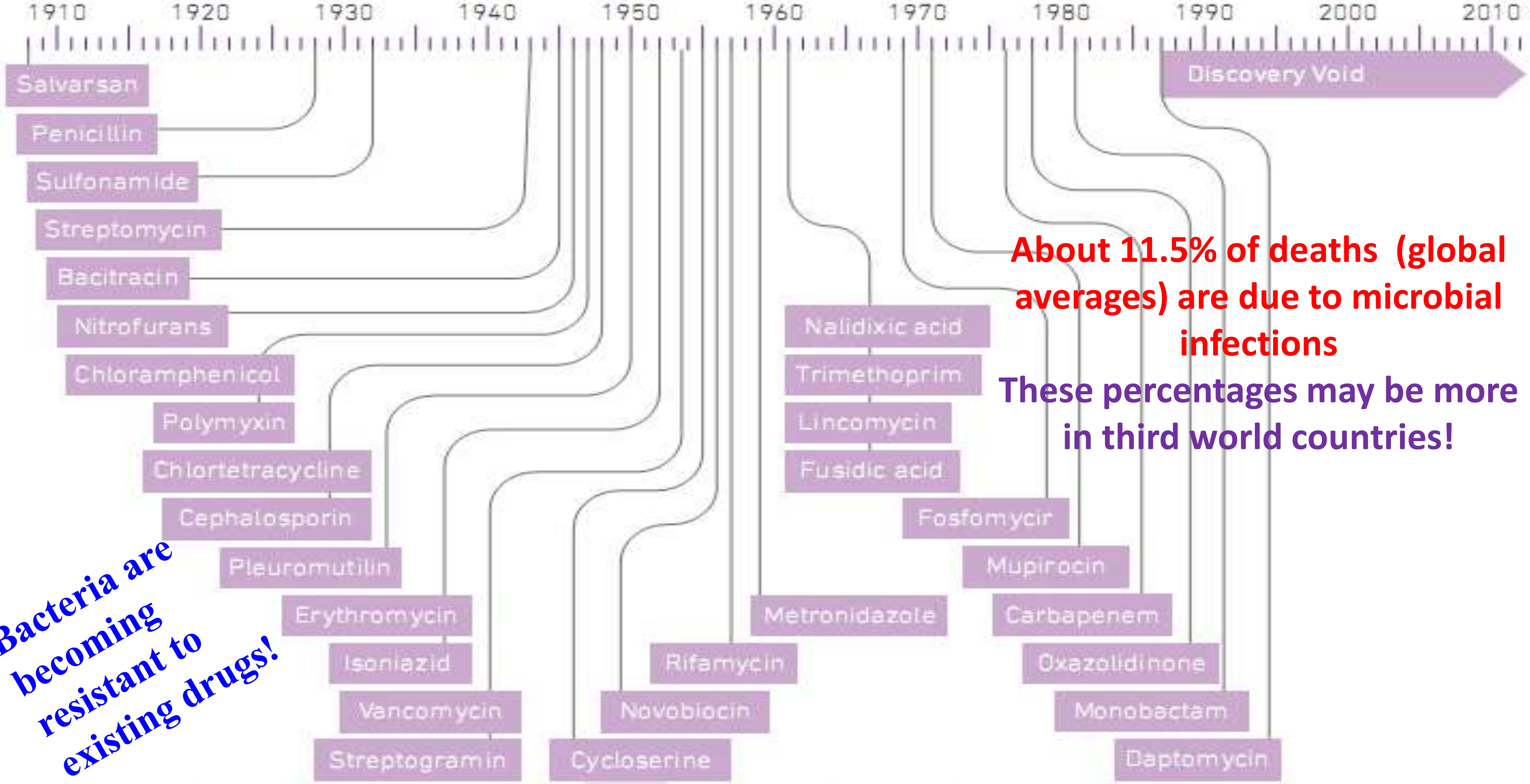
In a simplified view, a disease or a disorder can be traced to a protein going aberrant, lazy or overactive. Need activators for the former and inhibitors for the latter to cure disease / disorder. Most drugs are inhibitors.

If an essential protein is missing, it needs to be supplemented…gene therapy..insulin injections…or via some newer technologies of genome editing such as CRISPR/CAS9 etc..

Proteins thus far have been the most attractive choice for drug discovery. However, with advances in nanobiotechnology and drug delivery systems, DNA too could become a popular drug target.

Figure 1 Dates of discovery of distinct classes of antibacterial drugs

Illustration of the "discovery void." Dates indicated are those of reported initial discovery or patent.



About 11.5% of deaths (global averages) are due to microbial infections

These percentages may be more in third world countries!

Bacteria are becoming resistant to existing drugs!

Adapted from Silver 2011 (1) with permission of the American Society of Microbiology Journals Department.

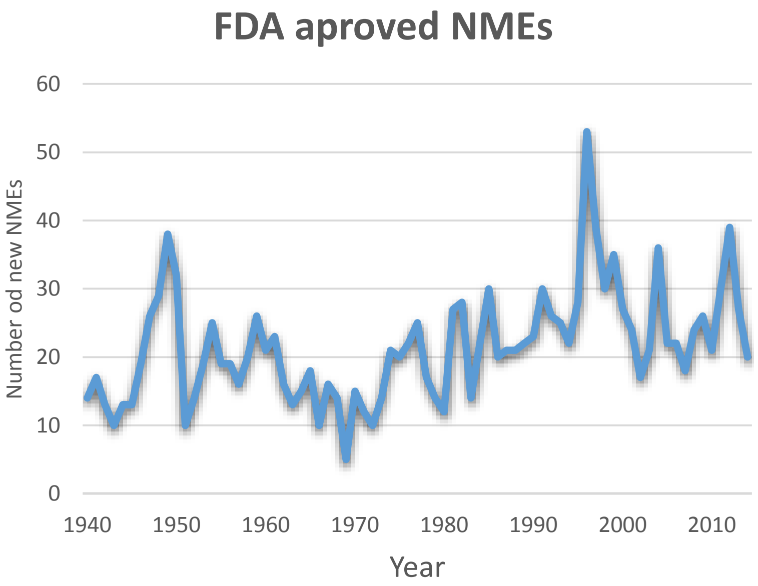
The fall from the bicycle did not kill but the scratch on the hand killed! That is what AMR can do!!

WHO-AMR Report 2014

Pharmaceutical R&D is Expensive

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

Total number of new molecular entities approved over the last 75 years: ~ 2500!



Drug Development is an Uphill Task
Of the new drugs approved by FDA
Only 35% were New Molecular Entities (NME).
Only 15% were deemed to provide significant improvement over existing medicines.

Drug discovery pipe-line is drying up / despite massive increase in genomic proteomic data.

Millions of molecules are available in databases and so many more are getting synthesized every day in organic chemistry laboratories all over the world.

Only 2500 drugs in 75 years?

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

COST & TIME INVOLVED IN DRUG DISCOVERY

Target Discovery

2.5yrs ↓ *4%*

Lead Generation

Lead Optimization

3.0yrs ↓ *15%*

Preclinical Development

1.0yrs ↓ *10%*

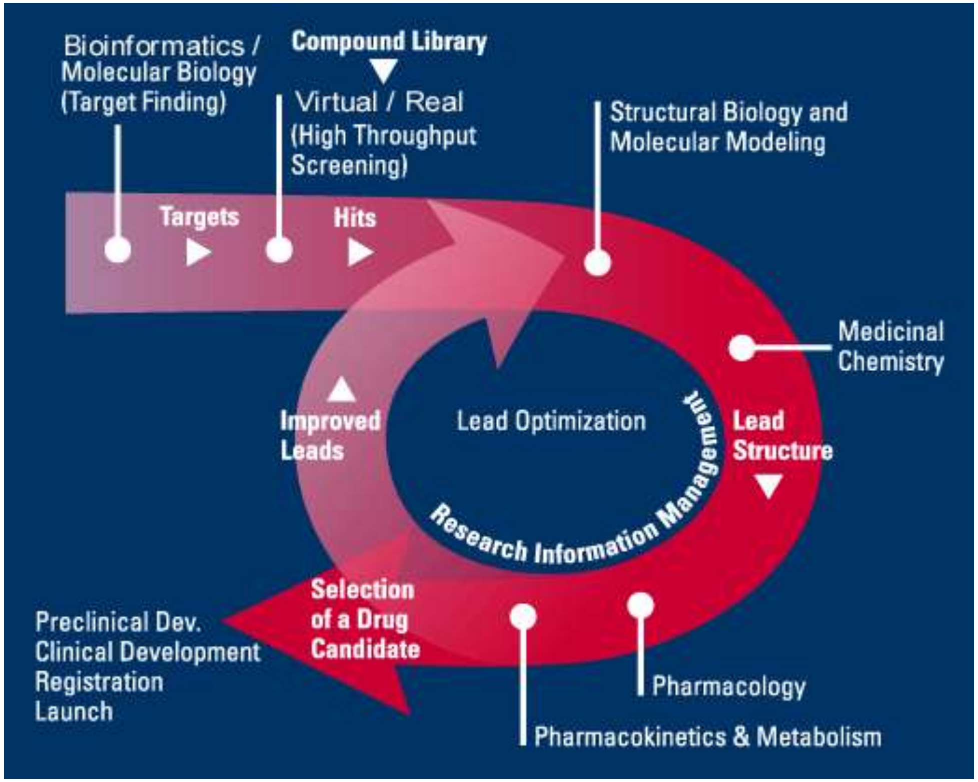
Phase I, II & III Clinical Trials

6.0yrs ↓ *68%*

FDA Review & Approval

1.5yrs ↓ *3%*

Drug to the Market


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

***In silico* interventions are poised to cut down the cost and time in drug discovery**

14 yrs **$1.4 billion (revised to $2.6 billion in 2016)**

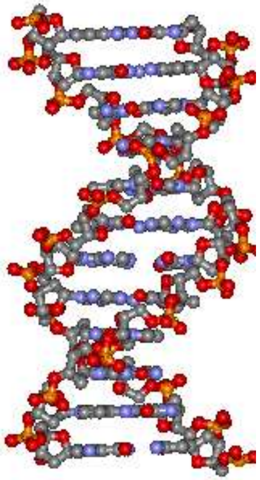
Source: PAREXEL's Pharmaceutical R&D Statistical Sourcebook, 2001, p96.; Hileman, Chemical Engg. News, 2006, 84, 50-1.

Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi
www.scfbio-iitd.res.in
A Centre of Excellence of the Department of Biotechnology, Govt. of India

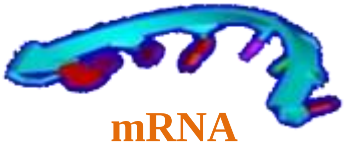
Genome

ChemGenome

A novel method to interpret the language of DNA and to identify protein coding genes and other functional units on genomic DNA



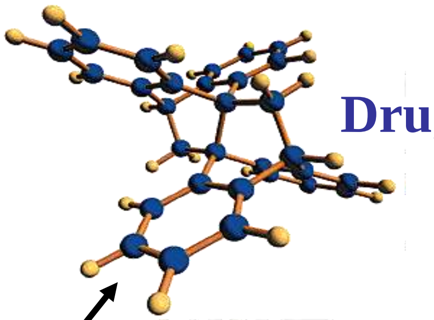
DNA



mRNA

...SEEARTINSCIENCE......
Polypeptide Sequence

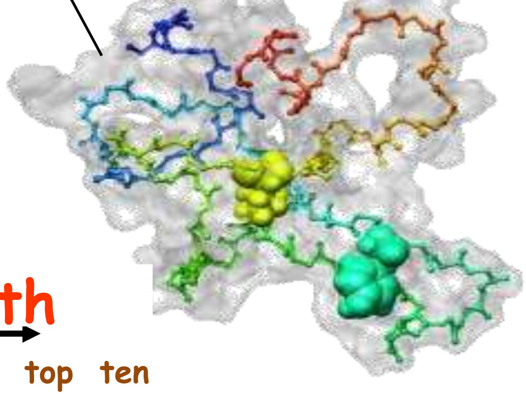
Snapshot of the Supercomputer @IITD used for the deployment of *Dhanvantari* Suite



Drug

Sanjeevini

A freely accessible state of the art software suite for structure based drug discovery



Protein

Bhageerath

Ranked among the top ten servers globally, Bhageerath predicts tertiary structures of proteins, tackling a grand challenge problem

"GENOME TO DRUG" (*DHANVANTARI*) PATHWAY ENVISAGES DELIVERING NOVEL DRUG MOLECULES/PERSONALIZED MEDICINE TO SOCIETY FROM GENOMIC / PROTEOMIC INFORMATION

Genome to Hit molecules: *Dhanvantari* – How does it work?

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.
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. The pre-S proteins are associated with virus attachment to the hepatocyte.**
3	ORF C	Core protein and HBeAg	HBcAg: forms the capsid. HBeAg: soluble protein and its biological function are still not understood. However, strong epidemiological associations with HBV replication and risk for hepatocellular carcinoma are known.
4	ORF X	HBx protein	Transactivator; required to establish infection in vivo. Associated with multiple steps leading to hepatocarcinogenesis.

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

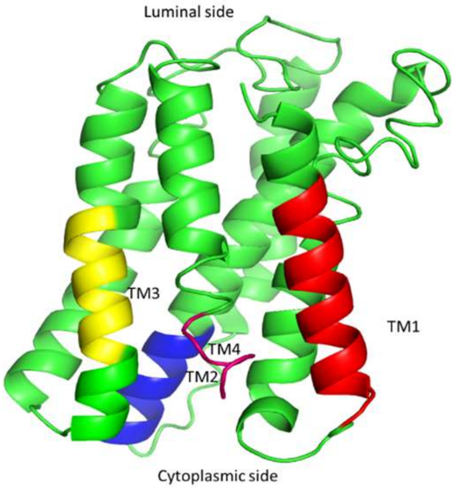
Gene 3 (BP: 157 to 837) predicted by the *ChemGenome 3.0* software encodes for the HBV surface protein (Gene Id: 944569)

(One could consider all the genes essential for viral replication but nonexistent in humans for *in silico* drug discovery)

>YP_009173871.1 small envelope protein [Hepatitis B virus]
MENITSGFLGPLLVLQAGFFLLTRILTIPQSLDSWWTSLNFLGGTTVCLGQNSQSPTSNHSPTSCPPTCPG
YRWMCLRRFIIFLFILLLCLIFLLVLLDYQGMLPVCPLIPGSSTTSTGPCRTCMTTAQGTSMYPSCCCTK
PSDGNCTCIPIPSSWAFGKFLWEWASARFSWLSLLVPFVQWFVGLSPTVWLSVIWMMWYWGPSLYSILS
PFLPLLPIFFCLWVYI

Input Amino acid sequence to *Bhageerath-H*

to obtain the structure



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

10 potential binding sites identified

Scan a million compounds library against the potential binding sites

RASPD/*Sanjeevini* calculation with an average cut off binding affinity to limit the number of candidates. (RASPD is a rapid empirical screening protocol which builds in Lipinski's rules, Wiener index etc. to extract binding energy without the compute-intensive docking)

RASPD output

Top 150 molecules were selected with binding energies above a *threshold* from one million molecule database corresponding to the first* predicted binding site.

***(One could consider other predicted binding sites based on some literature knowledge)**

Out of the 150 molecules, top 30 molecules are given as input to ParDOCK / *Sanjeevini* for atomic level binding energy calculations. Out of this 30, keeping a cut off value of -10 kcal/mol, top 5 molecules are seen to bind well to (small envelope protein) HBsAg and shortlisted for MD simulations. These molecules could be tested in the Laboratory.

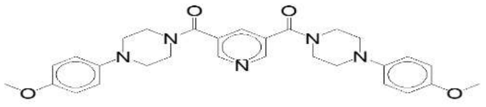
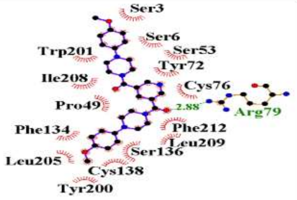
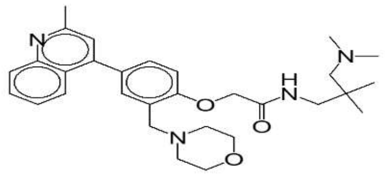
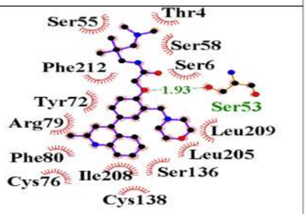
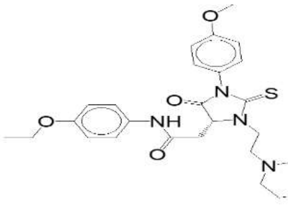
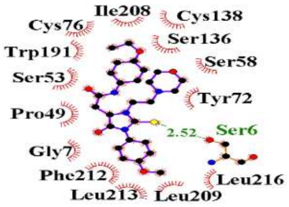
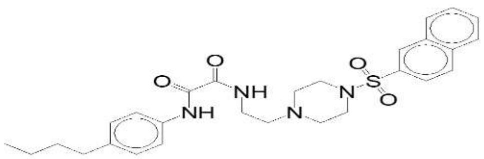
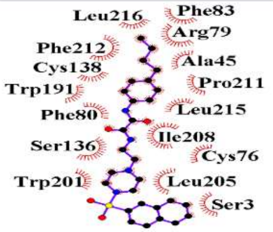
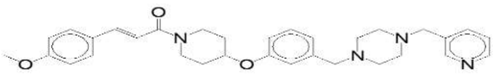
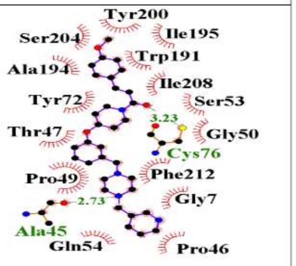
Sr. No.	ZINC ID	ParDOCK/Sanjeevini
1	ZINC00653293	-11.5
2	ZINC11787288	-11.4
3	ZINC20451377	-11.1
4	ZINC19809262	-10.8
5	ZINC19805326	-10.9
6	ZINC11910201	-10.8
7	ZINC03877668	-10.1
8	ZINC11913294	-10.1
9	ZINC01794178	-9.8
10	ZINC12050585	-9.1
11	ZINC04020431	-8.8
12	ZINC16193214	-8.7
13	ZINC01109335	-8.6
14	ZINC01139950	-8.2
15	ZINC02836173	-8.2
16	ZINC01092399	-8.1
17	ZINC05221544	-8.1
18	ZINC16667348	-8.1
19	ZINC03143011	-8.0
20	ZINC08680620	-8.0
21	ZINC01067619	-7.9
22	ZINC08892130	-7.9
23	ZINC19797618	-7.9
24	ZINC02880085	-7.7
25	ZINC19797529	-7.6
26	ZINC00793735	-7.3
27	ZINC20601870	-7.3
28	ZINC16248648	-7.2
29	ZINC08935093	-7.1
30	ZINC12576410	-6.9

Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi

www.scfbio-iitd.res.in

A Centre of Excellence of the Department of Biotechnology, Govt. of India

5 hit molecules suggested as inhibitors for small envelope protein (HBsAg) target of HBV

Name	Chemical Structure	Interactions with HBsAg
Molecule 1 (ZINC00653293)		 Ser3 Ser6 Ser53 Trp201 Tyr72 Ile208 Cys76 Pro49 2.88 Arg79 Phe212 Phe134 Leu209 Ser136 Leu205 Cys138 Tyr200
Molecule 2 (ZINC11787288)		 Ser55 Thr4 Ser58 Phe212 Ser6 1.93 Tyr72 Ser53 Arg79 Leu209 Phe80 Leu205 Cys76 Ile208 Ser136 Cys138
Molecule 3 (ZINC19809262)		 Ile208 Cys76 Cys138 Ser136 Trp191 Ser53 Ser58 Tyr72 Pro49 2.52 Ser6 Gly7 Phe212 Leu216 Leu213 Leu209
Molecule 4 (ZINC19805326)		 Leu216 Phe83 Arg79 Phe212 Cys138 Ala45 Pro211 Trp191 Leu215 Phe80 Ile208 Ser136 Cys76 Trp201 Leu205 Ser3
Molecule 5 (ZINC20451377)		 Tyr200 Ser204 Ile195 Trp191 Ala194 Ile208 Tyr72 Ser53 3.23 Thr47 Gly50 Cys76 Pro49 Phe212 2.73 Gly7 Ala45 Gln54 Pro46

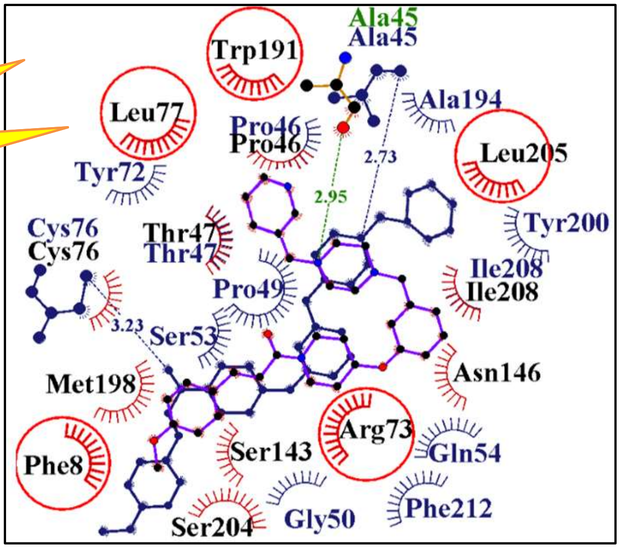
Done!

Experimental validation

R. Bhat, S. Kiruthika, R. Dash, A. S. Rathore,V. Perumal & B. Jayaram, A Novel Piperazine derivative that targets Hepatitis B Surface Antigen effectively inhibits Tenofovir Resistant Hepatitis B Virus, *Scientific Reports*, 2021. https://doi.org/10.1038/s41598-021-91196-1; Kiruthika, S.; Bhat, Ruchika; Jayaram, B.; V. Perumal, "A small molecule targeting Hepatitis B surface antigen inhibits clinically relevant drug-resistant hepatitis B virus", *Journal of Antimicrobial Chemotherapy,* 2022, accepted.

Molecule 5

Out of these five, Molecule 5 inhibits both Wild type and mutant strains



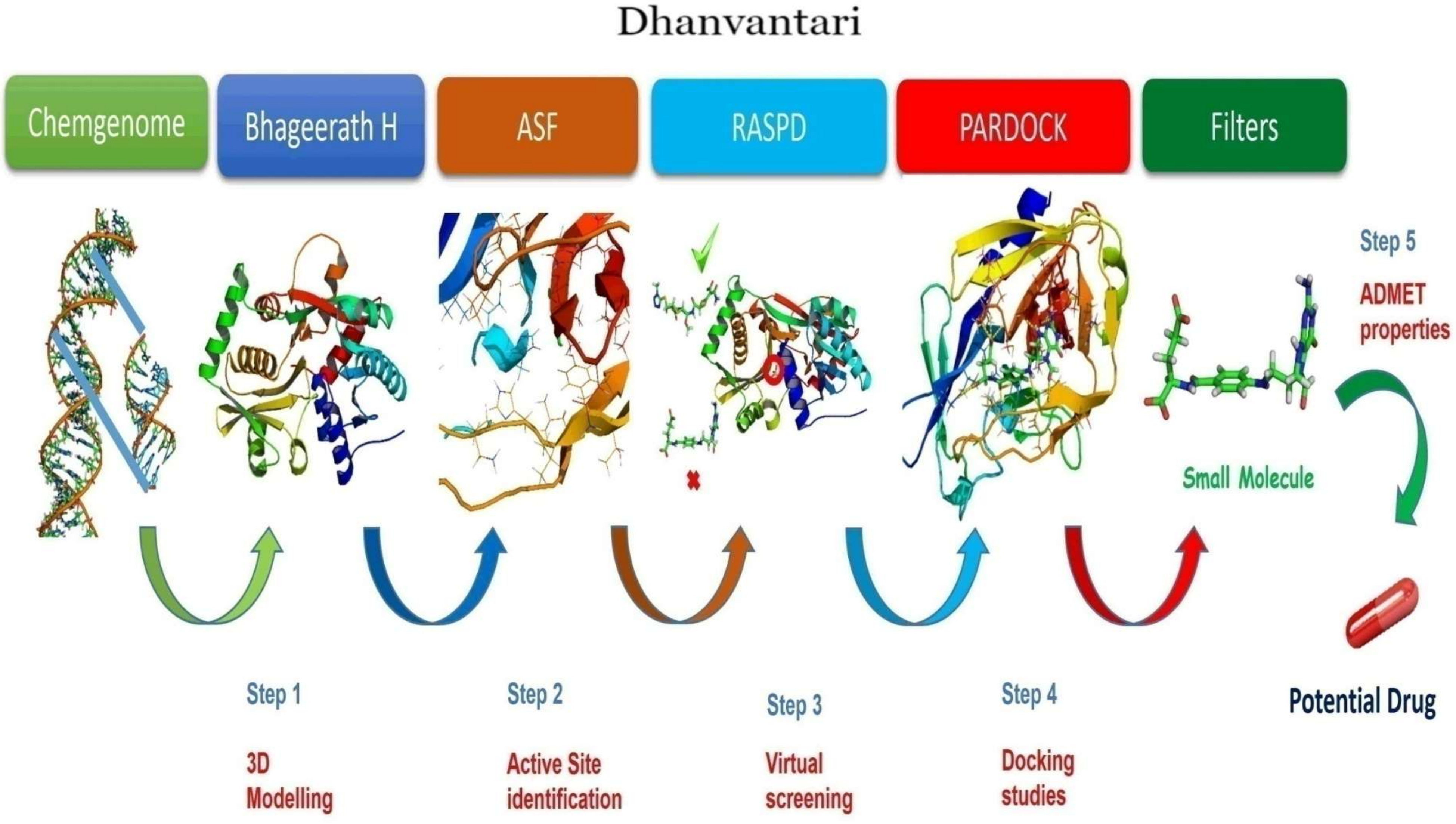
KD value by Surface Plasmon Resonance (SPR) for Molecule 5: 6.53×10^{-8} M

HBV Strain	Wild Type	Mutant Strains	
		rtM204I	CYEI
IC50 for Molecule 5 (µM)	20.84	5.561	11.39

A micromolar hit compound is guaranteed today. With some medicinal chemistry & toxicology, development of a nanomolar drug-like molecule is conceivable in an automated mode in the near future!

In silico Drug discovery assembly line developed at SCFBio

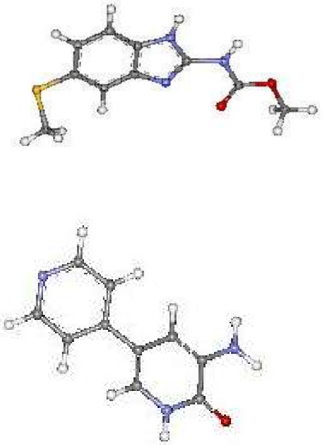
Ruchika Bhat, Rahul Kaushik, Ankita Singh, Debarati DasGupta, Abhilash Jayaraj, Anjali Soni, Ashutosh Shandilya, Vandana Shekhar, Shashank Shekhar, B. Jayaram, " A comprehensive automated computer-aided discovery pipeline from genomes to hit molecules" *Chemical Engineering Science*, 2020. https://doi.org/10.1016/j.ces.2020.115711



From Genome to Hits

Genome

X Teraflops
Chemgenome
BhageerathH
Sanjeevini



Hits

(A) B. Jayaram, Priyanka Dhingra, Goutam Mukherjee, Vivekanandan Perumal, "Genomes to Hits: The Emerging Assembly Line *In Silico*", Proceedings of the Ranbaxy Science Foundation 17th Annual Symposium on "*New Frontiers in Drug Design, Discovery and Development*" 2012, Chapter 3, 13-35. (B) Anjali Soni, K. M. Pandey, P. Ray, B. Jayaram, "Genomes to Hits *in Silico*: A Country Path Today, A Highway Tomorrow: A case study of chikungunya", *Current Pharmaceutical Design*, 2013, 19, 4687-4700, DOI: 10.2174/138161281131999990379. (C) Ruchika Bhat, Rahul Kaushik, Ankita Singh, Debarati DasGupta, Abhilash Jayaraj, Anjali Soni, Ashutosh Shandilya, Vandana Shekhar, Shashank Shekhar, B. Jayaram, " A comprehensive automated computer-aided discovery pipeline from genomes to hit molecules" *Chemical Engineering Science*, 2020. https://doi.org/10.1016/j.ces.2020.115711

•Genome Analysis - *ChemGenome*

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

•Protein Structure Prediction – *Bhageerath*

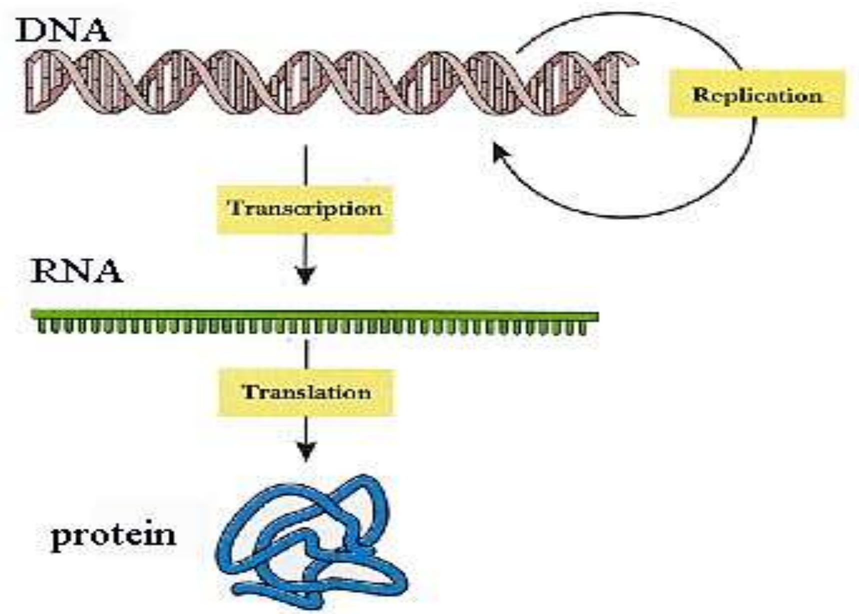
A *de novo* energy based protein structure prediction software

•Drug Design – *Sanjeevini*

A comprehensive active site/target directed lead molecule design protocol

**Details of the genome to hit pathway,
the scientific challenges overcome &
the questions pending answers →**

What is happening inside the cells?



Essentials:

1. DNA is made of 4 bases:
   A, G, C, T.

Watson-Crick pairing states
A pairs with T(U) and T(U) with A.
G pairs with C and C with G**.**

2. Proteins are made of 20 Amino acids.

Genetic code maps the correspondence between bases and amino acids.

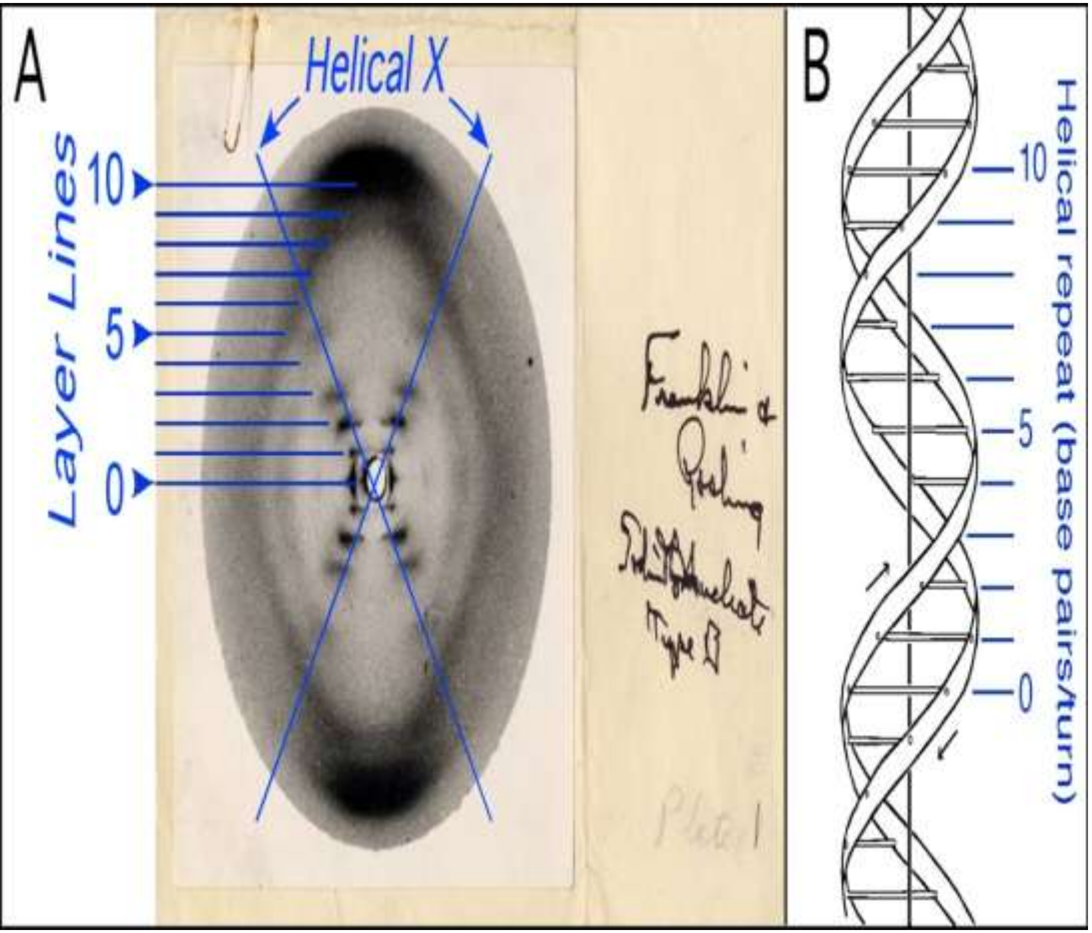
A depiction of gene expression (the central dogma), summarized as **DNA (gene) makes RNA & RNA makes proteins**, the two steps being called transcription and translation.

DNA carries genes which code for several types of RNAs such as mRNA, tRNA, rRNA, micro RNA etc.. Only mRNA gets converted into proteins.

RNA viruses pose an exception to central dogma in that RNA of virus gets converted to DNA within the host with the help of reverse transcriptase enzyme of the virus. The DNA of the virus now in the host, follows the central dogma using host cell machinery.

Double helical DNA

X-ray diffraction photograph of a DNA fiber at high humidity (Franklin and Gosling, 1953). Interpretation of the helical-X and layer lines added in blue.



The Nobel Prize in Physiology or Medicine 1962

"for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material"

Francis Crick

MRC, UK

b. 1916 (UK)

James Watson

Harvard U., USA

b. 1928 (IUSA)

Maurice Wilkins

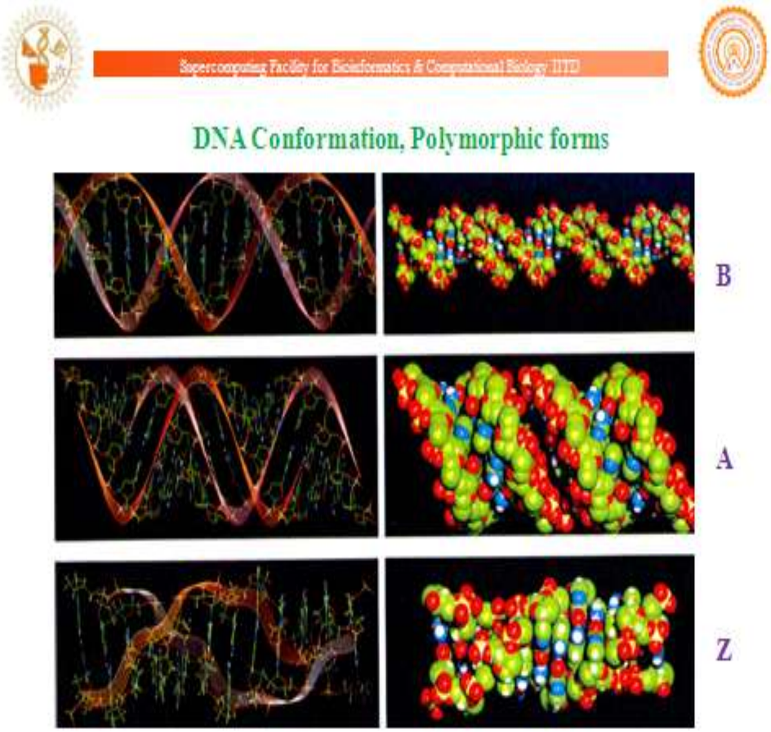
London U., UK

•b. 1916 (new Zealand)

A little bit about DNA before discussing Genomic language

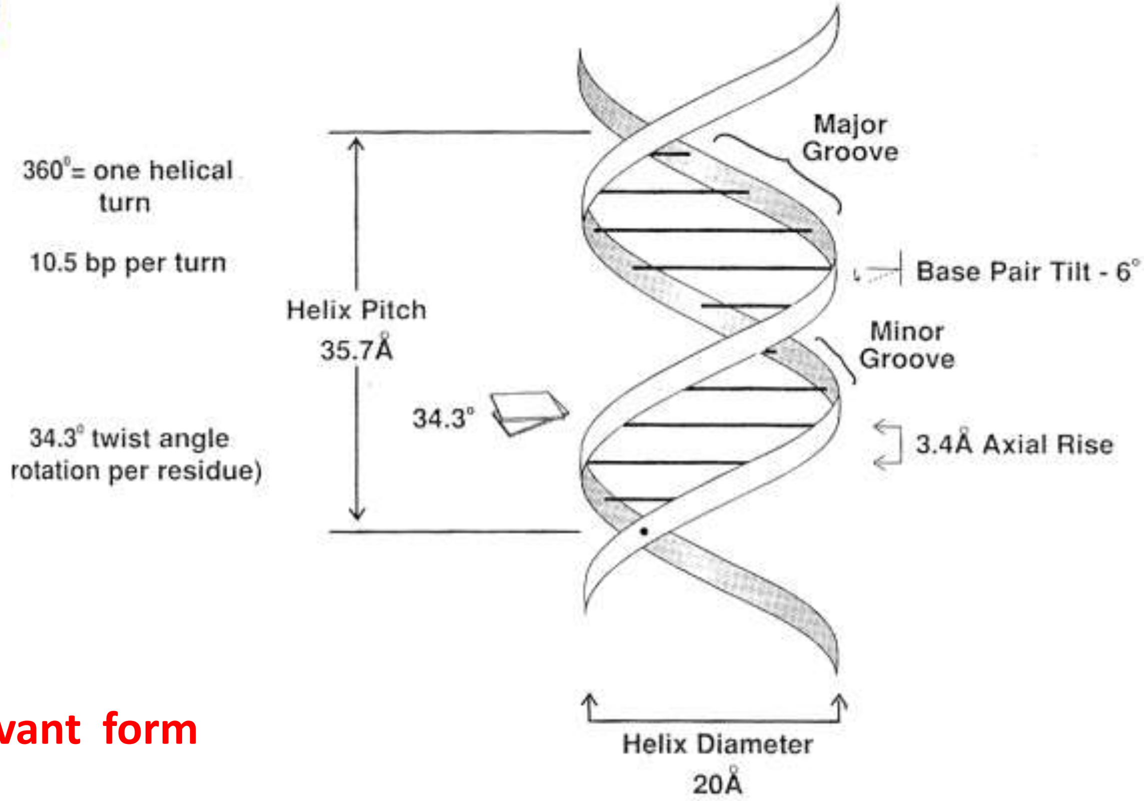
Structure types and helical parameters*

Structure type	Pitch	Helical Symmetry	Axial rise	Turn angle	Minor groove(w)	Major groove(w)	Minor depth (d)	Major groove(d)
A DNA	28.2	11	2.56	32.7	11.0	2.7	2.8	13.5
B DNA	**33.8**	**10**	**3.38**	**36.0**	**5.7**	**11.7**	**7.5**	**8.5**
Z DNA	45.0	6	3.70	-30.0	8.8	2.0	3.7	13.8



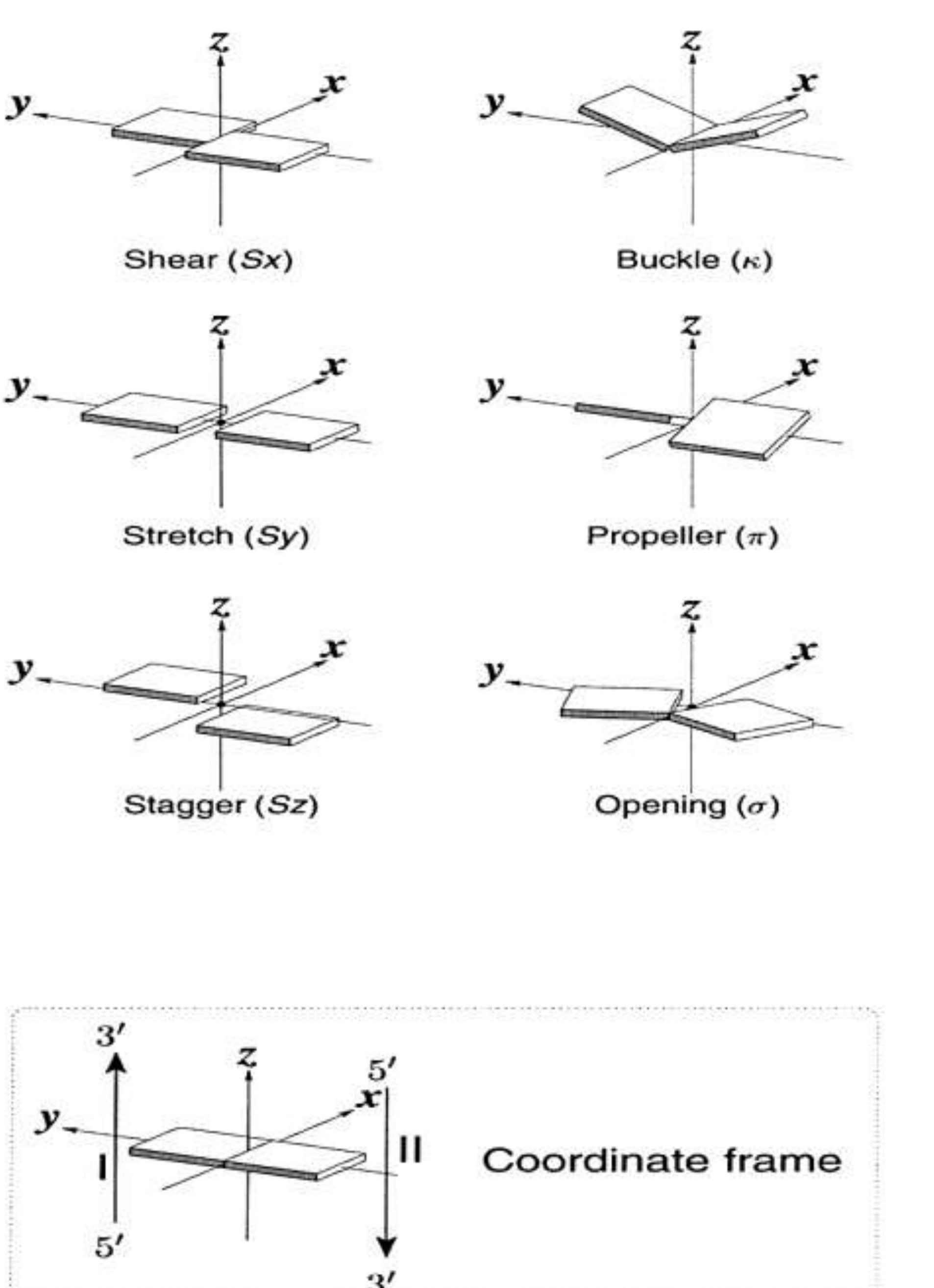
Supercomputing Facility for Bioinformatics & Computational Biology IITD

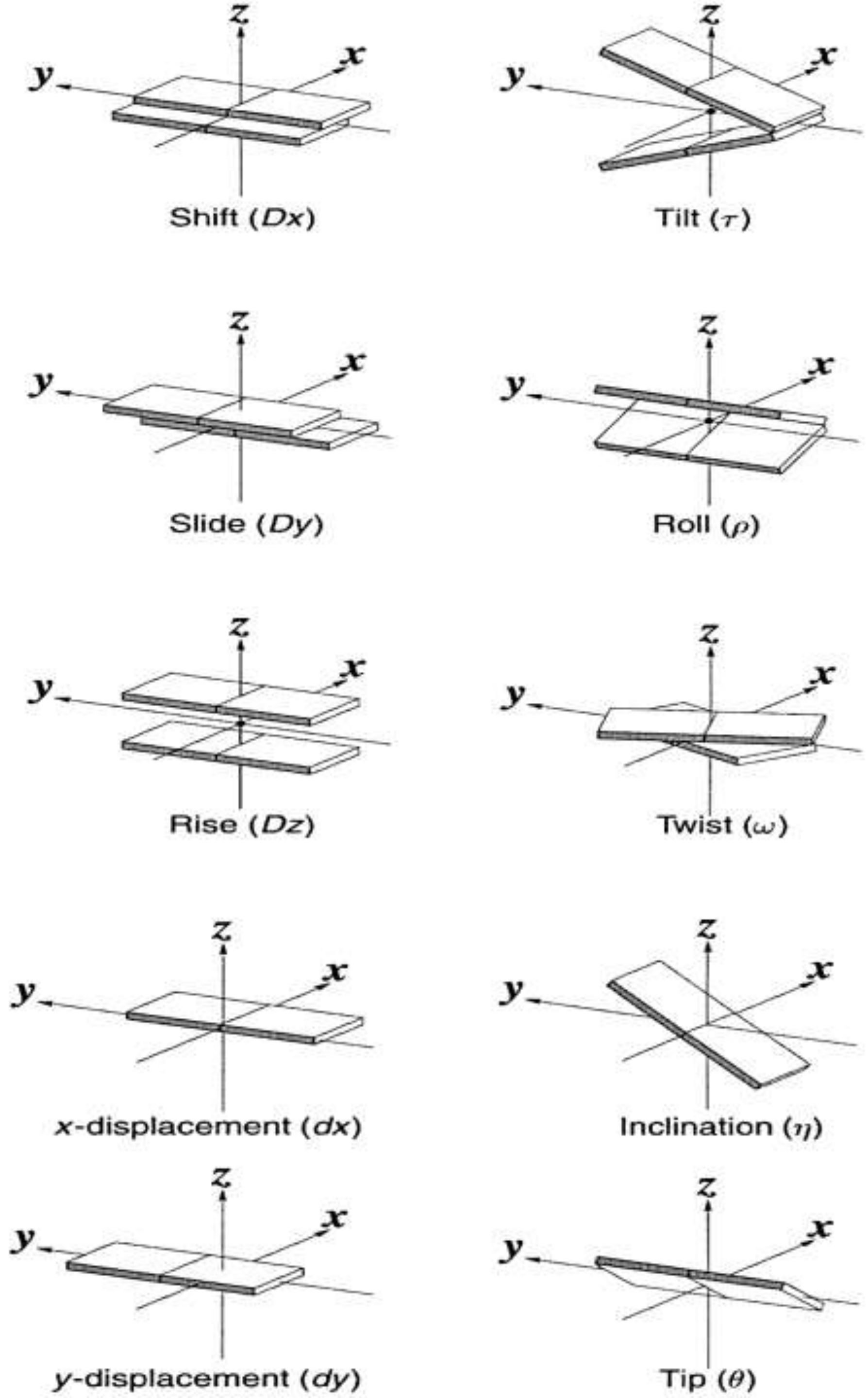
DNA Conformation, Polymorphic forms



B DNA is physiologically the most relevant form

Pradeep Pant, Saher Afshan Shaikh, B. Jayaram, "Design and Characterization of Symmetric Nucleic Acids via Molecular Dynamics Simulations", *Biopolymers*, 2016, DOI: 10.1002/bip.23002.

