From Gene to Drug: A Proof of Concept for a Plausible Computational Pathway

Sandhya Shenoy, Jayaram, B, Latha, N, Pooja Narang, Tarun Jain, Kumkum Bhushan, Saher Afshan Shaikh, Surojit Bose, Pankaj Sharma, Poonam Singhal, Gandhimathi, A., Praveen Agrawal, Vidhu Pandey, Samrat Dutta, Gurvisha Sandhu, Anuj Gupta, Shashank Shekhar and Shailesh Tripathi

Department of Chemistry &
Supercomputing Facility for Bioinformatics & Computational Biology,
Indian Institute of Technology, Hauz Khas, New Delhi-110016, India.

Abstract

The last decade has witnessed an exponential growth of sequence information in the field of biological macromolecules such as proteins and nucleic acids and their interactions with other molecules. Computational analyses for structure-function predictions based on such information are increasingly becoming an essential and integral part of modern biology. With rapid advances in the area, there is a growing need to develop efficient versatile bioinformatics software packages, which are hypothesis driven. 'Gene to Drug' is an attempt in this pursuit and an integration of heterologous applications of different technologies developed in-house and their translation into in silico products that cater to a majority of bioinformatics applications. This paper briefly presents the science behind Gene to Drug and how grid services and high performance computing platforms can be harnessed to bridge the gap between biomolecular sequence, structure and function.

1. Introduction

The world wide genome sequencing efforts and the concurrent developments in scientific software implementations on massively parallel computer architectures grant us the opportunity to dream that all genes and proteins in a cell would be identified, their spatial and temporal connections understood, their functions established and their structures determined so that drug design could be undertaken against suitable targets to develop individualized medicine almost in an automated way (Genome → Gene → Protein → Drug).

Currently without the help of any database, it is difficult to ascertain whether a given DNA sequence is a gene, and if it is a gene, what is the likely three-dimensional structure of its protein product. Also, the present drug design softwares fall short of expectations even if the structures of drug targets (proteins/DNA) are known.

Pursuing the dream of creating new science and converting it into technology and products useful to society, we have been developing a wide array of "Gene to Drug" software tools (www.scfbio-iitd.res.in), the major contributions being *ChemGenome* for genome analysis, *Bhageerath* for protein structure prediction and *Sanjeevini* for DNA/protein directed drug design.

2. Softwares Developed

2.1 ChemGenome

A Novel Genome Analysis Software Suite (http://www.scfbio-iitd.res.in/chemgenome2)

Based on DNA energetics, a physico-chemical model has been developed for whole genome analysis to identify protein coding regions in a genome. Accuracy as verified on 331 prokaryotic genomes exceeds 90% with a gene prediction rate of 95% in 206 genomes. Also, relatively high accuracy is observed in precise gene prediction (i.e. both start and stop site identification) where experimentally validated reliable data sets are available. In addition, analysis of all the 16 chromosomes of *S.cerevisiae*, a eukaryote, shows the generality of

Proceedings of the Sixth International Conference on Intelligent Systems Design and Applications (ISDA'06)
0-7695-2528-8/06 $20.00 © 2006 IEEE

the methodology without the need for genome dependent database training.

The physiochemical properties of DNA considered in *ChemGenome* [1, 2] namely hydrogen bonding energy and stacking energy together with the rule of conjugates [3] seem to embed sufficient information for gene identification. The physicochemical model employs a three-dimensional (3-D) vector to represent double-helical deoxyribonucleic acid (DNA) base sequences, with each dimension denoting one facet of DNA recognition by proteins. Each of the 64 trinucleotides is assigned three coordinates, *x, y, z*, in the interval of -1 to +1 *(x, y, z* € [-1, +1], corresponding to the three proposed chemical properties of DNA. For a given DNA sequence (genome segment), the resultant vector is found by accumulating the *x, y,* and *z* components of the individual codons ($X=\sum x$, $Y=\sum y$, $Z= \sum z$). The orientation of this resultant vector from the origin is given by the direction cosines. Results on 331 genomes indicate that the vectors for genes and nongenes segregate quite nicely and are separated by a universal plane passing through the unit sphere [1], (Fig. 1). The relatively high sensitivity of the method, coupled with the database independent and hypothesis driven nature of the algorithm should make it a useful tool for studies in structural genomics.

