3. Genomes to Hits: The Emerging Assembly Line *In Silico*

B. Jayaram[1,2,3], Priyanka Dhingra[1,2], Goutam Mukherjee[1,2], Vivekanandan Perumal[3]

[1]*Department of Chemistry* & [2]*Supercomputing Facility for Bioinformatics & Computational Biology* & [3]*Kusuma School of Biological Sciences, Indian Institute of Technology, Hauz Khas, New Delhi-110016, India*

Introduction

The genomic sequences of pathogens and of humans afford us an opportunity to dream that *in silico* suggestions of candidate drugs will be forthcoming almost in an automated assembly line within hours with high levels of affinity, specificity and low toxicity within the foreseeable future. The challenges to overcome to facilitate this include, (i) higher levels of accuracies in genome annotation, (ii) evolution of algorithms for automated identification of druggable targets in the genome or proteome, (iii) generation of accurate tertiary/quaternary structures of protein targets and, (iv) design of small molecules with high levels of affinity and selectivity to the target and with proper ADMET profiles. We describe, in this chapter the progress achieved in each of the above areas and the conceivability of a "Genome to hit" assembly line *in silico* (Figure 1).

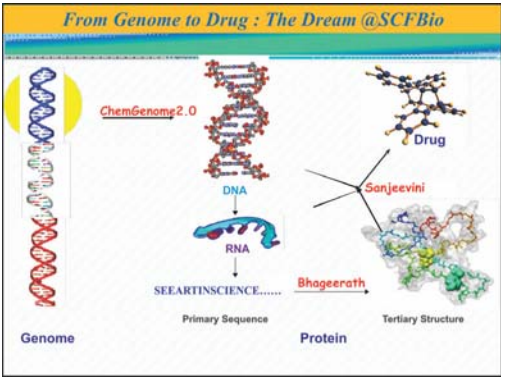


Figure 1 Genome to drug: an assembly line in the making at www.scfbio-iitd.res.in.

Genome Annotation

An organism's genome is an information resource, which can be best understood if properly annotated. The annotation bridges the gap between the genomic sequence and the biology of the organism [1]. The announcement of the human genome [2-4] changed the outlook of current biomedical research. Advances in technologies have made possible sequencing of the genomes of various organisms reliably and rapidly and this should be a routine matter soon. As on 23rd March, 2011, the total number of available complete prokaryotic genomes in NCBI (http://www.ncbi.nlm.nih.gov/) is 1491 and work is in progress on 4252 prokaryotic genomes. From a total of 1096 eukaryotic genome sequencing projects, work is in progress on 656 while 40 are complete and 400 are in assembly phase. In case of viruses, 2574 complete viral genomes are available in NCBI database. The flow of information from the sequencing projects has now shifted the focus to high quality genome annotation/analysis. The first step towards annotating a genome is to separate the genes coding for proteins from non-genes. Although there is no substitute for molecular biology for determining the exact locations of genes and control sequences in a genome, diverse computational methods have been shown to have reasonably successful predictive power [5,6]. Most of the available computational methods are knowledge based and use techniques like hidden Markov models or machine learning methods. The accuracies of these models are limited by the availability of samples of experimentally validated genes, and as typically seen in a newly sequenced genome can lead to suboptimal levels of prediction. An alternative to the knowledge based methods for gene prediction is an *ab initio* model. ChemGenome 2.0 (Figure 2) www.scfbio-iitd.res.in/chemgenome/chemgenomenew.jsp is an *ab initio* method for gene prediction in prokaryotes. It examines physicochemical properties and geometrical structures of codons [5,6] to

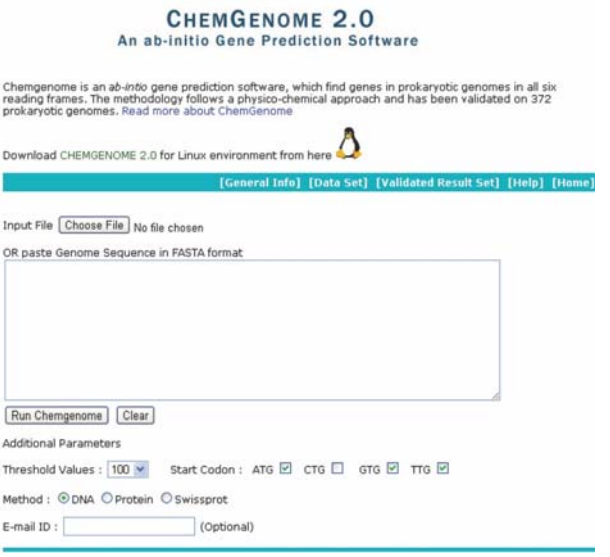


Figure 2 A front-end of the freely accessible ChemGenome 2.0 software (www.scfbio-iitd.res.in/chemgenome/chemgenomenew.jsp)

ask the fundamental question, "What is a gene?" and extracts an answer from the structural and thermodynamic properties of DNA sequences, more specifically, the hydrogen bond, the stacking and the solvation energies computed from molecular dynamics generated structures.

The methodology has been validated on 372 prokaryotic genomes and the sensitivity, specificity and correlation coefficients averaged over 356208 genes and an equal number of frame-shifted genes (non-genes) are 97.5%, 97.20% & 94.25%, respectively. A promoter prediction methodology developed on the same lines is being integrated with ChemGenome 2.0 to further reduce false positives [7, 8].

The physico-chemical model for gene evaluation and gene prediction for prokaryotic genomes has been web-enabled and is freely accessible at the SCFBio website (Figure 2).

Protein Tertiary Structure Prediction

Protein annotation is the next crucial step in the pipeline after the genome annotation. Despite the large genome sizes, only a fraction of the genome codes for proteins, particularly in higher organisms. For example, *Escherichia coli* str. K-12 substr. MG1655 (NC_000913) has 4.6Mb genome with 4494 genes, out of which 4145 are protein coding genes (covering ~ 85% of the genome), *Haemophilus influenzae* Rd KW20 chromosome (NC_000907) has 1.83Mb genome with 1798 genes out of which 1657 are protein coding genes. Human genome is ~3300Mb with ~20,500 protein coding genes and ~3000 non-protein coding RNA genes [9,10] which is about 2% of the genome. Of these proteins, only a small fraction has an experimentally determined 3D structure and well characterized functional annotation.

There are over 525,997 sequence entries in UniProtKB/Swiss-Prot[11], X-ray and NMR structures are available for only 71,794 sequence entries in the RCSB (Protein Data Bank) [12].