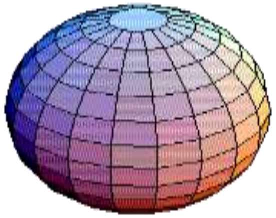




Intra-base pair & inter-base pair parameters

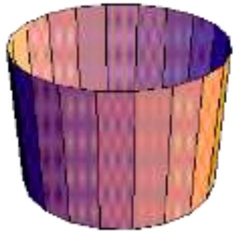
DNA is a dynamic molecule!!

Small Ion Matters (DNA in aq. Medium)



Sphere

No Condensation (Total Dissociation)



Cylinder

Partial neutralization of the charge

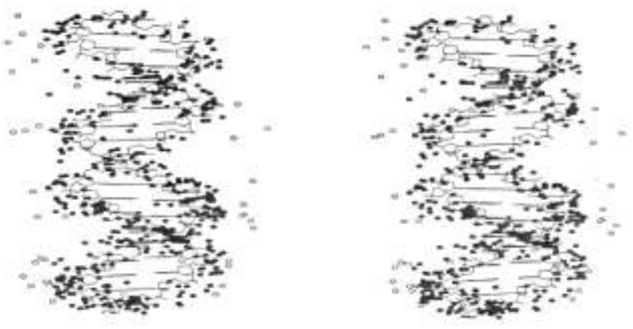
Plane

Total condensation

Manning Theory:

Net charge on DNA phosphates $Q_{phos} = \frac{1}{N\varsigma} \sim -0.24$

$$\varsigma = \frac{e^2}{\varepsilon kTb}$$



Counterion Condensation in Nucleic Acid Systems:

A microscopic view

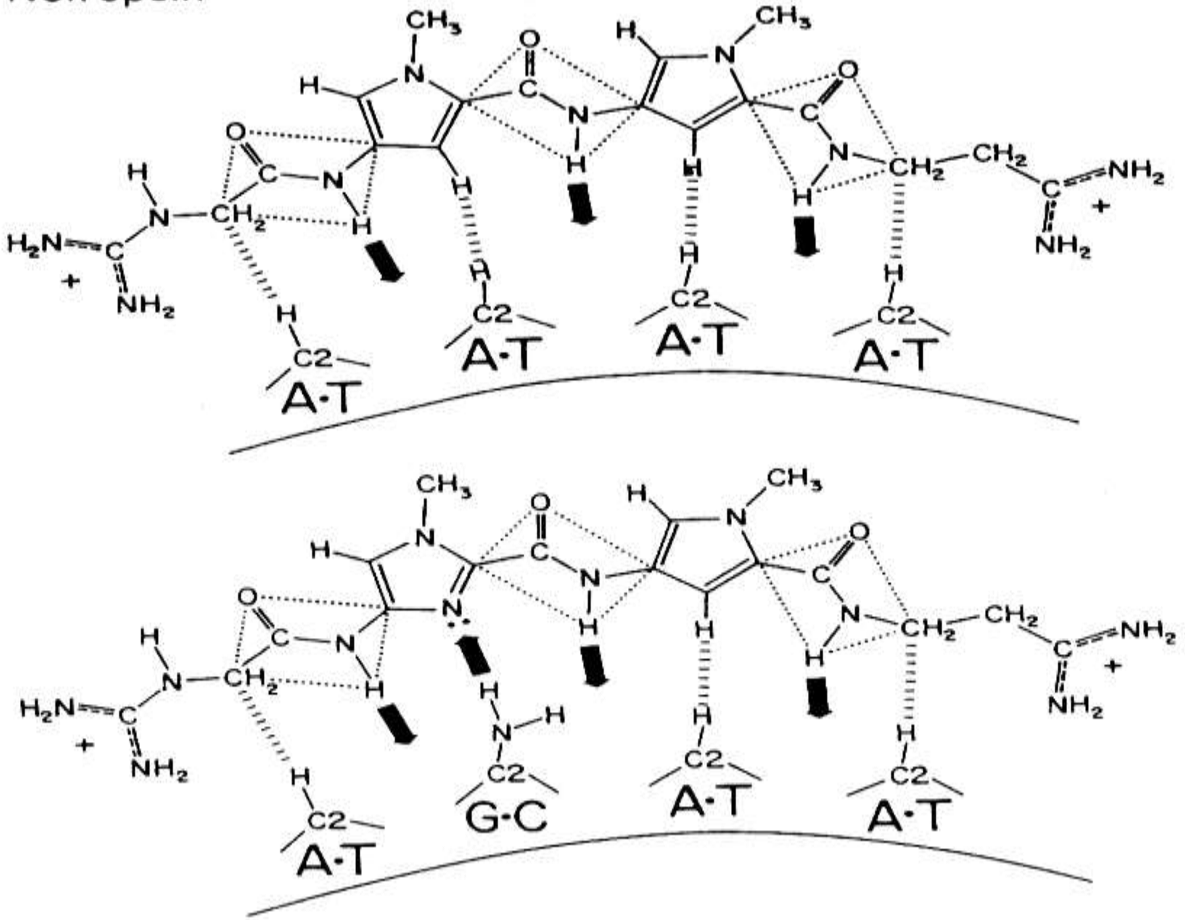
Jayaram et al., *Macromolecules*, 23, 3156 (1990);

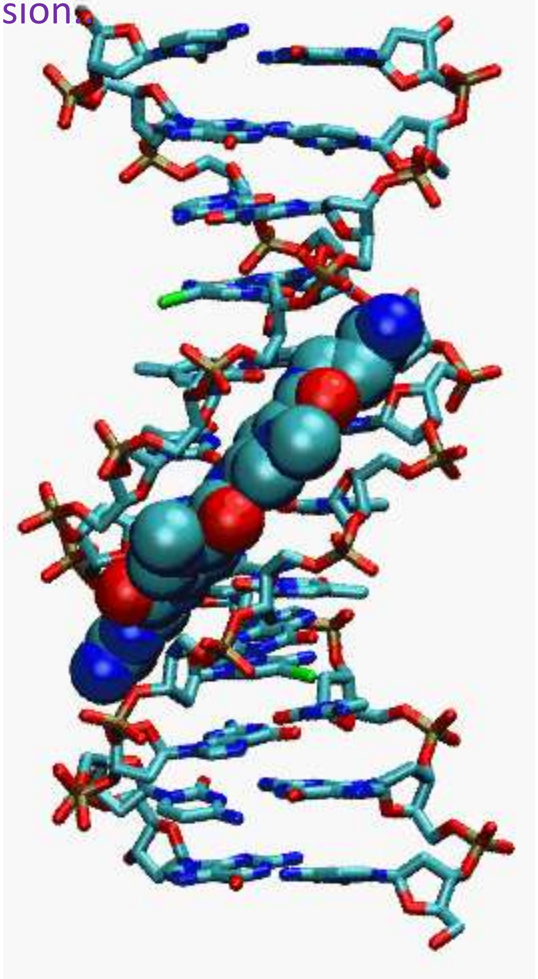
M. Young, B. Jayaram, and D. L. Beveridge, *J. Am. Chem. Soc.*, 1997, *119*, 59-69.

B. Jayaram and D. L. Beveridge, *Annu. Rev. Biophys. Biomol. Struc.*, 1996, *25*, 367-394.

Making Drugs against DNA: DNA-Drug: Minor groove interactions: Make a molecule that fits well in the grooves of DNA (Steric complementarity) and additionally makes hydrogen bonds with the base pairs (electrostatic complementarity) to cure cancer etc. or to control gene expression.

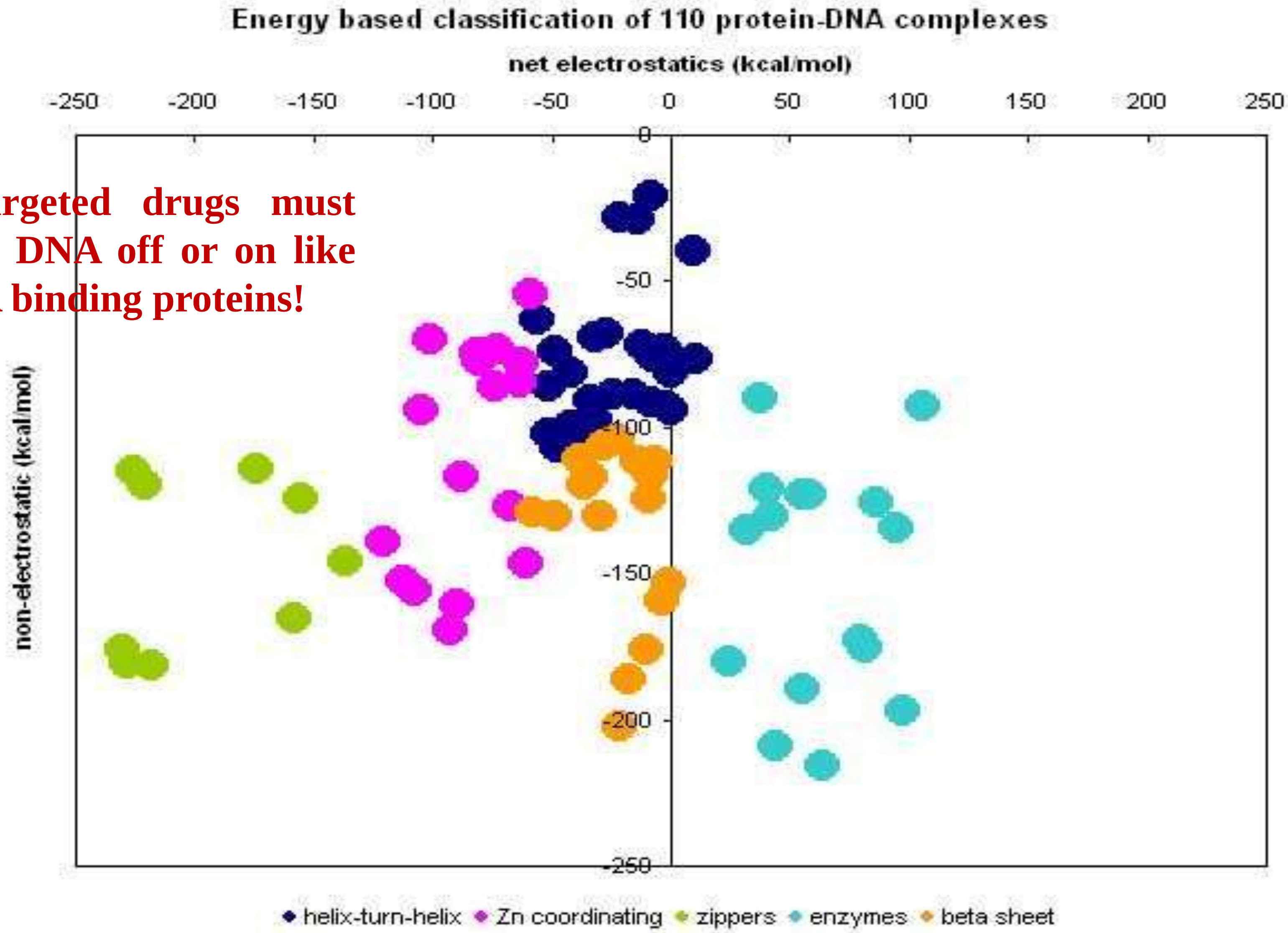
Netropsin





B. Jayaram , K. A. Sharp, and B. Honig, "The electrostatic potential of B DNA", *Biopolymers*, 1989, *28*,975-993; DNA exhibits sequence specific groove potentials.

DNA targeted drugs must turn the DNA off or on like the DNA binding proteins!

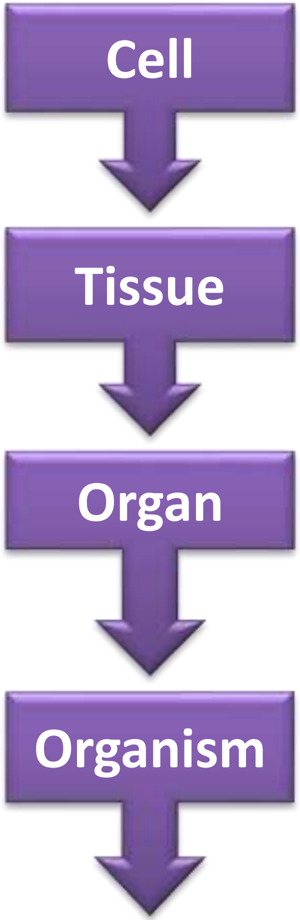


Energy components convey the signature of the DNA binding motif

Jayaram & Jain, *Annu Rev. Biophys. Biomol Struc.*, 2004, *33,* 343-61

Methdology: **B. Jayaram, K. McConnell, S. B. Dixit, D. L. Beveridge, "Free Energy Analysis of Protein-DNA Binding: The EcoRI-Endonuclease Complex", *J. Computational Physics*, 1999, *151*, 333-357.**

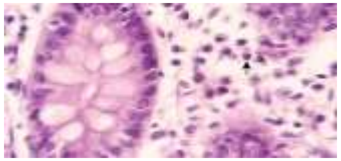
Back to Genome


Cell

Tissue

Organ

Organism



source:http://phenomena.nationalgeographic.com/2013/02/28/genomes-for-the-curious/

The Nobel Prize in Chemistry 1958 was awarded to Frederick Sanger *"for his work on the structure of proteins, especially that of insulin"*.

By structure above is meant, covalent connectivity viz. sequence in today's parlance.

The Nobel Prize in Chemistry 1980 was divided, one half awarded to Paul Berg *"for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA"*, the other half jointly to Walter Gilbert and Frederick Sanger *"for their contributions concerning the determination of base sequences in nucleic acids"*.

Protein Sequencing

Frederick Sanger, 1958

DNA Sequencing

Frederick Sanger, 1980

Source: www.nobelprize.org

ftp://ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/CHR_X/

DEFINITION Homo sapiens chromosome X

ORIGIN

```
  1 ccaggatggt ccttctcctg aaggttaatc cataggcaga tgaatcggat attgattcct
 61 gttcttggaa taatctagag gatctttaga atccattggg attcataatc acagctatgc
121 cgatgccatc atcaccggct tagccctttc tgaaaacaca gtcatcatct acccccattg
181 gaatcacgat gcaaaaaacc tgtcccaaag cggtggtttc ctatgtgatt cttgcatcca
241 ggacaaatga cagtcagcag agaggcgccc tgttccatct tttggtttga tccagttaaa
301 ggcacacacg tgagcaccca acgtttgcca actcagcact gggcagagcc tggcctctga
361 ggaaattggc atcttcgtaa tcaatatatt attatgtttt attgaaatgt aagtcattgc.....
```

Question: Can you infer the meaning of the sequences on the left just by reading them without looking at definitions or using some software?

Answer: No body can today….

DEFINITION Homo sapiens chromosome Y

ORIGIN

```
  1 ggtttcacca agttggccag gctggtctcg aactcctgac ctcaggtgat ctgtccacct
 61 cggtgtccca aagtgctggg attacaggtg tgaaccacca cacccagcct catgtaatac
121 ttaaaaatga actacaggtg gattacaaac ctgaatatca aagaaaactt ttttttttga
181 aaaatagagg gaaatgtctt ataacctcag agttaggagg tttttcttag atacaataca
241 aaaagcataa ccacgcccat agtcccagct actcaggagg ctgaggcata agaatcactt
301 gagctcgaga ggtggaggtt gcagtgagcc gagatcctgc cattgcactc cagctgaggc
361 tacagagtga gagtataaaa aaaaaaaaaa aagcataacc tttaaaaatg ggttagccta.....
```

What is the language of DNA that proteins understand and we don't?

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)

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

Today's Computational Challenge!

Genome assembly and genome annotation (understanding what each base pair does after correctly assembling the genome)

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**

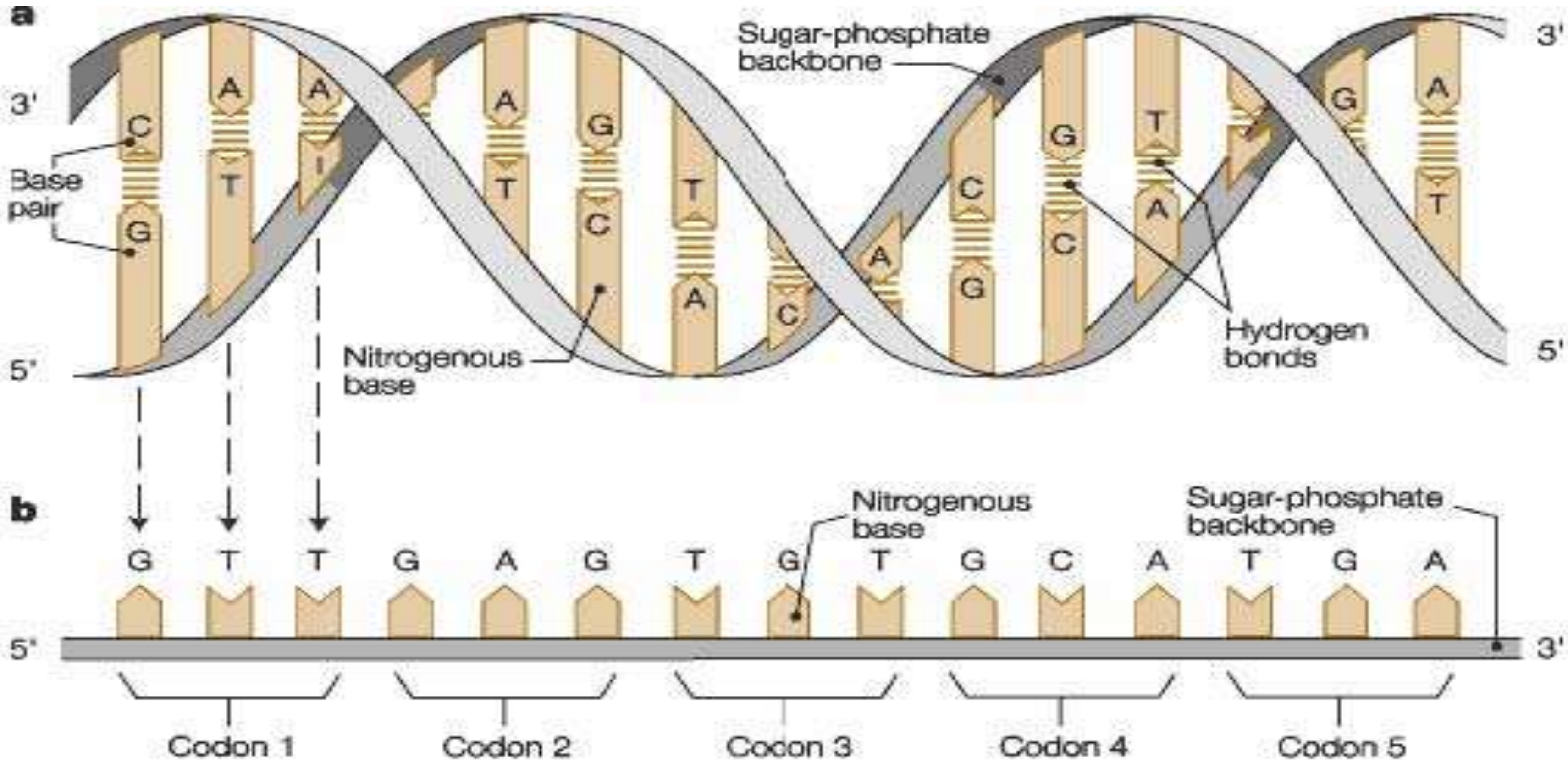
Most methods today are based on sophisticated mathematical and statistical techniques but rely heavily on sparse experimental data for training the models to do predictions. These methods are typically organism specific.

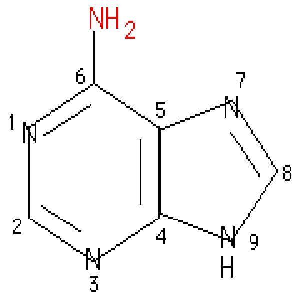
There is no universally applicable model!

Finding genes in Arabidopsis Thaliana
(Thale cress)

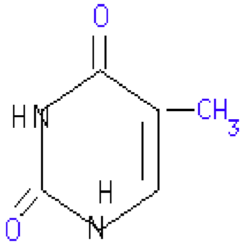
Software	Method	Sensitivity	Specificity
GeneMark.hmm http://www.ebi.ac.uk/genemark/	**5th-order Markov model**	**0.82**	**0.77**
GenScan http://genes.mit.edu/GENSCAN.html	**Semi Markov Model**	**0.63**	**0.70**
MZEF http://rulai.cshl.org/tools/genefinder/	**Quadratic Discriminant Analysis**	**0.48**	**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 it is remarkable that these methods perform so well with limited experimental data to train on, more research, new methods, new ways of looking at genomic DNA are required!

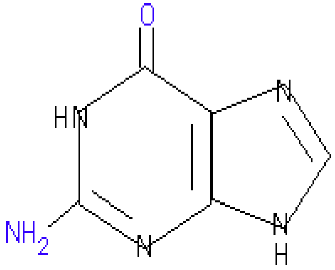




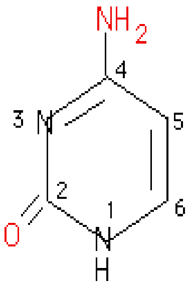
Adenine



Thymine



Guanine

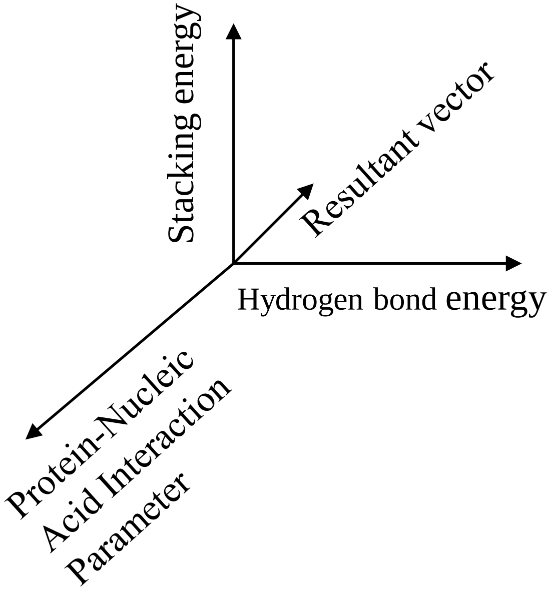


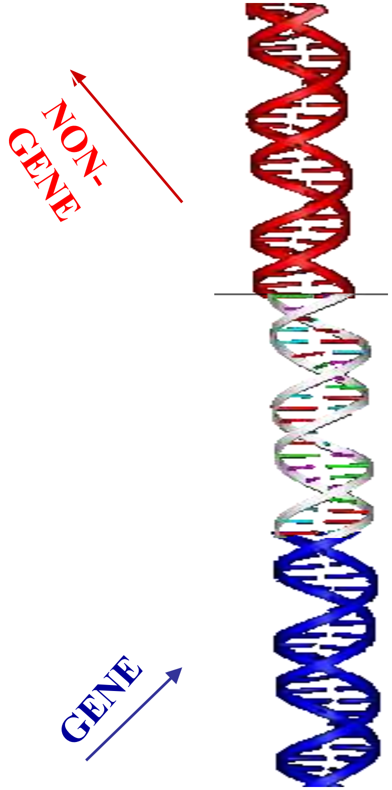
Cytosine

A universal model must factor in the chemical nature of the bases not just their alphabets

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. , 2006, 46(1), 78-85.*

i……l
j…..m
k…..n

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

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

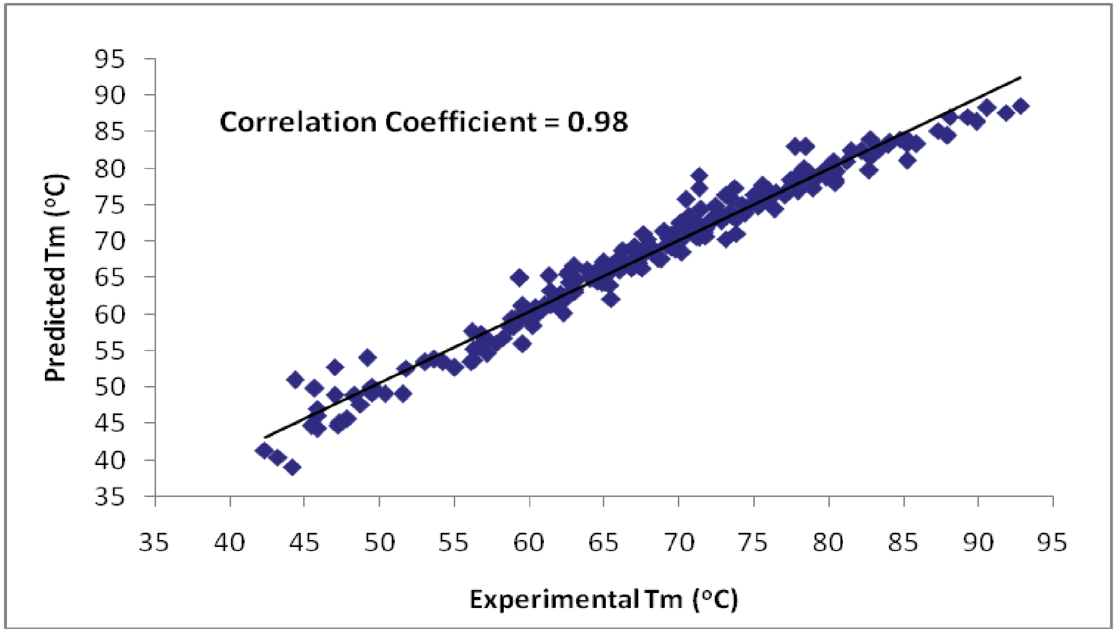
Hydrogen bond & Stacking energies for all 32 unique trinucleotides were calculated from long [*]*Molecular Dynamics Simulation Trajectories on 39 sequences encompassing all possible tetranucleotides in the [#]ABC database* and the data was averaged out from the multiple copies of the same trinucleotide. 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., *Biophys J, 2004,* 87, 3799-813; [#]Dixit et al., *Biophys J, 2005,* 89, 3721-40; Lavery et al., *Nucl. Acid Res., 2009,* 38, 299-313; Passi et al., *Nucl. Acids Res., 2014,* 42, 12272-12283 .

A self-consistent set of molecular dynamics derived hydrogen bonding, stacking and solvation energies for dinucleotides

Dinucleotide	Hydrogen Bond (kcal)	Stacking Energy (kcal)	Strength Parameter (E) (kcal)	Solvation (kcal/mol)
AA	-5.44	-26.71	-32.15	-171.84
AC	-7.14	-27.73	-34.87	-171.11
AG	-6.27	-26.89	-33.16	-174.93
AT	-5.35	-27.20	-32.55	-173.70
CA	-7.01	-27.15	-34.16	-179.01
CC	-8.48	-26.28	-34.76	-166.76
CG	-8.05	-27.93	-35.98	-176.88
CT	-6.27	-26.89	-33.16	-174.93
GA	-7.80	-26.78	-34.58	-167.60
GC	-8.72	-28.13	-36.85	-165.58
GG	-8.48	-26.28	-34.76	-166.76
GT	-7.14	-27.73	-34.87	-171.11
TA	-5.83	-26.90	-32.73	-174.35
TC	-7.80	-26.78	-34.58	-167.60
TG	-7.01	-27.15	-34.16	-179.01
TT	-5.44	-26.71	-32.15	-171.84

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 and B. Jayaram, "A phenomenological model for predicting melting temperatures of DNA sequences", *PLoS ONE, 2010, 5(8):* e12433. doi:10.1371/journal.pone.0012433
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

The Nobel Prize in Physiology or Medicine 1968

Robert W. Holley, Har Gobind Khorana and **Marshall W. Nirenberg**

"for their interpretation of the genetic code and its function in protein synthesis“

Why degeneracy in genetic code? What is the molecular basis for wobble?

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’ for gene identification

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**

Stacking & hydrogen bonding explain it!

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

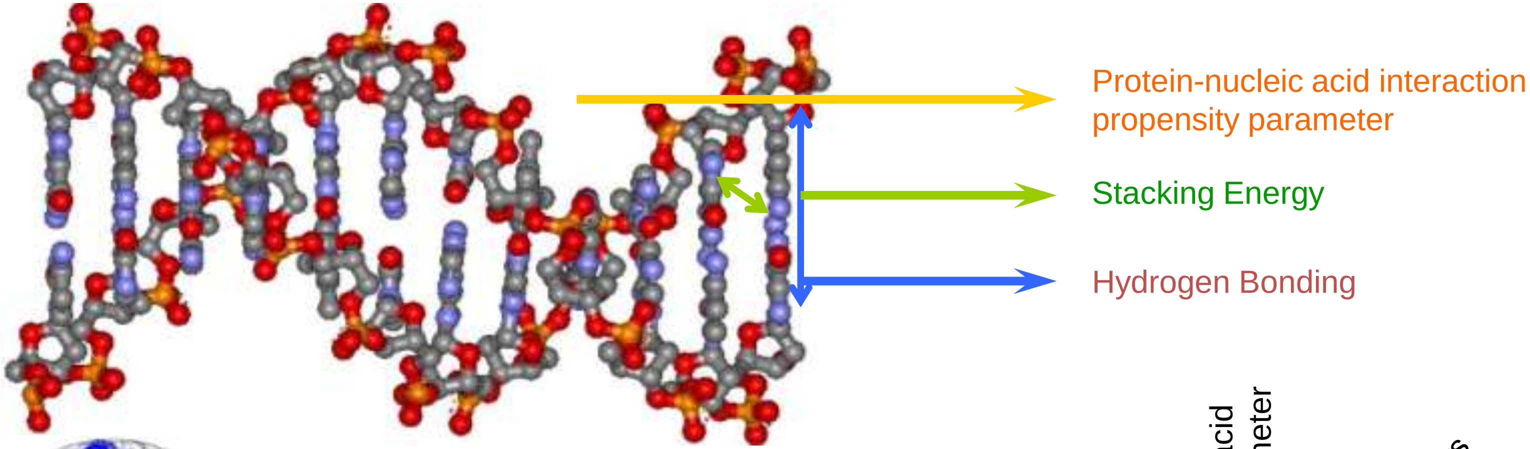
With 5 h-bonds middle pyrimidine renders third base inconsequential; middle purine requires third base.

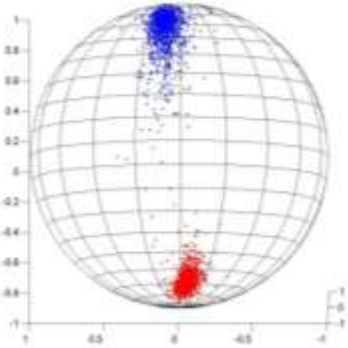
B. Jayaram, "Beyond Wobble: The Rule of Conjugates", *J. Molecular Evolution*, 1997, *45*, 704-705.

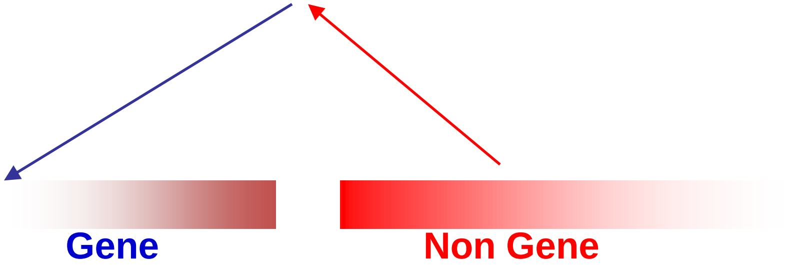
RULE: +1 if G is the first base, C at the 1st base and (T/A) on 2nd base; -1 if C on 1st base and (G/C) on 2nd base

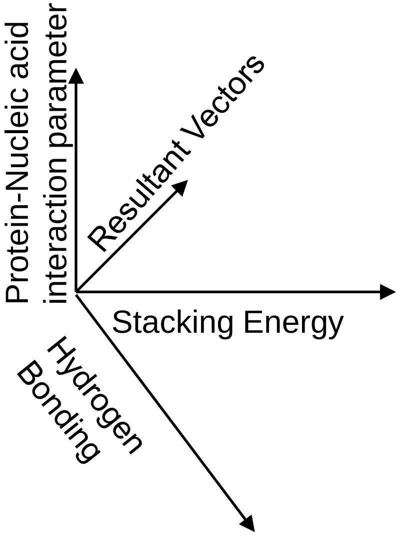
ChemGenome

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

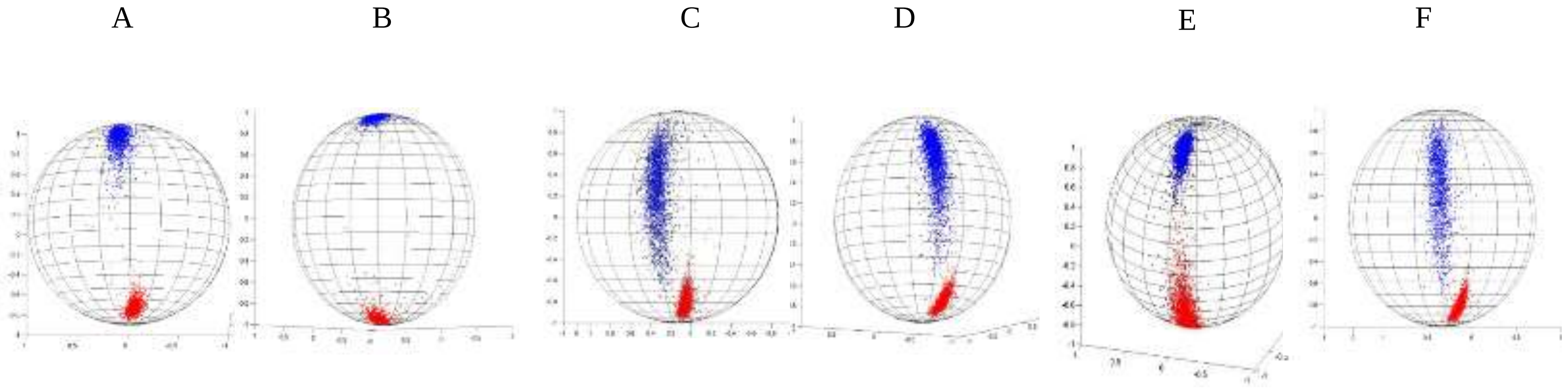








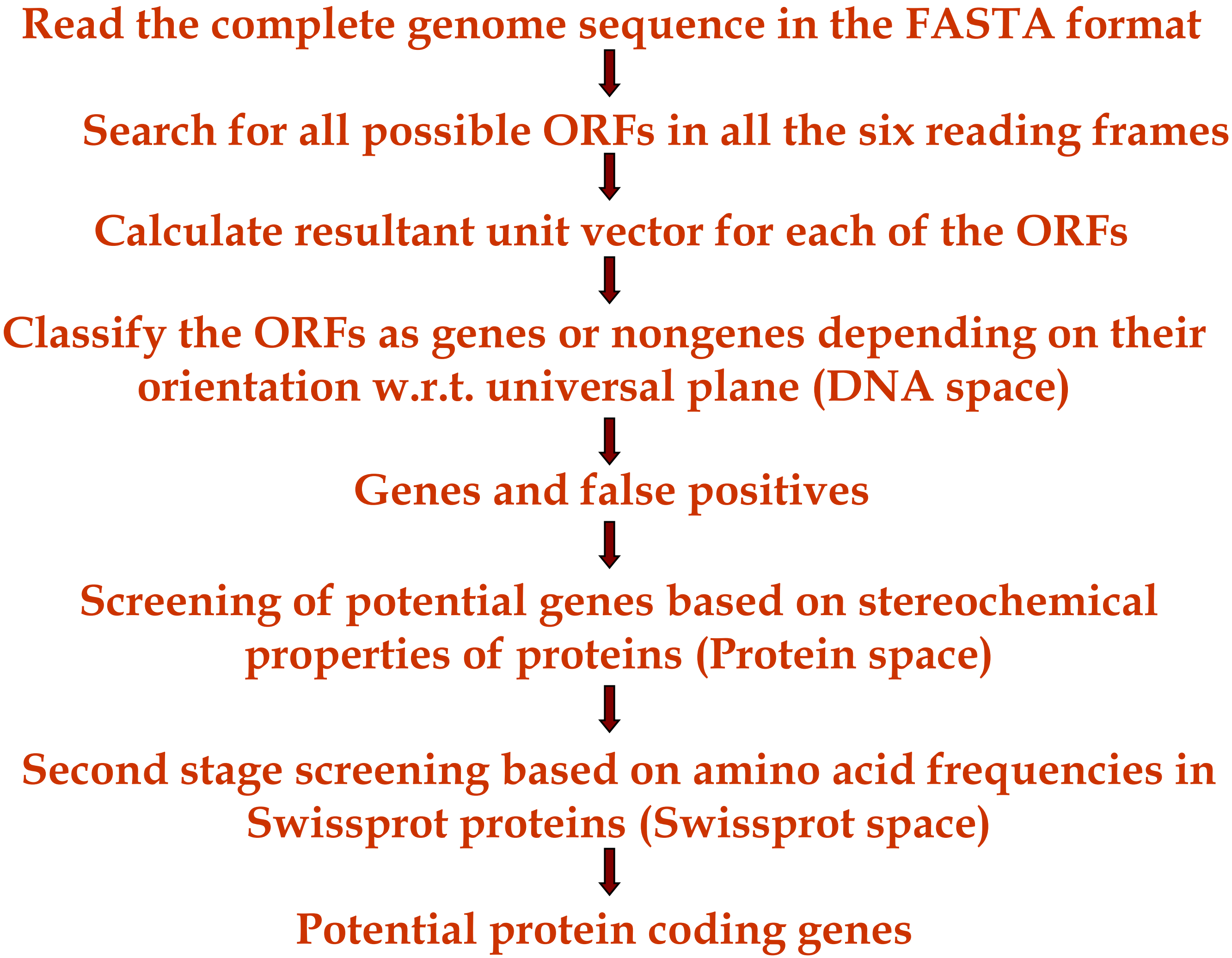
Distinguishing Genes (blue) from Non-Genes (red) in ~ 900 Prokaryotic Genomes



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 & 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

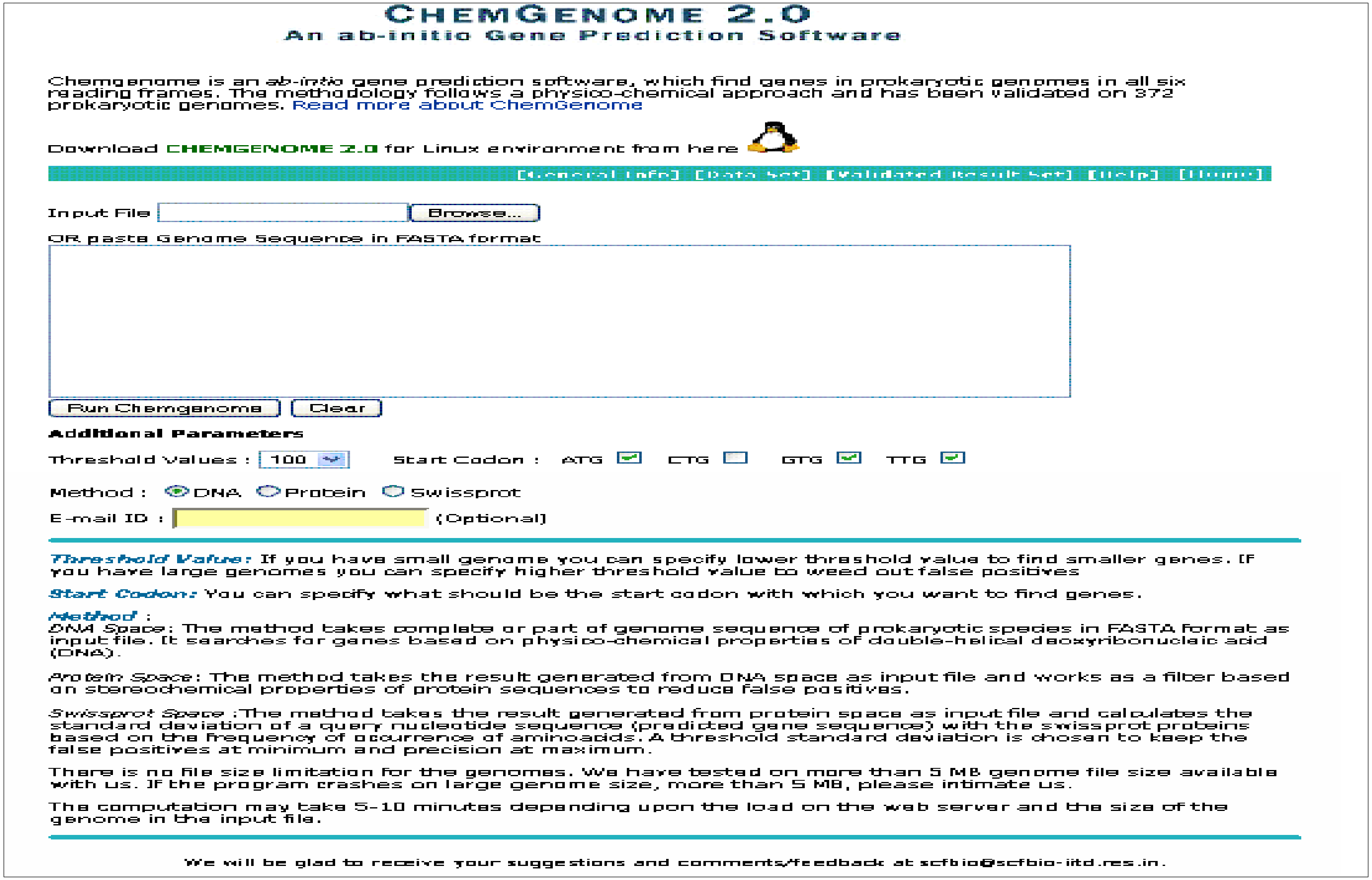
Screening of potential genes based on stereochemical properties of proteins (Protein space)

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

Potential protein coding genes

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.

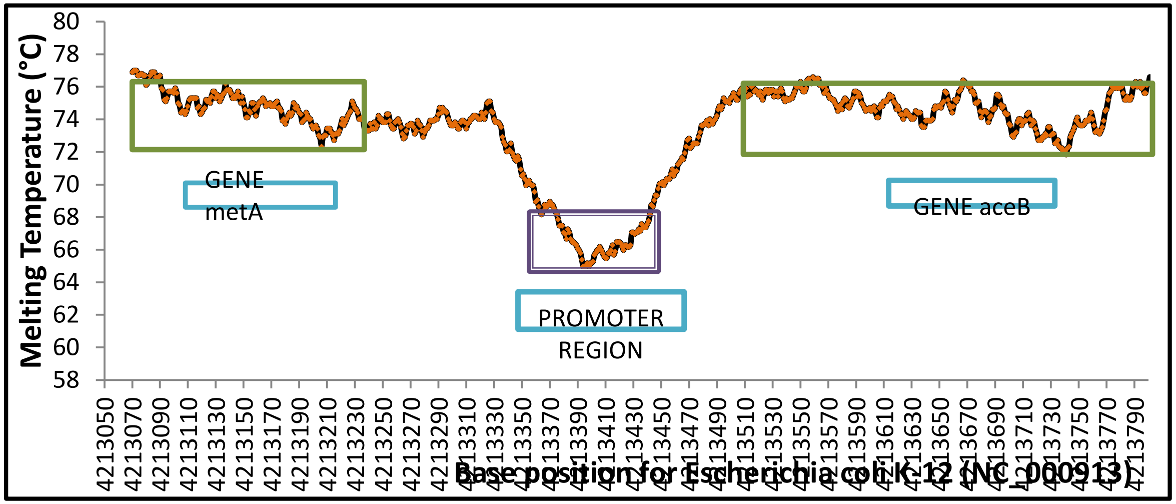
The computation may take 5-10 minutes depending upon the load on the web server and the size of the genome in the input file.

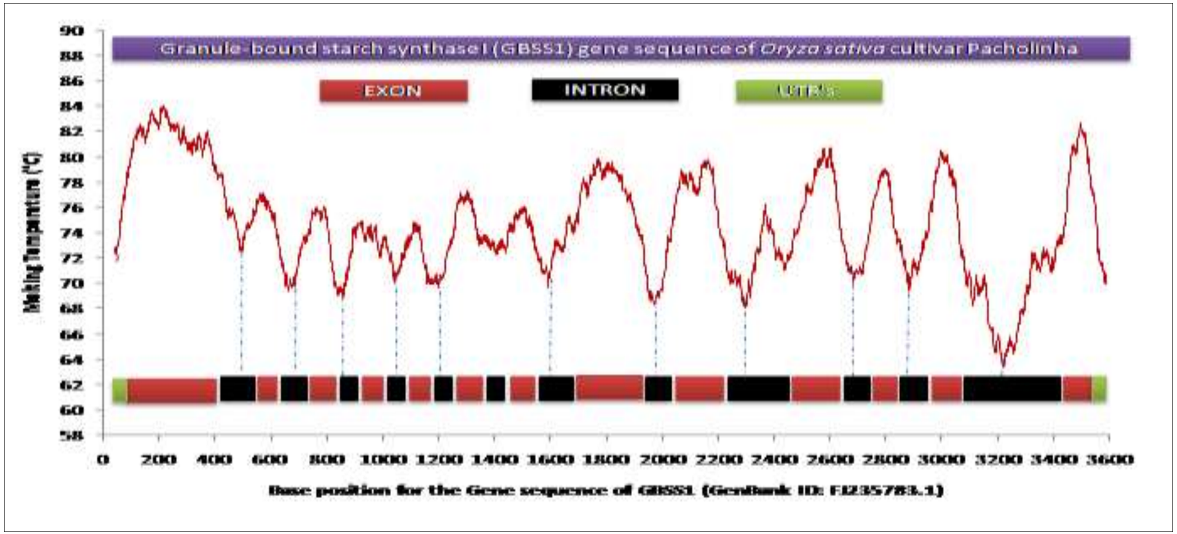
We will be glad to receive your suggestions and comments/feedback at scfbio@scfbio-iitd.res.in.