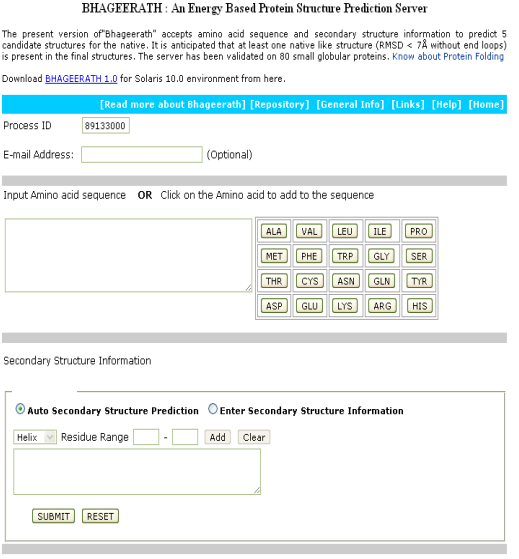


Figure 3 Front-end of the Bhageerath web server, freely accessible at http://www.scfbio-iitd.res.in/bhageerath/index.jsp

The increasing gap between sequences and structures makes it simply impossible to solve the structures of all existing proteins experimentally. The knowledge of the 3D structure of a protein can usher in tools for structure based drug discovery. Thus, a reliable computational method for protein tertiary structure prediction is desperately needed. Various computational methodologies have been developed for the prediction of tertiary structures of proteins over the past few years. These include (a) comparative modeling, (b) fold recognition or threading, (c) *ab initio* or *de novo* methods. Comparative modeling and fold recognition methods are database driven and their prediction accuracies depend on the sequence similarities realized in known structures. These methods are extremely popular, reliable and fast for protein tertiary structure prediction when a close sequence homolog exists in the database. *Ab initio* or *de novo* methods are used for predicting structure of protein sequences with no close structural homologs. We have developed two different computational protocols for predicting tertiary structures of soluble proteins. Bhageerath[13] [Figure 3] is an energy based software suite for predicting tertiary structures of small globular proteins. The protocol comprises eight different modules which use physicochemical properties of proteins and *ab initio* methodology to predict five candidates for the native from the input query sequence [14-20]. The methodology has been validated on 80 small globular proteins with $\leq$ 100 amino acids. For each of these proteins a structure within 3-7Å RMSD (root mean square deviation) from the native is predicted within few minutes to hours on a 280 processor cluster (~2 Teraflops of computing capacity).

Bhageerath-H (Figure 4)[21] is a homology *ab initio* hybrid server for protein tertiary structure prediction. The protocol identifies regions having local sequence similarity with database to generate 3D fragments which are patched with *ab initio* modeled fragments to put together complete structure of proteins. For sequences with available sequence homologs, Bhageerath-H software predicts a structure within 5 Å RMSD from the native. Work is in progress to further improve the prediction accuracies of the softwares. Figure 5 shows the performance of Bhageerath-H in CASP9 (9th Community Wide Experiment on Critical Assessment of Techniques for Protein Structure Prediction) and the progress achieved in prediction accuracy post CASP9.

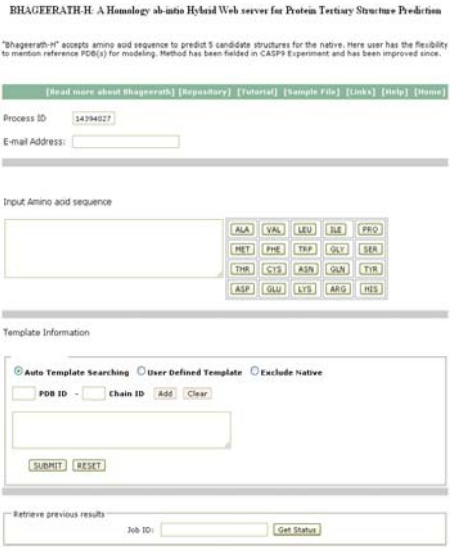


Figure 4 Front-end of Bhageerath-H web server, freely accessible at http://www.scfbio-iitd.res.in/bhageerath/bhageerath_h.jsp

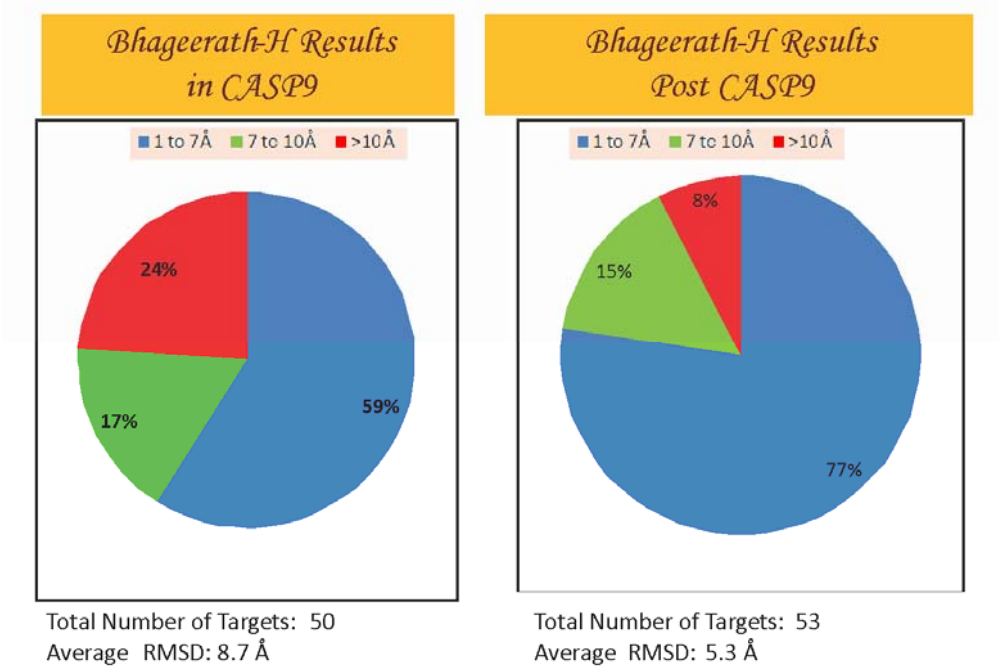


Figure 5 Performance of Bhageerath-H in CASP9 and Post CASP9.

Drug Design

The ability of biomolecules to bind to their substrates in a highly specific manner is an important characteristic in many biological processes. One of the major challenges for the CADD (Computer Aided Drug Design) techniques is to achieve this specificity i.e. specific binding of a small molecule to the biomolecular target *in vivo* at minimum cost and time, while maintaining novelty of the scaffolds and proper ADMET profiles. The two main steps in computer based drug design protocols are target identification and lead optimization. The knowledge of the structure of a biomolecule, such as a protein / nucleic acid assists in understanding the molecular level mechanism of action of drugs and lead optimization. Currently there are about 218 targets available for ~6000 FDA (U.S. Food and Drug Administration) approved drugs. A significant number of these targets are proteins and of the 218 targets, 3D structures are available for only 130. These targets can be classified into eight major biochemical classes (Table 1)[22].