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**

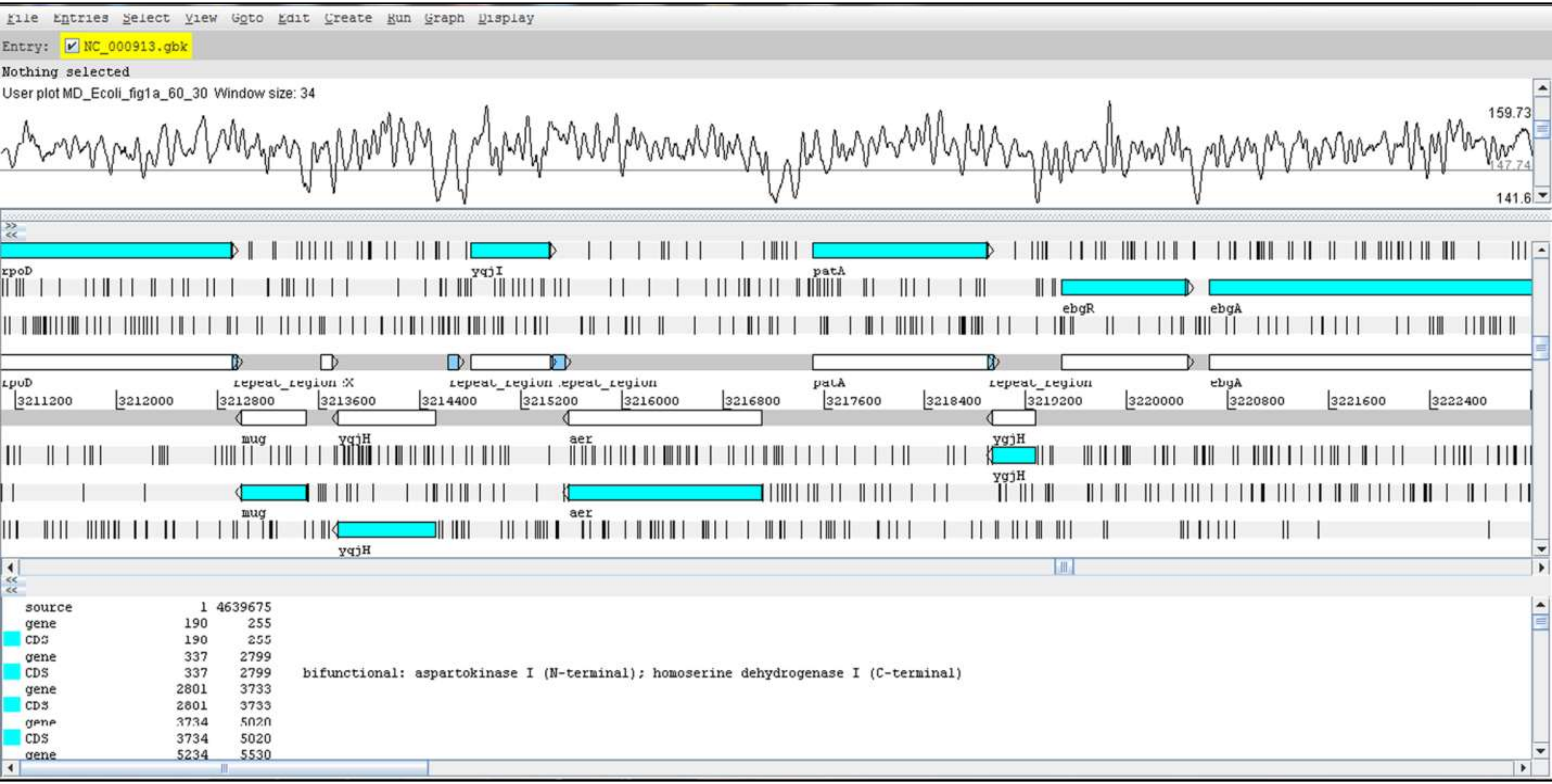
The physico-chemical model (Chemgenome) performs as well as any other sophisticated knowledge based methods. It is a simple three parameter model, transferable across organisms and is amenable to further systematic improvements.





Chemgenome methodology enables detection of not only protein coding regions but also promoters, introns & exons etc.. Garima Khandelwal and B. Jayaram, ”A phenomenological model for predicting melting temperatures of DNA sequences”, ***PLoS ONE*, 2010,** *5(8):* e12433. doi:10.1371/journal.pone.0012433

Nucleotide stability profile of Genomic DNA

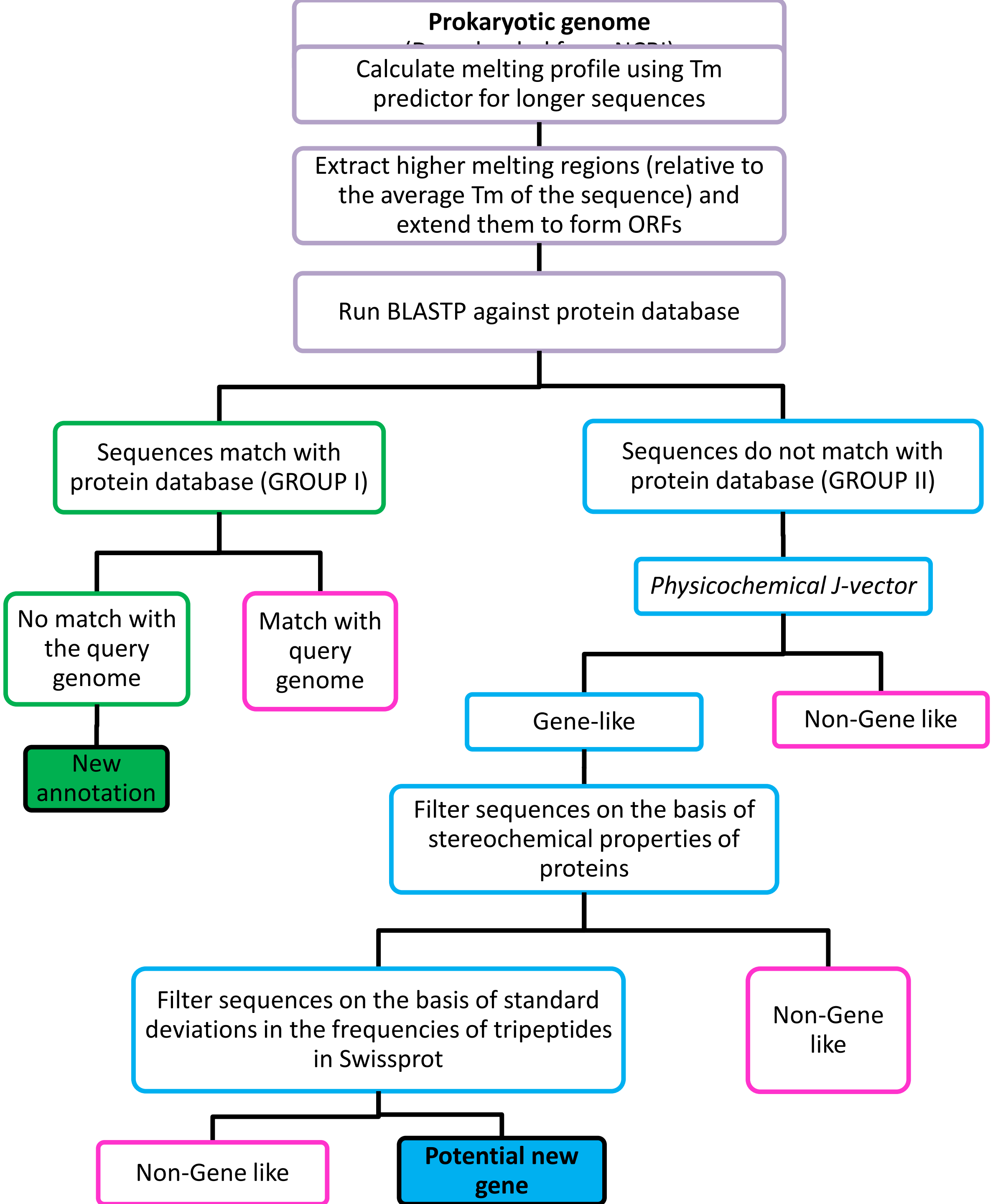


```
source             1 4639675
gene             190     255
CDS              190     255
gene             337    2799
CDS              337    2799   bifunctional: aspartokinase I (N-terminal); homoserine dehydrogenase I (C-terminal)
gene            2801    3733
CDS             2801    3733
gene            3734    5020
CDS             3734    5020
gene            5234    5530
```

Nucleotide stability profile (Top panel) for a stretch (1-12200 bases) of Escherichia coli K-12 (NCBI ID: NC_000913) genome, along with the sequence annotations plotted using Artemis software, depicting lower thermodynamic stability for non-genic regions. The blocks in green (middle panel) depict annotated CDS from the genome with their functional annotation information (bottom panel)

Garima Khandelwal, Rebecca Lee, B. Jayaram, D. L. Beveridge, "A Statistical Thermodynamic Model for Investigating the Stability of DNA Sequences from Oligonucleotides to Genomes", *Biophys. J., 2014, 106 (11),* 2465-2473; DOI: 10.1016/j.bpj.2014.04.029.

DNA Energetics helps in identifying new genes even in ‘annotated’ genomes!



Prokaryotic genome
(Downloaded from NCBI)

Calculate melting profile using Tm predictor for longer sequences

Extract higher melting regions (relative to the average Tm of the sequence) and extend them to form ORFs

Run BLASTP against protein database

Sequences match with protein database (GROUP I)

No match with the query genome

New annotation

Match with query genome

Sequences do not match with protein database (GROUP II)

Physicochemical J-vector

Gene-like

Non-Gene like

Filter sequences on the basis of stereochemical properties of proteins

Filter sequences on the basis of standard deviations in the frequencies of tripeptides in Swissprot

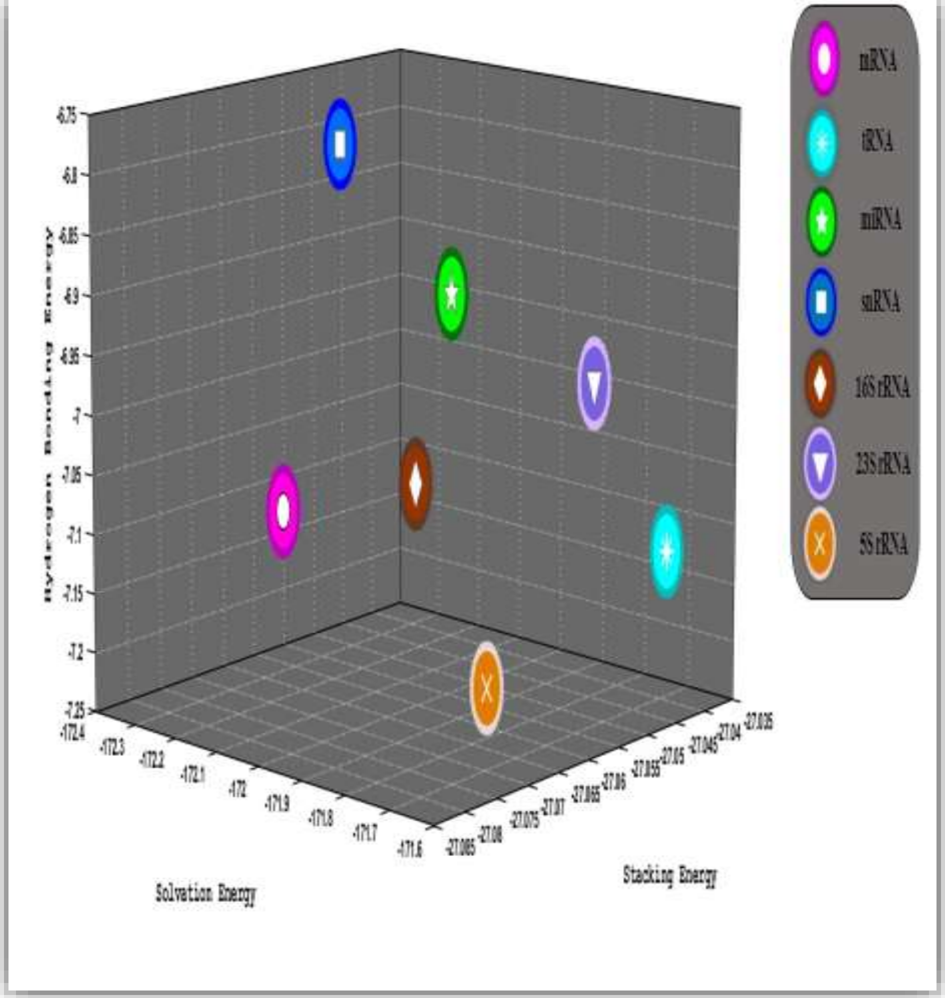
Non-Gene like

Non-Gene like

Potential new gene

Garima Khandelwal, Jalaj Gupta, B. Jayaram, “Predicting New Genes in Prokaryotes”, *J. Bio Sci., 2012, 37,* 433-444.

Physico-chemical fingerprinting of RNA genes



We advance here a novel concept for characterizing different classes of RNA genes on the basis of physico-chemical properties of DNA sequences.

Data consists of ~7.6 million RNA genes comprising ~7.3 million mRNA (magenta, circle), 255524 tRNA (cyan, star), 5250 miRNA (green, pentagon), 3747 snRNA (blue, square), 13997 16S rRNA (brown, diamond), 13745 23S rRNA (purple, triangle) and 12907 5S rRNA (orange, cross) genes for 9282 prokaryotes and eukaryotes available at NCBI.

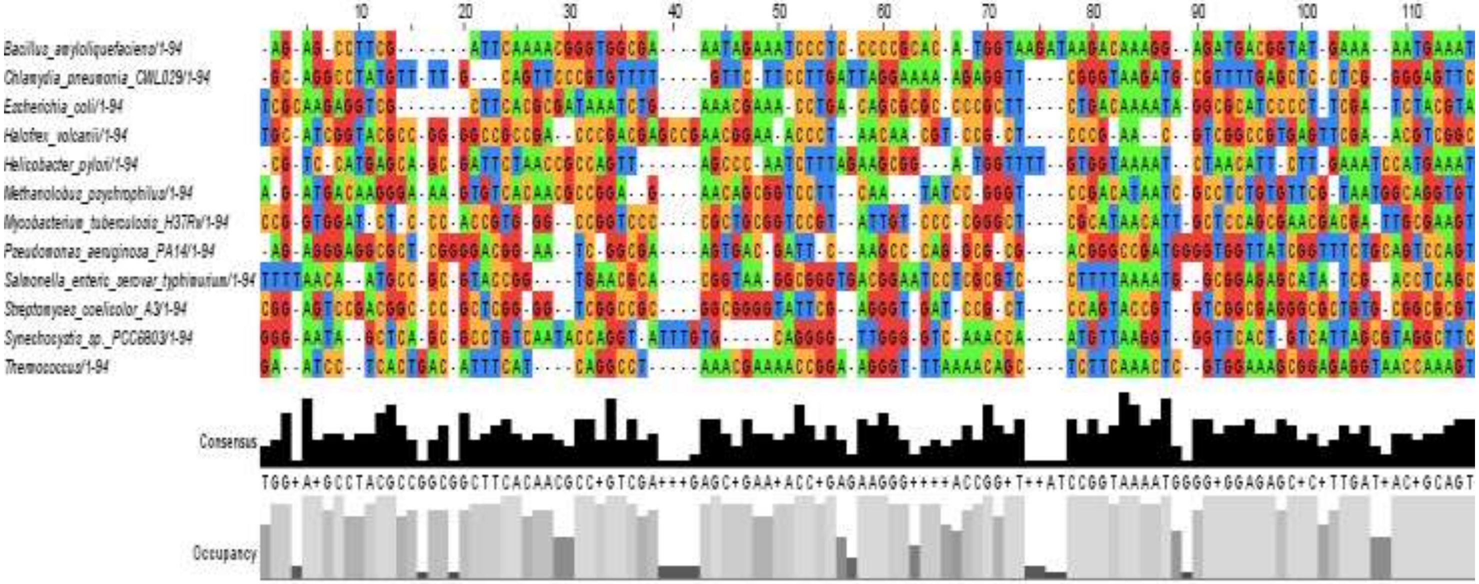
Ankita Singh, Akhilesh Mishra, Ali Khosravi, Garima Khandelwal, B. Jayaram, "Physico-chemical fingerprinting of RNA Genes", *Nucleic Acids Research*, 2016, DOI: 10.1093/nar/gkw1236

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

DNA is talking. What is the frequency to tune into…FM xx.x?

Transcriptional start sites

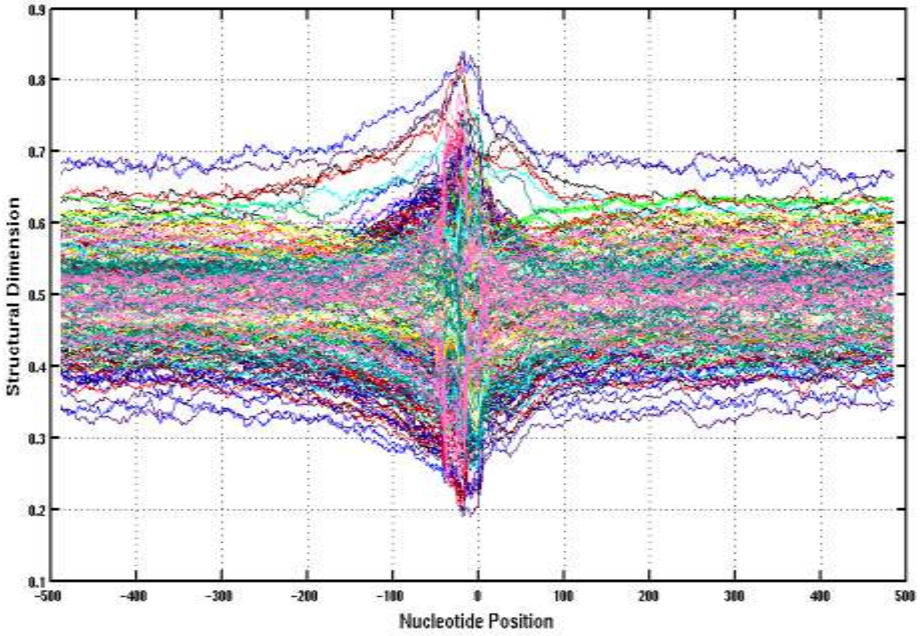
With almost no universal consensus promoter sequence in prokaryotes, recruitment of RNA polymerase (RNAP) to precise locations has remained an unsolved puzzle.


10
20
30
40
50
60
70
80
90
100
110
Bacillus_amyloliquefaciens/1-94
Chlamydia_pneumonia_CWL029/1-94
Escherichia_coli/1-94
Halofex_volcanii/1-94
Helicobacter_pylori/1-94
Methanolobus_psychrophilus/1-94
Mycobacterium_tuberculosis_H37Rv/1-94
Pseudomonas_aeruginosa_PA14/1-94
Salmonella_enteric_serovar_typhimurium/1-94
Streptomyces_coelicolor_A3/1-94
Synechocystis_sp._PCC6803/1-94
Thermococcus/1-94
Consensus
TGG+A+GCCTACGCCGGCGGCTTCACAACGCC+GTCGA+++GAGC+GAA+ACC+GAGAAGGG++++ACCGG+T++ATCCGGTAAAATGGGG+GGAGAGC+C+TTGAT+AC+GCAGT
Occupancy

Muliple sequence alignment of twelve sequences, of 90 nucleotide (-75 to +25 of TSS) length, randomly selected one from each organism using Tea-Coffee multiple sequence alignment tool using default parameters. The alignment was viewed by JalView software.

Sequences vary significantly from consensus

Towards a universal structural and energetic model for prokaryotic promoters



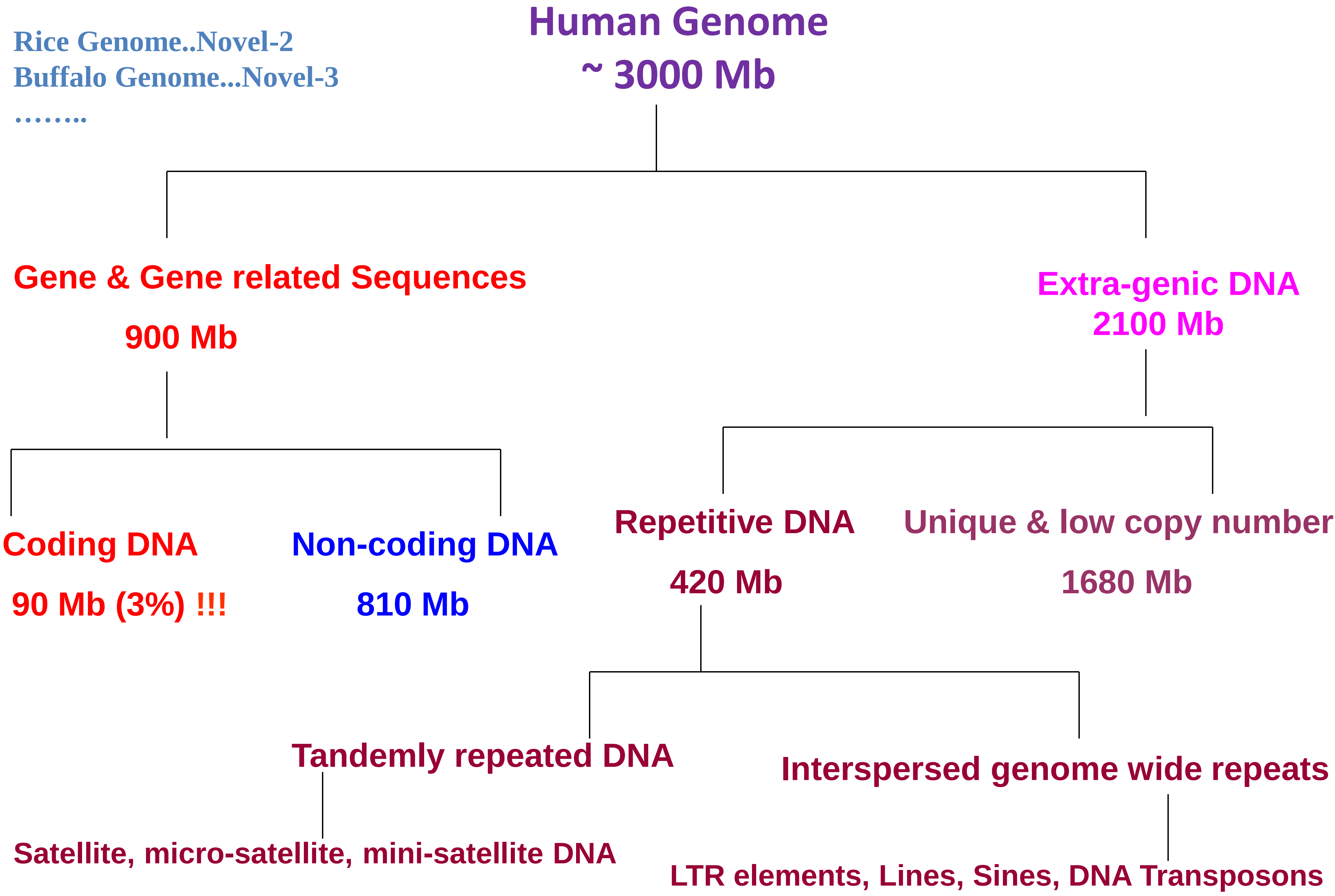
Normalized values of thirty one structural and energy parameters of all the twelve organisms vs nucleotide position with respect to TSS. Each organism was given single colour for all the 31 parameters. The plot represents 372 lines (31 x 12). Methodology was tested on 12 organisms (prokaryotes) belonging to Archaebacteria and Eubacteria comprising 16519 TSSs. A clear peak and cleft is observed at TSS.

A. Mishra, P. Siwach, P. Misra, B. Jayaram, M. Bansal, W.K. Olson, K.M. Thayer, D.L.Beveridge,"Towards a universal structural and energetic model for prokaryotic promoters", *Biophysical Journal, 2018*. DOI:10.1016/j.bpj.2018.08.002

DNA is talking!

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

Rice Genome..Novel-2
Buffalo Genome...Novel-3
……..



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

Unique & low copy number
1680 Mb

Tandemly repeated DNA

Satellite, micro-satellite, mini-satellite DNA

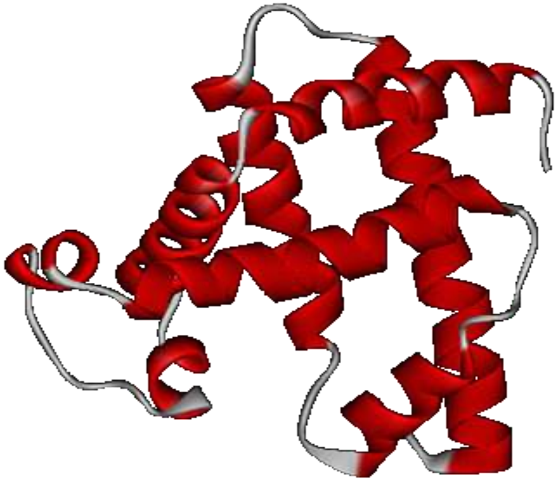
Interspersed genome wide repeats

LTR elements, Lines, Sines, DNA Transposons

The Grand Challenge of Protein Tertiary Structure Prediction

The Bhageerath Pathway

-------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
LEU HIS----



Is protein structure so important?

Seeing is believing! Proteins are the nanobiomachines which carry out the functions - coded in the genomes - to keep the organisms alive. How do they work?

How do cells sense their environment?

Nobel Prize in Chemistry 2012: **Robert J. Lefkowitz** and **Brian K. Kobilka** "for studies of G-protein-coupled receptors"

How are proteins synthesized?

Nobel Prize in Chemistry 2009: **Venkatraman Ramakrishnan, Thomas A. Steitz** and **Ada E. Yonath** "for studies of the structure and function of the ribosome"

How is mRNA made from DNA?

Nobel Prize in Chemistry 2006: **Roger D. Kornberg** "for his studies of the molecular basis of eukaryotic transcription"

How are ions and water transported in and out of cells?

Nobel Prize in Chemistry 2003: **Peter Agre and Roderick MacKinnon** "for discoveries concerning channels in cell membranes"

NMR for structure determination of proteins

Nobel Prize in Chemistry 2002: **John B. Fenn, Koichi Tanaka and Kurt Wüthrich** "for the development of methods for identification and structure analyses of biological macromolecules"

How is ATP synthesized?

Nobel Prize in Chemistry 1997: **Paul D. Boyer, John E. Walker and Jens C. Skou**
"for their elucidation of the enzymatic mechanism underlying the synthesis of adenosine triphosphate (ATP)“ and "for the first discovery of an ion-transporting enzyme, Na+, K+ -ATPase"

How does photosynthesis occur?

Nobel Prize in Chemistry 1988: **Johann Deisenhofer, Robert Huber** and **Hartmut Michel** "for the determination of the three-dimensional structure of a photosynthetic reaction centre"

Electron microscopy for structure determination of macromolecular assemblies: How do proteins recognize DNA?

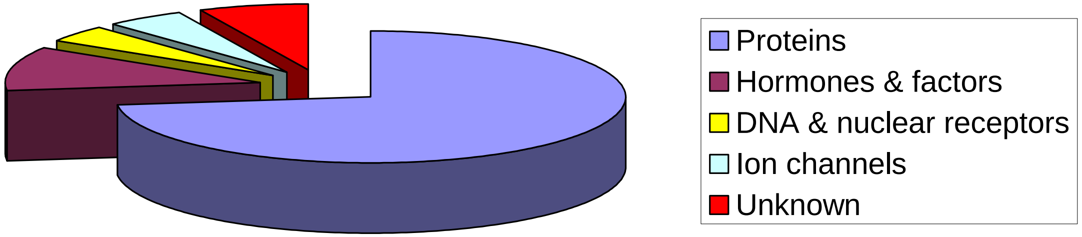
Nobel Prize in Chemistry 1982: **Aaron Klug** "for his development of crystallographic electron microscopy and his structural elucidation of biologically important nucleic acid-protein complexes"

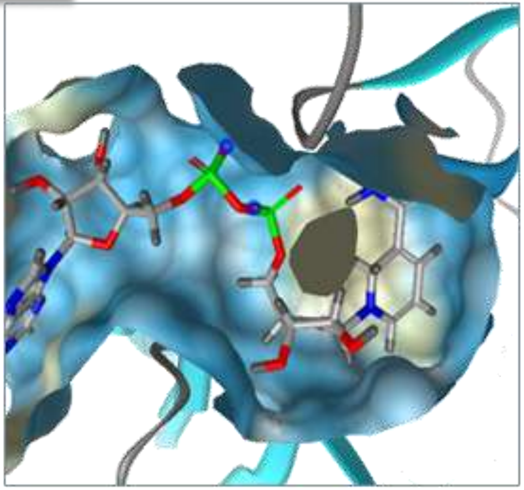
X-ray for structure determination of proteins: How is oxygen taken up by the body?

Nobel Prize in Chemistry 1962: **Max Ferdinand Perutz** and **John Cowdery Kendrew** "for their studies of the structures of globular proteins“

Source: www.nobelprize.org

WHY FOLD PROTEINS ?





“Proteins”- Majority of Drug Targets

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

	Experimental Approaches		**Computational approaches**	
	X-ray crystallography	**NMR spectroscopy**	**Comparative methods**	***De Novo* methods**
Time	months	months	Minutes to Hours	Hours to Days
Accuracy	very high	very high	Depends on similarity of template	Moderate
Limitation	Prone to failure as crystallizing a protein is still an art and many proteins (e.g. membrane) cannot be crystallized	Prone to failure, and is only applicable to small proteins (<150 amino acids)	Require a homologous template with at least 30% similarity. The accuracy is significantly reduced when the similarity is low.	Sampling and scoring limitations

Why Fold Proteins?

Motivation: Necessity for protein structure prediction and new algorithms

Non-redundant Reference Proteomes
10477 Organisms

- 993 Eukaryota
- 372 Archaea
- 2590 Virus
- 6522 Bacteria

www.uniprot.or g

Number of Protein Structures in PDB
136717 with large redundancy

Non-redundant Proteins in PDB
42263 Proteins

- 7681 Human Proteins (Out of 20214 reviewed proteins in UniProtKB)

www.rcsb.or g

Number of Protein Families in Pfam: 16712

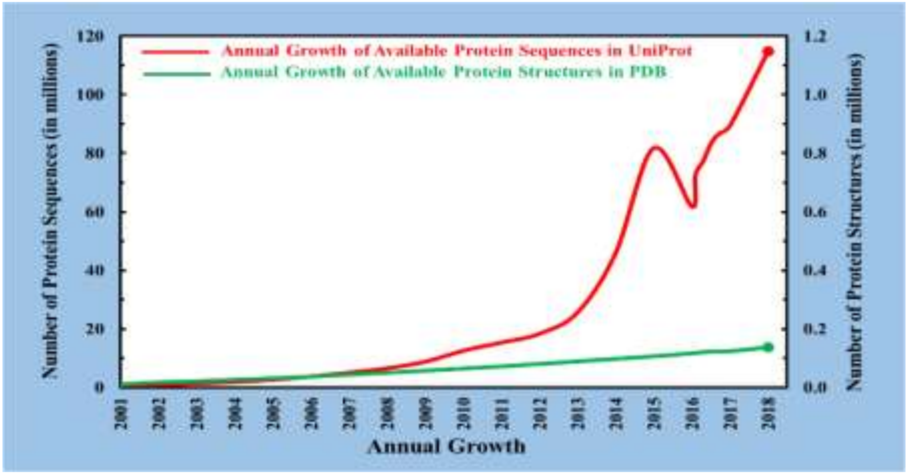
Number of Protein Families with Structural Information: 7990 (48%)

Number of Protein Families in Human Proteome: 5732

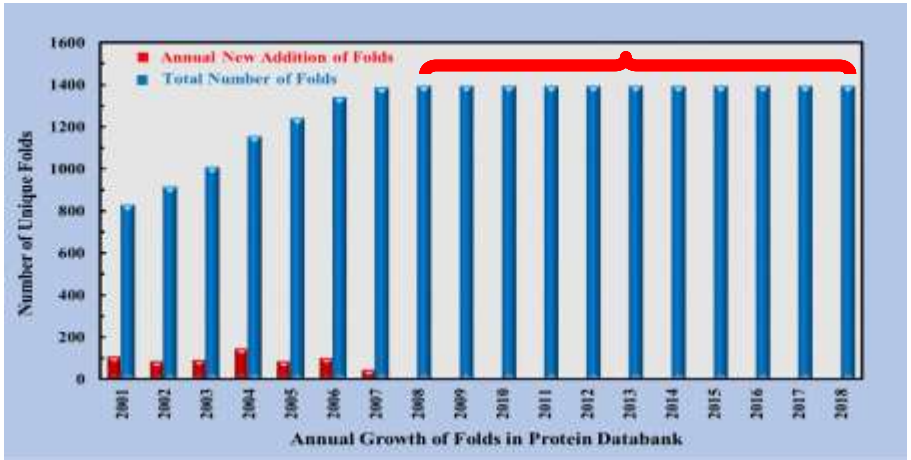
Number of Protein Families in Human Proteome with Structural Information: 2254 (39%)

www.pfam.xfam.org

How to bridge this gap?



Are all the possible folds already explored?



Data Source: www.rcsb.org; www.uniprot.org

Most of the drug targets are proteins and to initiate structure based drug discovery research protein structures are essential

Not enough experimental structures..
Urgency for good computational predictions!

Organism	**# Unique sequences**	**# Unique structures in RCSB**
Mycobacterium tuberculosis	4471 (uniprot)	380
Plasmodium falciparum 3D7 strain	5626	113
Plasmodium vivax	5392	53
Chikungunya virus	9	--
H1N1 influenza virus strain	15	7
Oryza Sativa	28,555 (ncbi)	24
Homo sapiens	26204 (uniprot) 37276 (ncbi)	5532

If you have the structure, you can hope to do structure based drug discovery and cure the disease

Proposal: Let us Create A Computational Protein (Data Bank) Structural Repository

*Also, if you know the rules of making structures from sequences, you can create designer structures (designer proteins) for specific functions (such as biocatalysts etc.) from amino acid sequences (the inverse folding problem). These synthetic biopolymers will be highly efficient & environment friendly.

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 a 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.

$H\Psi = E\,\Psi$ (Schrodinger) + $F = ma$ (Newton) on Supercomputers

The Nobel Prize in Chemistry 2013

"for the development of multiscale (QM/MM) models for complex chemical systems"

Martin Karplus
Harvard, USA
Univ. Strasbourg, France
b. 1930 (Austria)

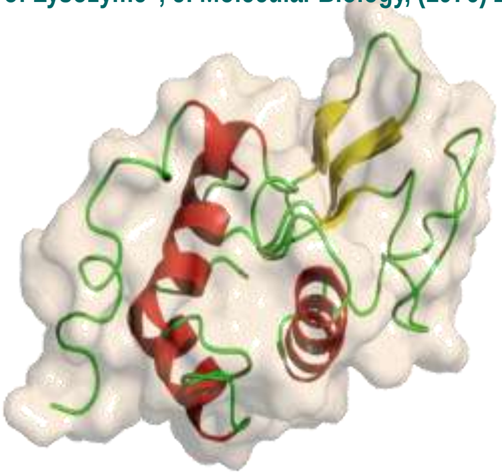
Michael Levitt
Stanford Univ., USA
b. 1947 (SA)

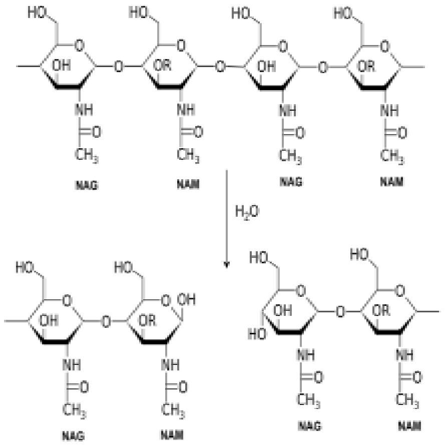
Arieh Warshel
USC, USA
b. 1940 (Israel)

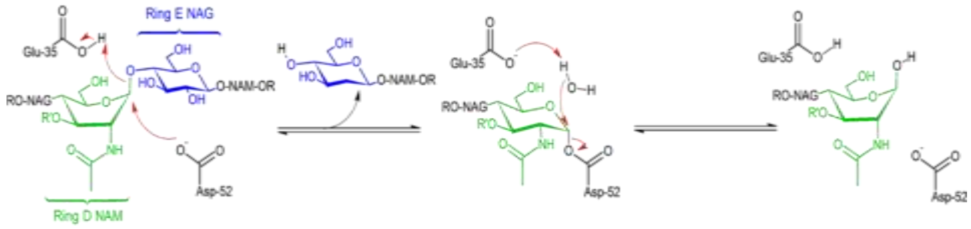
"Experiments in cyber space (in silico), without test tubes!"

First report of QM/MM

A. Warshel & M. Levitt, "Theoretical Studies of Enzymic Reactions; Dielectric, Electrostatic & Steric Stabilization of the Carbonium ion in the reaction of Lysozyme", J. Molecular Biology, (1976) 103, 227-249.





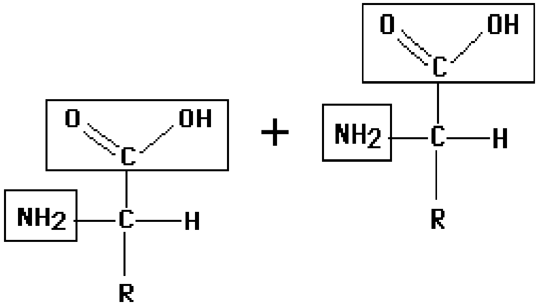


R = cell wall oligosaccharide chain
R' = cell wall peptide side chain

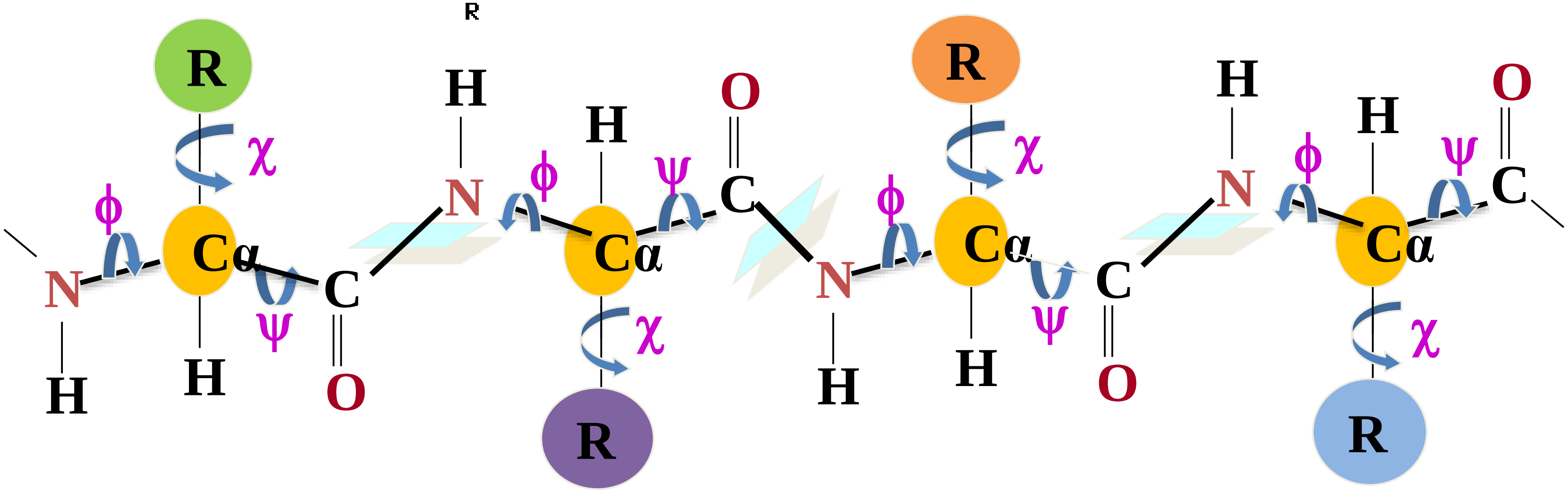
Computational methods are evolving to tackle complex many body problems and form the basis of sampling and scoring in *de novo* folding attempts

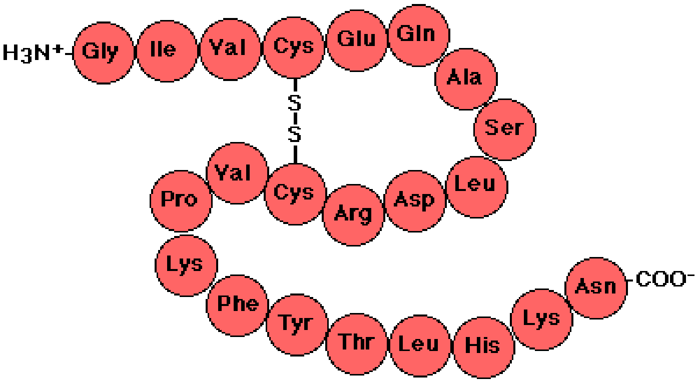
Some online software tools available for protein tertiary structure prediction

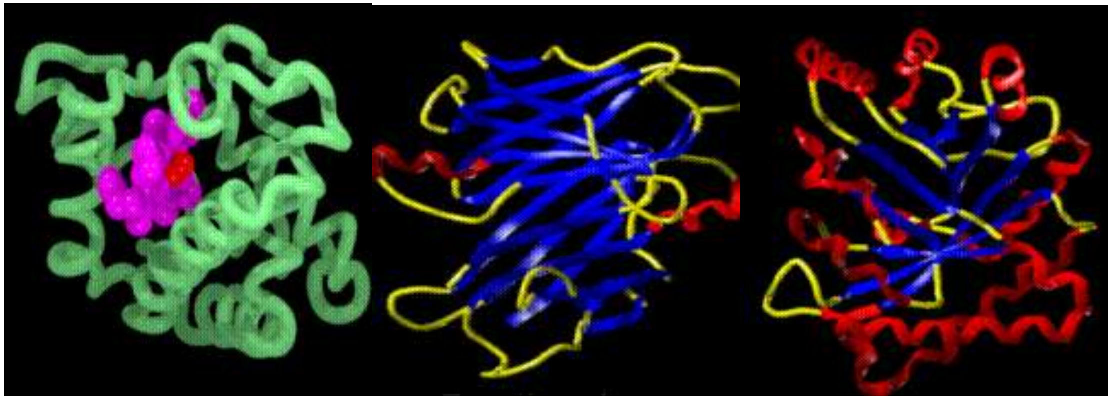
SN	Softwares	URLs	Description
1.	**CPHModels3.0**	http://www.cbs.dtu.dk/services/CPHmodels/	Protein homology modeling server
2.	**SWISS-MODEL**	http://swissmodel.expasy.org/SWISS-MODEL.html	Homology based methodology
3.	**Modeller**	http://salilab.org/modeller/	modeling by satisfaction of spatial restraints
4.	**3D-JIGSAW**	http://3djigsaw.com/	Homology based methodology
5.	**3D-PSSM**	http://www.sbg.bio.ic.ac.uk/~3dpssm/index2.html	Threading approach using 1D and 3D profiles
6.	**ROBETTA**	http://robetta.bakerlab.org	De novo Automated structure prediction
7.	**PROTINFO**	http://protinfo.compbio.washington.edu/	simulated annealing based methodology
8.	**SCRATCH**	http://scratch.proteomics.ics.uci.edu/	recursive neural networks, evolutionary information, fragment libraries and energy
9.	**I-TASSER**	http://zhanglab.ccmb.med.umich.edu/I-TASSER/	Based on threading approach
10.	**BHAGEERATH-H**	http://www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp	A Homology ab-initio Hybrid Web server for Protein Tertiary Structure Prediction



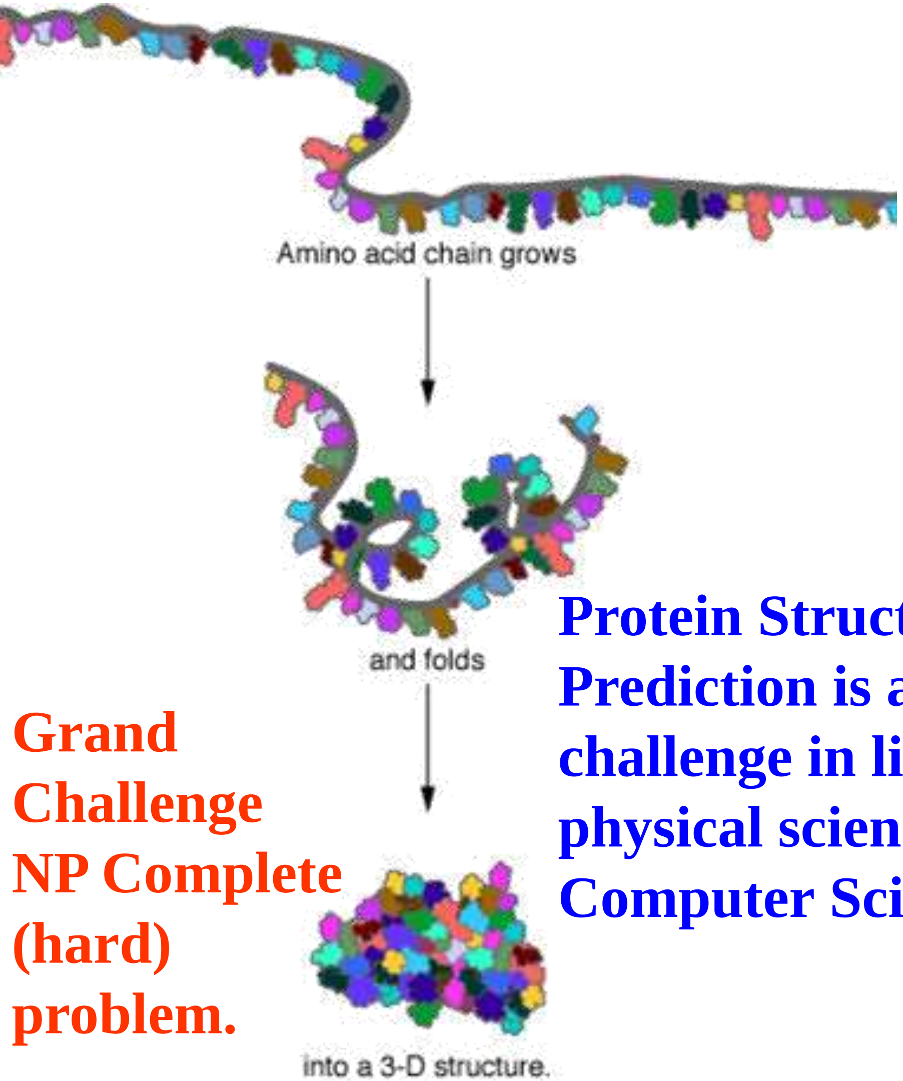
Specify the ϕ, ψ around each Cα-R
→ protein is folded!