Table 1: Biomolecular target for FDA approved drugs[22]

Therapeutic Target Class	Number of Known Targets	Number of 3D Structures Available
Enzyme	85	59
Receptors	56	25
Nuclear receptors	13	12
DNA, RNA and Ribosomes	8	6
Ion channels	11	6
Antibody targets	14	7
Transporters	11	4
Unknown/Miscellaneous targets	20	11

Structure of the protein target molecule can be determined either experimentally or computationally (as discussed above). The next step after target identification and structure determination is the detection of the ligand binding site. Most of the experimentally determined structures have some information on ligand binding site. In the absence of such information, detection of ligand binding site is required. We have developed an in-house fully automated active site finder program for proteins (http://www.scfbio-iitd.res.in/dock/ActiveSite_new.jsp). The active site finder, taking
3D structure of the protein target as input, detects 10 potential binding sites with 100% accuracy in capturing the actual binding (active) site[23]. The next step in CADD protocol is lead generation and optimization. A candidate molecule can either be sketched using publicly available drawing tools or searched from the small molecule libraries such as NRDBSM[22] which embeds Lipinski's rules[24a-c] or the one developed by us more recently consisting of one million small molecules(http://www.scfbio-iitd.res.in/software/nrdbsm/moleculesearchnew.jsp). For a given biomolecular target, these databases of small molecules can be scanned for identifying hit molecules based on their physico-chemical parameters and functional groups. Assessment of the candidate molecules is performed by calculating the binding energies with the target, one of the major bottlenecks being the computational time. The calculation of binding energy of a small ligand with a protein target using docking and scoring methods can take minutes, which approximates to around 12000 to 15000 hours for a dataset of one million molecules. We have developed an in-house methodology christened RASPD (A rapid identification of hit molecules for target proteins via physico-chemical descriptors) method (http://www.scfbio-iitd.res.in/software/drugdesign/raspd.jsp), for calculation of binding energy of target protein in significantly reduced amount of time[25].

The protocol screens the million compound libraries to suggest ~ 10^2 to 10^3 molecules within a minute. With the reduced dataset of hit molecules generated using RASPD, one can perform atomic level docking and scoring against the target. We have developed an in-house docking protocol "ParDOCK", an all-atom energy based Monte Carlo algorithm for protein ligand docking (http://www.scfbio-iitd.res.in/dock/pardock.jsp) [26]. ParDOCK uses BAPPL[27] for atomic level scoring of non-metallo protein ligand complexes (http://www.scfbio-iitd.res.in/software/drugdesign/bappl.jsp). The accuracy of this scoring function to predict binding free energy is high with ±1.02 kcal/mol average error and a correlation coefficient of 0.92 between the predicted and experimental binding energies for 161 protein-ligand complexes. For metallo-protein-ligand complexes with zinc metal ions, binding free energy can be calculated using BAPPL-Z (http://www.scfbio-iitd.res.in/software/drugdesign/bapplz.jsp), which shows a correlation coefficient of 0.88 for the predicted binding free energy against the experiment on a dataset of 90 zinc containing metalloprotein– ligand complexes[28]. Preddicta[20,29] methodology is used for docking and scoring candidate molecules to DNA. All of the above mentioned modules work in conduit and can be freely accessed from the drug design software christened "SANJEEVINI"[22-34] (Figure 6).

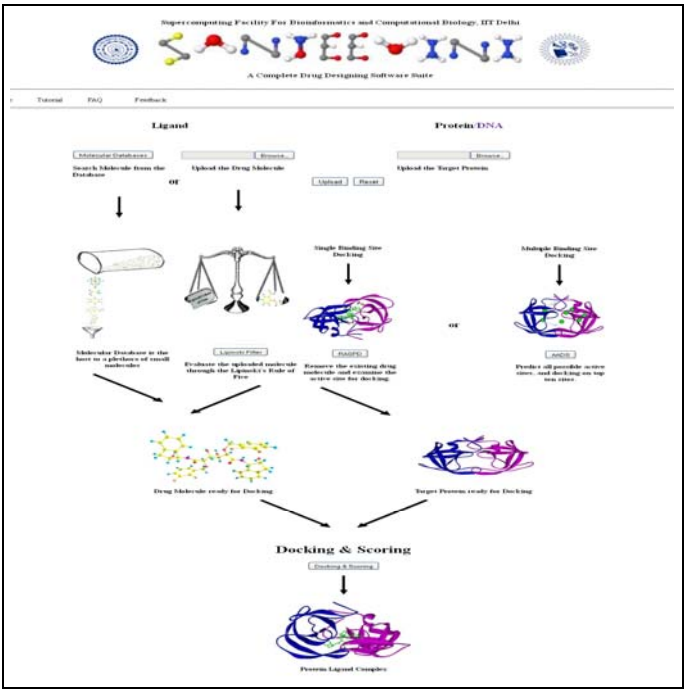


Figure 6 Front-end of SANJEEVINI, an automated drug design software freely accessible at http://www.scfbio-iitd.res.in/sanjeevini/sanjeevini.jsp.

Case Study

Here, we demonstrate the application of the above discussed *ChemGenome*, *Bhageerath* and *Sanjeevini* softwares to design lead-like molecules against Hepatitis-B virus infections.

Hepatitis B Virus

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 [35]. HBV is classified under family *Hepadnaviridae*. The HBV genome has a partially double stranded DNA genome of 3.2kb and contains 4 overlapping reading frames that encode 7 proteins[36]. The HBV genome lacks non-coding regions [37]. The HBV polymerase has reverse transcriptase activity and HBV replicates via an RNA intermediate. Eight HBV genotypes (A-H) have been described and the genotypes differ by >8% in the nucleotide sequence [37,38].

Chronic HBV Infection

The age at infection is an important factor that determines progression to chronicity; 95% of neonates, 30% of children (below six years) and less than 5% of adults exposed to HBV become chronic carriers [36]. Spontaneous clearance of chronic HBV infection occurs in a very small proportion (<1%/year); majority of those with chronic HBV infection remain infected for life [39].

HBV and Hepatocellular Carcinoma (HCC)

Chronic HBV infection increases the risk of developing HCC by over 100 fold [40]. The HBV DNA levels in the serum [41], presence of HBeAg[42], co-infection with hepatitis C virus (HCV)[43] and hepatitis D virus (HDV)[44] and alcohol consumption have been linked to progression to HCC. A few HBV proteins, genome integration and the inflammatory response represent some of the complex molecular mechanisms underlying HBV-related HCCs [45]. The lack of early diagnostic markers, the associated morbidity and mortality pose major challenges in the management of HCC. Liver transplantation remains the best of the available methods for the management of HCC. However, reinfection of the transplanted liver is almost inevitable [46]. A tyrosine kinase inhibitor, sorafenib has been promising and offers a ray of hope for HCC patients [47].

HBV Proteins

The proteins encoded by the HBV genome and their functions are summarized in Table 2.

Table 2: HBV proteins and their function

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

Antiviral Therapy for Chronic HBV Infection

Apart from interferon-alpha and peginterferon-alpha2a, five nucleoside/nucleotide analogues have been approved for anti-HBV therapy (Table 3) by United States Food and Drug Administration (US-FDA). In fact many of these nucleoside/nucleotide analogues have been approved in the last 5 years. Major challenges with currently available anti-HBV therapy include (a) rapid emergence of drug resistance (b) side effects of anti-HBV therapy, especially with interferon and (c) the lack of long lasting suppression of HBV replication in a good proportion of individuals is associated with virological breakthrough within a few months after therapy [55].

Table 3: 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	Nucleotide analogue
Tenofovir	Nucleotide analogue
Entecavir	Nucleoside analogue
Telbivudine	Nucleoside analogue

The Need for Newer Anti-HBV Drugs

The inadequacies of the currently available anti-HBV therapies necessitate the development of newer therapies. Nucleoside/nucleotide analogues that target the HBV polymerase represent majority of the available therapeutic options for HBV. Resistance to nucleoside analogues have been reported in over 65% of patients on long-term treatment [56]. The structure of most HBV proteins has not been elucidated; however, the availability of software for *in silico* prediction of protein structure, identification of active sites in proteins and automated drug design make feasible the identification of drugs that target viral proteins. It would be particularly interesting to target proteins other than the viral polymerase.

The "Genome to hit" assembly line *in silico* for HBV is given in a flowchart format below. Precore/core protein structure was determined, and 24 hit molecules were identified with good binding characteristics. Similar *in silico* experiment can be carried out with other proteins. The ensuing molecular suggestions can be tested in the laboratory, followed by further optimization, *in silico* and in *vitro* testing iteratively as steps towards lead generation.

Input the HBV Genome sequence to ChemGenome 2.0:

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

↓

ChemGenome 2.0 output

Five protein coding regions identified. Gene 2 (BP: 1814 to 2452) predicted by the ChemGenome 2.0 software encodes for the HBV precore/core protein (Gene Id:

944568). We considered this protein for 3D structure prediction. The same can be done for all of the predicted genes which do not share sequence similarity with human genes. In case of DNA targeted drug discovery, one could consider the regulatory regions of the genes exclusive to pathogen and perform the docking and scoring using DNA Ligand Docking software(26,32) http://www.scfbio-iitd.res.in/dock/dnadock.jsp

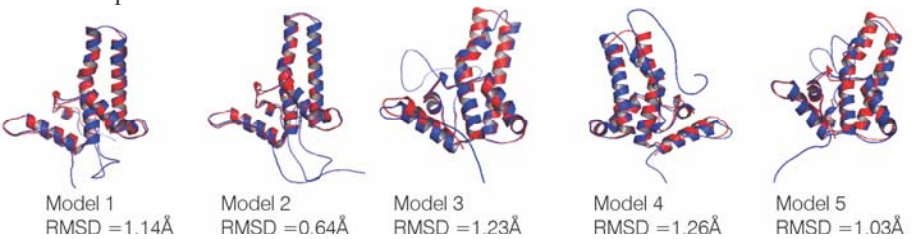
The gene sequence is translated into amino acid sequence using DNA converter (http://www.scfbio-iitd.res.in/chemgenome/genetranslator.jsp), which is then used as an input to the *Bhageerath*-H (or *Bhageerath*) for 3D structure prediction.

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

Proteins with experimentally known (X-ray/NMR) 3D structures can be used directly for active site detection and docking studies. In the absence of a known structure, computational softwares like Bhageerath-H or Bhageerath can be used. For the purposes of illustration, we have given the amino acid sequence of the protein with known crystal structure for Bhageerath-H prediction. We have tested the accuracy of the predicted structures by calculating the RMSD from the crystal structure (PDB id 1QGT). For proteins with no available structural homolog one can consider the top five energy ranked structures for further studies.

Bhageerath-H Output

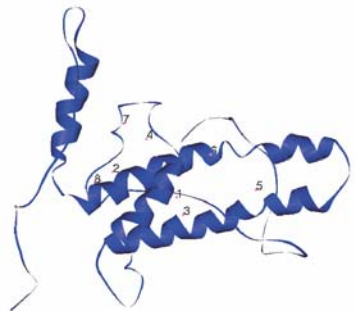
Top five energy ranked structures. The structure shown in red color is the native and that in blue is the predicted structure.



All the structures modeled by Bhageerath-H could be considered for active site identification. Here we used Model 2 as an input to the Active Site Finder.

Binding site information for the top five potential binding sites

Hydrogen Bond Donors	Hydrogen Bond Acceptors	logP of the Residues	Molar Refractivity of the Residues	ln(Volume)
11	8	31.879	353.481	8.065
9	9	36.335	499.108	7.936
10	8	32.219	444.193	7.56
13	11	35.441	353.861	7.615
5	6	25.157	413.511	7.32



Top 5 binding sites predicted in the previous step were selected and docked with Entecavir (One of the FDA approved drugs for HBV) to extract the active site information. The predicted binding sites could also be used to derive this information which is then used as input for RASPD calculation with an average cut off binding affinity to limit the number of candidates.

RASPD output

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

Out of the 2057 molecules, 40 molecules chosen randomly are given as input to ParDOCK for atomic level binding energy calculations. One could use all the 2057 molecules for atomic level docking and scoring. Out of this 40 with a cut off of (-7.0 kcal/mol), 24 molecules are seen to bind well to precore/core protein target. These molecules could be tested in the laboratory.

Atomic level docking and scoring output for the 24 hit molecules

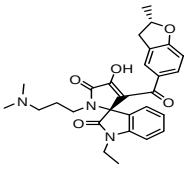
MoleculeID	Binding Energy (kcal/mol)	
0001398	-10.14	Cavity-1
0004693	-8.78	
0007684	-10.05	
0007795	-9.06	
0008386	-8.38	
0520933	-8.21	
0587461	-10.22	
0027252	-8.39	Cavity-2
0036686	-8.33	
0051126	-8.73	
0104311	-9.3	
0258280	-7.8	
0000645	-7.89	Cavity-3
0001322	-8.23	
0001895	-9.49	
0002386	-8.53	
0003092	-8.35	
0001084	-8.68	Cavity-4
0002131	-8.07	
0540853	-11.08	
1043386	-10.14	
0003623	-9.23	
0004704	-7.33	Cavity-5
0097895	-8.04	

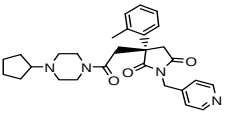
IUPAC Names of the 24 hit molecules

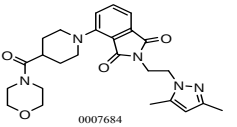
MoleculeID	IUPAC NAME
0001398	(3S)-1'-(3-dimethylaminopropyl)-1-ethyl-3'-hydroxy-4'-[(2S)-2-methyl2,3-dihydrobenzofuran-5-carbonyl]spiro[indoline-3,5'-pyrrole]-2,2'-dione
0004693	(3S)-3-[2-(4-cyclopentylpiperazin-1-yl)-2-oxo-ethyl]-3-(o-tolyl)-1-(4- pyridylmethyl)pyrrolidine-2,5-dione
0007684	2-[2-(3,5-dimethylpyrazol-1-yl)ethyl]-4-[4-(morpholine-4-carbonyl)-1- piperidyl]isoindoline-1,3-dione
0007795	(5S)-1-(3-dimethylaminopropyl)-4-(2,8-dimethylimidazo[3,2-a]pyridine-3-carbonyl)-3-hydroxy-5-(4-pyridyl)-5H-pyrrol-2-one
0008386	N,N-diethyl-2-[[(6S)-2-oxo-1-phenethyl-4-(2-thiazol-4-ylacetyl)-1,4- diazepan-6-yl]oxy]acetamide
0520933	(3R)-3-[2-[4-(furan-2-carbonyl)piperazin-1-yl]-2-oxo-ethyl]-1-(2- methoxyethyl)-3-(2-methoxyphenyl)pyrrolidine-2,5-dione
0587461	3-[(3S)-chroman-3-carbonyl]-9-methoxy-10-(thiomorpholine-4- carbonyl)-1,2,4,5-tetrahydropyrido[1,6-d][1,4]diazepin-7-
0027252	(5R)-1-(3-dimethylaminopropyl)-4-(2,8-dimethylimidazo[3,2-a]pyridine-3-carbonyl)-3-hydroxy-5-(4-pyridyl)-5H-pyrrol-2-one
0036686	2-[[(6S)-4-[3-(3,4-dimethoxyphenyl)propanoyl]-2-oxo-1-phenethyl-1,4-diazepan-6-yl]oxy]-N,N-diethyl-acetamide
0051126	[2-[4-[2-(dioxoBLAHyl)ethyl]piperazin-1-yl]-2-oxo-ethyl]BLAHdione
0104311	N,N-diethyl-2-[[(6S)-1-[(4-methoxyphenyl)methyl]-2-oxo-4-(4- pyrrolidin-1-ylbenzoyl)-1,4-diazepan-6-yl]oxy]acetamide
0258280	N,N-diethyl-2-[[(6S)-4-[(3S)-1-(2-methoxyethyl)-5-oxo-pyrrolidine-3- carbonyl]-2-oxo-1-phenethyl-1,4-diazepan-6-yl]oxy]acetamide
0000645	4-[(3R)-3-[4-(furan-2-carbonyl)piperazine-1-carbonyl]-1-piperidyl]-2- [[(2S)-tetrahydrofuran-2-yl]methyl]isoindoline-1,3-dione
0001322	(4R)-1-isopropyl-4-[3-(morpholine-4-carbonyl)-1-phenethyl-6,7- dihydro-4H-pyrazolo[4,5-c]pyridine-5-carbonyl]pyrrolidin-2-
0001895	(3R)-3-[2-[4-(2,4-dimethylphenyl)piperazin-1-yl]-2-oxo-ethyl]-1-(2- methoxyethyl)-3-(3-methoxyphenyl)pyrrolidine-2,5-dione