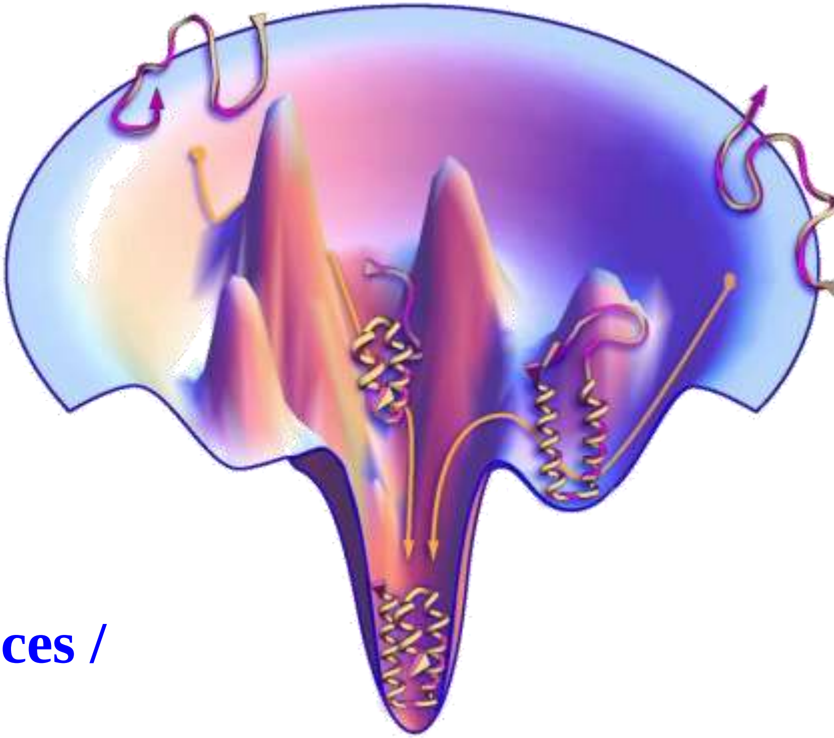






Protein Folding Problem



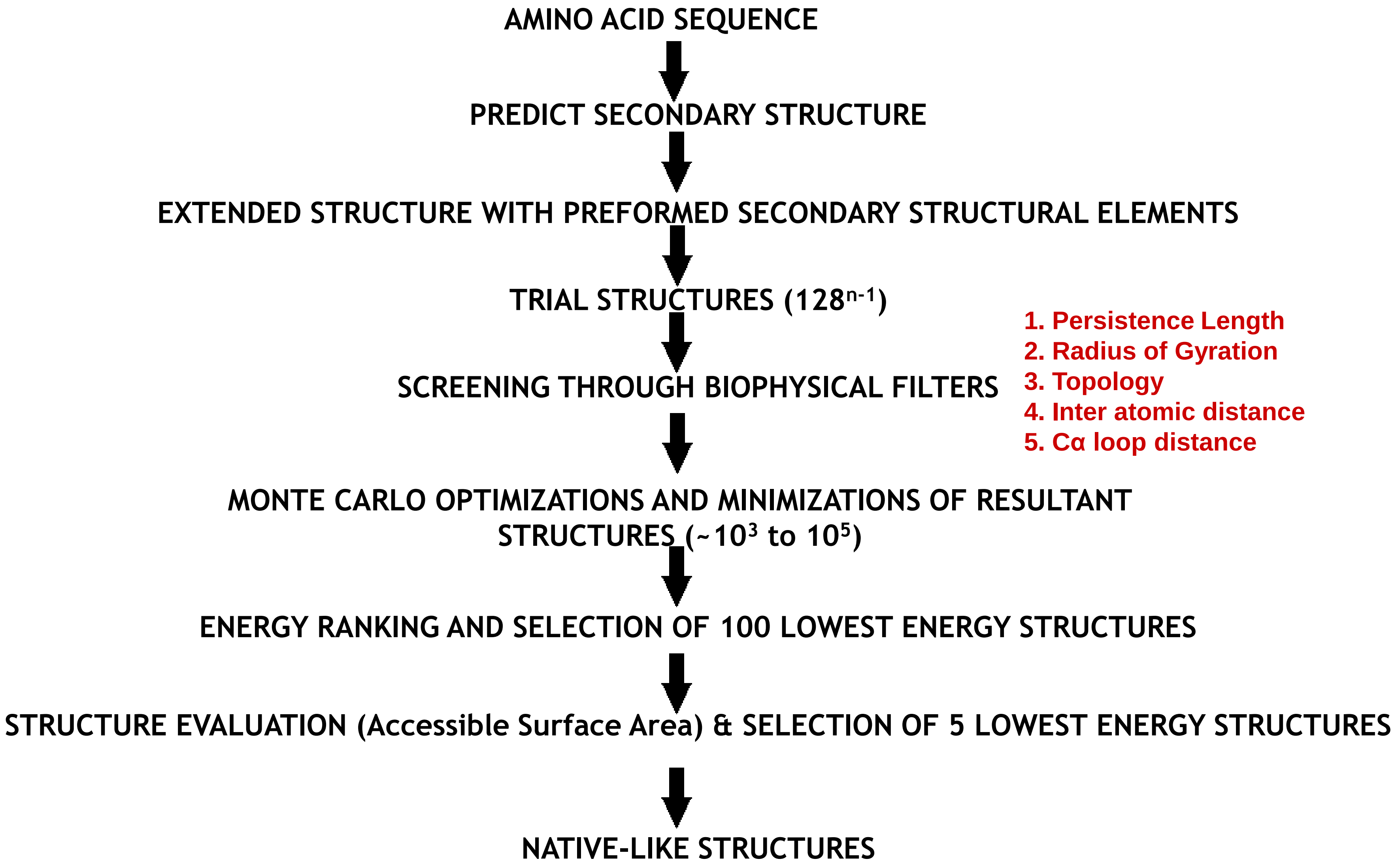


Protein Structure Prediction is an *open* challenge in life sciences / physical sciences and Computer Science.

Grand Challenge NP Complete (hard) problem.

"Native structure" at the bottom of the free energy Funnel. <u>Thermodynamic hypothesis of Anfinsen</u>

From Sequence to Structure: *Bhageerath* Pathway



AMINO ACID SEQUENCE

PREDICT SECONDARY STRUCTURE

EXTENDED STRUCTURE WITH PREFORMED SECONDARY STRUCTURAL ELEMENTS

TRIAL STRUCTURES (128^{n-1})

SCREENING THROUGH BIOPHYSICAL FILTERS

1. Persistence Length
2. Radius of Gyration
3. Topology
4. Inter atomic distance
5. Cα loop distance

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

ENERGY RANKING AND SELECTION OF 100 LOWEST ENERGY STRUCTURES

STRUCTURE EVALUATION (Accessible Surface Area) & SELECTION OF 5 LOWEST ENERGY STRUCTURES

NATIVE-LIKE STRUCTURES

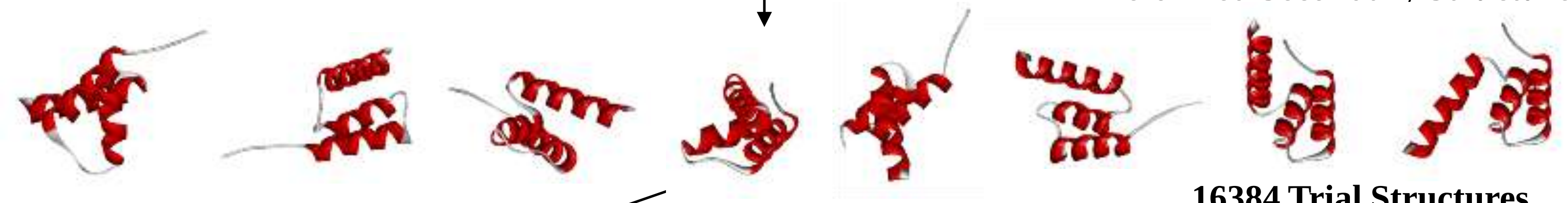
Bhageerath Strategy: A Case Study of Mouse C-Myb

DNA Binding (52 AA)

LIKGPWTKEEDQRVIELVQKYGPKRWSVIAKHLKGRIGKQCRERWHNHLNPE

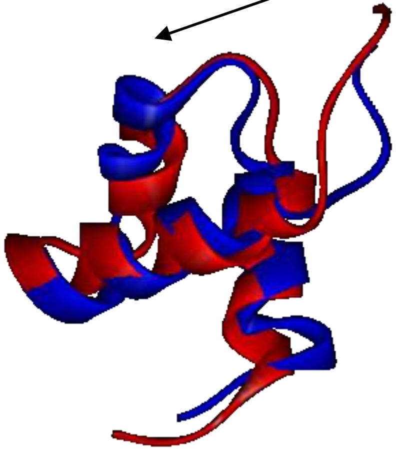
Sequence

Preformed Secondary Structure



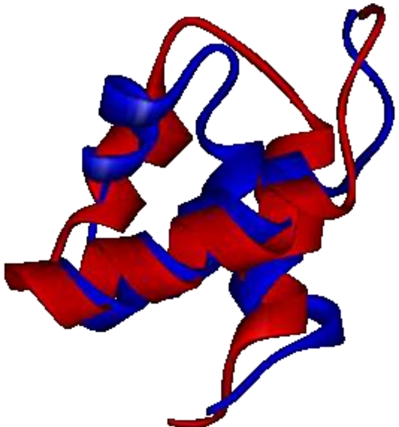
16384 Trial Structures

Biophysical Filters & Clash Removal
10632 Structures



RMSD=2.87, Energy Rank=1774

Energy Scans



RMSD=4.0 Å, Energy Rank=4

Blue: Native; Red: Predicted

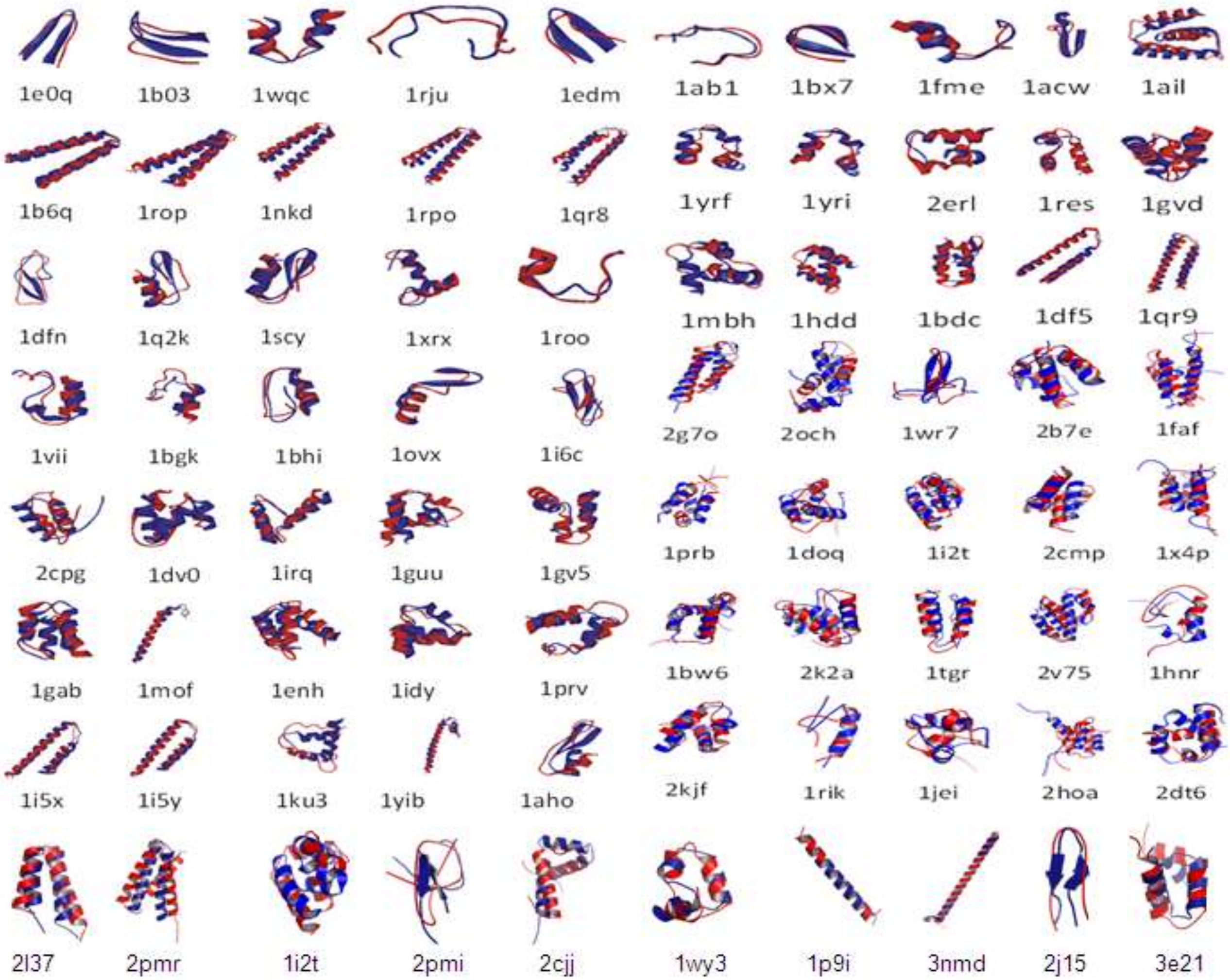
Bhageerath structures for 80 small globular proteins

🟥

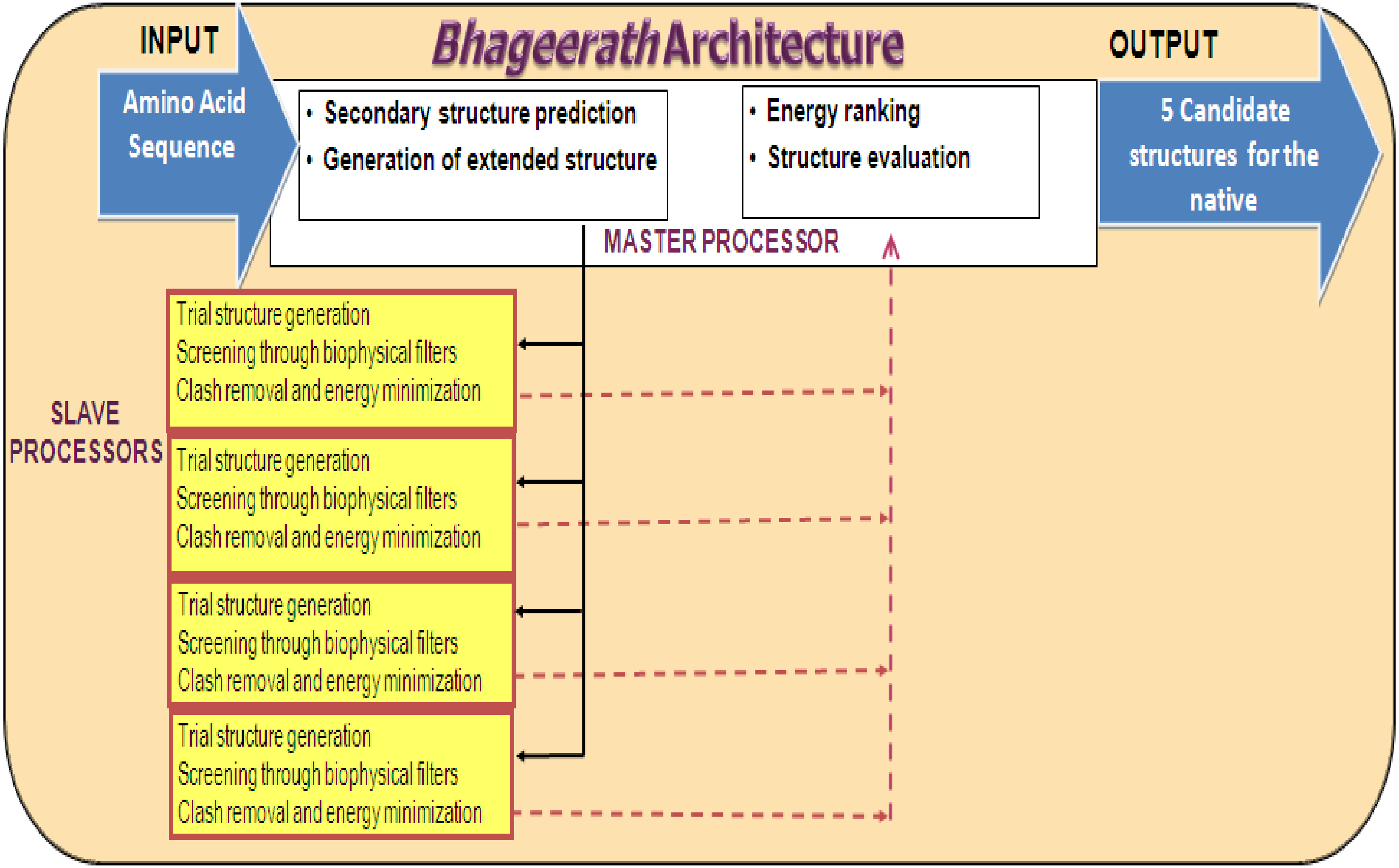
Bhageerath predicted Structure

🟦

Native Structure



***Bhageerath* protocol can predict one or more candidate structures within an RMSD of 5Ǻ from the native for small globular proteins with less than five secondary structural elements (≤100 amino acids).** **P. Narang, K. Bhushan, S. Bose and B. Jayaram, "A computational pathway for bracketing native-like structures for small alpha helical globular proteins", *Phys. Chem. Chem. Phys.*, 2005, *7*, 2364-2375.**



Bhageerath Architecture

INPUT

Amino Acid Sequence

- Secondary structure prediction
- Generation of extended structure

- Energy ranking
- Structure evaluation

MASTER PROCESSOR

OUTPUT

5 Candidate structures for the native

SLAVE PROCESSORS

Trial structure generation
Screening through biophysical filters
Clash removal and energy minimization

Trial structure generation
Screening through biophysical filters
Clash removal and energy minimization

Trial structure generation
Screening through biophysical filters
Clash removal and energy minimization

Trial structure generation
Screening through biophysical filters
Clash removal and energy minimization

All the (eight) modules of the protocol are currently implemented on a dedicated 280 AMD Opteron 2.4 GHz processor cluster (~3 teraflops).

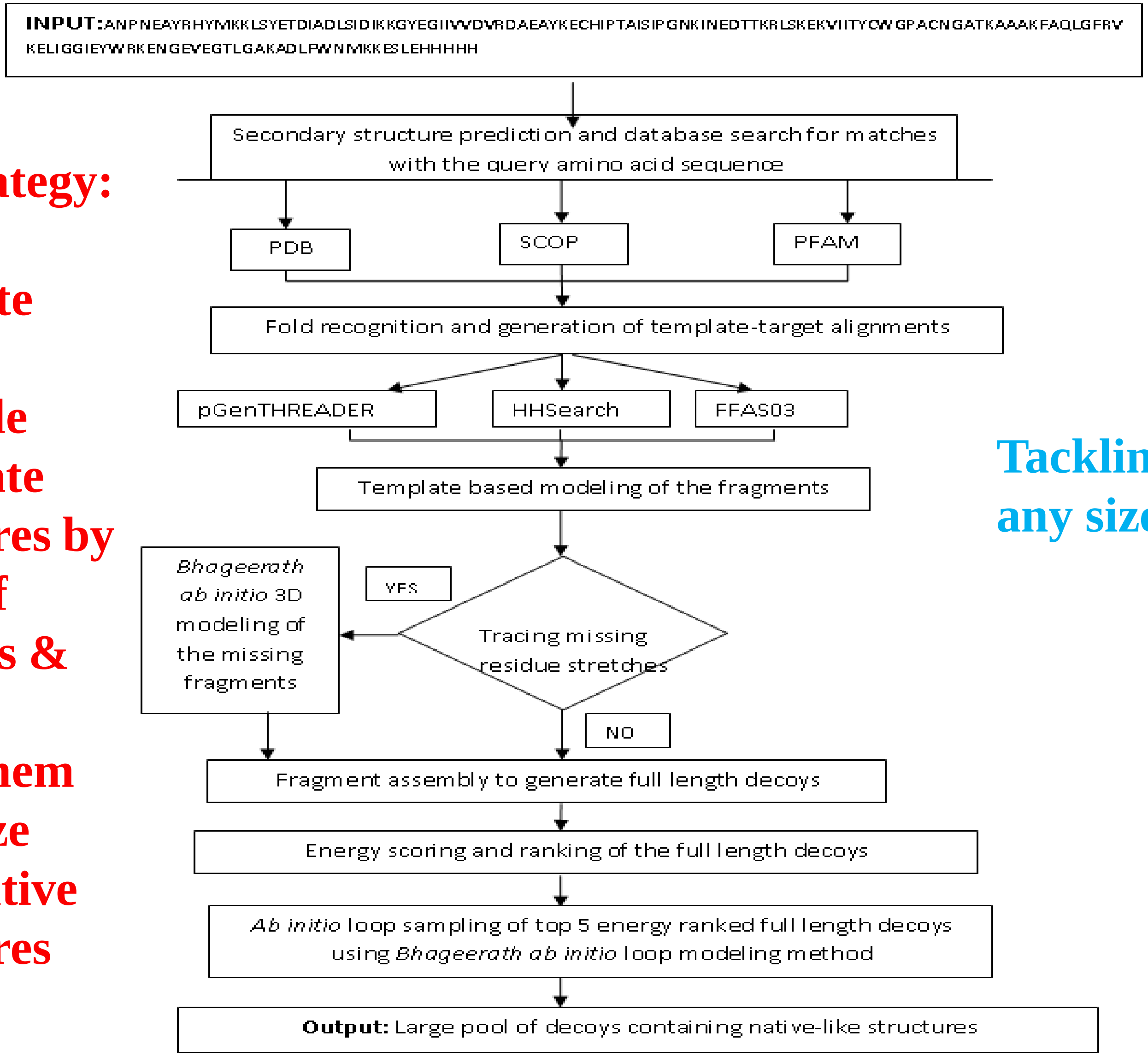
Jayaram et al., *Bhageerath, Nucl. Acid Res.*, 2006, 34, 6195-6204

Bhageerath-H Strgen: an exhausitive homology/ *ab initio* hybrid method for protein conformational sampling

Overall strategy:

(1) Generate several plausible candidate structures by a mix of methods &

(2) Score them to realize near-native structures



INPUT: ANPNEAYRHYMKKLSYETDIADLSIDIKKGYEGIIVVDVRDAEAYKECHIPTAISIPGNKINEDTTKRLSKEKVIITYCWGPACNGATKAAAKFAQLGFRV KELIGGIEYWRKENGEVEGTLGAKADLPWNMKKESLEHHHHH

Secondary structure prediction and database search for matches with the query amino acid sequence

PDB

SCOP

PFAM

Fold recognition and generation of template-target alignments

pGenTHREADER

HHSearch

FFAS03

Template based modeling of the fragments

Bhageerath ab initio 3D modeling of the missing fragments

YES

Tracing missing residue stretches

NO

Fragment assembly to generate full length decoys

Energy scoring and ranking of the full length decoys

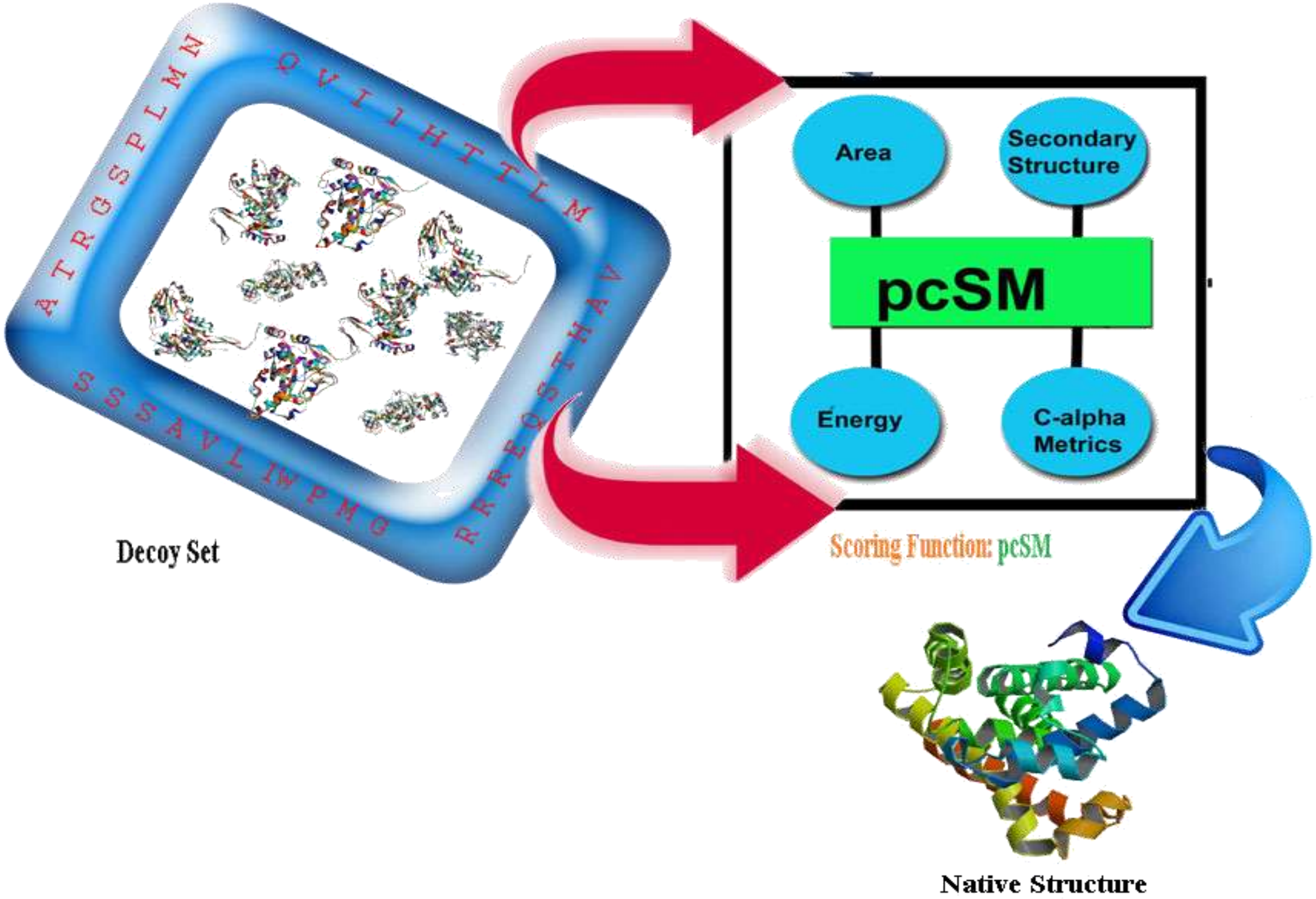
Ab initio loop sampling of top 5 energy ranked full length decoys using *Bhageerath ab initio* loop modeling method

Output: Large pool of decoys containing native-like structures

Tackling proteins of any size and fold

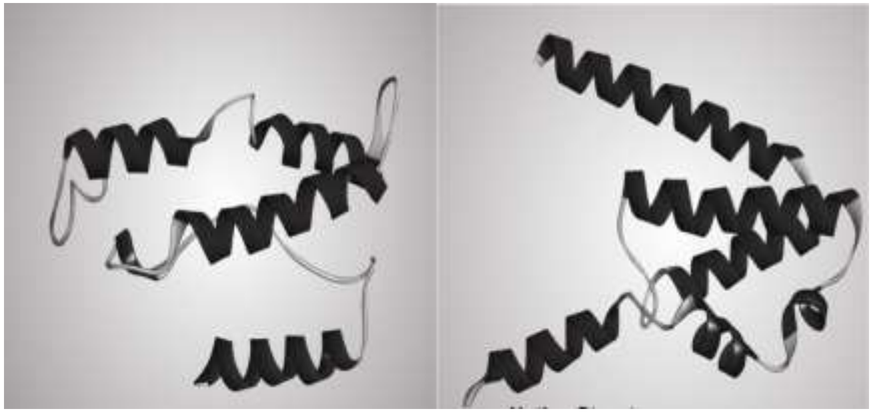
Priyanka Dhingra, B. Jayaram, "A homology/*ab initio* hybrid algorithm for sampling near-native protein conformations", *J. Comput. Chem., 2013,* DOI: 10.1002/jcc.23339.

pcSM (A physico-chemical scoring function)

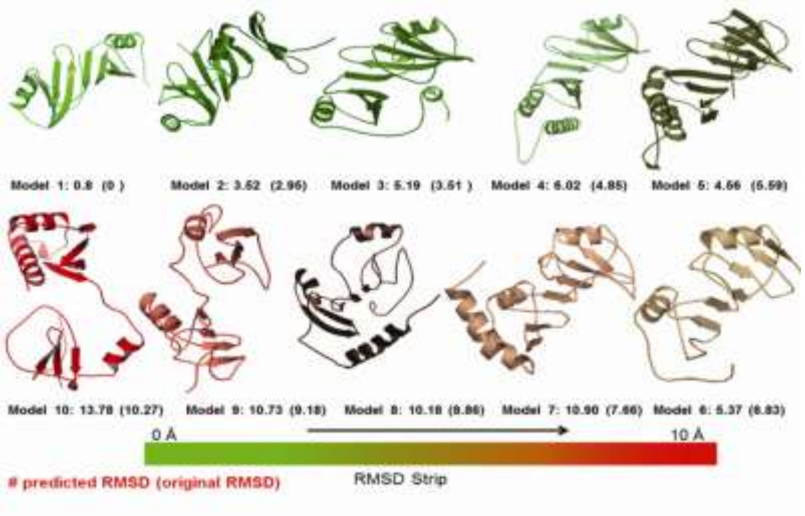


Avinash Mishra, Satyanarayan Rao, Aditya Mittal and B. Jayaram, "Capturing Native/Native-like Structures with a Physico-Chemical Metric (pcSM) in Protein Folding", ***BBA proteins & proteomics, 2013,*** **1834(8), 1520-31; DOI: 10.1016/j.bbapap.2013.04.023.**

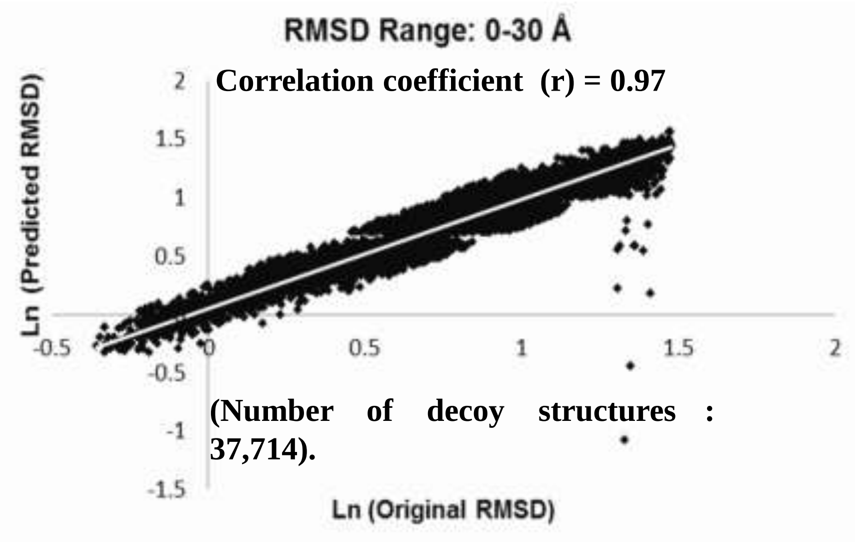
D2N: Distance to the native



Who is the Native ?



Case Study: Predicted vs Original RMSD of T0644 target from CASP10 dataset, original RMSD is given in parenthesis.



Correlation coefficient (r) = 0.97

(Number of decoy structures : 37,714).

It is also possible to assess the distance of a predicted structure from the native (to within 2 Å in the absence of experimental structures !

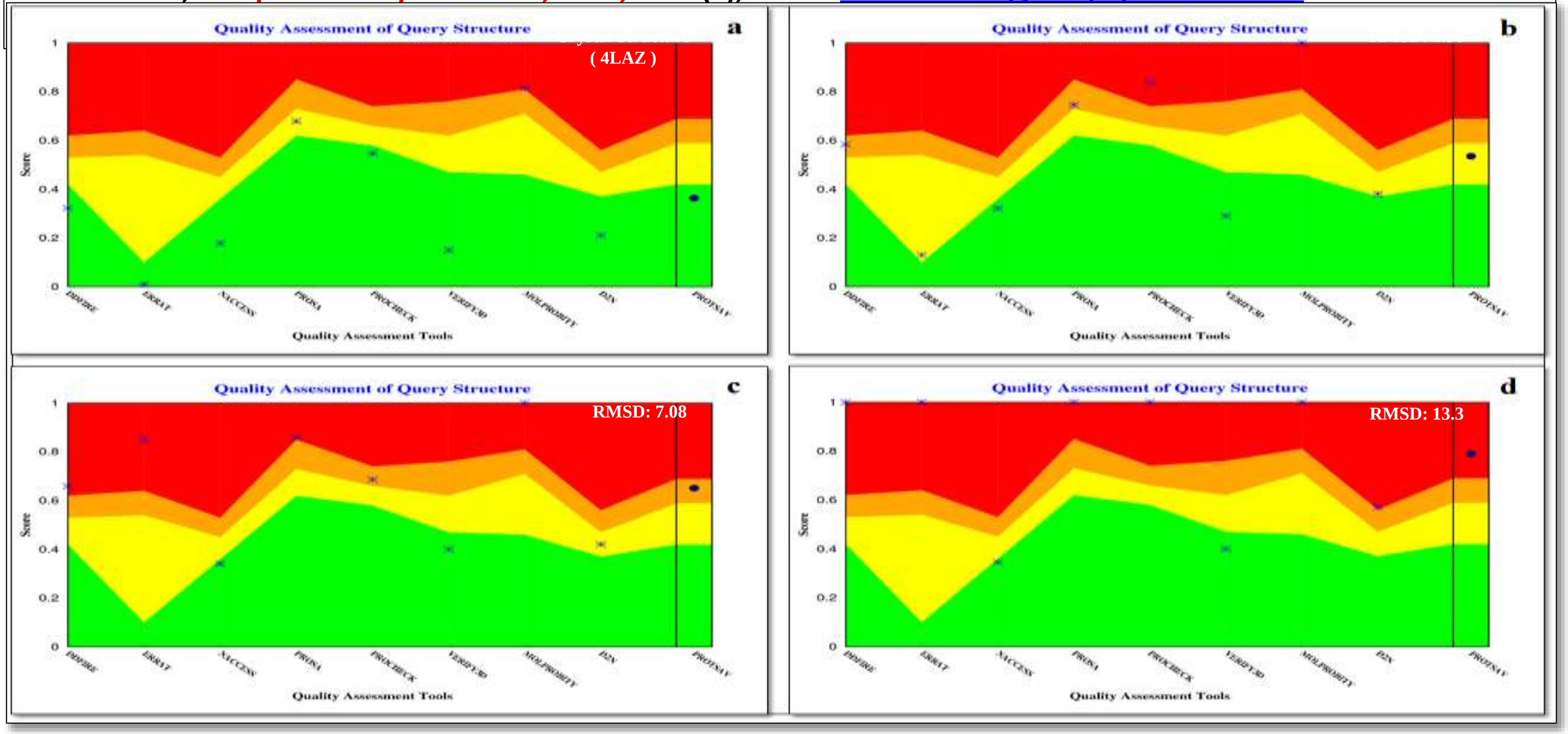
Avinash Mishra, Prashant Rana, Aditya Mittal and B. Jayaram, “D2N:Distance to the native”, ***BBAP&P, 2014,*** **1844 (10), 1798-1807; doi:10.1016/j.bbapap.2014.07.010.**

Availability: http://www.scfbio-iitd.res.in/software/d2n.jsp

Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi
A Centre of Excellence of the Department of Biotechnology, Govt. of India

ProTSAV: *Protein Tertiary Structure Analysis & Validation Server*

ProTSAV (Protein Tertiary Structure Analysis & Validation Server) **is a meta-webserver designed to evaluate and validate protein structures. The goal of this work is to provide different scores of protein structures from different validation tools and a final score derived from them to furnish better insight into the quality of the predicted protein structure to the user . The Meta-server is freely accessible in public domain at: http://www.scfbio-iitd.res.in/software/proteomics/protsav.jsp; Ankita Singh, Rahul Kaushik, Avinash Mishra, Asheesh Shankar Sharma, B. Jayaram, "PROTSAV: A Protein Tertiary Structure Analysis & Validation Metaserver",** ***BBA proteins & proteomics, 2016,*** **1864(1), 11-19. DOI: 10.1016/j.bbapap.2015.10.004**



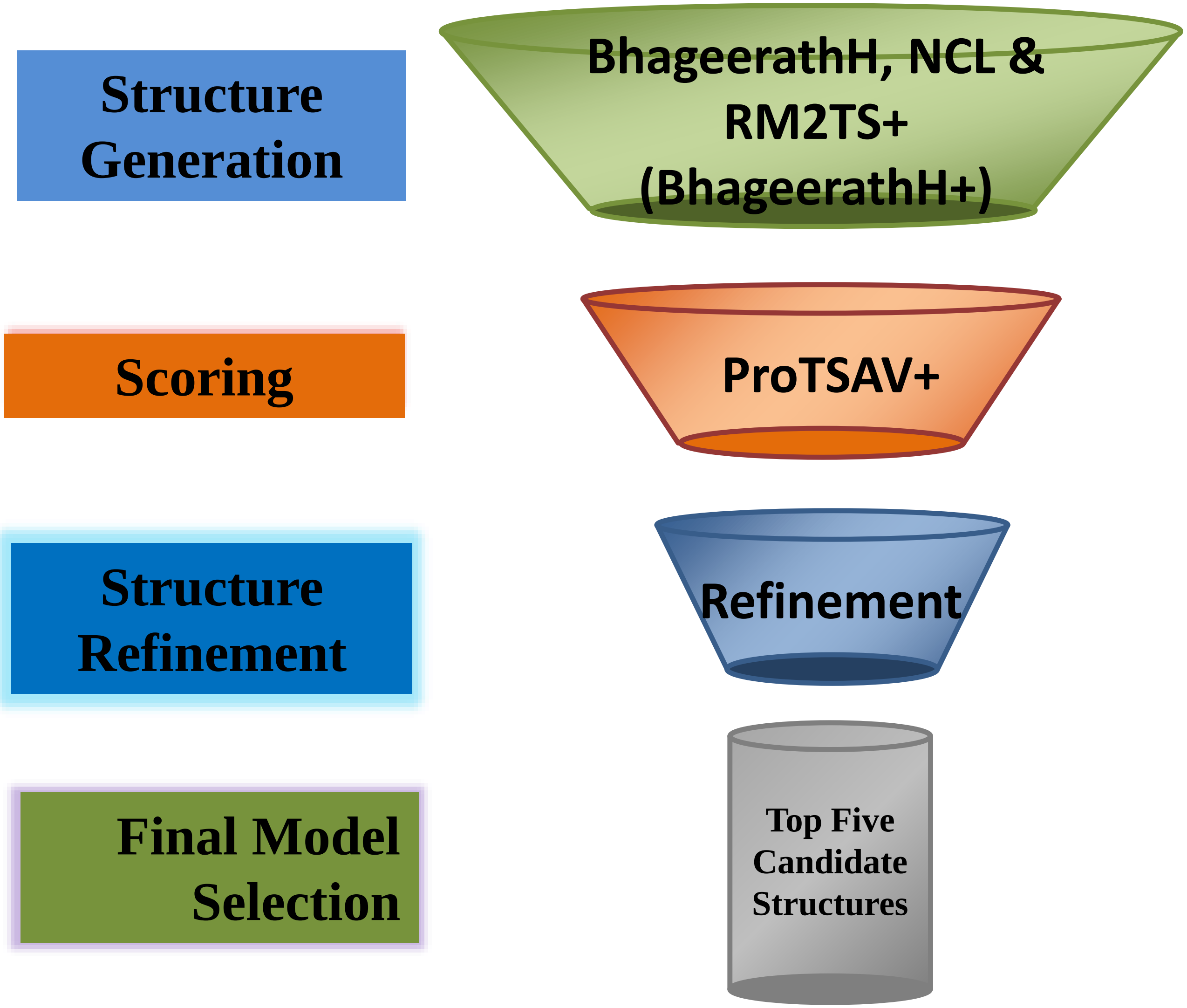
Any module predicts submitted query structure within a range of 0-2 Å rmsd.

Any module predicts submitted query structure within a range of 2-5 Å rmsd.

Any module predicts submitted query structure within a range of 5-8 Å rmsd.

Any module predicts submitted query structure above 8 Å rmsd.