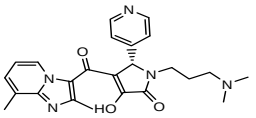
0002386	[(3S)-chroman-3-yl]-[7-[(4-pyrimidin-2-ylpiperazin-1-yl)methyl]-3,5-dihydro-2H-1,4-benzoxazepin-4-yl]methanone
0003092	ethyl 4-[(2S)-6-[2-(1,3-dioxoisoindolin-2-yl) acetyl]6-azaspiro[2.5] octane-2-carbonyl]piperazine-1-carboxylate
0001084	2-[[4-allyl-5-(4-pyridyl)-1,2,4-triazol-3-yl]sulfanyl]-1-(4-benzhydrylpiperazin-1-yl)ethanone
0002131	(3R)-1-(2-methoxyethyl)-3-(o-tolyl)-3-[2-oxo-2-[(3S)-3-(2-phenylacetyl)-1-piperidyl]ethyl]pyrrolidine-2,5-dione
0540853	2-(2-methoxyethyl)-4-[4-[4-[(2S)-tetralin-2-yl]piperazine-1-carbonyl]-1- piperidyl]isoindoline-1,3-dione
1043386	methyl 3-[3-(3,5-dimethylpyrazol-1-yl)propanoyl]-7-oxo-9-[2-(2-pyridyl) ethoxy]-1,2,4,5-tetrahydropyrido[1,2-d][1,4]diazepine-10-carboxylate
0003623	(3R)-3-[2-[4-(2-methoxyphenyl)piperazin-1-yl]-2-oxo-ethyl]-1-methyl-3-(4-phenylphenyl)pyrrolidine-2,5-dione
0004704	(3R)-1-[2-(1,3-dihydroisobenzofuran-5-yl)-1,3-dioxo-isoindolin-4-yl]-N-[[(2S)-1,4-dioxan-2-yl]methyl]-N-methyl-piperidine-3-carboxamide
0097895	N,N-diethyl-2-[[(6R)-1-[(4-methoxyphenyl)methyl]-2-oxo-4-(4-pyrrolidin-1-ylbenzoyl)-1,4-diazepan-6-yl]oxy]acetamide