The “Bhageerath-H+” Pathway



Structure Generation

BhageerathH, NCL & RM2TS+ (BhageerathH+)

Scoring

ProTSAV+

Structure Refinement

Refinement

Final Model Selection

Top Five Candidate Structures

BhageerathH+ was fielded in CASP12 (May-July, 2016)

CASP12

12th Community Wide Experiment on the Critical Assessment of Techniques for Protein Structure Prediction

http://predictioncenter.org/

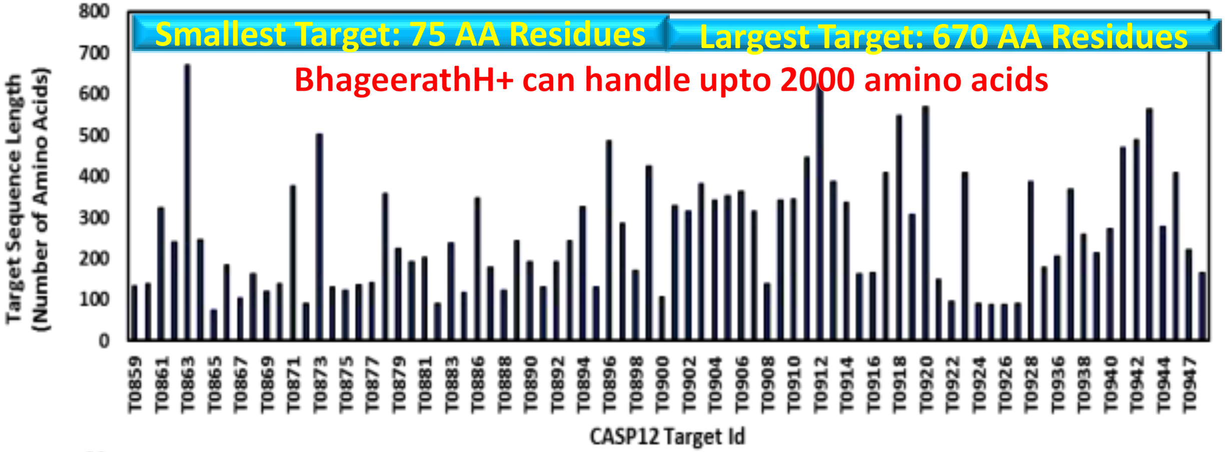
Start Date: 2nd May, 2016
End Date: 12th July, 2016

Total Number of Participants: 189
80 (Servers) + 109 (Human Group)

Total Number of Countries Involved: 28

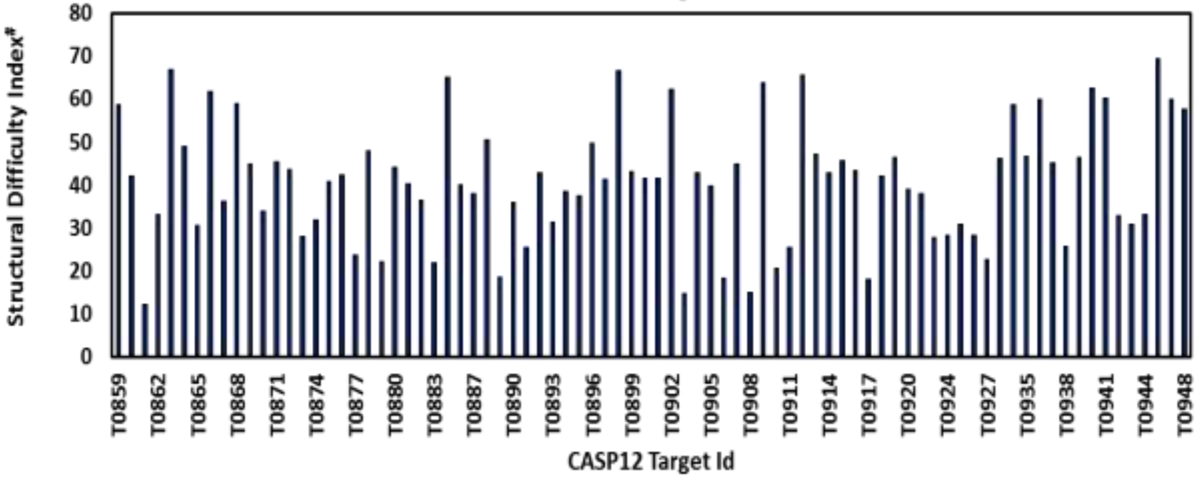
Total Number of Unique Targets: 82

Total Number of Domains: 96
38 (TBM*) + 39 (FM*) + 19 (TBM/FM*)


Smallest Target: 75 AA Residues

Largest Target: 670 AA Residues

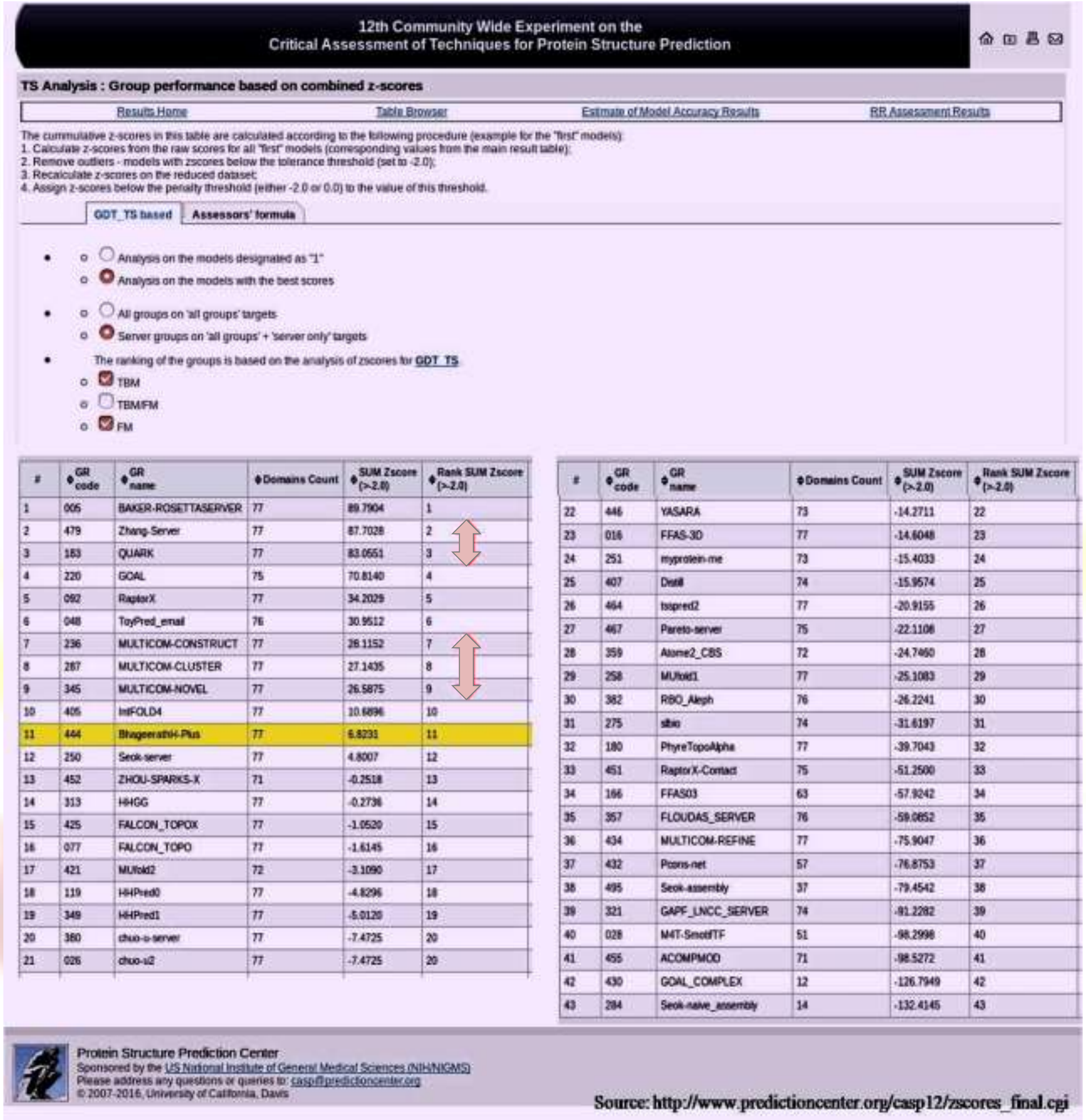
BhageerathH+ can handle upto 2000 amino acids



*TBM: Template Based Modeling *FM- Free Modeling

#Kaushik, R. and Jayaram, B. (2016). Structural difficulty index: A reliable measure for modelability of protein tertiary structures. *PEDS, 29(9)*, 391-97

BhageerathH+ Prediction Official Ranking among the Participating Servers



12th Community Wide Experiment on the
Critical Assessment of Techniques for Protein Structure Prediction

TS Analysis : Group performance based on combined z-scores

The cummulative z-scores in this table are calculated according to the following procedure (example for the "first" models):
1. Calculate z-scores from the raw scores for all "first" models (corresponding values from the main result table);
2. Remove outliers - models with zscores below the tolerance threshold (set to -2.0);
3. Recalculate z-scores on the reduced dataset;
4. Assign z-scores below the penalty threshold (either -2.0 or 0.0) to the value of this threshold.

- ○ Analysis on the models designated as "1"
  ● Analysis on the models with the best scores
- ○ All groups on 'all groups' targets
  ● Server groups on 'all groups' + 'server only' targets
- The ranking of the groups is based on the analysis of zscores for GDT_TS
  - ☑ TBM
  - □ TBM/FM
  - ☑ FM

#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)
1	005	BAKER-ROSETTASERVER	77	89.7904	1
2	479	Zhang-Server	77	87.7028	2
3	183	QUARK	77	83.0551	3
4	220	GOAL	75	70.8140	4
5	092	RaptorX	77	34.2029	5
6	048	ToyPred_email	76	30.9512	6
7	236	MULTICOM-CONSTRUCT	77	28.1152	7
8	287	MULTICOM-CLUSTER	77	27.1435	8
9	345	MULTICOM-NOVEL	77	26.5875	9
10	405	IntFOLD4	77	10.6896	10
11	444	BhageerathH-Plus	77	6.8231	11
12	250	Seok-server	77	4.8007	12
13	452	ZHOU-SPARKS-X	71	-0.2518	13
14	313	HHGG	77	-0.2736	14
15	425	FALCON_TOPOX	77	-1.0520	15
16	077	FALCON_TOPO	77	-1.6145	16
17	421	MUfold2	72	-3.1090	17
18	119	HHPred0	77	-4.8296	18
19	349	HHPred1	77	-5.0120	19
20	380	chuo-u-server	77	-7.4725	20
21	026	chuo-u2	77	-7.4725	20

#	GR code	GR name	Domains Count	SUM Zscore (>-2.0)	Rank SUM Zscore (>-2.0)
22	446	YASARA	73	-14.2711	22
23	016	FFAS-3D	77	-14.6048	23
24	251	myprotein-me	73	-15.4033	24
25	407	Distill	74	-15.9574	25
26	464	tsspred2	77	-20.9155	26
27	467	Pareto-server	75	-22.1108	27
28	359	Atome2_CBS	72	-24.7460	28
29	258	MUfold1	77	-25.1083	29
30	382	RBO_Aleph	76	-26.2241	30
31	275	slbio	74	-31.6197	31
32	180	PhyreTopoAlpha	77	-39.7043	32
33	451	RaptorX-Contact	75	-51.2500	33
34	166	FFAS03	63	-57.9242	34
35	357	FLOUDAS_SERVER	76	-59.0852	35
36	434	MULTICOM-REFINE	77	-75.9047	36
37	432	Pcons-net	57	-76.8753	37
38	495	Seok-assembly	37	-79.4542	38
39	321	GAPF_LNCC_SERVER	74	-91.2282	39
40	028	M4T-SmotifTF	51	-98.2998	40
41	455	ACOMPMOD	71	-98.5272	41
42	430	GOAL_COMPLEX	12	-126.7949	42
43	284	Seok-naive_assembly	14	-132.4145	43

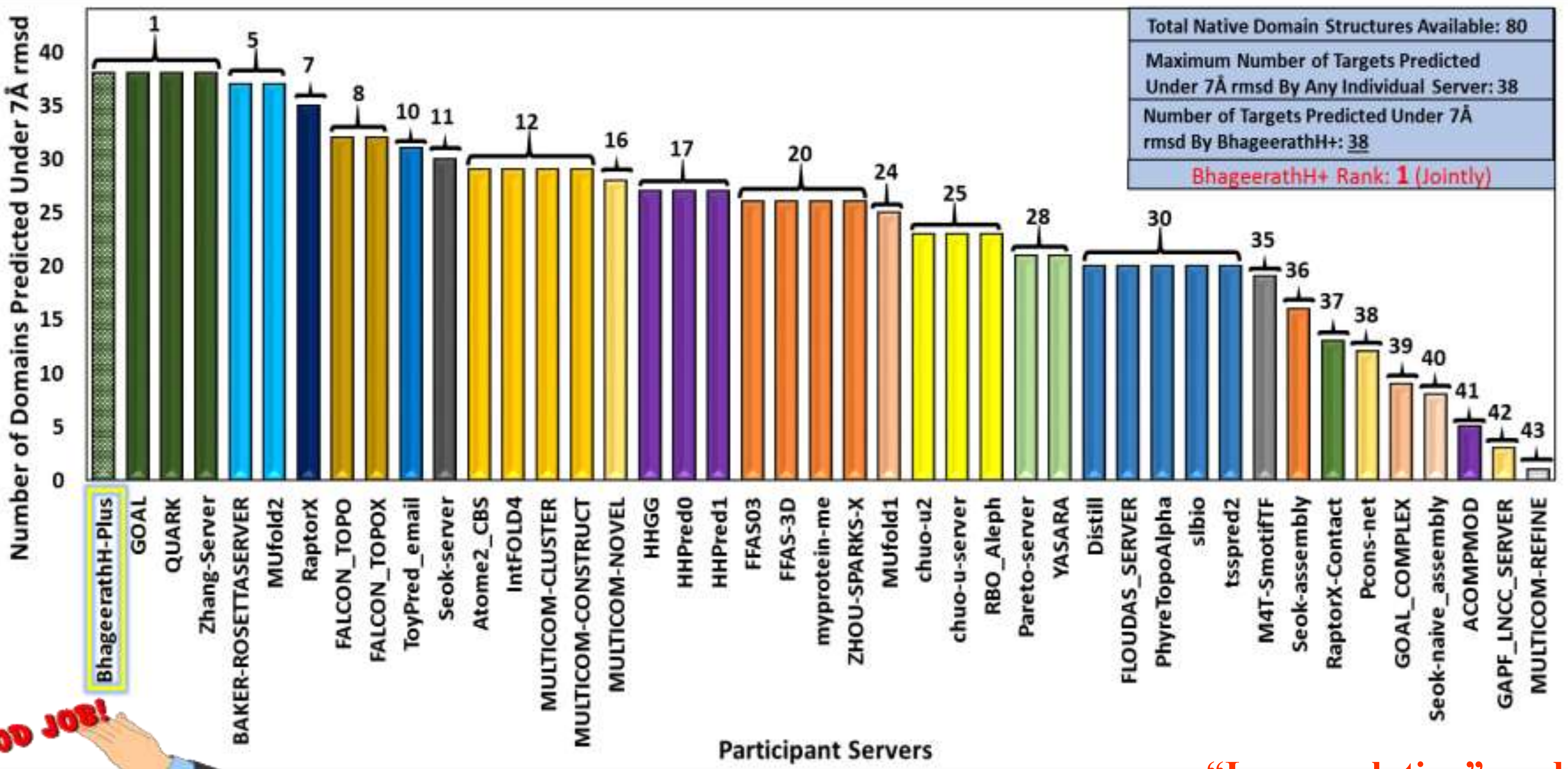
Protein Structure Prediction Center
Sponsored by the US National Institute of General Medical Sciences (NIH/NIGMS)
Please address any questions or queries to: casp@predictioncenter.org
© 2007-2016, University of California, Davis

Source: http://www.predictioncenter.org/casp12/zscores_final.cgi

A Snap Shot of CASP12 Official Result Web Page

BhageerathH+ in CASP12: A closer look

BhageerathH+ Prediction for Low Resolution Model Structures (under 7Å rmsd)



Total Native Domain Structures Available: 80

Maximum Number of Targets Predicted Under 7Å rmsd By Any Individual Server: 38

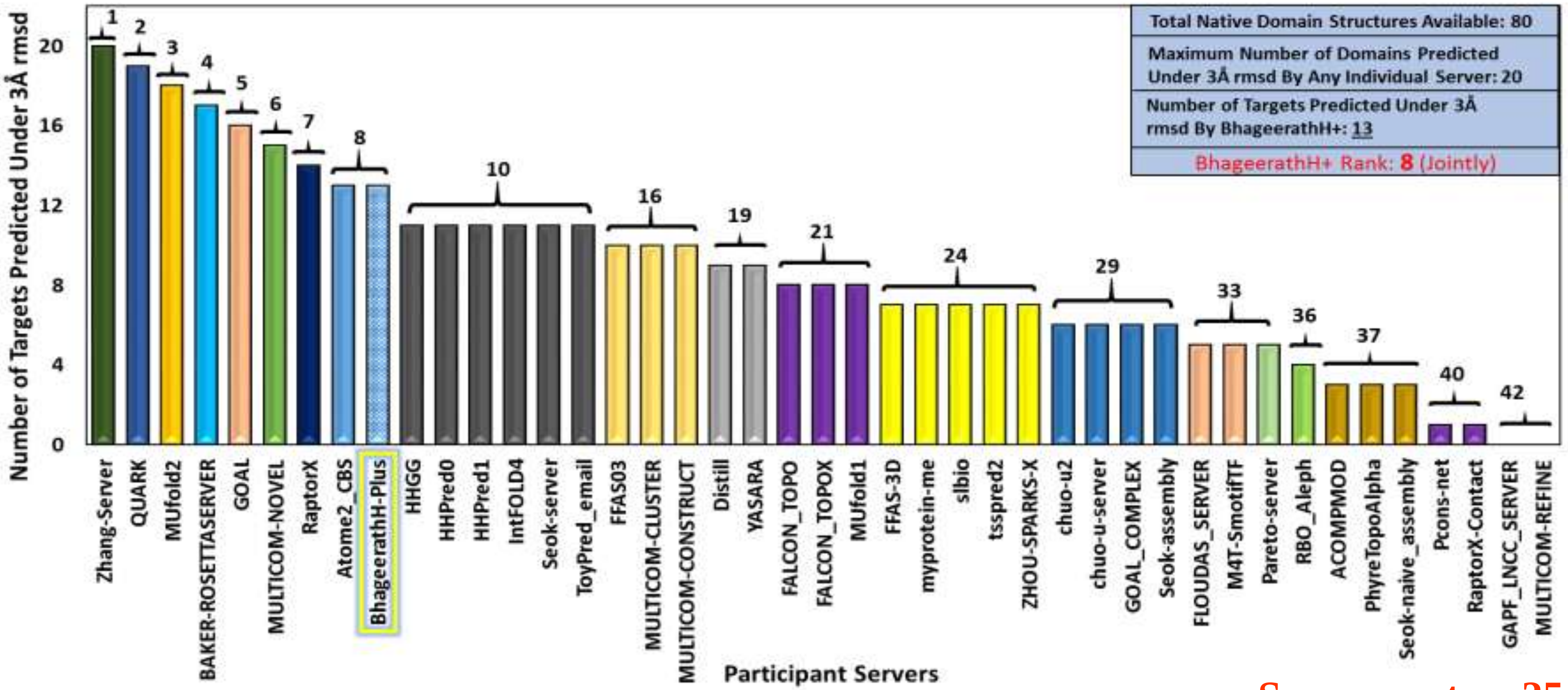
Number of Targets Predicted Under 7Å rmsd By BhageerathH+: 38

BhageerathH+ Rank: **1** (Jointly)

"Low resolution" models
Success rate ~ 48%

BhageerathH+ in CASP12

BhageerathH+ Prediction for High Resolution Model Structures (under 3Å rmsd)



Total Native Domain Structures Available: 80

Maximum Number of Domains Predicted Under 3Å rmsd By Any Individual Server: 20

Number of Targets Predicted Under 3Å rmsd By BhageerathH+: 13

BhageerathH+ Rank: **8** (Jointly)

Work in progress: Bhageerath models need refinement.

Success rate ~ 25%

To initiate computer aided structure based drug discovery..one needs < 3 Å RMSD structures ➔

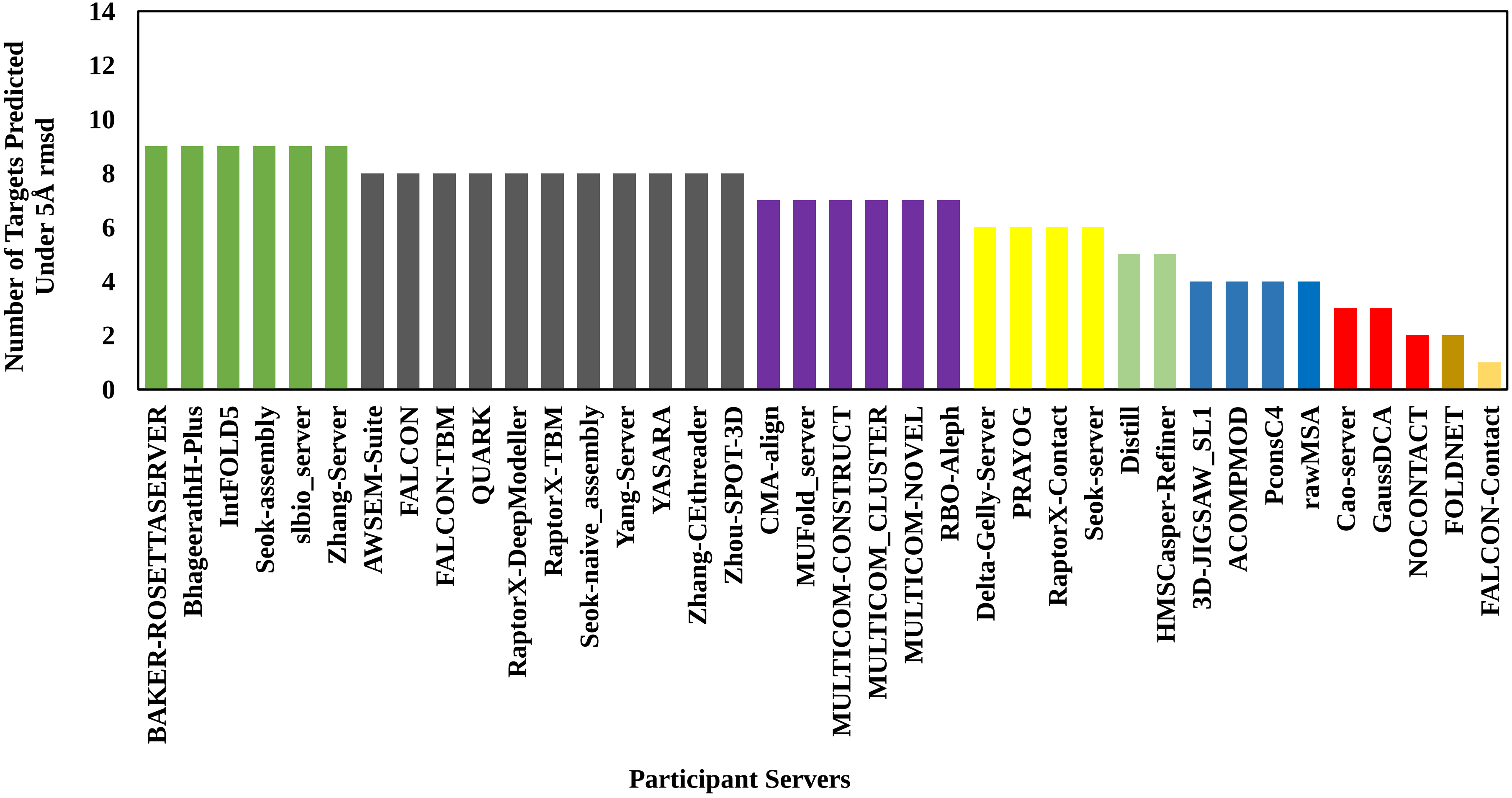
Let us innovate and improve the methods for structure generation and refinement !

BhageerathH+ in CASP12

Assessment of *BhageerathH*$^+$ on CASP12 Targets in terms of GDT-TS
(Rank in parenthesis)

Server Name	GDT ≥ 25	GDT ≥ 50	GDT ≥ 75	Server Name	GDT ≥ 25	GDT ≥ 50	GDT ≥ 75
Zhang-Server	33 (1)	23 (1)	8 (7)	ROSETTA	25 (23)	19 (16)	6 (16)
QUARK	32 (2)	22 (6)	9 (2)	FFAS-3D	25 (23)	18 (19)	6 (16)
BhageerathH+	32 (2)	23 (1)	8 (7)	chuo-u-server	25 (23)	14 (26)	3 (26)
MULTICOM-CLUST	31 (4)	21 (10)	8 (7)	chuo-u2	25 (23)	14 (26)	3 (26)
MULTICOM-CONSTR	31 (4)	19 (16)	8 (7)	ZHOU-SPARKS-X	24 (27)	17 (23)	5 (20)
IntFOLD4	31 (4)	23 (1)	7 (12)	Pareto-server	22 (28)	12 (29)	3 (26)
RaptorX	31 (4)	23 (1)	7 (12)	PhyreTopoAlpha	21 (29)	9 (31)	3 (26)
ToyPred_email	31 (4)	22 (6)	7 (12)	RaptorX-Contact	21 (29)	1 (41)	0 (40)
MULTICOM-NOVEL	30 (9)	21 (10)	10 (1)	MUfold1	17 (31)	13 (28)	4 (24)
GOAL	30 (9)	23 (1)	9 (2)	Pcons-net	17 (31)	4 (37)	0 (40)
Seok-server	30 (9)	20 (15)	7 (12)	YASARA	16 (33)	11 (30)	3 (26)
HHGG	29 (12)	21 (10)	9 (2)	Atome2_CBS	15 (34)	8 (32)	6 (16)
HHPred0	29 (12)	21 (10)	9 (2)	Seok-assembly	10 (35)	6 (34)	2 (33)
HHPred1	29 (12)	21 (10)	9 (2)	MUfold2	8 (36)	7 (33)	2 (33)
Distill	29 (12)	18 (19)	8 (7)	ACOMPMOD	8 (36)	5 (35)	2 (33)
RBO_Aleph	29 (12)	16 (24)	4 (24)	Seok-naive_assembly	7 (38)	3 (38)	1 (39)
FALCON_TOPO	28 (17)	22 (6)	5 (20)	MULTICOM-REFINE	7 (38)	1 (41)	0 (40)
FALCON_TOPOX	28 (17)	22 (10)	5 (20)	GOAL_COMPLEX	6 (40)	3 (38)	3 (26)
myprotein-me	28 (17)	18(19)	3 (26)	GAPF_LNCC	6 (40)	0 (43)	0 (40)
slbio	28 (17)	18 (19)	6 (16)	M4T-SmotifTF	5 (42)	5 (35)	2 (33)
tsspred2	28 (17)	19 (16)	5 (20)	FFAS03	3 (43)	3 (38)	2 (33)
FLOUDAS	26 (22)	16 (24)	2 (33)				

“Bhageerath-H+” in CASP13
May-July, 2018



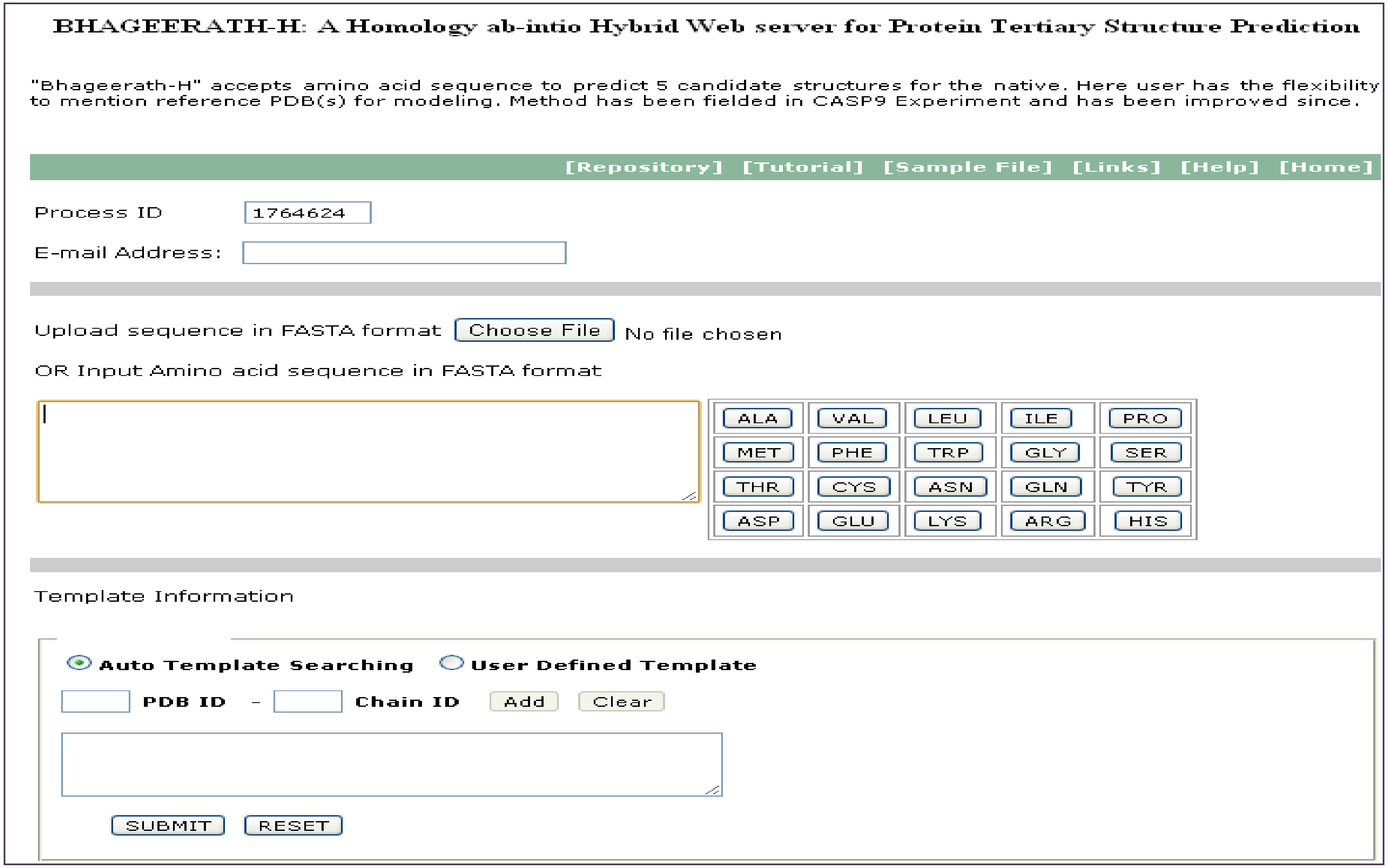
# Groups: 207
# Servers: 87
# From India: 1
(Only IITD)

BhageerathH^{+} methodology with latest improvements was fielded in the recently concluded CASP13 experiment (1st May – 17th July, 2018). _BhageerathH^{+}_ succeeded in predicting 9 targets out of 15 (whose experimental structures (native information) is released in PDB) with rmsds under 5Å and this is as good as the performance of any other participant server.

Bhageerath-H: A Freely Accessible Web Server

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

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



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.

Process ID 1764624

E-mail Address:

Upload sequence in FASTA format

No file chosen

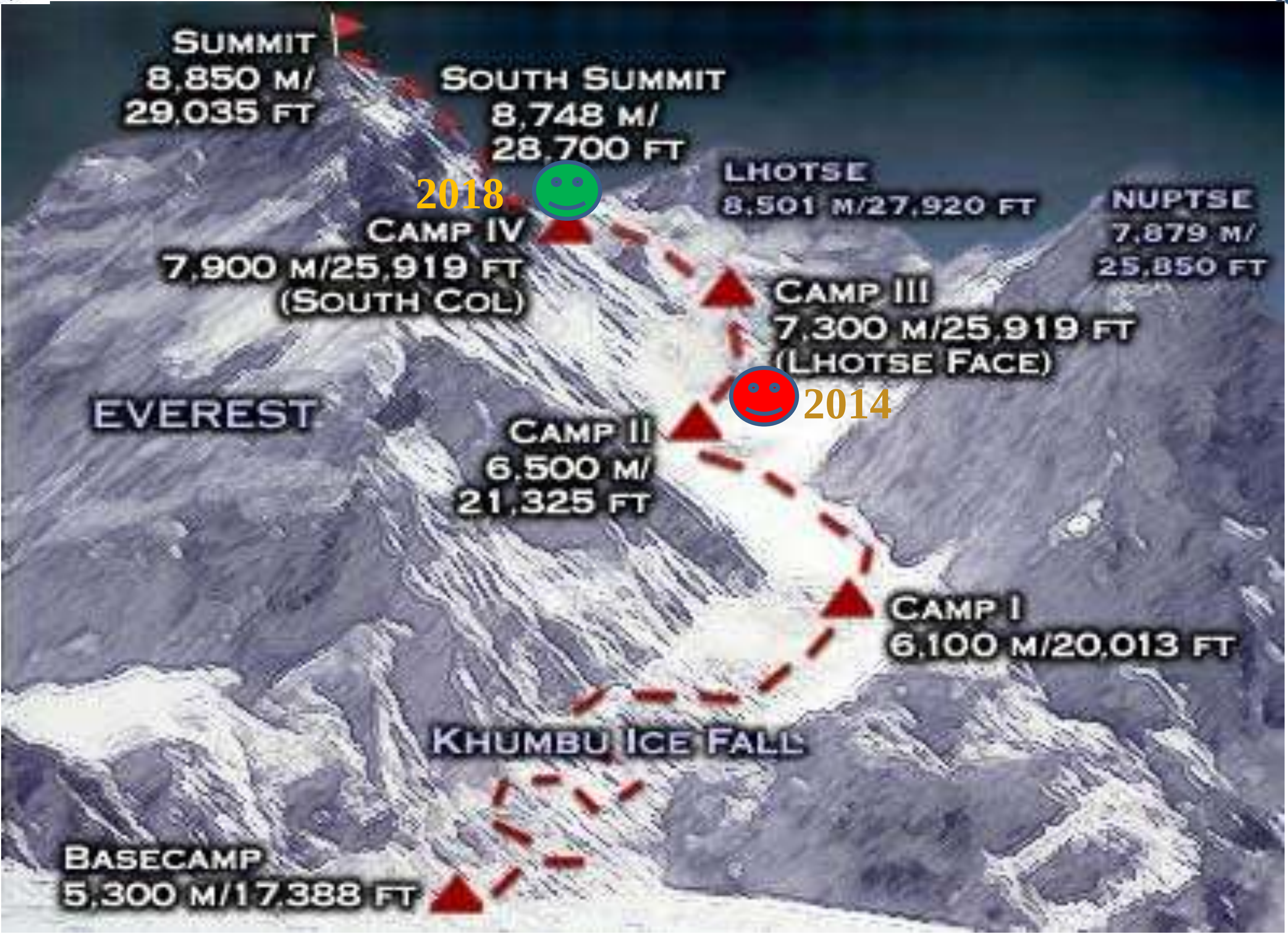
OR Input Amino acid sequence in FASTA format

Template Information

● Auto Template Searching ○ User Defined Template

PDB ID - Chain ID

B Jayaram, Priyanka Dhingra, Avinash Mishra, Rahul Kaushik, Goutam Mukherjee, Ankita Singh and Shashank Shekhar, "Bhageerath-H: A homology/ ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins", *BMC Bioinformatics, 2014,* Volume 15 Suppl 12, S8.



Protein Structure Prediction: Path to Mt. Everest is worked out! Near CAMP IV !!

Protein Folding – Unsolved
Protein-DNA recognition – Unsolved
Protein-RNA recognition – Unsolved
Protein – Protein recognition – Unsolved

Non-covalent recognition beyond hydrogen bonds (W&C; LP) is an unsolved problem

Are we looking at Proteins right? Here is the conventional wisdom.

Physico-chemical properties of amino acids

***Text book classifications of amino acids*.** The side chains of the proteins differ in size, shape, charge, hydrogen bonding capacity, and chemical reactivity.

They can be grouped as follows:

(a) Aliphatic side chains – Gly (G), Ala (A), Val (V), Leu (L), Ile (I) and Pro (P);

(b) Hydroxyl aliphatic side chains - Ser (S) and Thr (T);

(c) Aromatic side chains – Phe (F), Tyr (Y), and Trp (W);

(d) Basic side chains - Lys (K), and Arg (R) and His (H);

(e) Acidic side chains - Asp (D) and Glu (E);

(f) Amide side chains - Asn (N) and Gln (Q);

(g) Sulphur side chains - Cys (C) and Met (M).

(i) Charged amino acids: K, R, H, D, E

(ii) Polar amino acids: S, C, T, Y, N, Q, W

(iii)Non polar (hydrophobic) amino acids: G, A, V, I, L, F, P, M

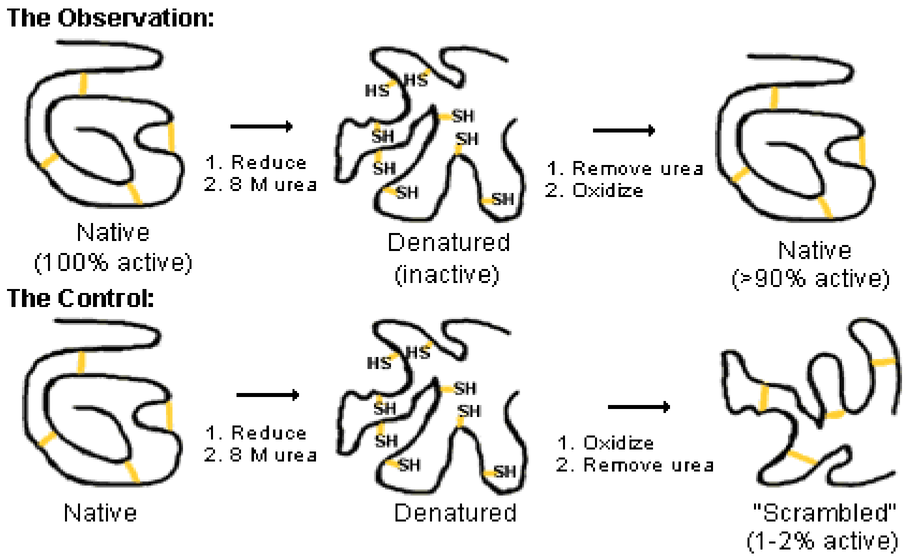
Other properties of amino acids.

Average masses, volumes and surface areas of each amino acid

1-letter code	**3-letter code**	**Chemical formula**	**Average (Daltons)**	**Residue Volume Å^3**	**Surface Area Å^2**
A	Ala	C_3H_5ON	71.0788	88.6	115
R	Arg	$C_6H_{12}ON_4$	156.1875	173.4	225
N	Asn	$C_4H_6O_2N_2$	114.1038	111.1	150
D	Asp	$C_4H_5O_3N$	115.0886	114.1	160
C	Cys	C_3H_5ONS	103.1388	108.5	135
E	Glu	$C_5H_7O_3N$	129.1155	138.4	190
Q	Gln	$C_5H_8O_2N_2$	128.1307	143.8	180
G	Gly	C_2H_3ON	57.0519	60.1	75
H	His	$C_6H_7ON_3$	137.1411	153.2	195
I	Ile	$C_6H_{11}ON$	113.1594	166.7	175
L	Leu	$C_6H_{11}ON$	113.1594	166.7	170
K	Lys	$C_6H_{12}ON_2$	128.1741	168.6	200
M	Met	C_5H_9ONS	131.1926	162.9	185
F	Phe	C_9H_9ON	147.1766	189.9	210
P	Pro	C_5H_7ON	97.1167	112.7	145
S	Ser	$C_3H_5O_2N$	87.0782	89.0	115
T	Thr	$C_4H_7O_2N$	101.1051	116.1	140
W	Trp	$C_{11}H_{10}ON_2$	186.2132	227.8	255
Y	Tyr	$C_9H_9O_2N$	163.1760	193.6	230
V	Val	C_5H_9ON	99.1326	140.0	155

What we do know so far? Sequence to structure….Yes

Anfinsen's experiments /results on RNAase A



Anfinsen proposed his "Thermodynamic Hypothesis", which states that there is sufficient information contained in the protein sequence to guarantee correct folding from any of a large number of unfolded states.

The Nobel Prize in Chemistry 1972

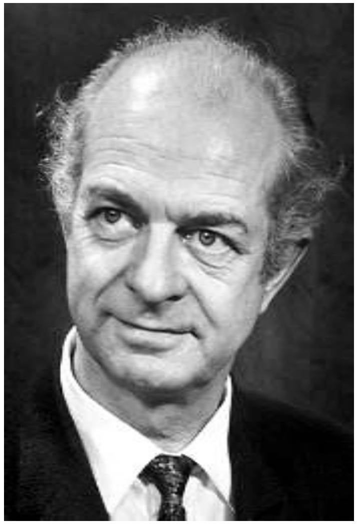
to Christian B. Anfinsen *"for his work on ribonuclease, especially concerning the connection between the amino acid sequence and the biologically active conformation"*, the other half jointly to Stanford Moore and William H. Stein *"for their contribution to the understanding of the connection between chemical structure and catalytic activity of the active centre of the ribonuclease molecule"*.

C. B. Anfinsen

S. Moore

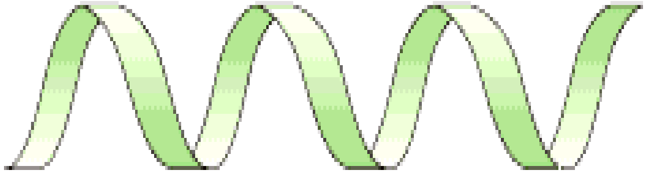
W. H. Stein

Secondary structure formation -- Yes

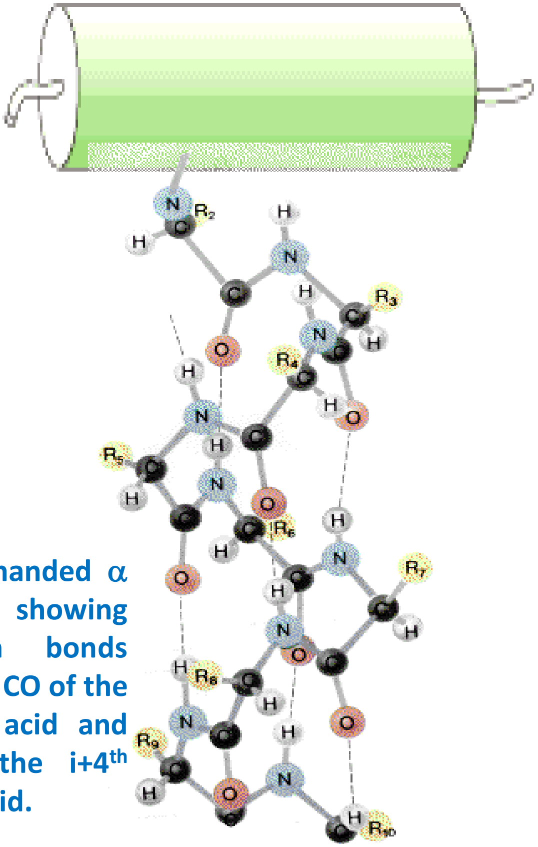


Linus Pauling
(1901-1994)

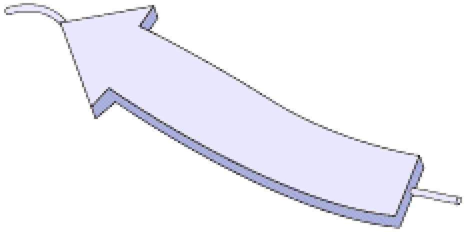
Nobel Prize for Chemistry in 1954 & Nobel Prize for Peace in 1962



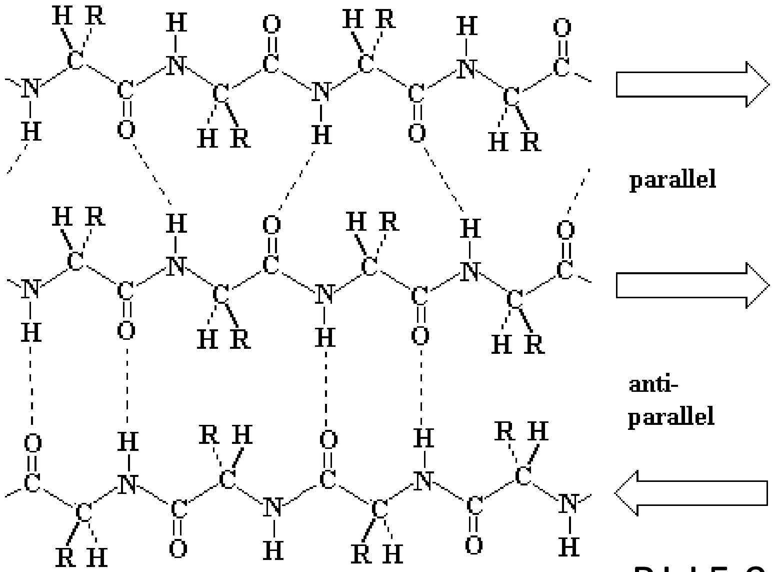
Alpha helix representations



A right handed α – helix showing hydrogen bonds between CO of the i^{th} amino acid and NH of the $i+4^{th}$ amino acid.



Beta sheet representations



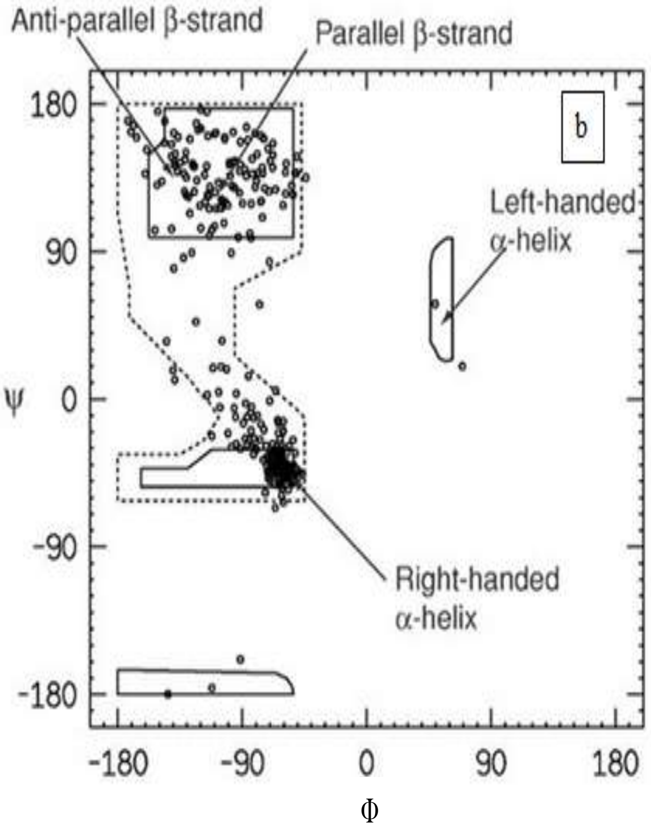
Hydrophobicity … Yes;
Only ~ 15% of the phi, psi space is populated… Yes

Walter Kauzmann (1916-2009)

Oil and water don't mix. "Conventional thinking today: Hydrophobic (nonpolar) residues in (away from water) and hydrophilic (polar) residues out (facing solvent water) in the structure of a protein"

G.N.Ramachandran (1922-2001)

The Ramachandran Plot



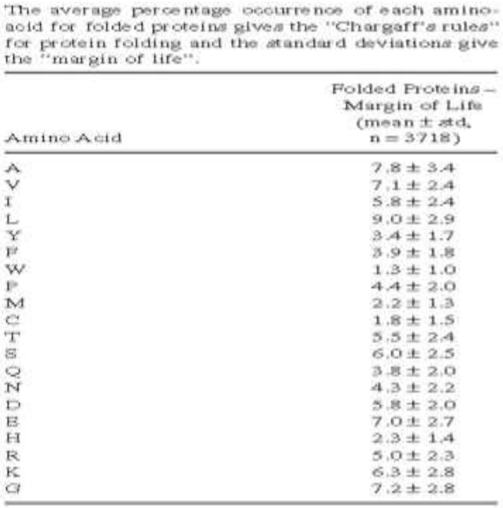
White space in the map above is the sterically disallowed region

Prevailing concepts on proteins are not sufficient to build the tertiary structures of proteins from their sequences

In search of rules of protein folding

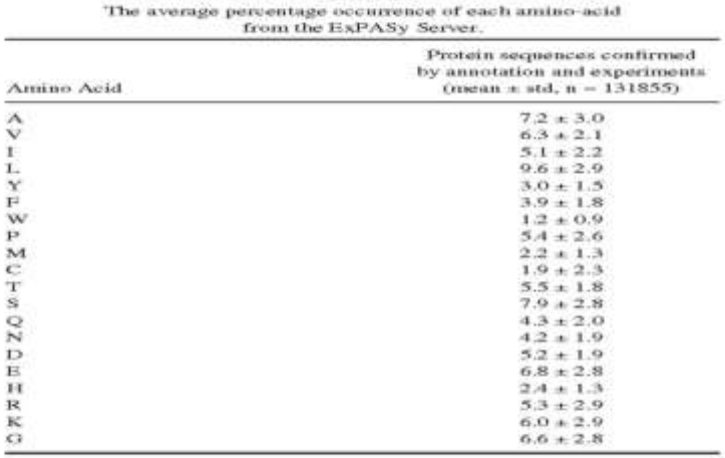
Margin of Life: Amino acid compositions in proteins have a tight distribution

Stoichiometry hypothesis



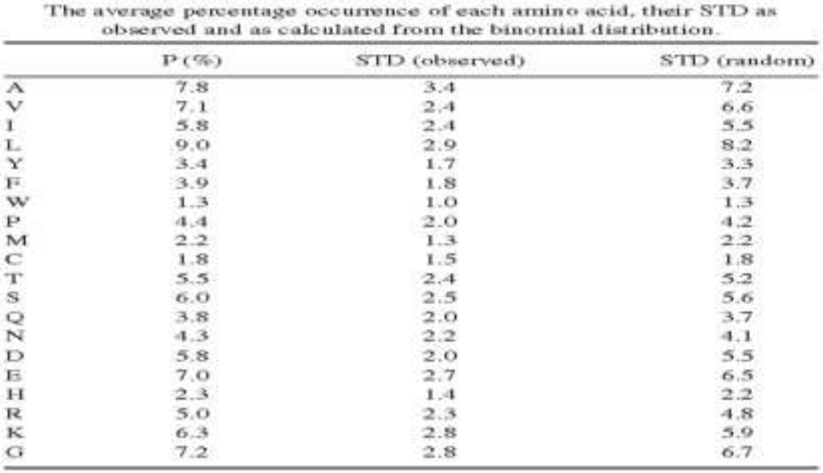
The average percentage occurrence of each amino-acid for folded proteins gives the "Chargaff's rules" for protein folding and the standard deviations give the "margin of life".

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



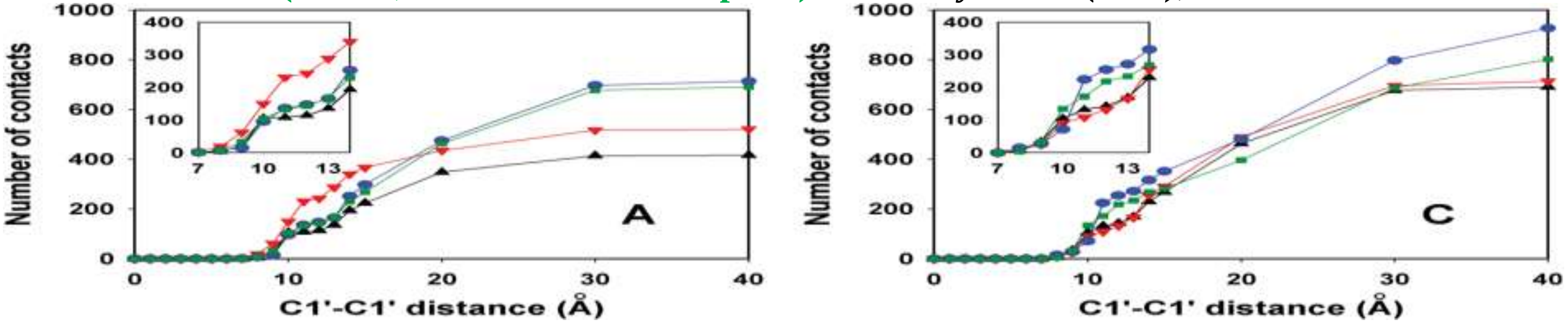
The average percentage occurrence of each amino acid, their STD as observed and as calculated from the binomial distribution.

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

Neighborhood Analysis (C1’) of 18 DNA Double Helices

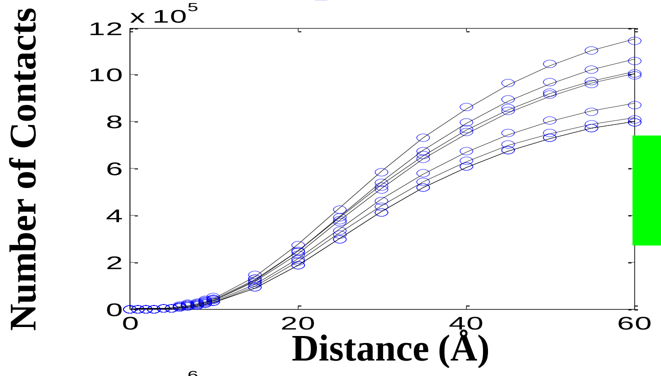
(85 A-T, 113 C-G nucleotide pairs) *Galzitskaya et al. (2011), JBSD*



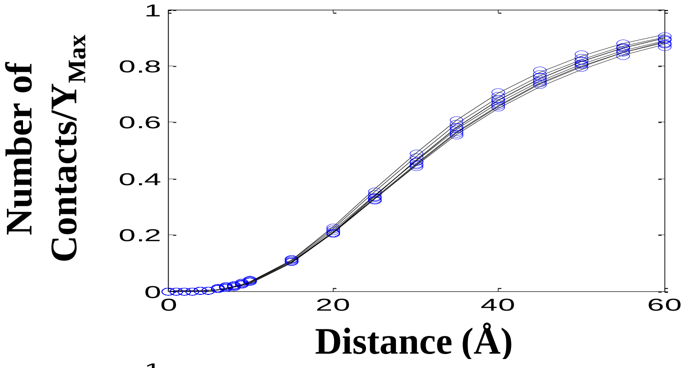
In search of rules of protein folding: **Neighborhood Analysis: 3718 Protein Crystal Structures:**

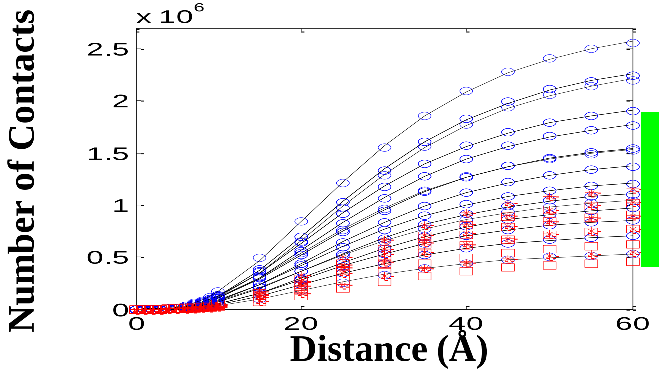
Unlike nucleic acid bases which show preferential interactions, amino acids show secularity!

Cα spatial distributions show universality: $Y = Y_{Max}(1-e^{-kX})^n$

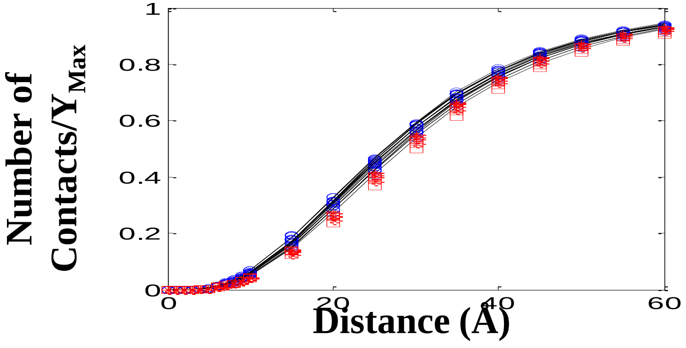


R, K vs
R, K, D, E





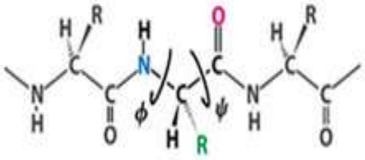
A, V, I, L vs
each other &
N, Q (red)

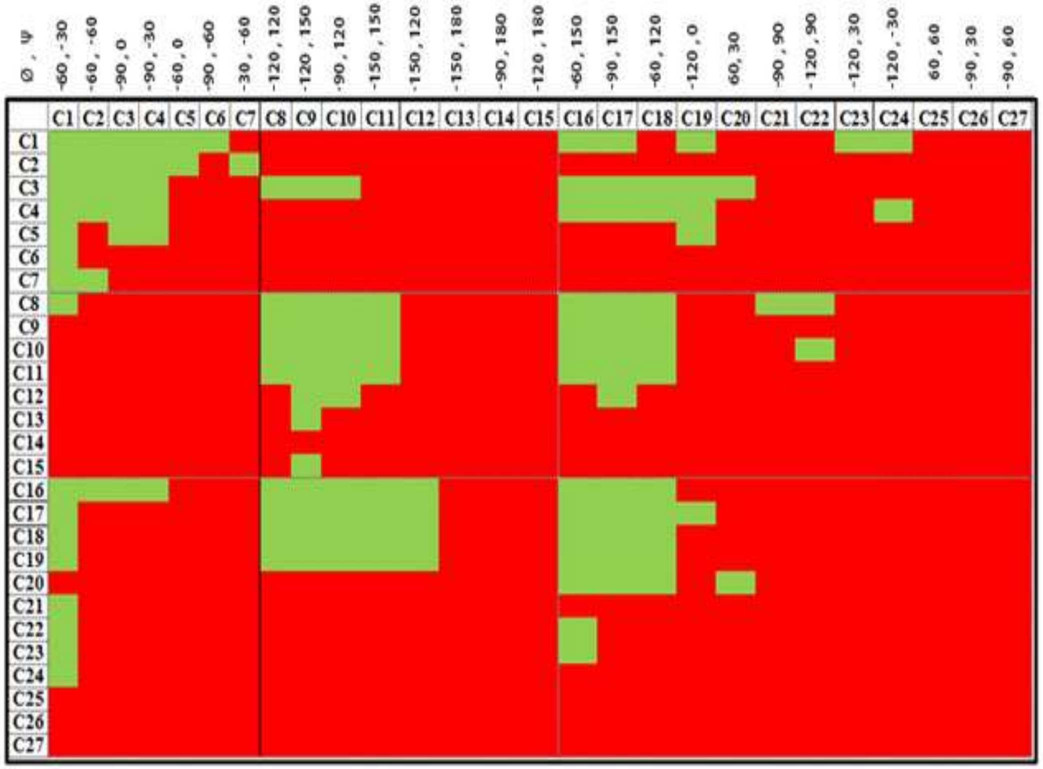


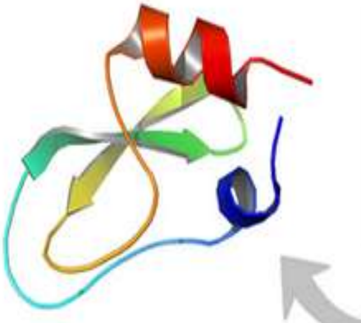
A. Mittal, B. Jayaram et al. *J. Biomol. Struc. Dyn.*, 2010, *Vol. 28 (2)*, 133-142; 2011, *28(4)*, 443 -454; 2011, *28(4)*, 669-674.

From Ramachandran Maps to Tertiary Structures of Proteins : The missing Link

Key: Construct higher order Ramachandran maps

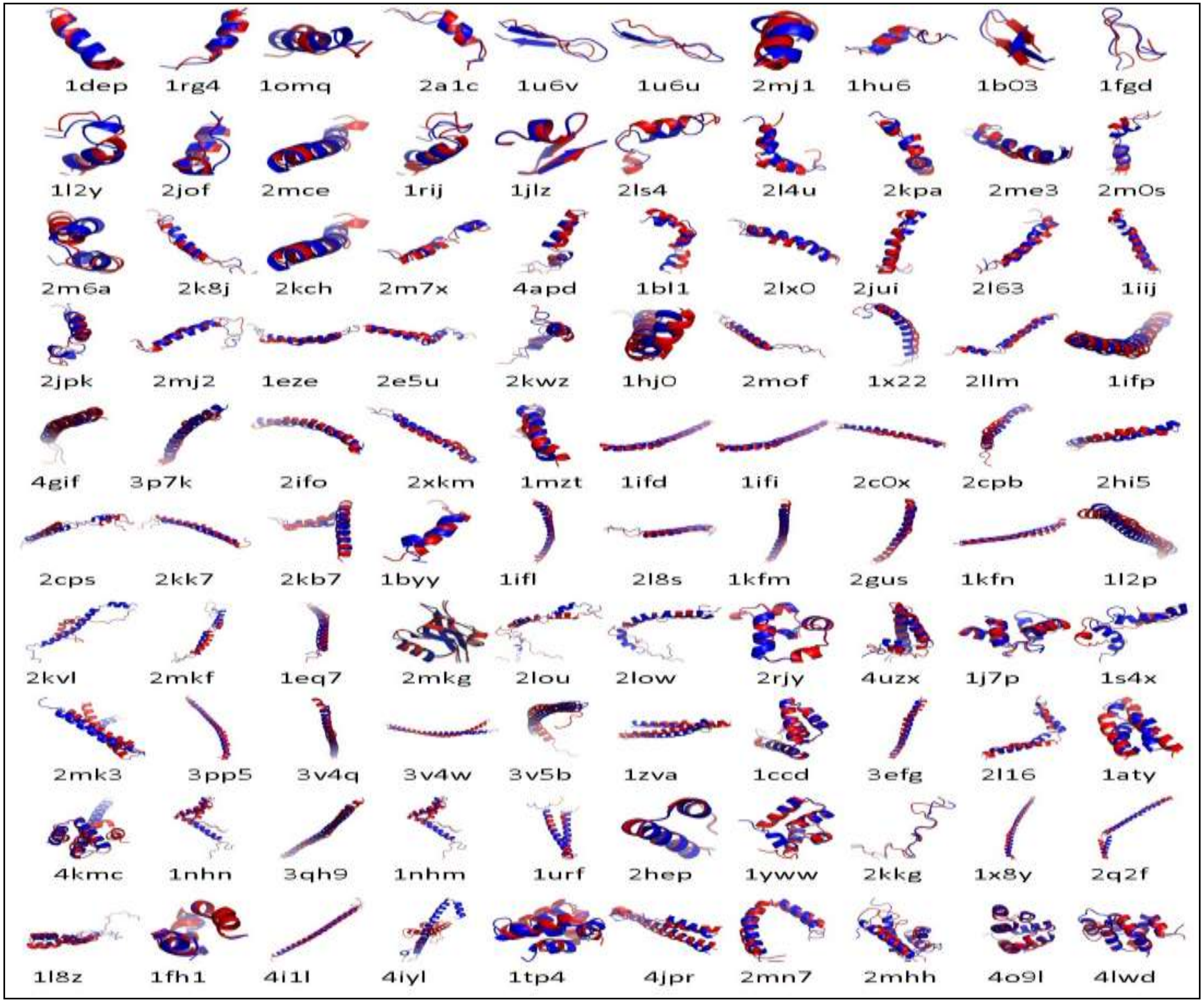






Debarati DasGupta, Rahul Kaushik, and B. Jayaram "From Ramachandran Maps to Tertiary Structures of Proteins", *J. Phys. Chem. B, 2015,* 119(34), 11136-11145. DOI: 10.1021/acs.jpcb.5b02999

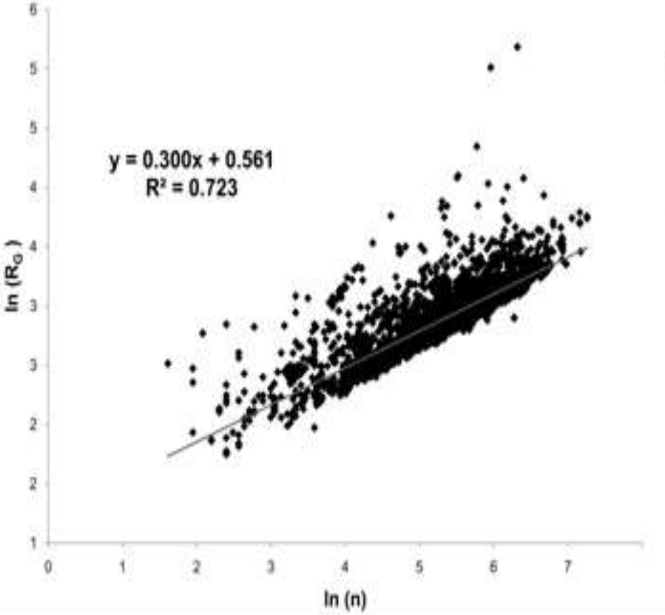
Debarati Das Gupta, Rahul Kaushik, B. Jayaram, "Protein folding is a convergent problem!", *Biochemical and Biophysical Research Communications, 2016, 480 (4)*, 741-44.



In 90 out of 100 cases a structure within 5 Å from native is generated

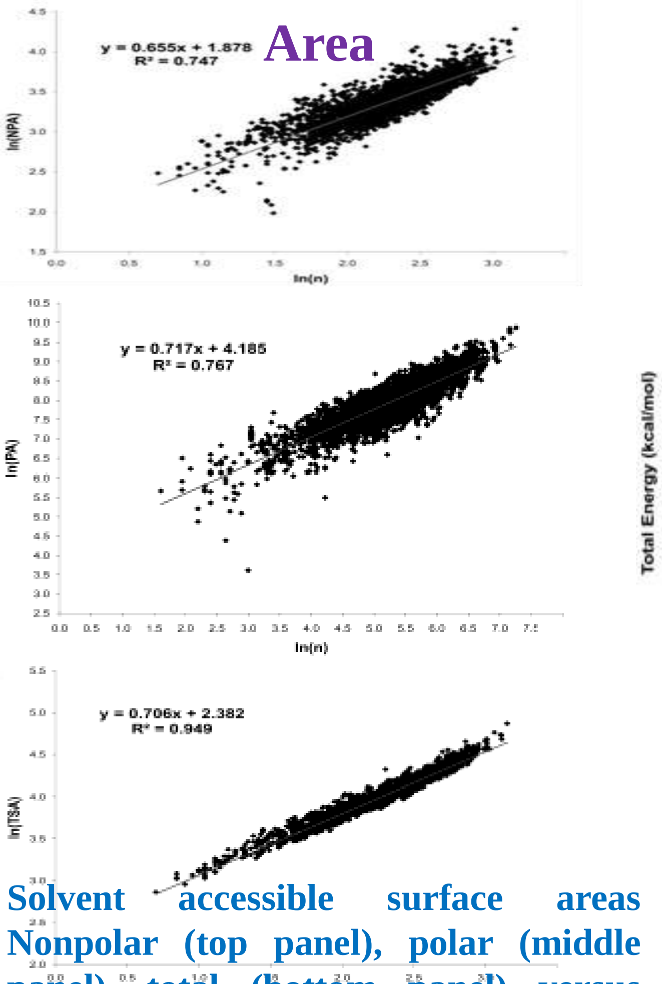
B. Jayaram, Aditya Mittal, Avinash Mishra, Chanchal Acharya, Garima Khandelwal "Universalities in Protein Tertiary Structures: Some New Concepts", in *Biomolecular Forms and Functions*, 2013, World Scientific Publishing Co. Pte. Ltd., Singapore, Eds; Manju Bansal & N. Srinivasan, pp 210-219.

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.

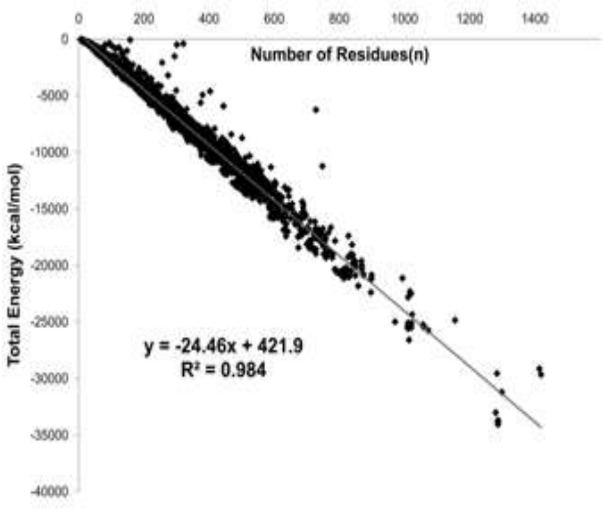
Area



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 whether residues are polar or nonpolar.

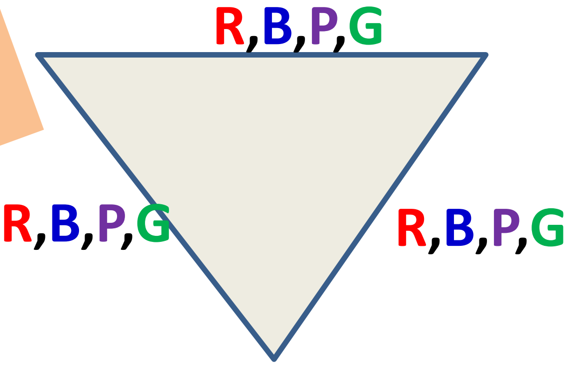
Energy



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

An invariant energy/residue metric appears to exist.

Let us reexamine the language of amino acids. May be we are missing something!



With 4 distinct colours to paint the 3 edges of a triangle, 64 coloured triangles are possible. By virtue of the symmetries of the triangle, only 20 of these are unique.

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

What has triangular symmetries got to do with amino acids?

There must be a chemical logic behind the evolution of 20 amino acids.

Let us hypothesize the following:

***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.

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.2008.2135.1 200

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.

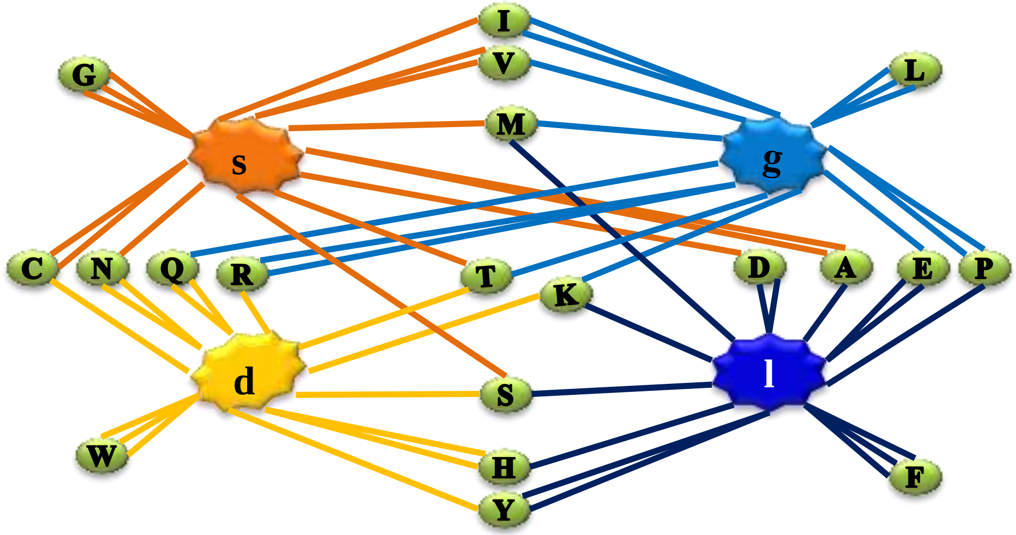
The most likely candidate for property (IV): Absence of branching. Linearity of the side chains / non-occurrence of bidentate forks with terminal hydrogens in the side chains. (a) This pools together 10 amino acids in the set {A, D, E, F, H, K, M, P, S, Y}. Side chains with single rings are treated as 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.

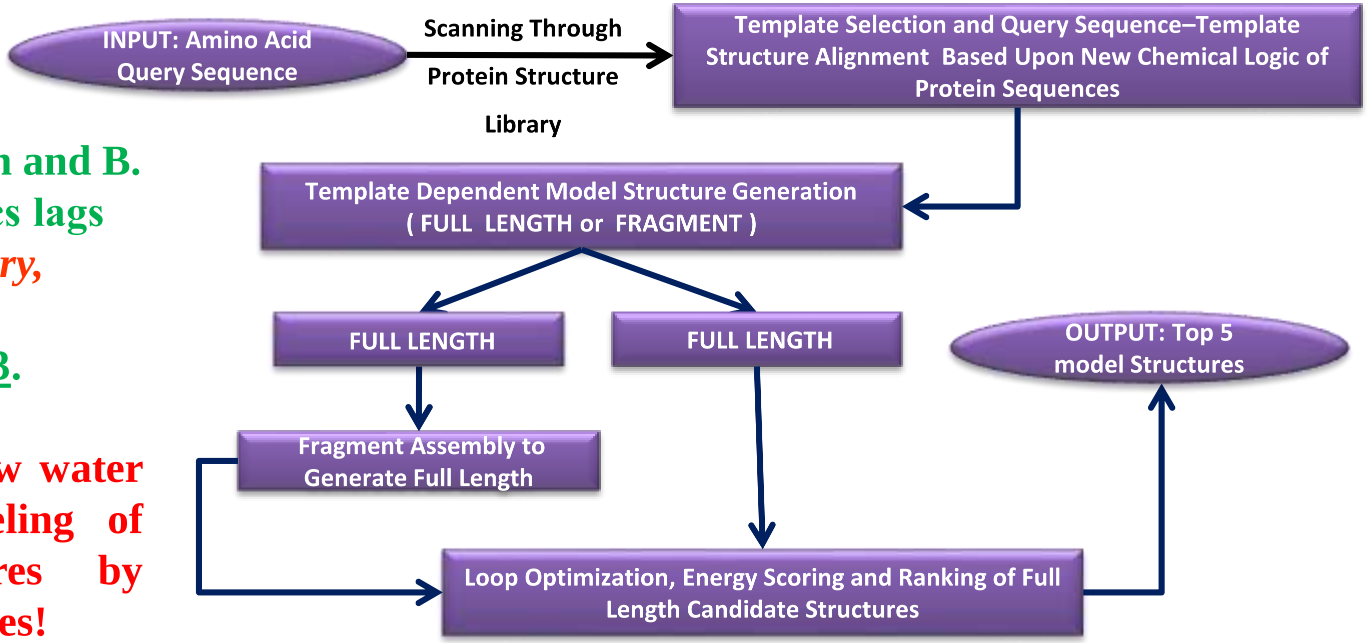
The 20 amino acids and some unique chemical properties of their side chains

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)	A new chemical logic based identities of amino acids
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	$g_0d_1s_0l_2$

Time to re-examine molecular recognition events involving proteins with a new chemical logic of AAs!



Amino acid chemical logic alignment based protein structure modeling



INPUT: Amino Acid Query Sequence

Scanning Through Protein Structure Library

Template Selection and Query Sequence–Template Structure Alignment Based Upon New Chemical Logic of Protein Sequences

Template Dependent Model Structure Generation (FULL LENGTH or FRAGMENT)

FULL LENGTH

FULL LENGTH

Fragment Assembly to Generate Full Length

Loop Optimization, Energy Scoring and Ranking of Full Length Candidate Structures

OUTPUT: Top 5 model Structures

Rahul Kaushik, Ankita Singh and B. Jayaram, "Where informatics lags chemistry leads", *Biochemistry, 2018*, 57(5): 503 – 505. DOI: 10.1021/acs.biochem.7b01073.

This methodology sets a new water mark for homology modeling of protein tertiary structures by unravelling hidden similarities!

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

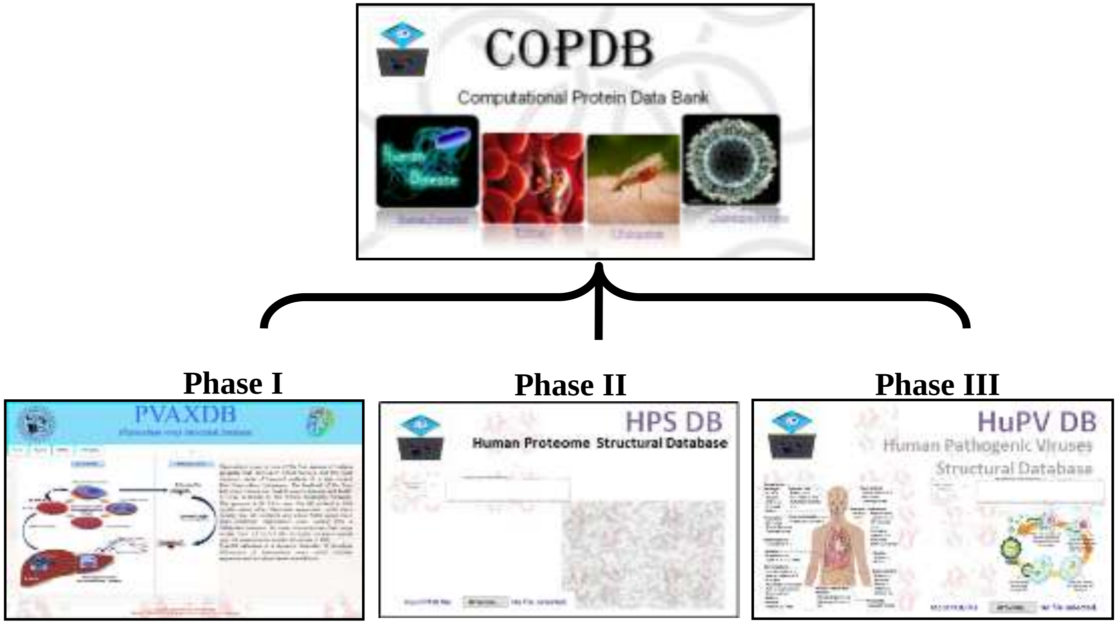
Suggested reading:

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

Ashutosh Shandilya, B. Jayaram, Sajeev Chacko, Indira Ghosh, "A Plausible mechanism for the antimalarial activity of artemisinin", *Scientific Reports, 2013, 3: 2513, doi:10.1038/srep02513* .

Computational Protein Databank (CoPDB)

A Comprehensive Organism Specific Proteome-wide Structural Repository of Soluble Proteins



Phase I

Phase II

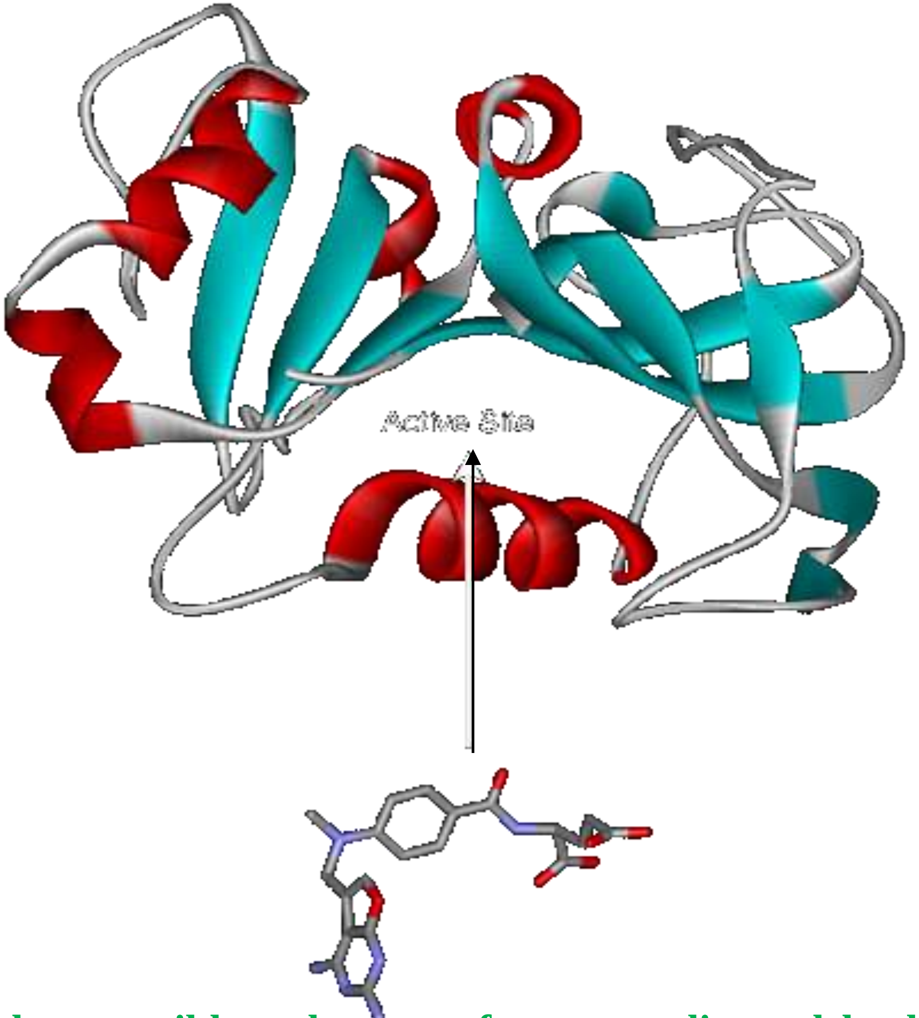
Phase III

Ankita Singh, Rahul Kaushik, Himani Kuntal, B. Jayaram, "PvaxDB: A comprehensive structural repository of *Plasmodium vivax* proteome", *Database, 2018,* doi/10.1093/database/bay021/4938395.

Target Directed Lead Molecule Design

Sanjeevini

महावीर विक्रम बजरंगी । कुमति निवार सुमति के संगी ॥
कंचन बरन बिराज सुबेसा । कानन कुंडल कुंचित केसा ॥
www.santabanta.com

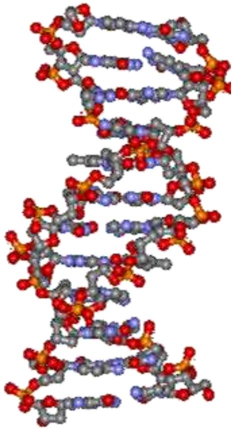


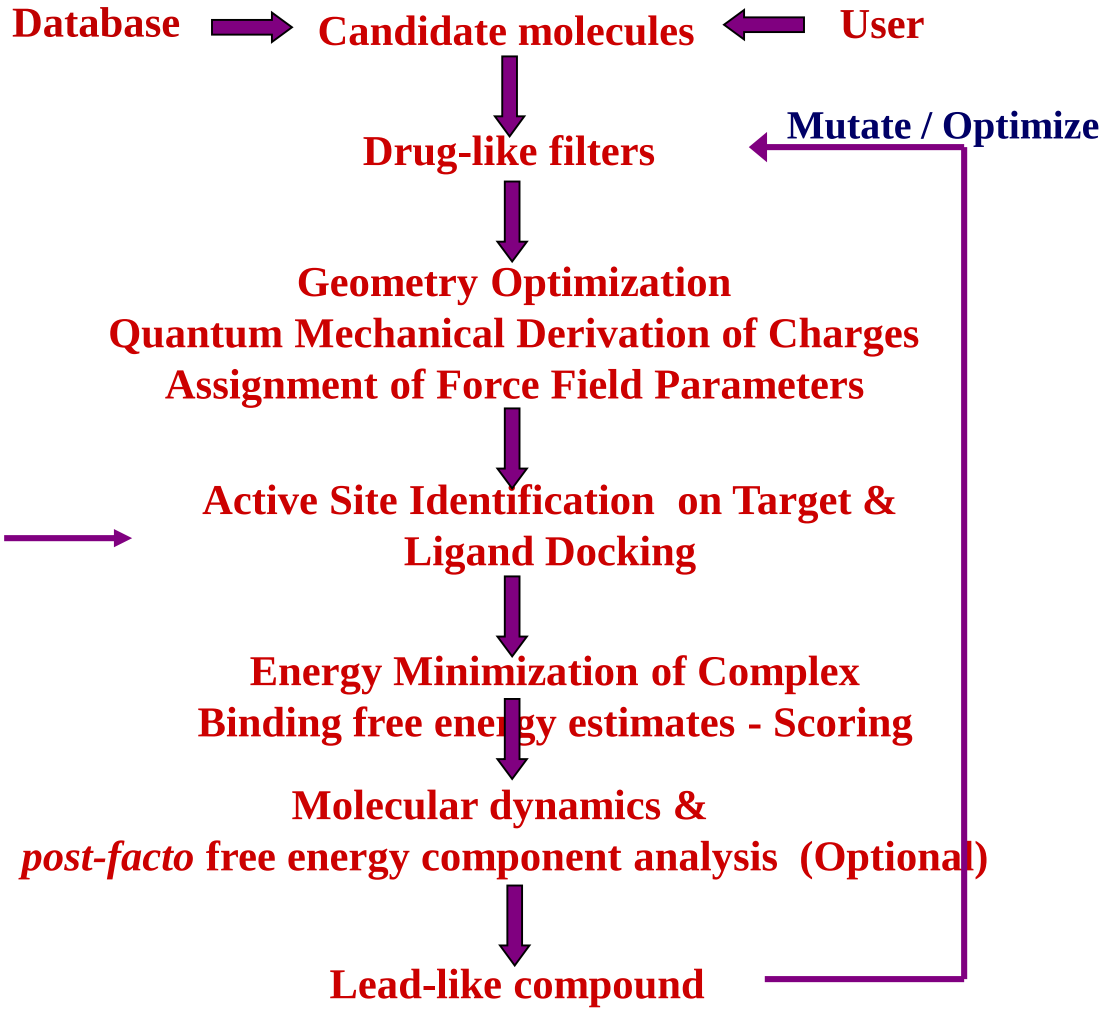
B. Jayaram, Tanya Singh et al., "Sanjeevini: a freely accessible web-server for target directed lead molecule discovery", *BMC Bioinformatics, 2012, 13,* S7. http://www.biomedcentral.com/1471-2105/13/S17/S7

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

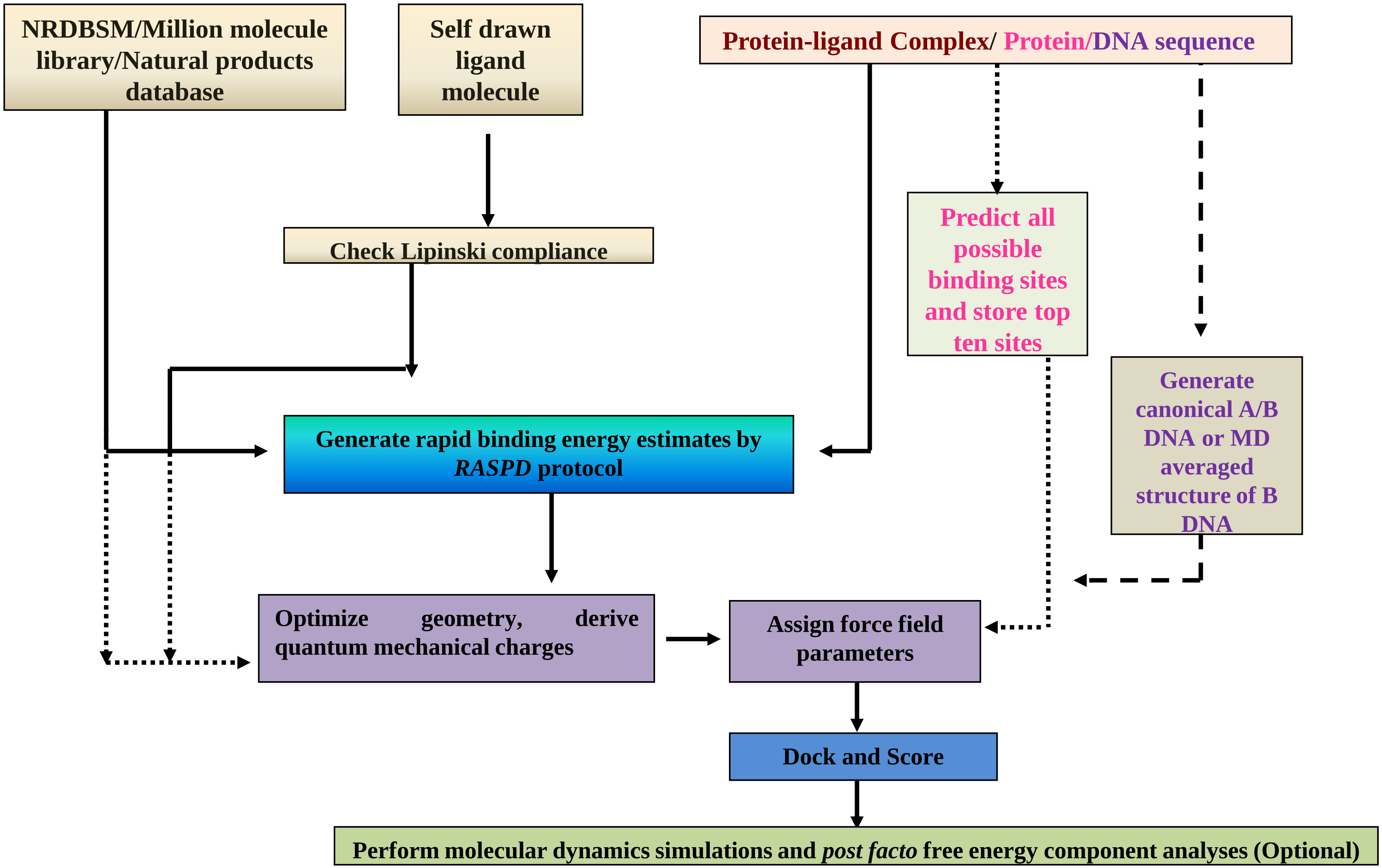
Active Site Identification on Target & Ligand Docking

Energy Minimization of Complex
Binding free energy estimates - Scoring

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

Lead-like compound

Jayaram et al., "Sanjeevini", ***Indian J. Chemistry-A. 2006,*** **45A, 1834-1837.**



NRDBSM/Million molecule library/Natural products database

Self drawn ligand molecule

Check Lipinski compliance

Protein-ligand Complex/ Protein/DNA sequence

Predict all possible binding sites and store top ten sites

Generate canonical A/B DNA or MD averaged structure of B DNA

Generate rapid binding energy estimates by *RASPD* protocol

Optimize geometry, derive quantum mechanical charges

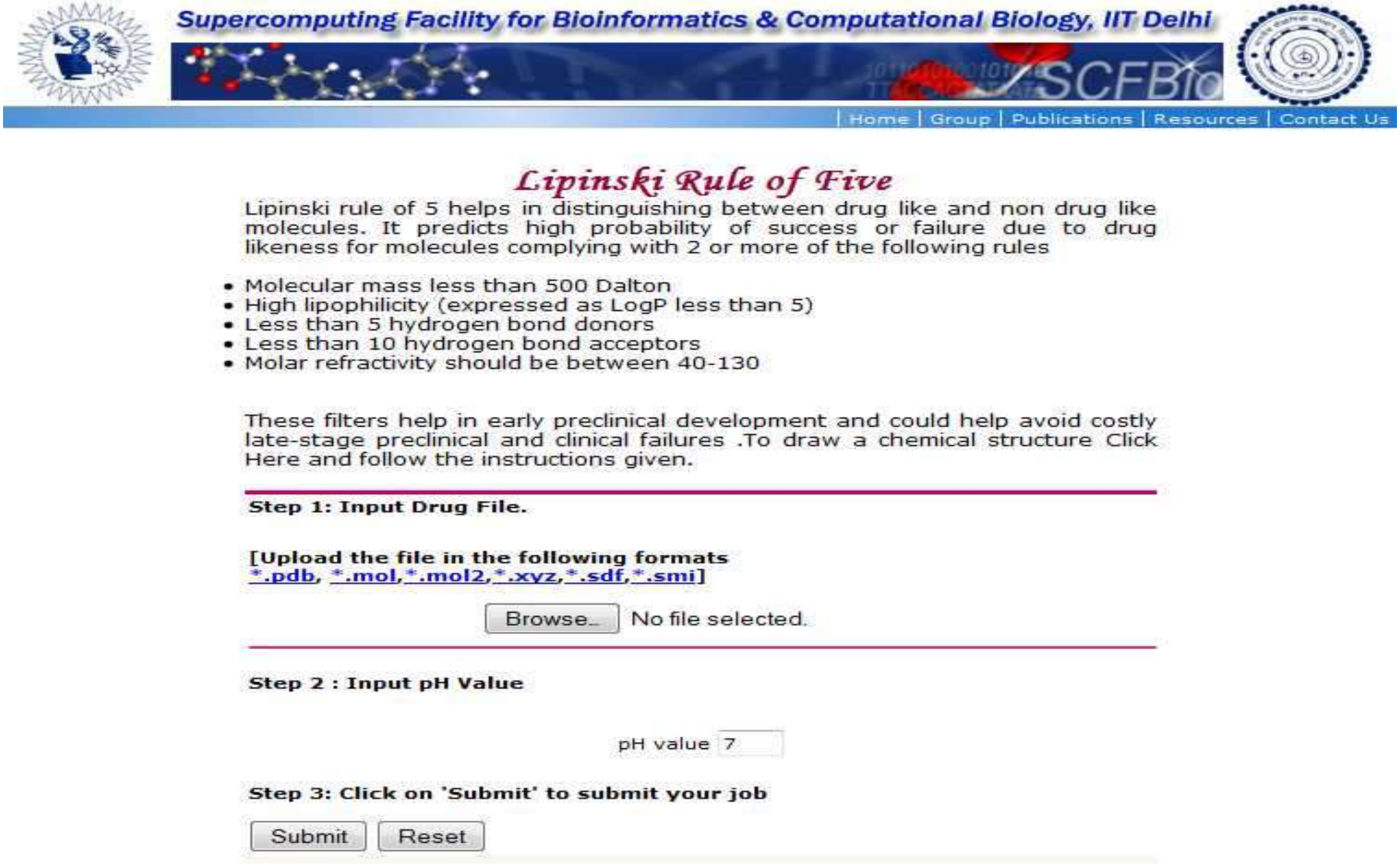
Assign force field parameters

Dock and Score

Perform molecular dynamics simulations and *post facto* free energy component analyses (Optional)

B. Jayaram, Tanya Singh, Goutam Mukherjee, Abhinav Mathur, Shashank Shekhar, Vandana Shekhar, "Sanjeevini: a freely accessible web-server for target directed lead molecule discovery", *BMC Bioinformatics 2012,* 13 (Suppl 17):S7.

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


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

Lipinski Rule of Five

Lipinski rule of 5 helps in distinguishing between drug like and non drug like molecules. It predicts high probability of success or failure due to drug likeness for molecules complying with 2 or more of the following rules

- Molecular mass less than 500 Dalton
- High lipophilicity (expressed as LogP less than 5)
- Less than 5 hydrogen bond donors
- Less than 10 hydrogen bond acceptors
- Molar refractivity should be between 40-130

These filters help in early preclinical development and could help avoid costly late-stage preclinical and clinical failures .To draw a chemical structure Click Here and follow the instructions given.

Step 1: Input Drug File.