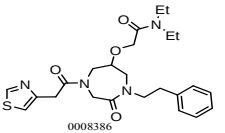
2D representations of 24 hit molecules

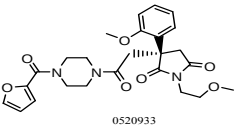

0001398

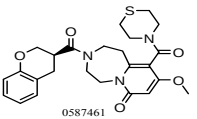

0004693

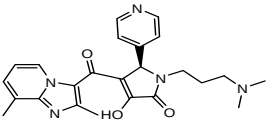

0007684

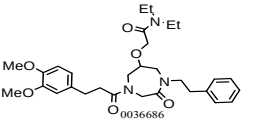

0007795

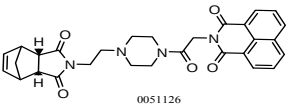

0008386

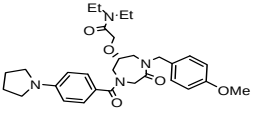

0520933

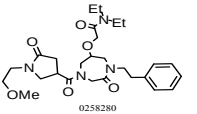

0587461

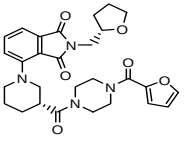

0027252

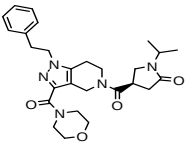

0036686

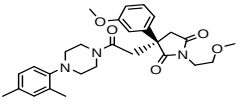

0051126

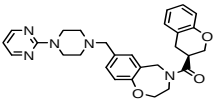

0104311

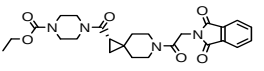

0258280

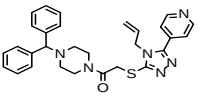

0000645

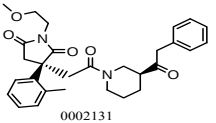

0001322

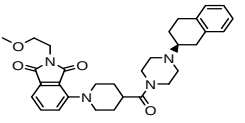

0001895

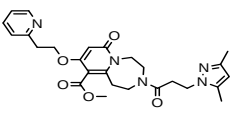

0002386

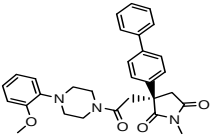

0003092

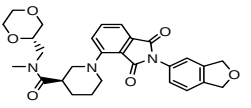

0001084

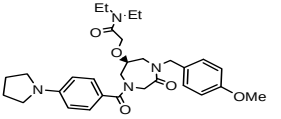

0002131


0540853


1043386


0003623


0004704


0097895

Acknowledgements

B. Jayaram is grateful to the Department of Biotechnology, Govt. of India, for Programme Support to SCFBio, IIT Delhi.

References

1. Stein, L. Genome annotation from sequence to biology, Nature Reviews 2001, 493.

2. Venter, J.C.; Adams, M.D.; Myers, E.W.; Li, P.W.; Mural, R.J.; Sutton, G.G.; Smith, H.O.; Yandell, M.; Evans, C.A.; Holt, R.A.; Gocayne, J.D.; Amanatides, P.; Ballew, R.M.; Huson, D.H.; Wortman, J.R.; Zhang, Q.; Kodira, C.D.; Zheng, X.H.; Chen, L.; Skupski, M.; Subramanian, G.; Thomas, P.D.; Zhang, J.; Gabor Miklos, G.L.; Nelson, C.; Broder, S.; Clark, A.G.; Nadeau, J.; McKusick, V.A.; Zinder, N.; Levine, A.J.; Roberts, R.J.; Simon, M.; Slayman, C.; Hunkapiller, M.; Bolanos, R.; Delcher, A.; Dew, I.; Fasulo, D.; Flanigan, M.; Florea, L.; Halpern, A.; Hannenhalli, S.; Kravitz, S.; Levy, S.; Mobarry, C.; Reinert, K.; Remington, K.; Abu-Threideh, J.; Beasley, E.; Biddick, K.; Bonazzi, V.; Brandon, R.; Cargill, M.; Chandramouliswaran, I.; Charlab, R.; Chaturvedi, K.; Deng, Z.; Di Francesco, V.; Dunn, P.; Eilbeck, K.; Evangelista, C.; Gabrielian, A.E.; Gan, W.; Ge, W.; Gong, F.; Gu, Z.; Guan, P.; Heiman, T.J.; Higgins, M.E.; Ji, R.R.; Ke, Z.; Ketchum, K.A.; Lai, Z.; Lei, Y.; Li, Z.; Li, J.; Liang, Y.; Lin, X.; Lu, F.; Merkulov, G.V.; Milshina, N.; Moore, H.M.; Naik, A.K.; Narayan, V.A.; Neelam, B.; Nusskern, D.; Rusch, D.B.; Salzberg, S.; Shao, W.; Shue, B.; Sun, J.; Wang, Z.; Wang, A.; Wang, X.; Wang, J.; Wei, M.; Wides, R.; Xiao, C.; Yan, C.; Yao, A.; Ye, J.; Zhan, M.; Zhang, W.; Zhang, H.; Zhao, Q.; Zheng, L.; Zhong, F.; Zhong, W.; Zhu, S.; Zhao, S.; Gilbert, D.; Baumhueter, S.; Spier, G.; Carter, C.; Cravchik, A.; Woodage, T.; Ali, F.; An, H.; Awe, A.; Baldwin, D.; Baden, H.; Barnstead, M.; Barrow, I.; Beeson, K.; Busam, D.; Carver, A.; Center, A.; Cheng, M.L.; Curry, L.; Danaher, S.; Davenport, L.; Desilets, R.; Dietz, S.; Dodson, K.; Doup, L.; Ferriera, S.; Garg, N.; Gluecksmann, A.; Hart, B.; Haynes, J.; Haynes, C.; Heiner, C.; Hladun, S.; Hostin, D.; Houck, J.; Howland, T.; Ibegwam, C.; Johnson, J.; Kalush, F.; Kline, L.; Koduru, S.; Love, A.; Mann, F.; May, D.; McCawley, S.; McIntosh, T.; McMullen, I.; Moy, M.; Moy, L.; Murphy, B.; Nelson, K.; Pfannkoch, C.; Pratts, E.; Puri, V.; Qureshi, H.; Reardon, M.; Rodriguez, R.; Rogers, Y.H.; Romblad, D.; Ruhfel, B.; Scott, R.; Sitter, C.; Smallwood, M.; Stewart, E.; Strong, R.; Suh, E.; Thomas, R.; Tint, N.N.; Tse, S.; Vech, C.; Wang, G.; Wetter, J.; Williams, S.; Williams, M.; Windsor, S.; Winn-Deen, E.; Wolfe, K.; Zaveri, J.; Zaveri, K.; Abril, J.F.; Guigó, R.; Campbell, M.J.; Sjolander, K.V.; Karlak, B.; Kejariwal, A.; Mi, H.; Lazareva, B.; Hatton, T.; Narechania, A.; Diemer, K.; Muruganujan, A.; Guo, N.; Sato, S.; Bafna, V.; Istrail, S.; Lippert, R.; Schwartz, R.; Walenz, B.; Yooseph, S.; Allen, D.; Basu, A.; Baxendale, J.; Blick, L.; Caminha, M.; Carnes-Stine, J.; Caulk, P.; Chiang, Y.H.; Coyne, M.; Dahlke, C.; Mays, A.; Dombroski, M.; Donnelly, M.; Ely, D.; Esparham, S.; Fosler, C.; Gire, H.; Glanowski, S.; Glasser, K.; Glodek, A.; Gorokhov, M.; Graham, K.; Gropman, B.; Harris, M.; Heil, J.;

Henderson, S.; Hoover, J.; Jennings, D.; Jordan, C.; Jordan, J.; Kasha, J.; Kagan, L.; Kraft, C.; Levitsky, A.; Lewis, M.; Liu, X.; Lopez, J.; Ma, D.; Majoros, W.; McDaniel, J.; Murphy, S.; Newman, M.; Nguyen, T.; Nguyen, N.; Nodell, M.; Pan, S.; Peck, J.; Peterson, M.; Rowe, W.; Sanders, R.; Scott, J.; Simpson, M.; Smith, T.; Sprague, A.; Stockwell, T.; Turner, R,; Venter, E; Wang, M.; Wen, M.; Wu, D.; Wu, M.; Xia, A.; Zandieh, A.; Zhu, X. The sequence of human genome, Science 2001, 1304.