[Upload the file in the following formats *.pdb, *.mol,*.mol2,*.xyz,*.sdf,*.smi]

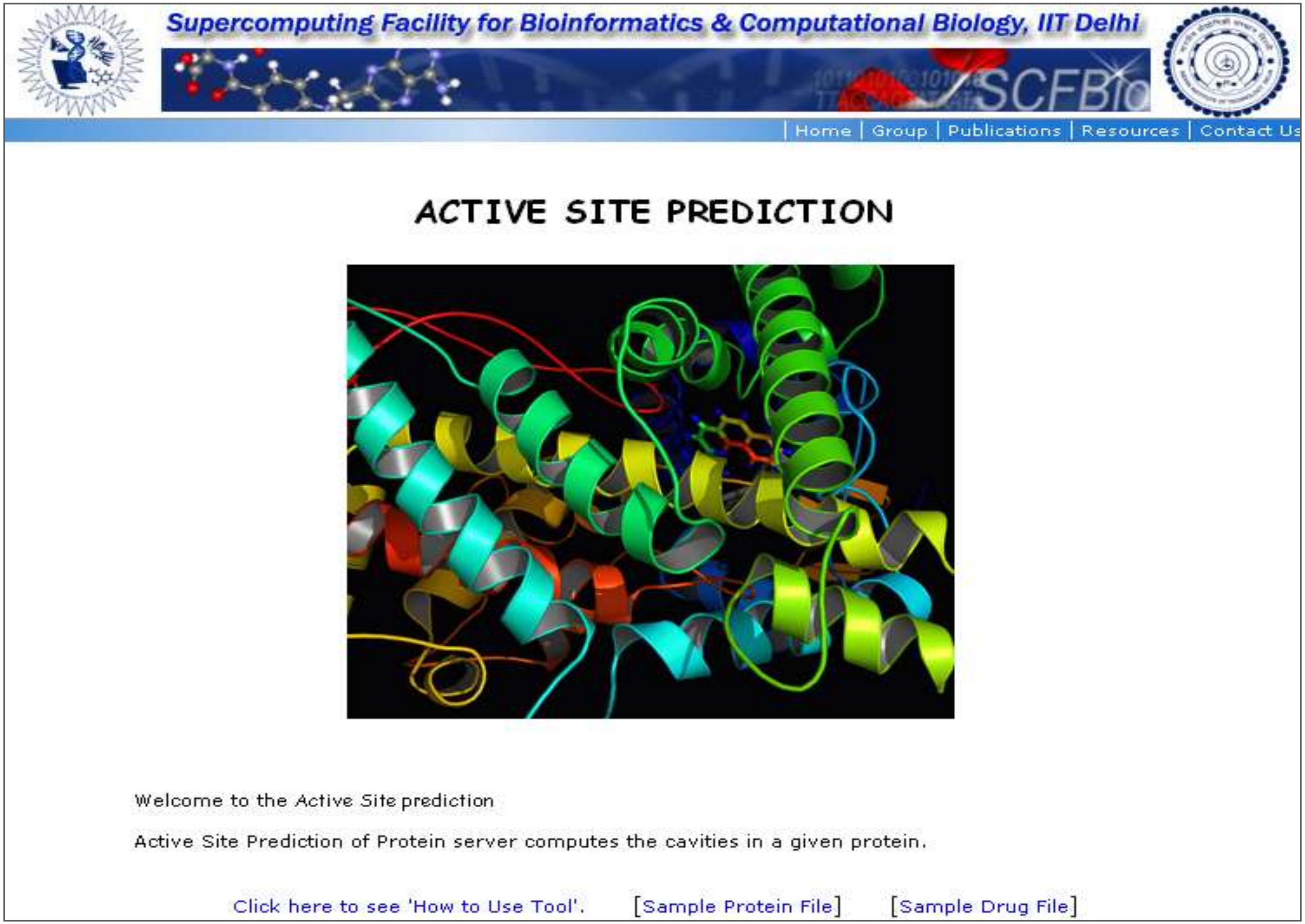
No file selected.

Step 2 : Input pH Value

pH value 7

Step 3: Click on 'Submit' to submit your job

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


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

ACTIVE SITE PREDICTION

Welcome to the Active Site prediction

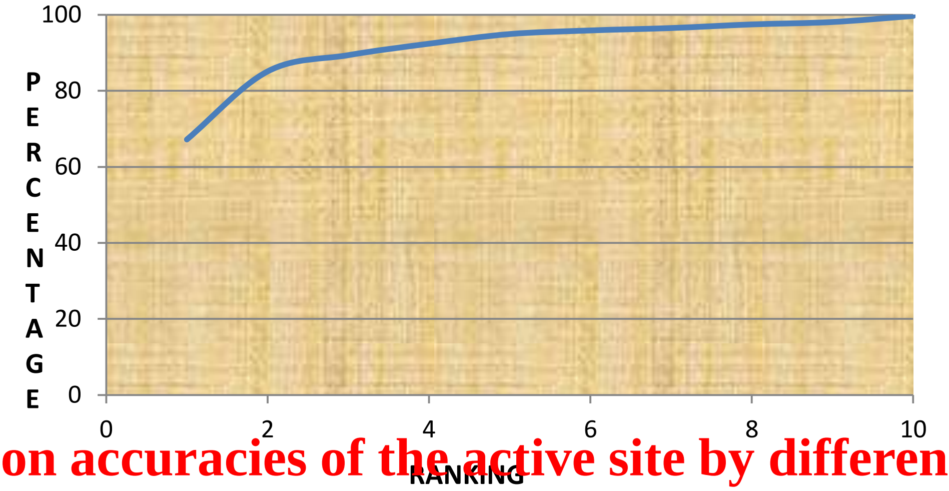
Active Site Prediction of Protein server computes the cavities in a given protein.

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

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

Rank of the cavity points vs. cumulative percentage prediction

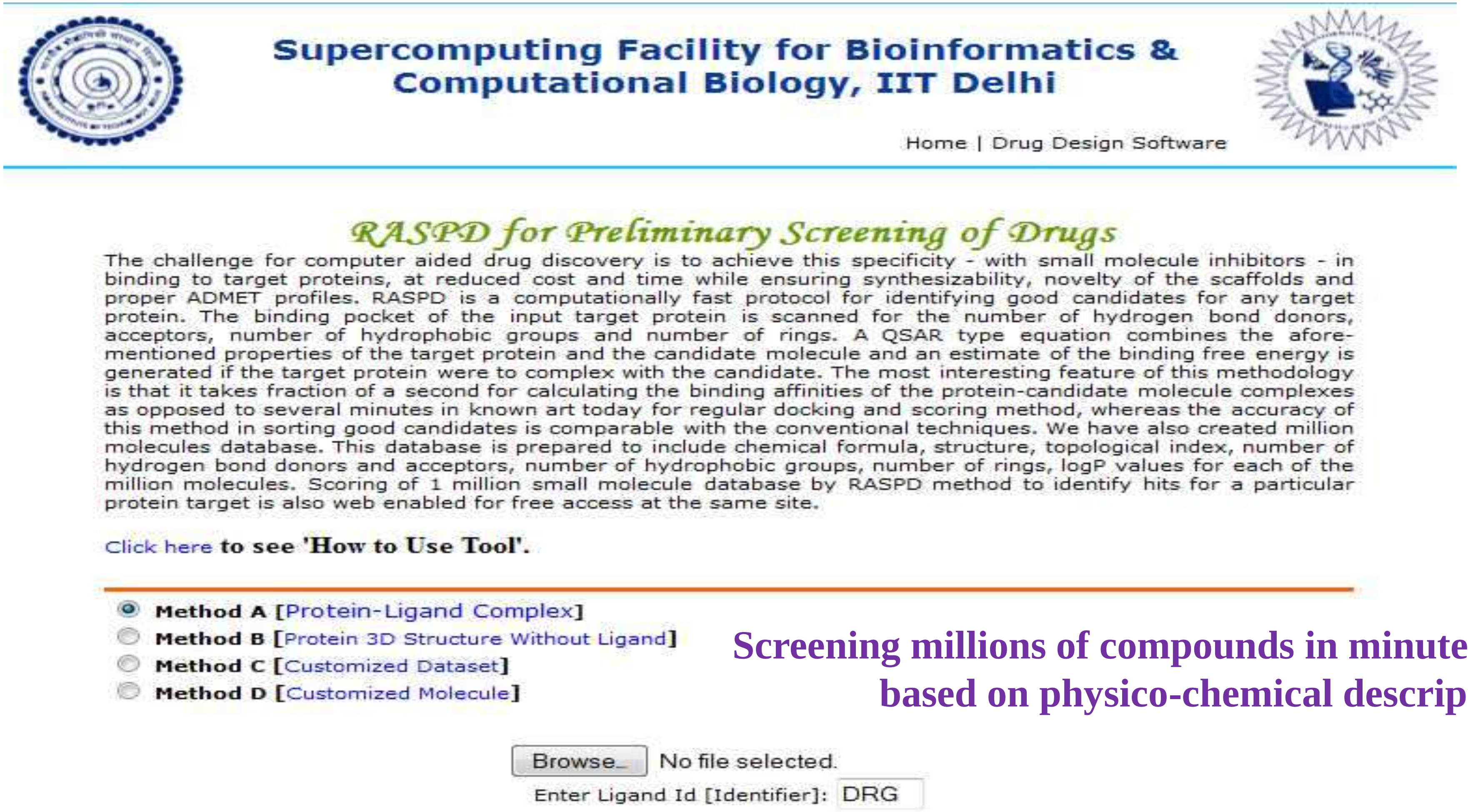
Top ten cavity points capture the active site 100 % of time in 640 protein targets



Prediction accuracies of the active site by different softwares

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

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



Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi

Home | Drug Design Software

RASPD for Preliminary Screening of Drugs

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

Click here to see 'How to Use Tool'.

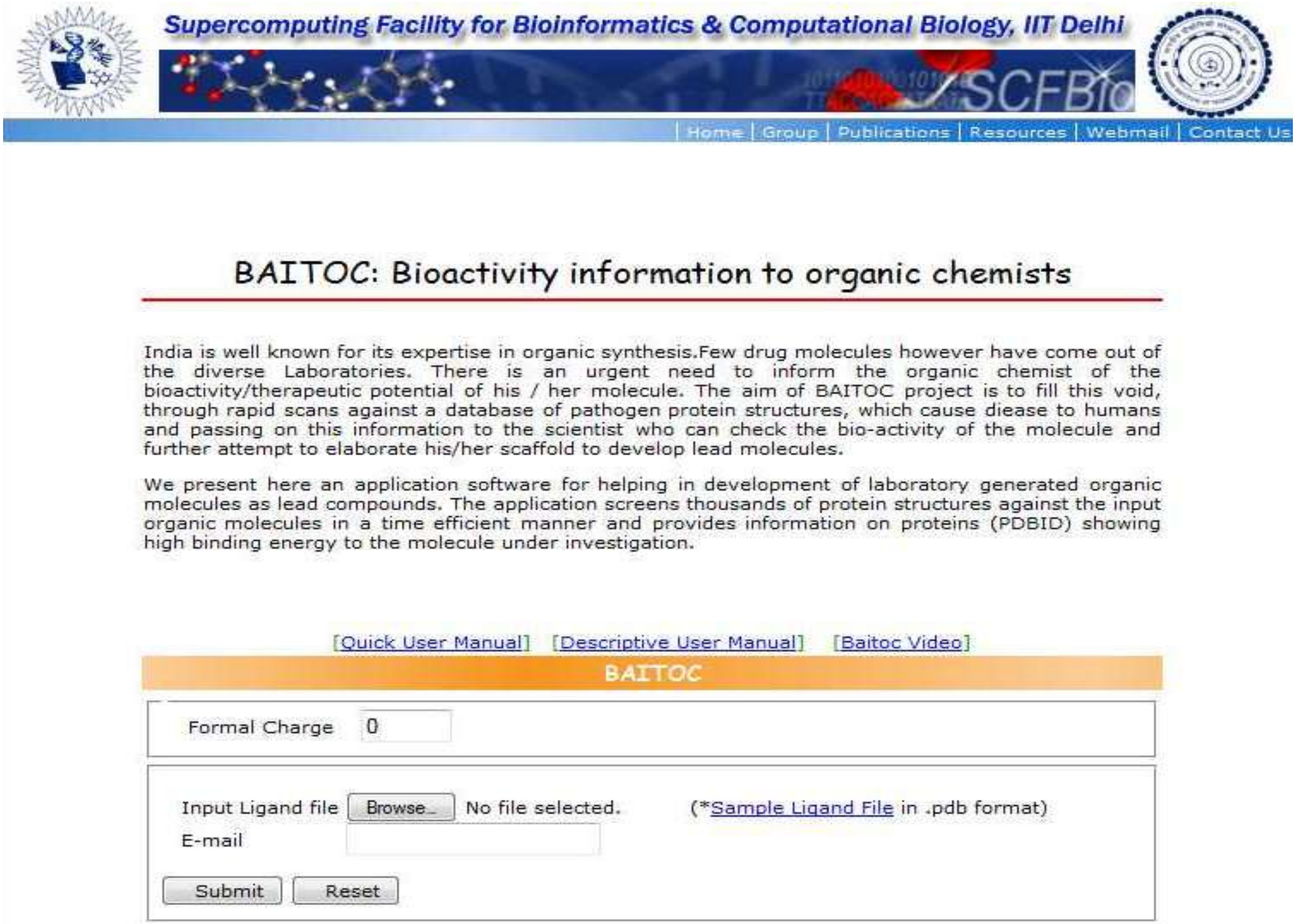
- ● **Method A** [Protein-Ligand Complex]
- ○ **Method B** [Protein 3D Structure Without Ligand]
- ○ **Method C** [Customized Dataset]
- ○ **Method D** [Customized Molecule]

Enter Ligand Id [Identifier]: DRG

Screening millions of compounds in minutes (!) based on physico-chemical descriptors

Goutam Mukherjee and B. Jayaram, "A Rapid Identification of Hit Molecules for Target Proteins via Physico-Chemical Descriptors", *Phys. Chem. Chem. Phys., 2013*, DOI:10.1039/C3CP44697B.

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


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

BAITOC: Bioactivity information to organic chemists

India is well known for its expertise in organic synthesis.Few drug molecules however have come out of the diverse Laboratories. There is an urgent need to inform the organic chemist of the bioactivity/therapeutic potential of his / her molecule. The aim of BAITOC project is to fill this void, through rapid scans against a database of pathogen protein structures, which cause diease to humans and passing on this information to the scientist who can check the bio-activity of the molecule and further attempt to elaborate his/her scaffold to develop lead molecules.

We present here an application software for helping in development of laboratory generated organic molecules as lead compounds. The application screens thousands of protein structures against the input organic molecules in a time efficient manner and provides information on proteins (PDBID) showing high binding energy to the molecule under investigation.

[Quick User Manual] [Descriptive User Manual] [Baitoc Video]

BAITOC

Formal Charge 0

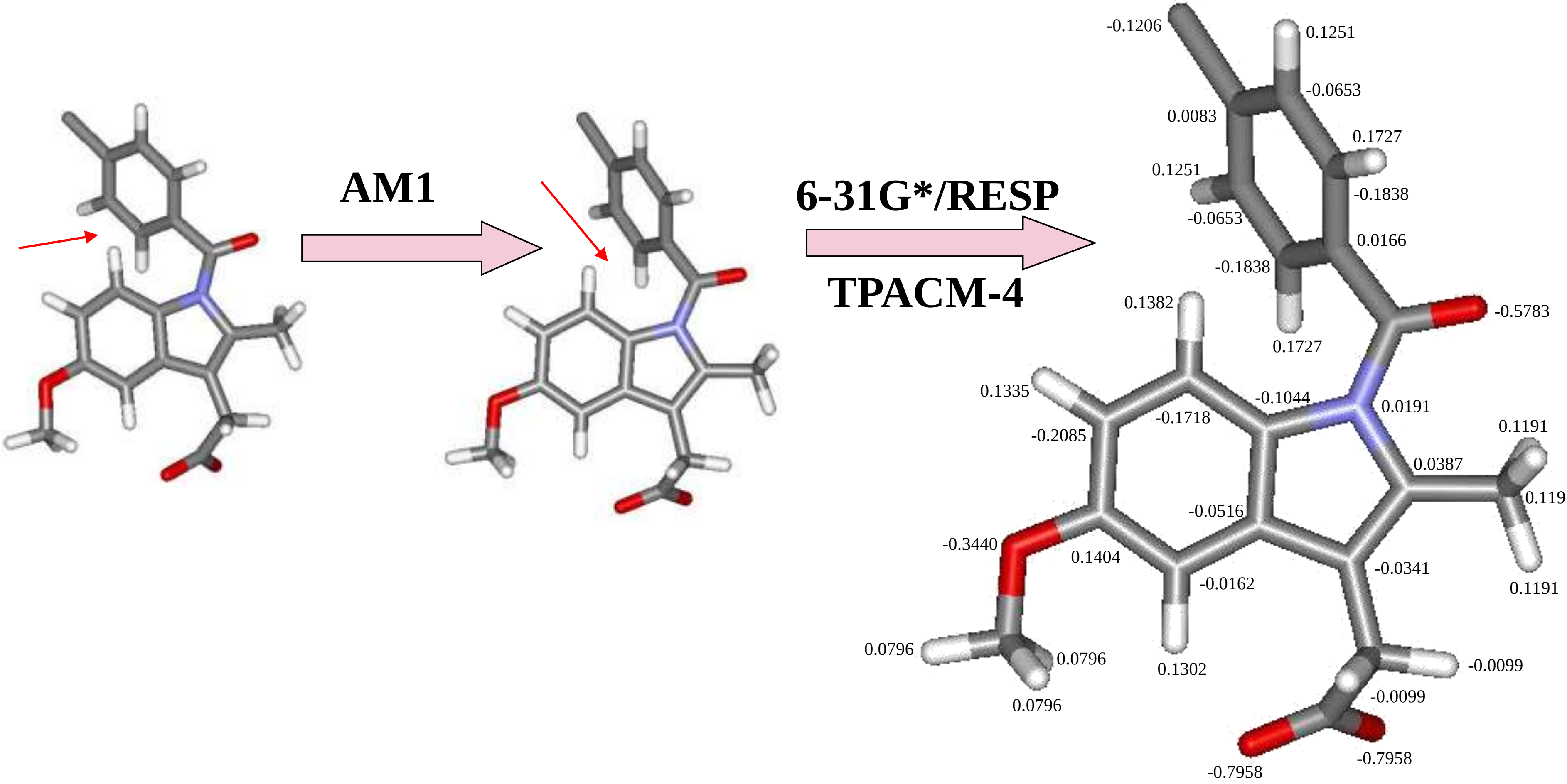
Input Ligand file Browse... No file selected. (*Sample Ligand File in .pdb format)

E-mail

Abhilash Jayaraj et al., 2018, manuscript in preparation.

Quantum Chemistry on Candidate drugs for Assignment of Force Field Parameters

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



AM1

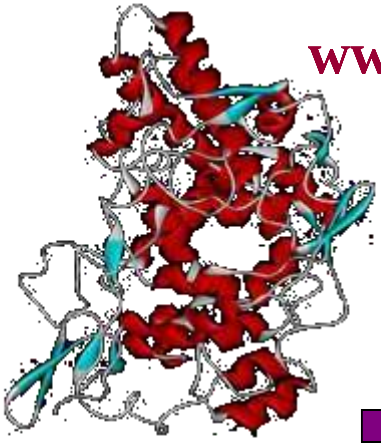
6-31G*/RESP

TPACM-4

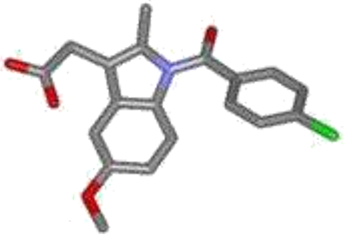
G. Mukherjee, N. Patra, P. Barua and B. Jayaram, *J. Computational Chemistry, 2011*, 32, 893-907.

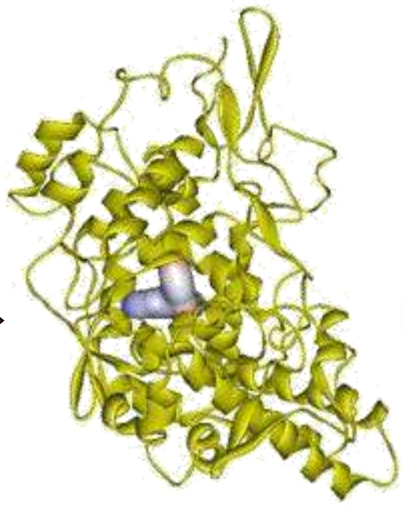
MONTE CARLO DOCKING OF THE CANDIDATE DRUG IN THE ACTIVE SITE OF THE TARGET

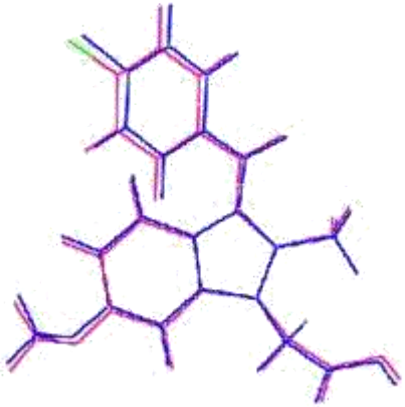
www.scfbio-iitd.res.in/dock/pardock.jsp



+







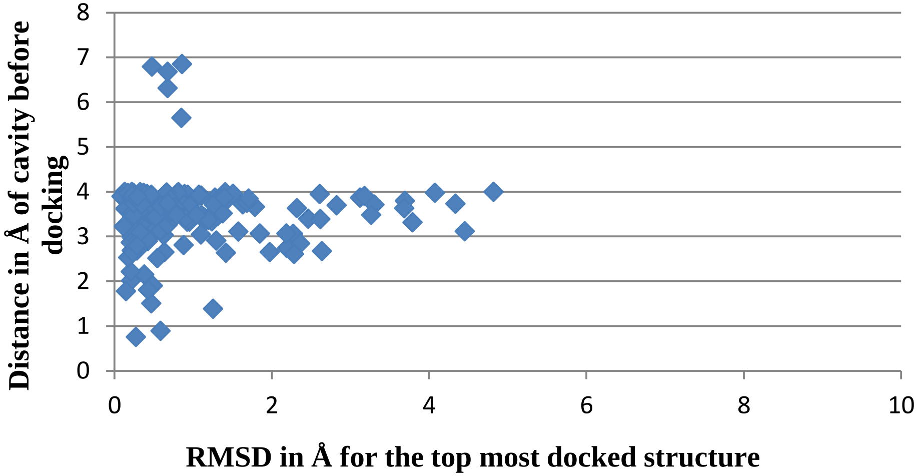
RMSD between the docked & the crystal structure is 0.2Å

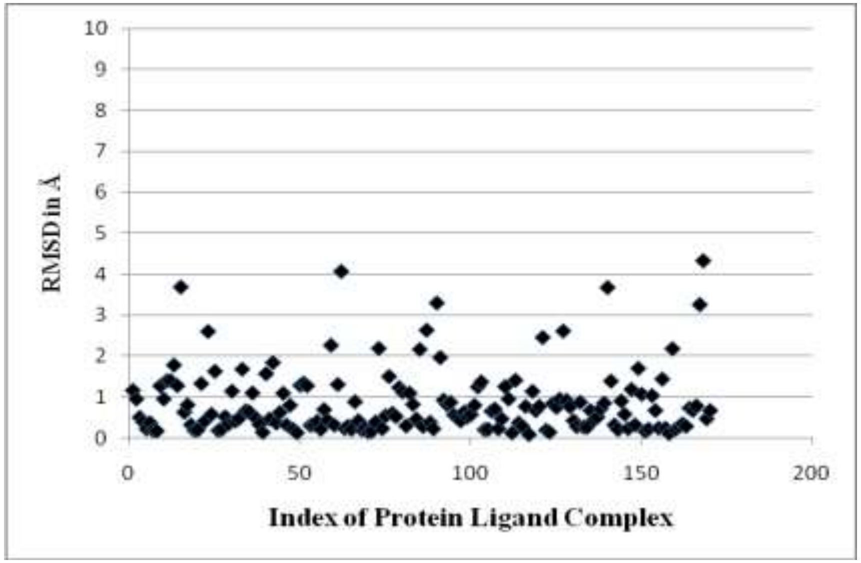
ENERGY MINIMIZATION

5 STRUCTURES WITH LOWEST ENERGY SELECTED

A. Gupta, A. Gandhimathi, P. Sharma, and B. Jayaram, "ParDOCK: An All Atom Energy Based Monte Carlo Docking Protocol for Protein-Ligand Complexes", ***Protein and Peptide Letters, 2007, 14(7)*****, 632-646.**

Docking Accuracies



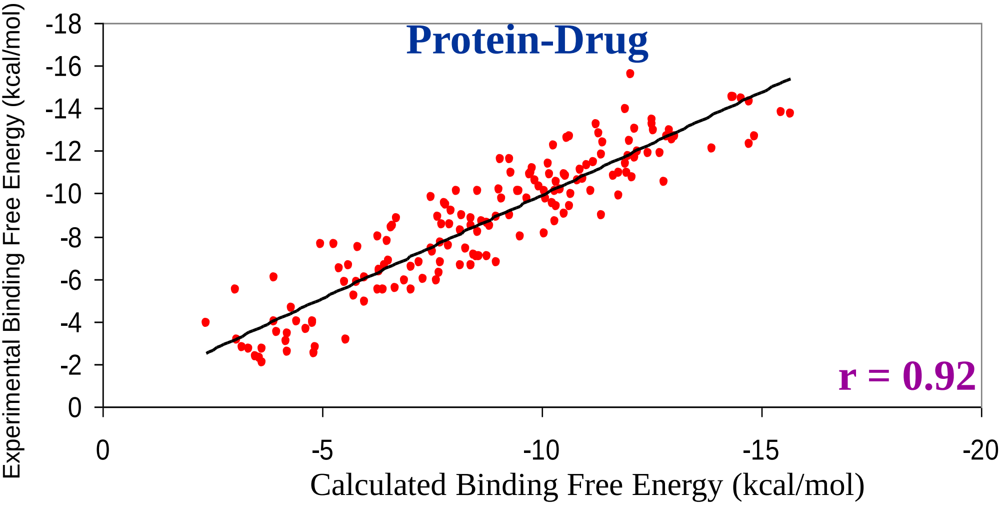


RMSD between the crystal structure and one of the top five docked structures

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

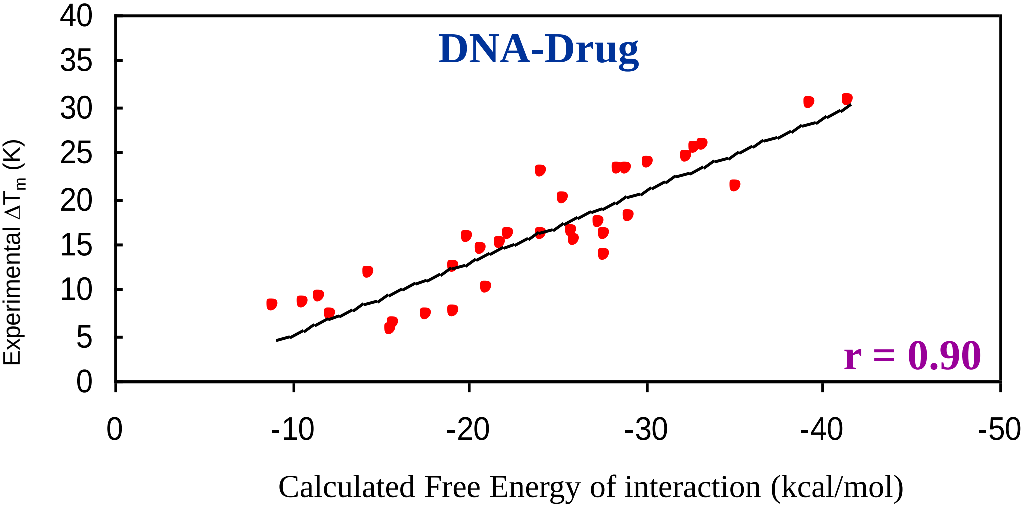
ENERGY BASED SCORING FUNCTION

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



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

T. Jain & B.Jayaram,
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

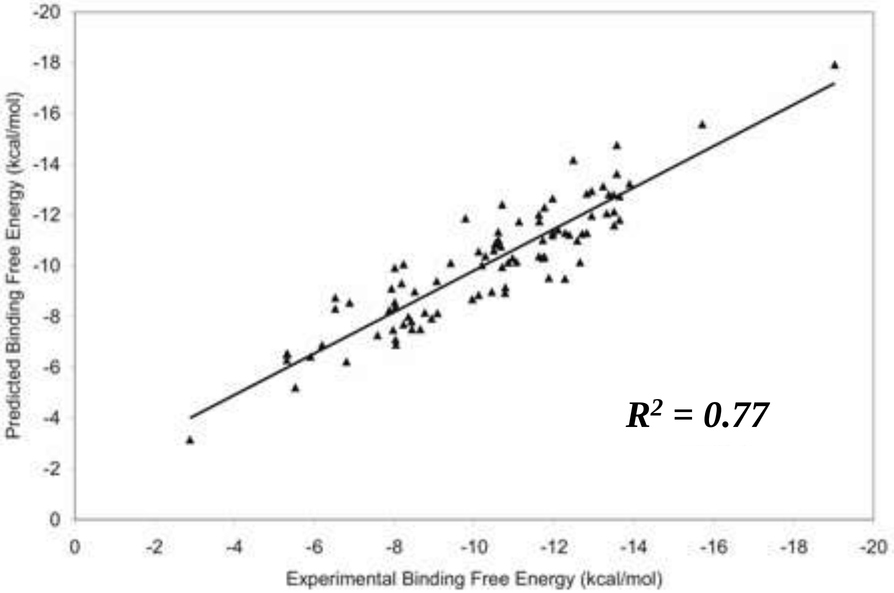
www.scfbio-iitd.res.in/software/drugdesign/preddicta.jsp

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

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

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.

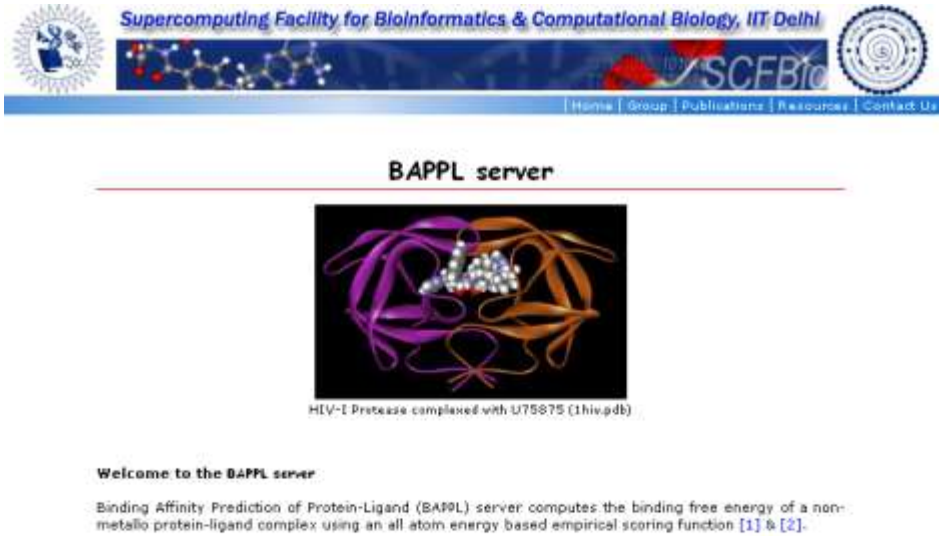
www.scfbio-iitd.res.in/software/drugdesign/bapplz.jsp

Tanya Singh, Olayiwola Adedotun Adekoya, B. Jayaram, "Understanding the binding of inhibitors of Matrix Metalloproteinases by molecular docking, quantum mechanical calculations, molecular dynamics simulations, and a MMGBSA/MMBappl study", *Mol. BioSyst.*, 2015, 11, 1041-1051.

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*

Some freely accessible web-servers from SCFBio for docking and scoring

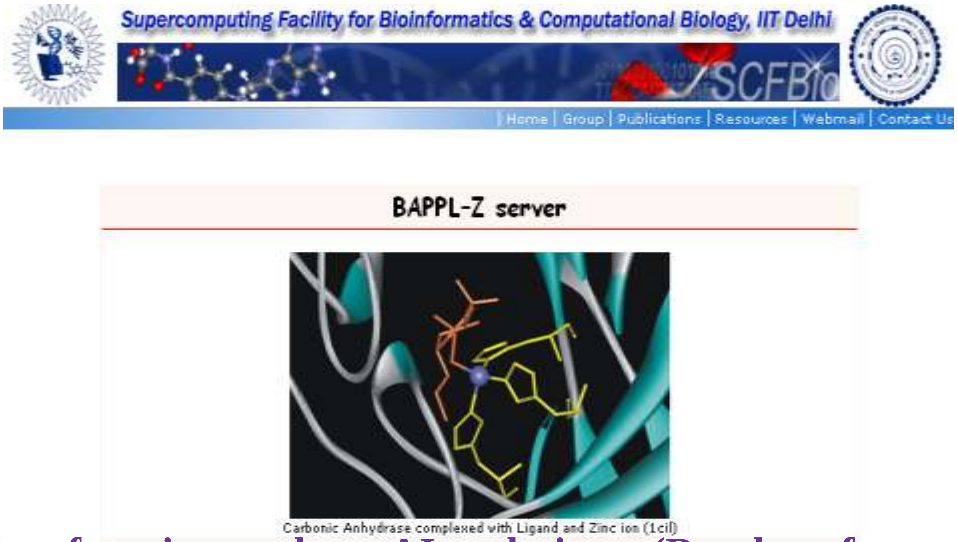

Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi
SCFBio
| Home | Group | Publications | Resources | Contact Us

BAPPL server

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

Welcome to the BAPPL server

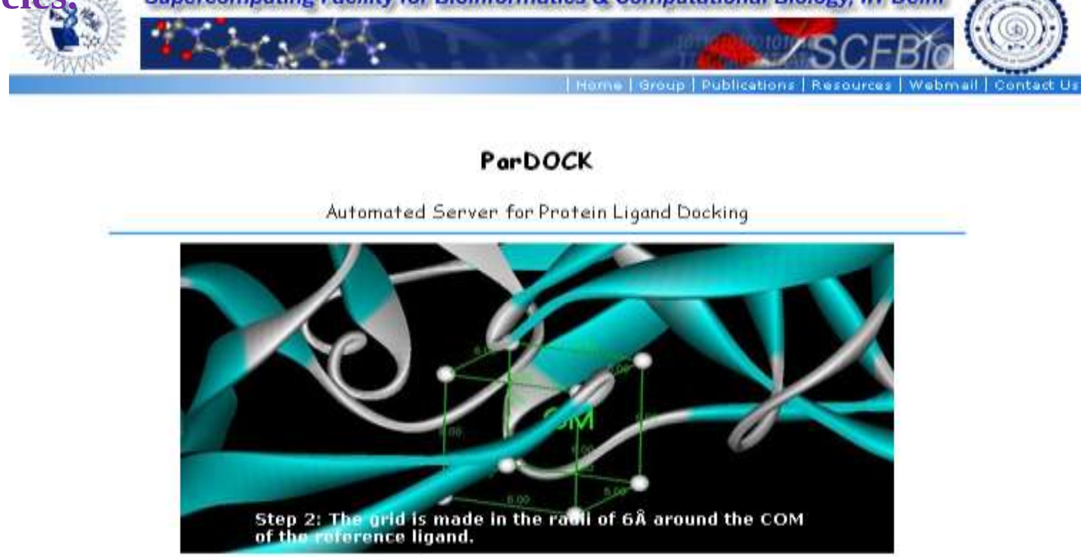
Binding Affinity Prediction of Protein-Ligand (BAPPL) server computes the binding free energy of a non-metallo protein-ligand complex using an all atom energy based empirical scoring function [1] & [2].


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

BAPPL-Z server

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

Bappl+ web-server with a physico-chemical scoring function and an AI technique (Random forest prediction) is coming up soon that can handle ligand binding to any protein with or without any metal ion with higher accuracies.

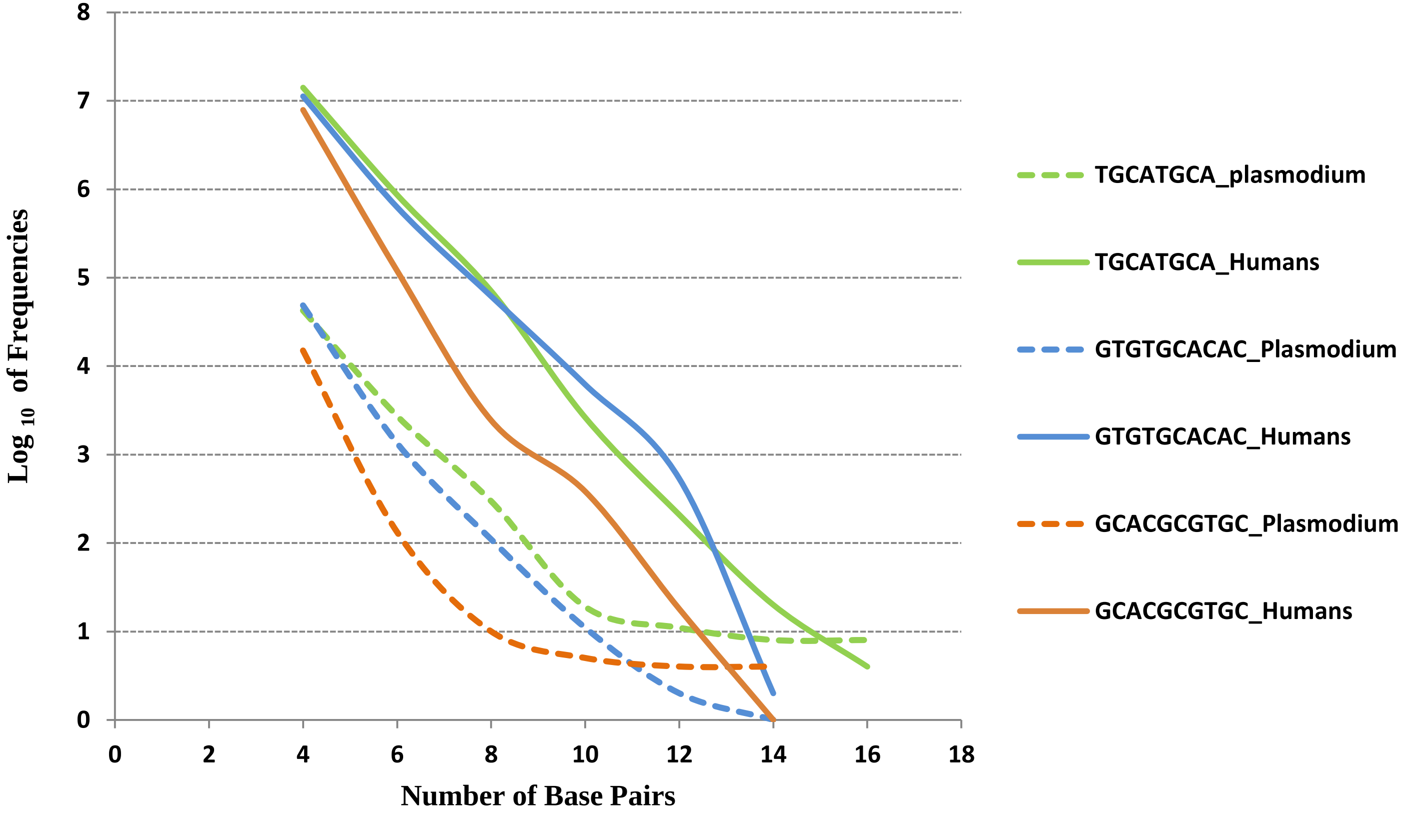

Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi
SCFBio
| Home | Group | Publications | Resources | Webmail | Contact Us
COM

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.

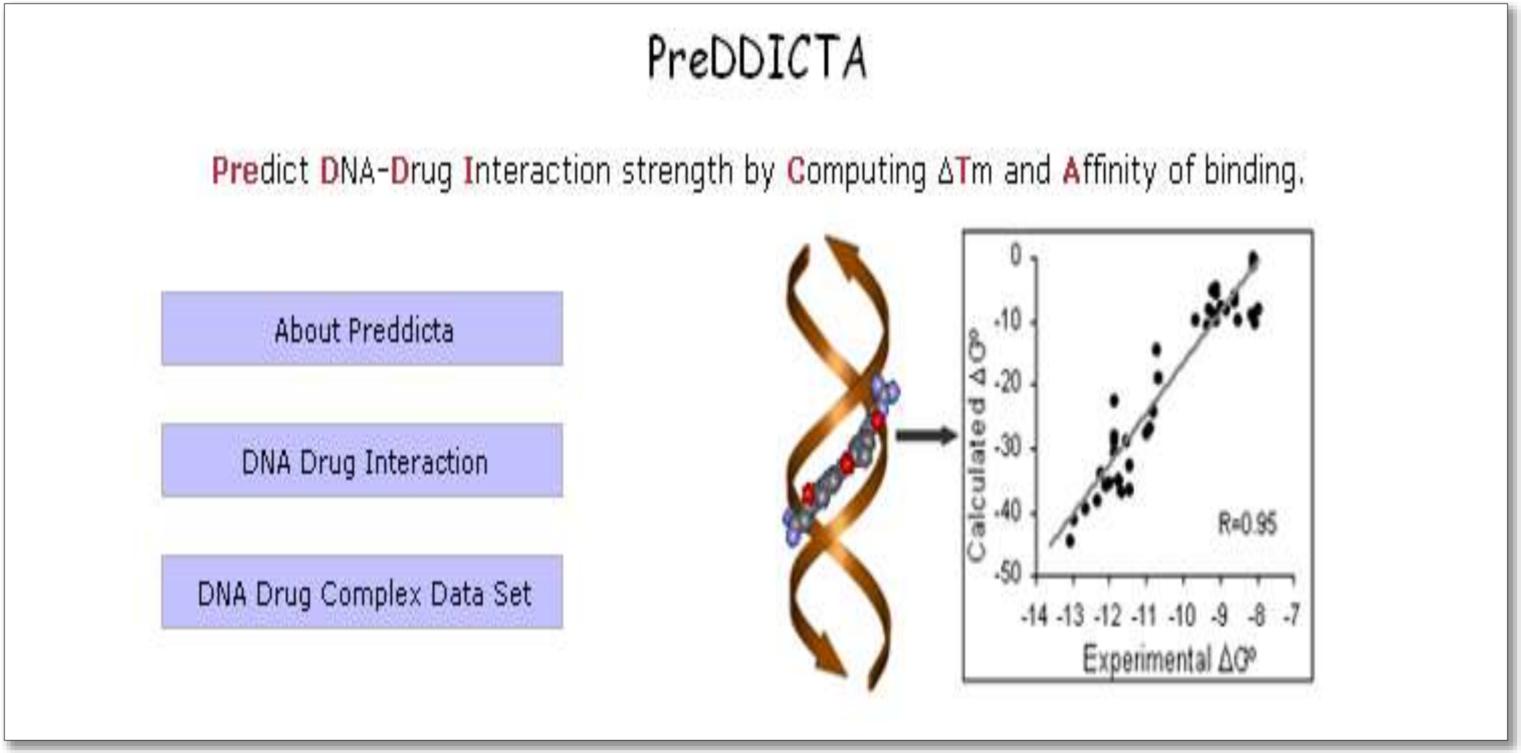
B. Jayaram, Tanya Singh, and Marcia Fenley, "DNA-Drug Interactions: A Theoretical Perspective" in "*Methods for Studying DNA/Drug Interactions*", Eds, Dr. Meni Wanunu & Prof. Yitzhak Tor, 2011, Ch-14, 317-338, CRC Press.



Logarithm of the frequencies of the occurrence of base sequences of lengths 4 to 18 base pairs in *Plasmodium falciparum (malarial parasite)* 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 (occuring only once) must be 18 to 20 bp long.

Methods and software for DNA targeted drug discovery for minor groove binders and intercalators

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



PreDDICTA

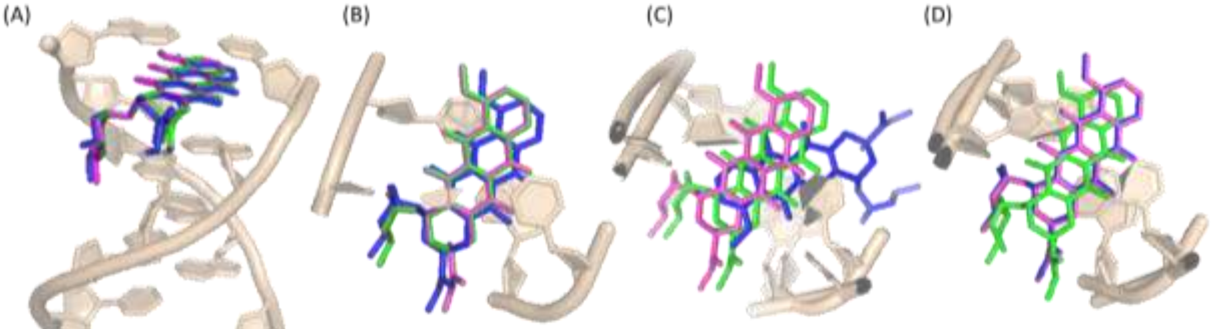
Predict DNA-Drug Interaction strength by Computing ΔTm and Affinity of binding.

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

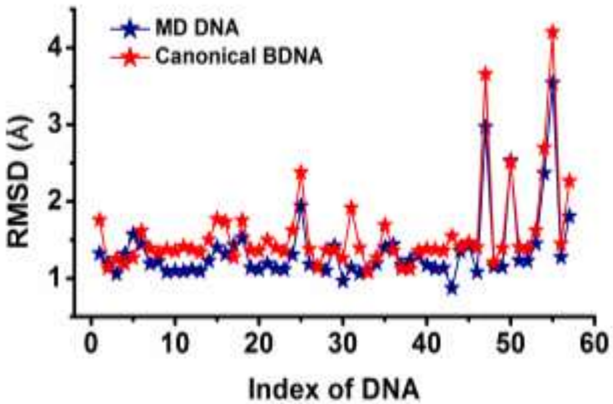
S.A. Shaikh and B. Jayaram, *J. Med. Chem., 2007,* 50, 2240-2244

Methods and software for DNA targeted drug discovery for minor groove binders and intercalators

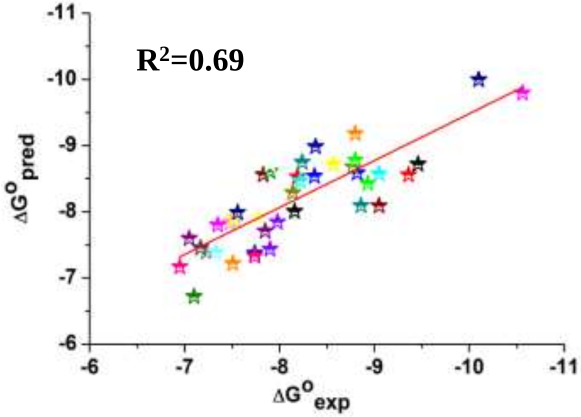
www.scfbio-iitd.res.in/intercalate



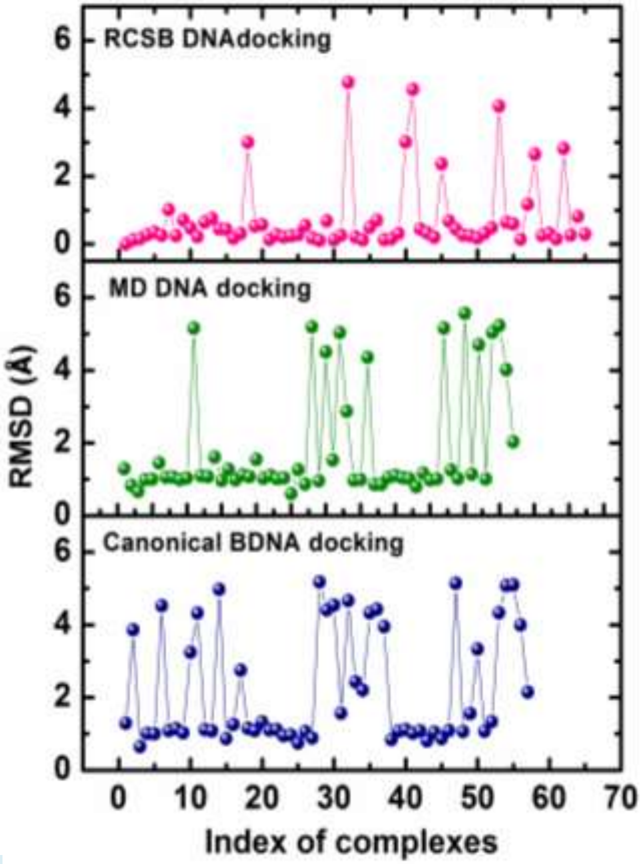
PDB -1D54; 🟩 crystal ligand 🟦 top ranked 🟪 best ranked



DNA structure generation accuracies on 58 systems



Correlation between the ΔG^{o}_{exp} and ΔG^{o}_{pred} for 43 complexes (in kcal/mol).

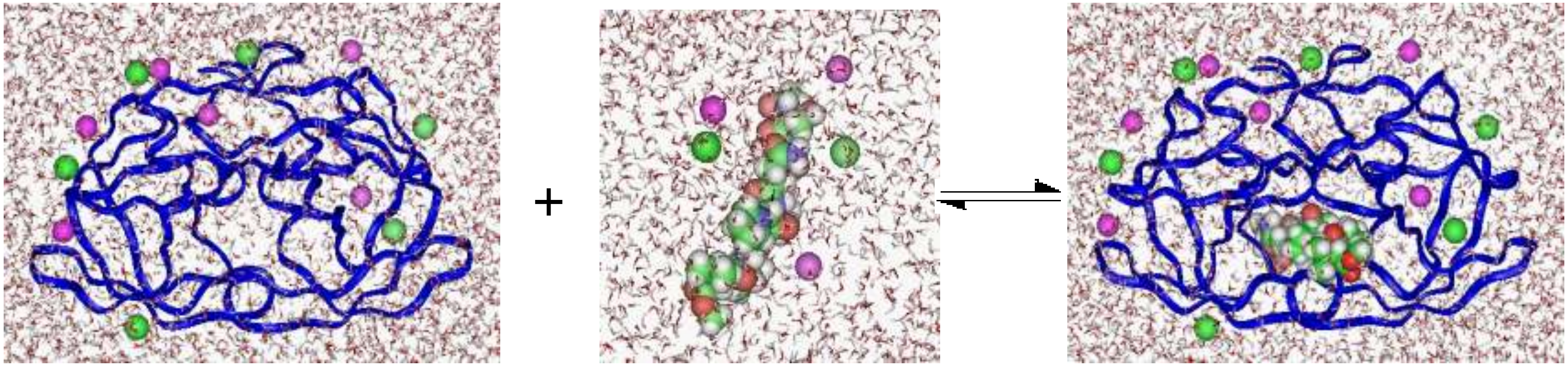


Docking accuracies on 65 drug-DNA complexes

Anjali Soni, Pooja Khurana, Tanya Singh, B. Jayaram, "A DNA Intercalation Methodology for an Efficient Prediction of Ligand Binding Pose and Energetics", *Bioinformatics 2017*, 33, 1488-96.

Binding Affinity Analysis

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



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

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

Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi
A Centre of Excellence of the Department of Biotechnology, Govt. of India

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***, *Current Pharmaceutical Design, 2007,* **13, 3454-3470. (Tackling side effects due to off-target binding computationally!)**

	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	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟧	🟩	🟦

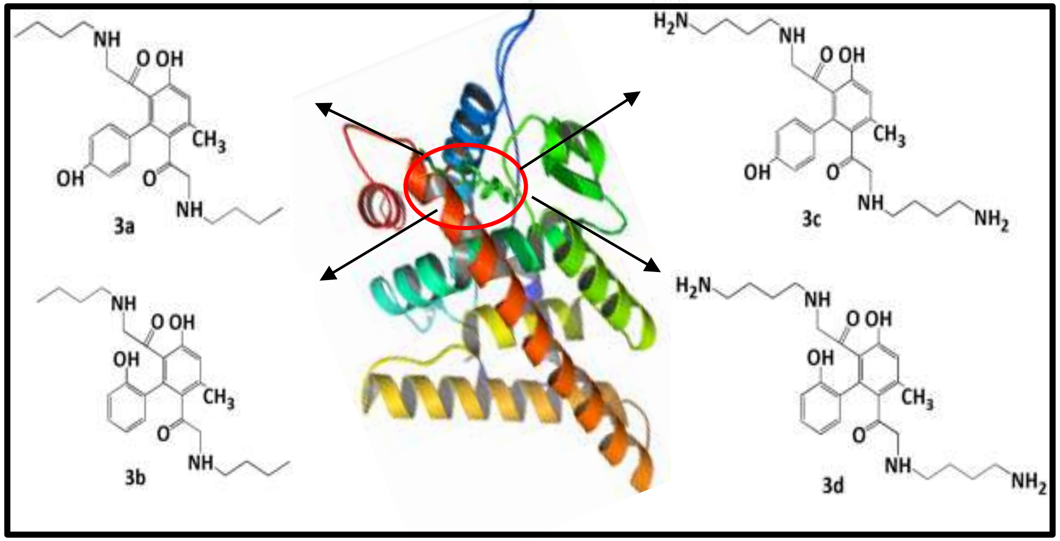
BLUE: HIGH BINDING AFFINITY **GREEN: MODERATE AFFINITY** **ORANGE: POOR AFFINITY**

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.

Sanjeevini's success stories-1

Anti cancer

A few designed, synthesized & tested biphenyl based molecules against estrogen receptor targeting breast cancer



Biphenyl compounds	ERα B.E (kcal/mol)	ERα IC_{50} (μM)	ERβ B.E (kcal/mol)	ERβ IC_{50} (μM)
3a	-7.00	146	-6.67	>150
3b	-7.50	2.27	-7.03	>100
3c	-6.81	68.5	-6.48	>150
3d	-6.01	4.27	-5.46	>150

3b is identified to be the most potent inhibitor with an IC_{50} of 2.27 μM.

MMGBSA (kcal/mol)

Complexes	ERα-3b	ERβ-3b
ΔG_{bind}	-23.63	-20.41

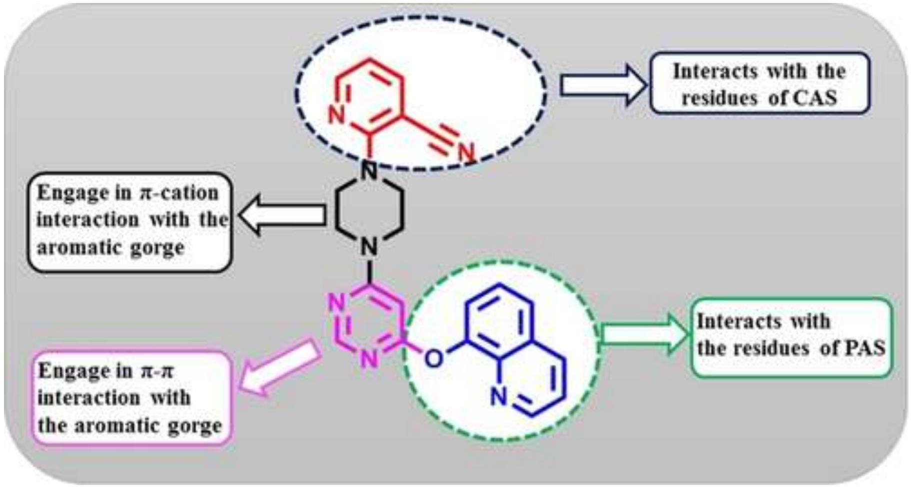
Soni, A.; Virmani, M.; Kaushik, S.; Bhatnagar, S.; Jayaram, B. Indian patent entitled "1,3-Diacetyl Biphenyl Analogs, and their Derivatives" with Application number: 3126/DEL/2012; *Chemical Biology & Drug Design, 2017,* DOI: 10.1111/cbdd.13126

Sanjeevini's success stories-2

Anti Alzheimer's

Target: Acetylcholinesterase

Compounds	IC50 value against AChE^a (μM)
10c	2.84 ± 0.25
10e	0.67 ± 0.13
11a	2.77 ± 0.67
11c	0.161 ± 0.04
11d	1.39 ± 0.14
12a	1.37 ± 0.44
12b	0.036 ± 0.12
12c	0.93 ±0.28
Donepezil	0.038 ± 0.34
Tacrine	0.13 ± 0.21

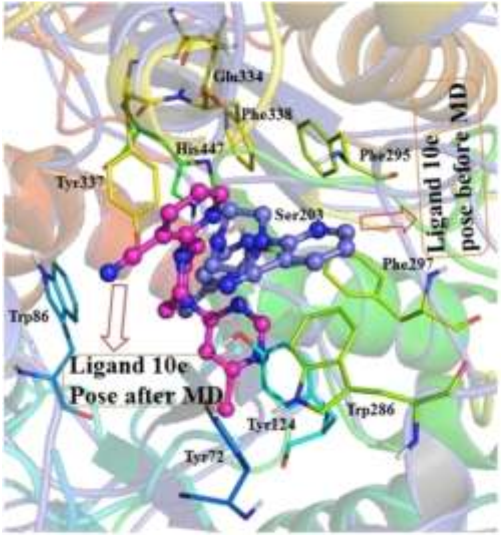


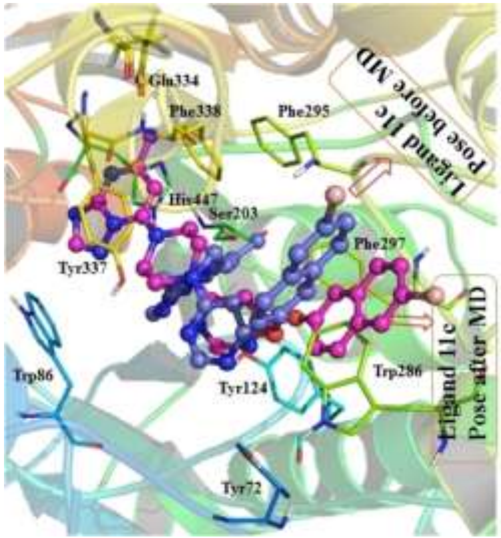
Interacts with the residues of CAS

Engage in π-cation interaction with the aromatic gorge

Interacts with the residues of PAS

Engage in π-π interaction with the aromatic gorge





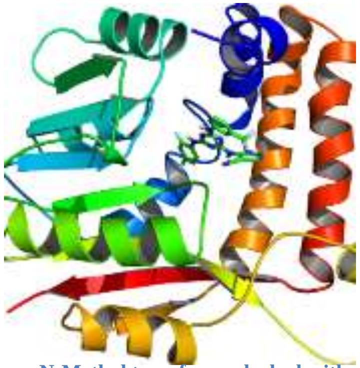
Jitendra Kumar,Asim Gill,Marziya Shaikh,Anju Singh, Ashutosh Shandilya, Ehtesham Jameel, Nitin Sharma, Nirotpal Mrinal, Nasimul Hoda, B. Jayaram, "Pyrimidine-Triazolopyrimidine and Pyrimidine-Pyridine Hybrids as Potential Acetylcholinesterase Inhibitors for Alzheimer's Disease", *Chemistry Select: Medicinal Chemistry & Drug Discovery*, 2017. DOI: 10.1002/slct.201702599

Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi
A Centre of Excellence of the Department of Biotechnology, Govt. of India

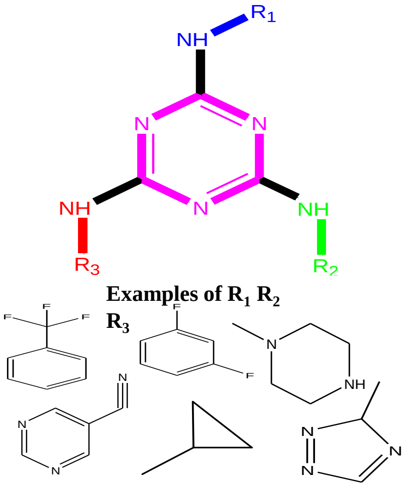
Sanjeevini's success stories-3

Anti malarials

Target: N-Methyl transferase

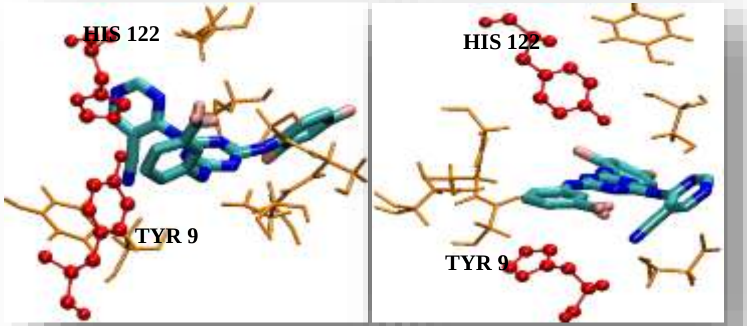


N-Methyl transferase docked with triazine derivative



Examples of R_1 R_2 R_3

IC_{50} = 0.8 µM to 10 µM and top two inhibitors showing 17- 34 fold selectivity for parasite cell over human cell. Other inhibitors also show more than 10 fold selectivity

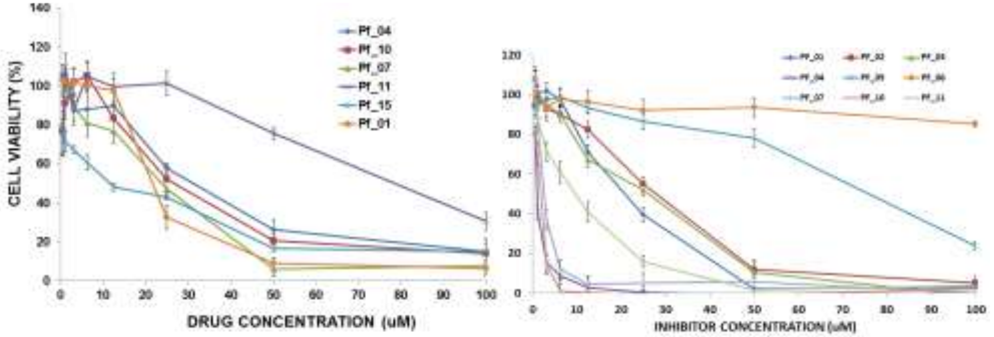


Ashutosh Shandilya, Nasimul Hoda, Sameena Khan, Ehtesham Jameel, Jitendra Kumar, B. Jayaram, "*De novo* lead optimization of triazine derivatives identifies potent antimalarial", *J. Mol. Graphics & Modelling, 2017*, 71, 96–103, DOI: 10.1016/j.jmgm.2016.10.022.

Computational predicted energies (kcal/mol) **Experimental binding affinities**

PF_SCF_1	-8.23	PF_SCF_8	-10.63
PF_SCF_2	-9.11	PF_SCF_9	-10.14
PF_SCF_3	-9.12	PF_SCF_10	-12.12
PF_SCF_4	-11.85	PF_SCF_11	-11.57
PF_SCF_5	-8.12	PF_SCF_12	-10.34
PF_SCF_6	-9.12	PF_SCF_13	-9.85
PF_SCF_7	-9.96	PF_SCF_14	-9.15

Pf_SCF_04	~0.8µM	Pf_SCF_02	~28 µM
Pf_SCF_13	~ 58 µM	Pf_SCF_14	~71 µM
Pf_SCF_05	~77 µM	Pf_SCF_06	~100 µM
Pf_SCF_10	~1.5µM	Pf_SCF_07	~2.4 µM
Pf_SCF_11	~10 µM	Pf_SCF_12	~ 10 µM
Pf_SCF_ 01	~21 µM	Pf_SCF_03	~26 µM



Parasite growth inhibition assay. (A) selected inhibitors from the docking analysis tested in the parasite growth inhibition assay using double dilution till 8 points (100 µM to 0.78 µM). (B) Cytotoxicity measurement for inhibitors exhibiting anti-plasmodium effect is shown.

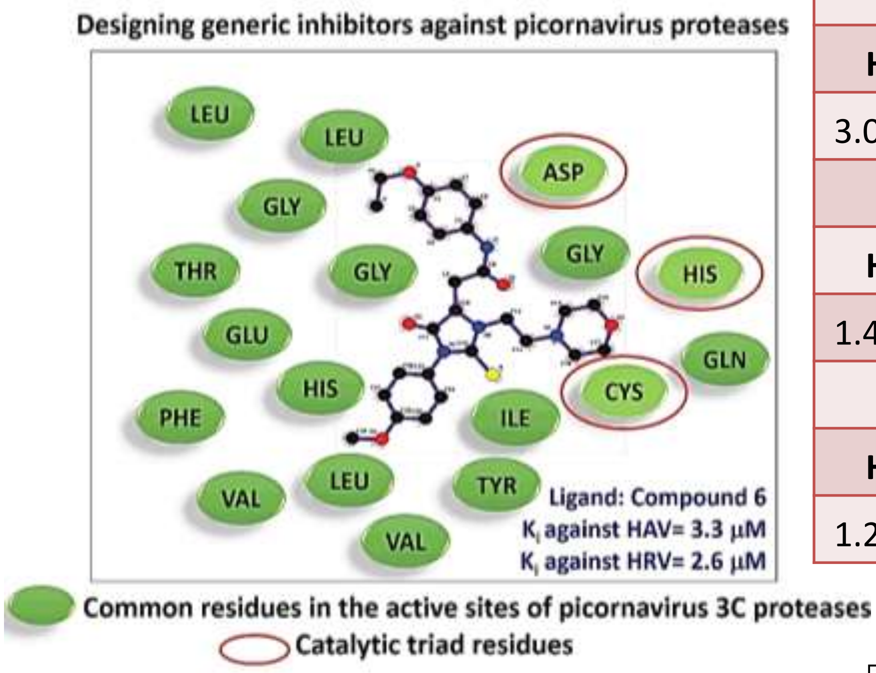
HIS 122 functions as a general base to abstract a proton from the hydroxyl group of TYR 9 to activate the residue. Pf_SCF_10 by translating between these two residues further sway them apart from each other to distort this mechanism of action and hence TYR 9 alone is insufficient to drive the methylation reaction.

Nanomolar compounds, Bioorganic Medicinal Chemistry Letters, 2018 (under revision).

Sanjeevini's success stories-4

Anti virals

Target: Picornavirus 3C Proteases



Ligand: Compound 6
K_i against HAV= 3.3 µM
K_i against HRV= 2.6 µM

🟩 Common residues in the active sites of picornavirus 3C proteases

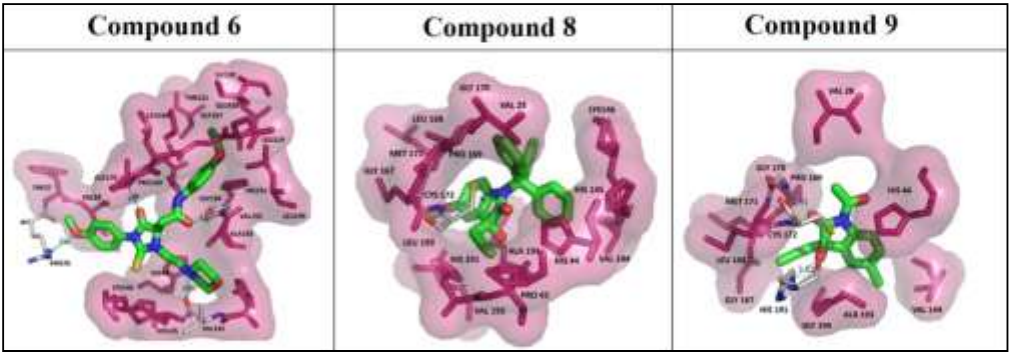
○ Catalytic triad residues

Experimental Ki values against Hepatitis A and Human Rhino Viruses

Compound 1		Compound 2		Compound 3	
HAV	HRV	HAV	HRV	HAV	HRV
3.0 ± 0.1	3.2 ± 0.1	8.6 ± 0.7	5.5 ± 0.3	2.5 ± 0.1	3.2 ± 0.2
Compound 4		**Compound 5**		**Compound 6**	
HAV	HRV	HAV	HRV	HAV	HRV
1.4 ± 0.1	1.7 ± 0.1	117.8 ± 22.3	N.D.	3.3 ± 0.2	2.6 ± 0.1
Compound 7		**Compound 8**		**Compound 9**	
HAV	HRV	HAV	HRV	HAV	HRV
1.2 ± 0.1	1.5 ± 0.1	2.1 ± 0.1	2.5 ± 0.1	1.6 ± 0.1	1.6 ± 0.1

**Ki values are in µM concentrations*

Kamalika Banerjee ‡, Ruchika Bhat ‡, V. U. Bhaskara Rao, Anshu Nain, Kartik Lakshmi Rallapalli, Sohona Gangopadhyay, R. P. Singh, Manidipa Banerjee*, B. Jayaram*, Towards development of generic inhibitors against the 3C proteases of picornaviruses, *The FEBS Journal*, Accepted, Nov. 2018. Doi: 10.1111/febs.14707.



Sanjeevini's success stories-5

Anti fungals

You may clap now!

A Collaboration with Harvard Medical School, USA

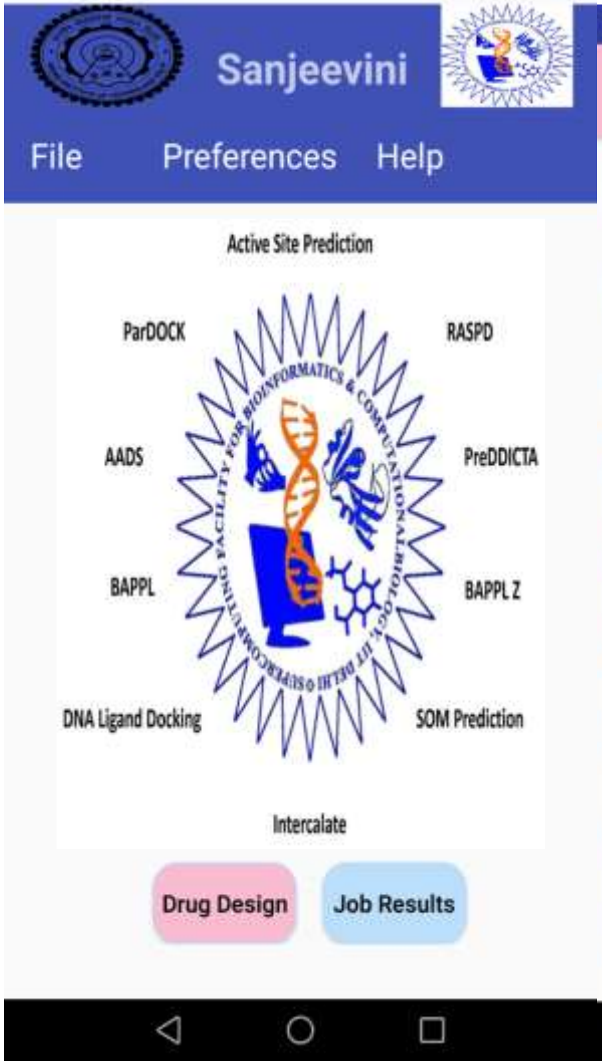
Joy L Nishikawa, Andras Boeszoermenyi, Luis A Vale-Silva, Ricardo Torelli, Brunella Posteraro, Yoo-Jin Sohn, Fei Ji, Vladimir Gelev, Dominique Sanglard, Maurizio Sanguinetti, Ruslan I Sadrayev, Goutam Mukherjee, Jayaram B., Sara J Buhrlage, Nathanael S Gray, Gerhard Wagner, Anders M Naar, Haribabu Arthanari, "Inhibiting fungal multidrug resistance by disrupting an activator-mediator interaction", *Nature, 2016,* 530, 485-489. doi:10.1038/nature16963.

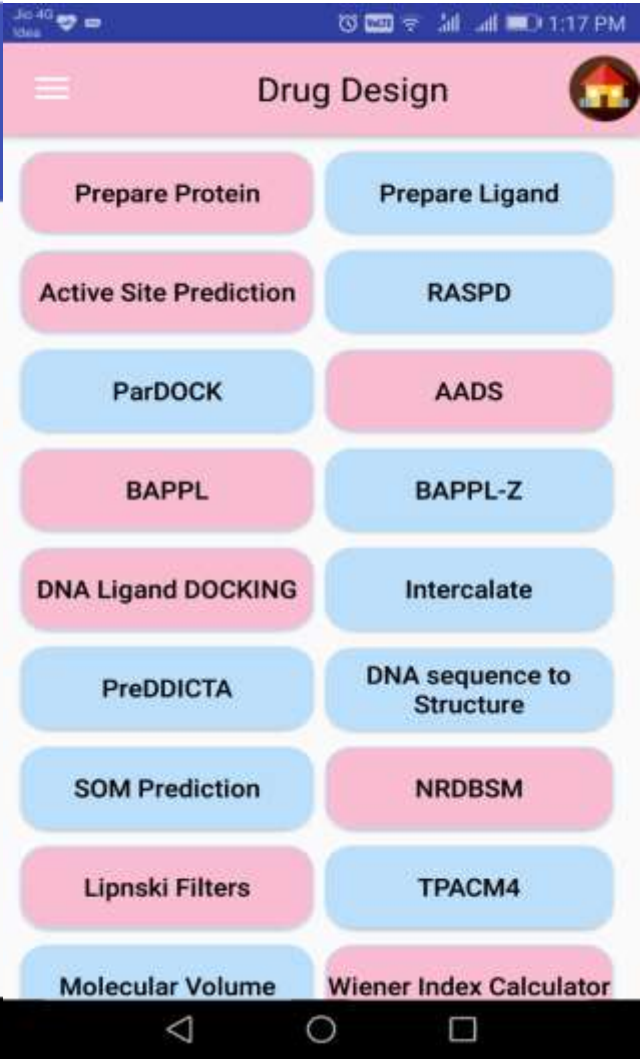
A Collaboration with National Cancer Institute (NCI), USA

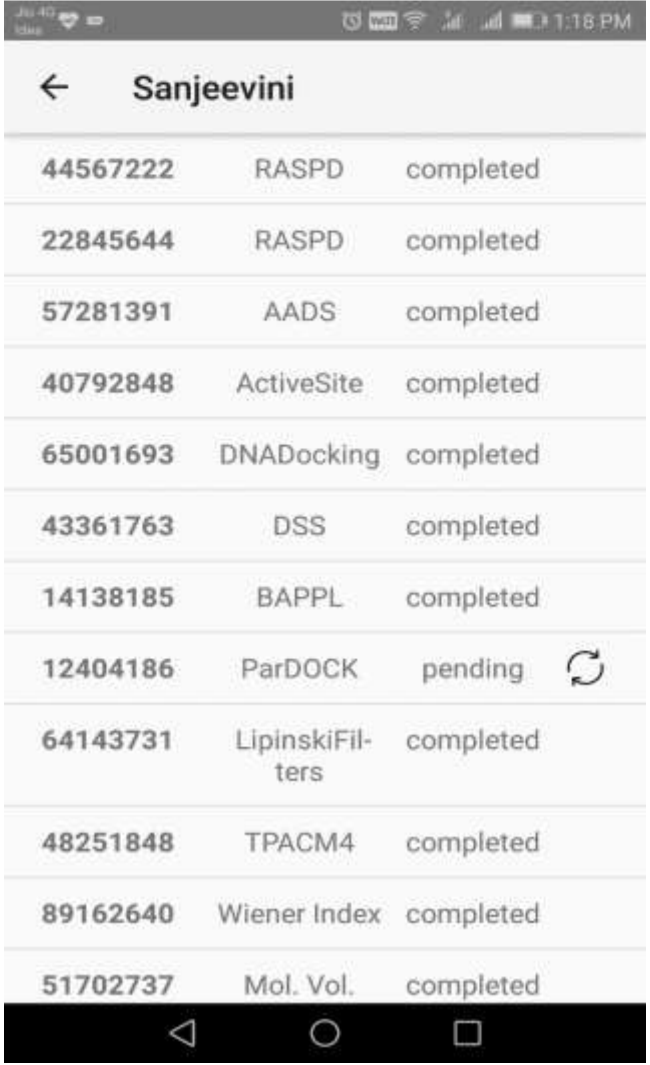
On Drug Repurposing: In Progress

+ Several more nanomolar compounds/publications with *Sanjeevini*

Sanjeevini application for Android devices (released on July 31st, 2017)



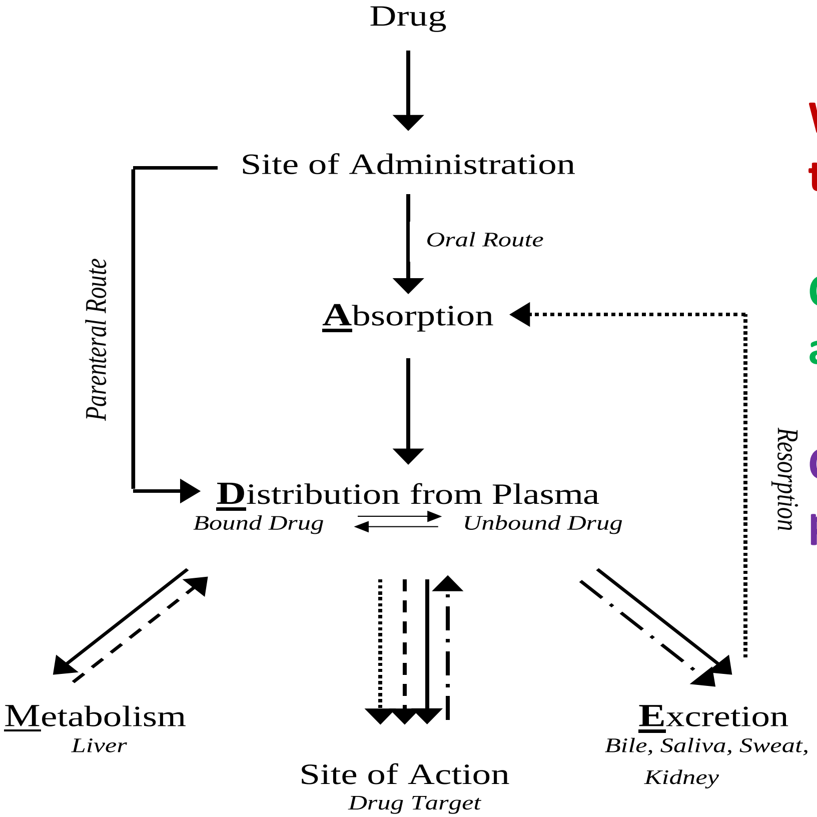




44567222	RASPD	completed
22845644	RASPD	completed
57281391	AADS	completed
40792848	ActiveSite	completed
65001693	DNADocking	completed
43361763	DSS	completed
14138185	BAPPL	completed
12404186	ParDOCK	pending
64143731	LipinskiFil-ters	completed
48251848	TPACM4	completed
89162640	Wiener Index	completed
51702737	Mol. Vol.	completed

- The application is freely accessible on google play store and can be installed by searching for “Sanjeevini – SCFBio - CADD”.
- Android application: https://play.google.com/store/apps/details?id=com.sanjeevini&hl=en
- Application webpage: http://www.scfbio-iitd.res.in/sanjapp/webSearch/Sanjeevini_webpage.html

Future of Drug Discovery: Towards a Molecular View of ADMET



What are the five rules of toxicity?

Can we make drugs which are least toxic to humans?

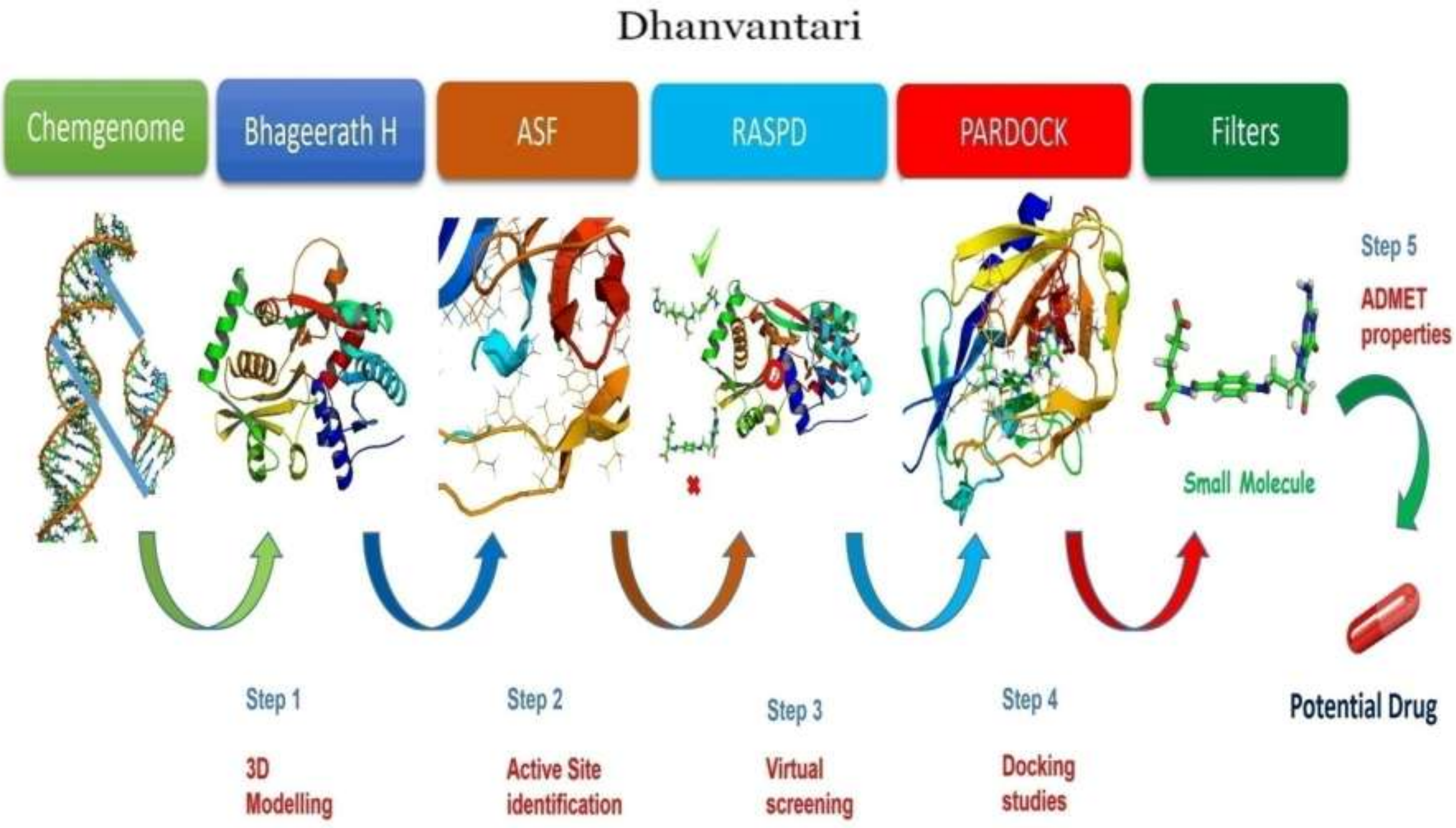
Can we automate this process?

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.

To summarize

In silico Drug discovery assembly line developed at SCFBio



SCFBio Team

~ 16 teraflops of computing; 200 terabytes of storage + Some smart dedicated student power

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

Present

Shashank Shekhar
Ankita Singh
Ruchika Bhat
Prof. Priyanka Siwach

Vandana Shekhar
Amita Pathak
Akhilesh Mishra
A. Mohan Rao

Dr. Abhilash Jayaraj
Pradeep Pant
Manpreet Singh
Puneeta

Former

Dr. Rahul Kaushik
Dr. Ashutosh Shandilya
Dr. Tanya Singh
Dr. Garima Khandelwal
Dr. Tarun Jain
Dr. N. Latha
Dr. Surjit Dixit
Pankaj Sharma
A.Gandhimathi
Neelam Singh
Dr. Sandhya Shenoy
Sahil Kapoor
Navneet Tomar
Varsha Singh
R. Nagarajan

Dr. Debarati DasGupta
Dr. Avinash Mishra
Dr. Goutam Mukherjee
Dr. Poonam Singhal
Dr. Saher Afshan Shaikh
Dr. Achintya Das
Dr. Nidhi Arora
Praveen Agrawal
Gurvisha Sandhu
Shailesh Tripathi
Rebecca Lee
Satyanarayan Rao
Surojit Bose
Ali Khosravi

Dr. Anjali Soni
Dr. Priyanka Dhingra
Dr. Pooja Narang
Dr. Kumkum Bhushan
Dr. Parul Kalra
Dr. E. Rajasekaran
Dr. Prashant S. Rana
Vidhu Pandey
Anuj Gupta
Dhrubajyoti Biswas
Bharat Lakhani
Pooja Khurana
Kritika Karri
Preeti Bisht

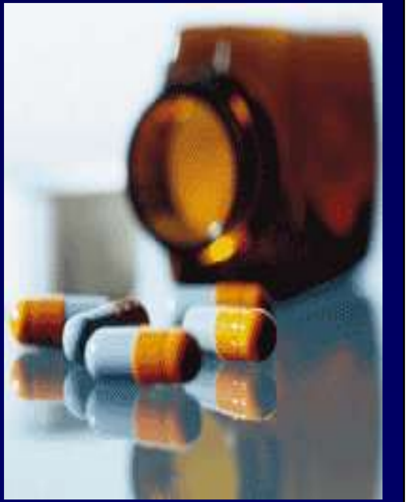
LeadInvent

Technologies

A start-up company formed by former students of Prof. BJ based on software developed at SCFBio

Novel Drug Discovery

Drug Design Solutions



Incubated at IIT Delhi (2007-2010)

DSIR Certified (2011)

Biospectrum Award 2011
Asia Pacific Emerging Company of the Year

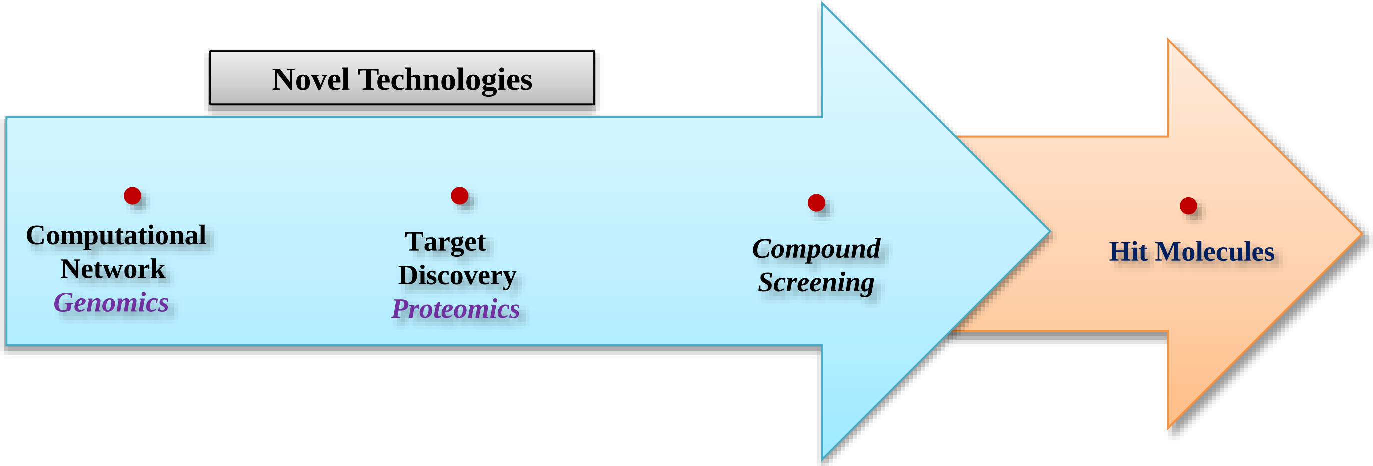
Mr. Pankaj Sharma
Mr. Surojit Bose
Mr. Praveen Aggarwal
Ms. Gurvisha Sandhu

www.leadinvent.com

Novo Informatics
Smart Minds.Smart Technologies.

Incubated at IIT Delhi (2011-2014)

Recipient of TATA NEN 2012 Award
Recipient of Biospectrum 2013 Award
Recipient of BioAsia 2014 Award



NI research pipeline

Sahil Kapoor
Avinash Mishra
Shashank Shekhar

A start-up company formed by former students of Prof. BJ based on software developed at SCFBio

Acknowledgements

Department of Biotechnology

Department of Science & Technology

National Supercomputing Mission

Ministry of Information Technology

Council of Scientific & Industrial Research

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

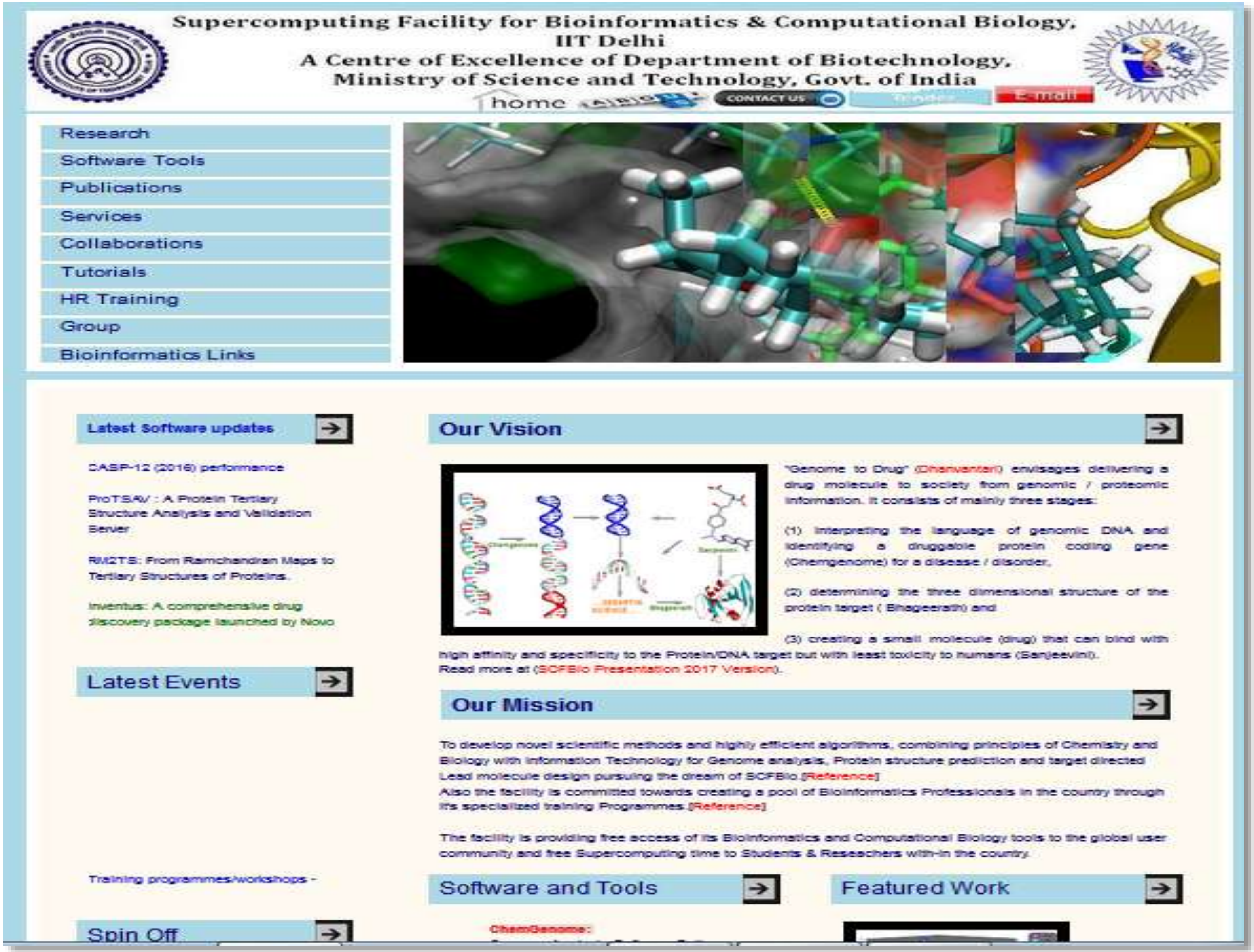
HCL Life Science Technologies

Dabur Research Foundation

NIIT

Indian Institute of Technology, Delhi

Supercomputing Facility for Bioinformatics & Computational Biology, IIT Delhi
A Centre of Excellence of the Department of Biotechnology, Govt. of India



Supercomputing Facility for Bioinformatics & Computational Biology,
IIT Delhi
A Centre of Excellence of Department of Biotechnology,
Ministry of Science and Technology, Govt. of India

Latest Software updates

CASP-12 (2016) performance

ProTSAV : A Protein Tertiary Structure Analysis and Validation Server

RM2TS: From Ramchandran Maps to Tertiary Structures of Proteins.

Inventus: A comprehensive drug discovery package launched by Novo

Latest Events

Training programmes/workshops -

Spin Off

Our Vision

"Genome to Drug" (Dhanvantari) envisages delivering a drug molecule to society from genomic / proteomic information. It consists of mainly three stages:

(1) interpreting the language of genomic DNA and identifying a druggable protein coding gene (Chemgenome) for a disease / disorder,

(2) determining the three dimensional structure of the protein target (Bhageerath) and

(3) creating a small molecule (drug) that can bind with high affinity and specificity to the Protein/DNA target but with least toxicity to humans (Sanjeevini).
Read more at (SCFBio Presentation 2017 Version).

Our Mission

To develop novel scientific methods and highly efficient algorithms, combining principles of Chemistry and Biology with Information Technology for Genome analysis, Protein structure prediction and target directed Lead molecule design pursuing the dream of SCFBio.[Reference]
Also the facility is committed towards creating a pool of Bioinformatics Professionals in the country through it's specialized training Programmes.[Reference]

The facility is providing free access of its Bioinformatics and Computational Biology tools to the global user community and free Supercomputing time to Students & Reseachers with-in the country.

Software and Tools

ChemGenome:

Featured Work

Visit Us at www.scfbio-iitd.res.in ***Thank You***