3. Lander, E.S.; Linton, L.M.; Birren, B.; Nusbaum, C.; Zody, M.C.; Baldwin, J.; Devon, K.; Dewar, K.; Doyle, M.; FitzHugh, W.; Funke, R.; Gage, D.; Harris, K.; Heaford, A.; Howland, J.; Kann, L.; Lehoczky, J.; LeVine, R.; McEwan, P.; McKernan, K.; Meldrim, J.; Mesirov, J.P.; Miranda, C.; Morris, W.; Naylor, J.; Raymond, C.; Rosetti, M.; Santos, R.; Sheridan, A.; Sougnez, C.; Stange-Thomann, N.; Stojanovic, N.; Subramanian, A.; Wyman, D.; Rogers, J.; Sulston, J.; Ainscough, R.; Beck, S.; Bentley, D.; Burton, J.; Clee, C.; Carter, N.; Coulson, A.; Deadman, R.; Deloukas, P.; Dunham, A.; Dunham, I.; Durbin, R.; French, L.; Grafham, D.; Gregory, S.; Hubbard, T.; Humphray, S.; Hunt, A.; Jones, M.; Lloyd, C.; McMurray, A.; Matthews, L.; Mercer, S.; Milne, S.; Mullikin, J.C.; Mungall, A.; Plumb, R.; Ross, M.; Shownkeen, R.; Sims, S.; Waterston, R.H.; Wilson, R.K.; Hillier, L.W.; McPherson, J.D.; Marra, M.A.; Mardis, E.R.; Fulton, L.A.; Chinwalla, A.T.; Pepin, K.H.; Gish, W.R.; Chissoe, S.L.; Wendl, M.C.; Delehaunty, K.D.; Miner, T.L.; Delehaunty, A.; Kramer, J.B.; Cook, L.L.; Fulton, R.S.; Johnson, D.L.; Minx, P.J.; Clifton, S.W.; Hawkins, T.; Branscomb, E.; Predki, P.; Richardson, P.; Wenning, S.; Slezak, T.; Doggett, N.; Cheng, J.F.; Olsen, A.; Lucas, S.; Elkin, C.; Uberbacher, E.; Frazier, M.; Gibbs, R.A.; Muzny, D.M.; Scherer, S.E.; Bouck, J.B.; Sodergren, E.J.; Worley, K.C.; Rives, C.M.; Gorrell, J.H.; Metzker, M.L.; Naylor, S.L.; Kucherlapati, R.S.; Nelson, D.L.; Weinstock, G.M.; Sakaki, Y.; Fujiyama, A.; Hattori, M.; Yada, T.; Toyoda, A.; Itoh, T.; Kawagoe, C.; Watanabe, H.; Totoki, Y.; Taylor, T.; Weissenbach, J.; Heilig, R.; Saurin, W.; Artiguenave, F.; Brottier, P.; Bruls, T.; Pelletier, E.; Robert, C.; Wincker, P.; Smith, D.R.; Doucette-Stamm, L.; Rubenfield, M.; Weinstock, K.; Lee, H.M.; Dubois, J.; Rosenthal, A.; Platzer, M.; Nyakatura, G.; Taudien, S.; Rump, A.; Yang, H.; Yu, J.; Wang, J.; Huang, G.; Gu, J.; Hood, L.; Rowen, L.; Madan, A.; Qin, S.; Davis, R.W.; Federspiel, N.A.; Abola, A.P.; Proctor, M.J.; Myers, R.M.; Schmutz, J.; Dickson, M.; Grimwood, J.; Cox, D.R.; Olson, M.V.; Kaul, R.; Raymond, C.; Shimizu, N.; Kawasaki, K.; Minoshima, S.; Evans, G.A.; Athanasiou, M.; Schultz, R.; Roe, B.A.; Chen, F.; Pan, H.; Ramser, J.; Lehrach, H.; Reinhardt, R.; McCombie, W.R.; de la Bastide, M.; Dedhia, N.; Blöcker, H.; Hornischer, K.; Nordsiek, G.; Agarwala, R.; Aravind, L.; Bailey, J.A.; Bateman, A.; Batzoglou, S.; Birney, E.; Bork, P.; Brown, D.G.; Burge, C.B.; Cerutti, L.; Chen, H.C.; Church, D.; Clamp, M.; Copley, R.R.; Doerks, T.; Eddy, S.R.; Eichler, E.E.; Furey, T.S.; Galagan, J.; Gilbert, J.G.; Harmon, C.; Hayashizaki, Y.; Haussler, D.; Hermjakob, H.; Hokamp, K.; Jang, W.; Johnson, L.S.; Jones, T.A.; Kasif, S.; Kaspryzk, A.; Kennedy, S.; Kent, W.J.; Kitts, P.; Koonin, E.V.; Korf, I.; Kulp, D.; Lancet, D.; Lowe, T.M.; McLysaght, A.; Mikkelsen, T.; Moran, J.V.; Mulder, N.; Pollara, V.J.; Ponting, C.P.; Schuler, G.; Schultz, J.; Slater, G.; Smit, A.F.; Stupka, E.; Szustakowski, J.; Thierry-Mieg, D.; Thierry-Mieg,

J.; Wagner, L.; Wallis, J.; Wheeler, R.; Williams, A.; Wolf, Y.I.; Wolfe, K.H.; Yang, S.P.; Yeh, R.F.; Collins, F.; Guyer, M.S.; Peterson, J.; Felsenfeld, A.; Wetterstrand, K.A.; Patrinos, A.; Morgan, M.J.; de Jong, P.; Catanese, J.J.; Osoegawa, K.; Shizuya, H.; Choi, S.; Chen, Y.J. International human genome sequencing consortium, Nature 2001, 860.

4. International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome, Nature 2004, 931.

5. Dutta, S.; Singhal, P.; Agrawal, P.; Tomer, R.; Kritee K.; Khurana, E; Jayaram, B. A physicochemical model for analyzing DNA sequences, J. Chem. Inf. Model 2006, 78.

6. Singhal, P.; Jayaram, B.; Dixit, S. B.; Beveridge, D. L. Molecular dynamics based physicochemical model for gene prediction in prokaryotic genomes, Biophysical J. 2008, 4173.

7. Khandelwal, G.; Jayaram, B. A phenomenological model for predicting melting temperatures of DNA sequences, PLos ONE 2010, e12433.

8. Khandelwal, G.; Jayaram, B.; Beveridge, D. L. Finger prints for functional units in genomes, Manuscript in preparation 2011.

9. Clamp, M.; Fry, B.; Kamal, M.; Xie, X.; Cuff, J.; Lin, M. F.; Kellis, M.; Lindblad-Toh, K.; Lander, E. S. Distinguishing protein-coding and noncoding genes in the human genome, PNAS 2007, 19428.

10. Wright, M. W.; Bruford, E. A. Naming 'junk': Human non-protein coding RNA (ncRNA) gene nomenclature, Human Genomics 2011, 90.

11. Boeckmann, B.; Bairoch, A.; Apweiler, R.; Blatter, M.C.; Estreicher, A.; Gasteiger, E.; Martin, M. J.; Michoud, K.; O'Donovan, C.; Phan, I.; Pilbout, S.; Schneider, M. The Swiss-Prot Protein knowledgebase and its supplement TrEMBL in 2003, Nucleic Acids Res 2003, 365.

12. www.pdb.org. Berman H. M.; Westbrook J.; Feng Z.; Gilliland G.; Bhat T.N.; Weissig H.; Shindyalov I.N.; Bourne P.E. The protein data bank. Nucleic Acids Res 2000, 235.

13. Jayaram, B.; Bhushan, K.; Shenoy, S.R.; Narang, P.; Bose, S.; Aggarwal, P.; Sahu, D.; Pandey, V. Bhageerath : An energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins, Nucleic Acids Res. 2006, 6195.

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

15. Narang, P.; Bhushan, K.; Bose, S.; Jayaram, B. Protein structure evaluation using an all-atom energy based empirical scoring function, J. Biomol. Struc. Dyn. 2006, 385.

16. Mittal, A.; Jayaram, B.; Shenoy, S. R.; Bawa, T. S. A stoichiometry driven universal spatial organization of backbones of folded proteins : are there Chargaff's rules for protein folding?, J. Biomol. Struc Dyn. 2010, 133.

17. Mittal, A.; Jayaram, B. Backbones of folded proteins reveal novel invariant amino acid neighborhoods, J. Biomol. Struc. Dyn. 2011, 443.

18. Mittal, A; Jayaram, B. The newest view on protein folding: Stoichiometric and spatial unity in structural and functional diversity, J. Biomol. Struc. Dyn. 2011, 669.

19. Shenoy, S. R.; Jayaram, B. Proteins: Sequence to structure and function – Current status, CPPS 2010, 498.

20. Jayaram, B.; Dhingra, P.; Lakhani, B.; Shekhar, S. Bhageerath: Attempting the near impossible – Pushing the frontiers of atomic models for protein tertiary structure prediction, J Chemical Sciences, in press 2012.

21. Lakhani, B.; Dhingra, P.; Jayaram, B. Bhageerath-H, An *ab initio* – homology hybrid server for protein tertiary structure prediction, Manuscript in preparation, 2012.

22. Shaikh, S.; Jain. T.; Sandhu, G.; Latha, N.; Jayaram, B. From drug target to leads - Sketching a physicochemical pathway for lead molecule design *in silico*, Current Pharmaceutical Design 2007, 3454.

23. Singh, T.; Biswas, D.; Jayaram, B. A robust active site identification protocol based on physico-chemical descriptors lining the cavities in proteins, J. Chem. Inf. Model 2011, 2515.

24. (a) Lipinski C A. Lead- and drug-like compounds: the rule-of-five revolution, Drug Discov. Today: Tech 2004, 337.

(b) Lipinski, C. A.; Lombardo, F.; Dominy, B. W.; Feeney, P. J. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv Drug Delivery Rev 2001, 3.

(c) Lipinski C. A. Drug-like properties and the causes of poor solubility and poor permeability, J. Pharmacol Toxicol. Methods 2000, 235.

25. Mukherjee, G.; Jayaram, B. A rapid scoring methodology based on physico-chemical descriptors of small molecules (RASPD) for identifying hits against a protein target, Manuscript in preparation.

26. Gupta, A.; Gandhimathi, A.; Sharma, P.; Jayaram, B. ParDOCK: An all atom energy based Monte Carlo docking protocol for protein-ligand complexes, Protein and Peptide Letters 2007, 632.

27. Jain, T.; Jayaram, B. All atom energy based computational protocol for predicting binding affinities of protein–ligand complexes, FEBS Letters 2005, 6659.

28. Jain, T.; Jayaram, B. Computational protocol for predicting the binding affinities of zinc containing metalloprotein–ligand complexes, Proteins: Structure, Function, and Bioinformatics, 2007, 1167.

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

30. Shaikh, S.; Jayaram. B. A swift all-atom energy-based computational protocol to predict DNA−ligand binding affinity and ΔTm, J. Med. Chem. 2007, 2240.

31. Jayaram, B; Latha, N.; Jain, T.; Sharma, P.; Gandhimathi, A.; Pandey, V. S. Sanjeevini: A comprehensive active site directed lead design software, Indian Journal of Chemistry-A 2006, 1834.

32. Kalra, P.; Reddy, V.; Jayaram, B. Free energy component analysis for drug design: A case study of HIV-1 protease−inhibitor binding, J. Med. Chem. 2001, 4325.

33. Mukherjee, G.; Patra, N.; Barua, P. J.; Jayaram, B. A fast empirical GAFF compatible partial atomic charge assignment scheme for modeling interactions of small molecules with biomolecular targets, J. Computational Chemistry 2011, 893.

34. Jayaram, B.; Singh, T.; Fenley, M. Drug-DNA interactions: A Theoretical Perspective, Editors, Dr. Meni Wanunu & Prof. Yitzhak Tor, 2011, Ch-14, 317, CRC Press.

35. Kao, J. H.; Chen, D. S. Global control of hepatitis B virus infection, Lancet Infect. Dis. 2002, 395.

36. Lee, W. M. Hepatitis B virus infection, N. Engl. J. Med. 1997, 1733.

37. Kidd-Ljunggren, K.; Miyakawa, Y.; Kidd, A. H. Genetic variability in hepatitis B viruses, J. Gen. Virol. 2002, 1267.

38. Arauz-Ruiz, P.; Norder, H.; Robertson, B. H.; Magnius, L. O. Genotype H: a new Amerindian genotype of hepatitis B virus revealed in Central America, J. Gen. Virol. 2002, 2059.

39. McMahon, B. The natural history of chronic hepatitis B virus infection, Hepatology 2009, S45.

40. Beasley, R. P. Hepatitis B virus. The major etiology of hepatocellular carcinoma, Cancer 1988, 1942.

41. Chen, C. J.; Yang. H. I. Su, J.; Jen, C. L.; You, S. L.; Lu, S. N.; Huang, G. T.; Iloeje, U. H.; REVEAL - HBV Study Group. Risk of hepatocellular carcinoma across a biological gradient of serum hepatitis B virus DNA level, JAMA 2006, 65.

42. Yang, H. I.; Lu, S. N.; Liaw, Y. F.; You, S. L.; Sun, C. A.; Wang, L. Y.; Hsiao, C. K.; Chen, P. J.; Chen, D. S.; Chen, C. J. Hepatitis B e antigen and the risk of hepatocellular carcinoma, N. Engl. J. Med. 2002, 168.

43. Fattovich, G.; Giustina, G.; Christensen, E.; Pantalena, M.; Zagni, I.; Realdi, G.; Schalm, S.W. Influence of hepatitis delta virus infection on morbidity and mortality in compensated cirrhosis type B, The European Concerted Action on Viral Hepatitis (Eurohep), Gut 2000, 420.

44. Benvegnu, L.; Fattovich, G.; Noventa, F.; Tremolada, F.; Chemello, L.; Cecchetto, A.; Alberti, A. Concurrent hepatitis B and C virus infection and risk of hepatocellular carcinoma in cirrhosis. A prospective study, Cancer 1994, 2442.

45. Brechot, C.; Kremsdorf, D.; Soussan, P.; Pineau, P.; Dejean, A.; Paterlini-Brechot, P.; Tiollais, P. Hepatitis B virus (HBV)-related hepatocellular carcinoma (HCC): molecular mechanisms and novel paradigms, Pathol. Biol. (Paris) 2010, 278.

46. Jiang, L.; Yan, L. N. Current therapeutic strategies for recurrent hepatitis B virus infection after liver transplantation, World J. Gastroenterol. 2010, 2468.

47. Guan, Y. S.; He, Q. Sorafenib: activity and clinical application in patients with hepatocellular carcinoma, Expert Opin. Pharmacother. 2011, 303.

48. Kann, M; Gerlich, W. Structure and molecular virology. In: Zuckerman A. J.; Thomas, H. C. Viral hepatitis, second edition. London, UK, Churchill Livingstone
1998, 77.

49. Mahoney, F. J. Update on diagnosis, management, and prevention of hepatitis B virus infection, Clin. Microbiol. Rev. 1999, 351.

50. Tiollais, P.; Pourcel, C.; Dejean, A. The hepatitis B virus, Nature 1985, 489.

51. Neurath, A. R.; Kent, S. B.; Strick, N.; Parker, K. Identification and chemical synthesis of a host cell receptor binding site on hepatitis B virus, Cell 1986, 429.

52. Hsu, Y. S.; Chien, R. N.; Yeh, C. T.; Sheen, I. S.; Chiou, H. Y.; Chu, C. M.; Liaw, Y. F. Long-term outcome after spontaneous HBeAg seroconversion in patients with chronic hepatitis B, Hepatology 2002, 1522.

53. Spandau, D. F.; Lee, C. H. Trans-activation of viral enhancers by the hepatitis B virus X protein, J. Virol. 1988, 427.

54. Zoulim, F.; Saputelli, J.; Seeger, C. Woodchuck hepatitis virus X protein is required for viral infection *in vivo*, J. Virol. 1994, 2026.

55. Liaw, Y. F. Antiviral therapy of chronic hepatitis B: opportunities and challenges in Asia, J. Hepatol. 2009, 403.

56. Lok, A. S.; Lai, C. L.; Leung, N.; Yao, G. B.; Cui, Z. Y.; Schiff, E. R.; Dienstag, J. L.; Heathcote, E. J.; Little, N. R.; Griffiths, D. A.; et. al. Long-term safety of lamivudine treatment in patients with chronic hepatitis B, Gastroenterology 2003, 1714.