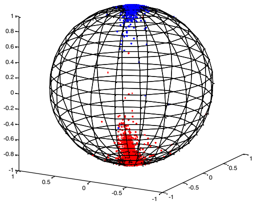


Figure 1a. Distribution of gene (north pole) and nongene (south pole) direction vectors for genome of *E.coli* (NC_000913), (sensitivity – 0.99; specificity – 0.98; correlation coefficient – 0.97).

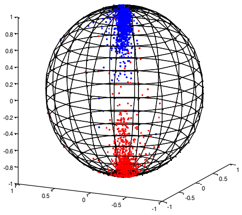


Figure 1b. Distribution of gene (north pole) and nongene (south pole) direction vectors for genome of *B. subtilis* (NC_ 000964), (sensitivity – 1.00; specificity – 0.84; correlation coefficient – 0.82).

2.2 Bhageerath

A Protein Structure Prediction Software Suite (http://www.scfbio-iitd.res.in/bhageerath/index.jsp)

Combining bioinformatics tools and *ab initio* methodologies, considerable progress has been made towards a pathway that is computationally expeditious for tertiary structure prediction of small proteins. The software suite *Bhageerath* comprises eight modules configured to function independently or in a conduit. Starting with amino acid sequence and secondary structure information (helix/sheet/loop) of a protein in the first module, multiple three dimensional atomic level structures [4-6] are generated sampling the conformational space of the loop dihedrals in the second; in the third module a set of biophysical filters (persistence length, radius of gyration etc.) are applied which are designed to screen the trial structures to reduce the sample size. The resultant structures are refined in the fourth module by a Monte Carlo sampling in dihedral space to remove steric clashes / overlaps in 3-D space. An atomic level energy optimization is carried out in the fifth module and the structures scored based on energy in the sixth. Module seven reduces the probable candidates based on the protein regularity index of the φ and ψ dihedral values and module eight further reduces the structures selected to 10 using topological equivalence criterion and accessible surface area. Thus millions of possible structures for a given protein sequence are brought down to 10 candidate structures with the possibility of bracketing the native in these 10. Results on a few small globular helical proteins (Fig. 2) have shown that the native like folds in the root mean square deviation (RMSD) range of 3-5 Å are captured in the best 10 structures energy-wise in all the cases without exception. The "needle in a hay stack" problem is thus reduced to choosing the best candidate from among the 10 lowest energy structures. Comparison with existing bioinformatics tools suggests the performance of the present methodology to be satisfactory and useful particularly when the database is deficient in sequence homologues.

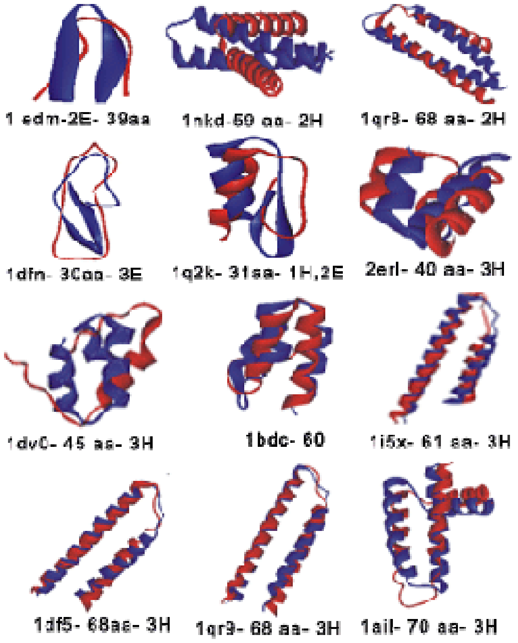


Figure 2. The lowest RMSD structure (red color) emerging from the *Bhageerath* pathway superimposed on the corresponding native (experimental/X-ray; blue color) structure for each of the globular test proteins. Results for other 38 systems are made available at the web site.

Preliminary work on alpha / beta systems has yielded encouraging results. Currently, the expected prediction time with Bhageerath web server for a 2-helix system (with one loop in between) is ~4-5 min while for a 3 helix system (with two loops in between), it is ~2-3 hours on a 32 processor cluster. Attempts to extend the methodology to larger systems with reduction in computational times are in progress.

2.3. Sanjeevini

Active-Site Directed Drug Lead Molecule Design Software Suite (http://www.scfbio-iitd.res.in/research/drugdesign.htm)

A comprehensive software suite, *Sanjeevini* [7, 8] has been developed for target directed drug design. The software is system-independent and accomplishes lead-like molecule design based on binding affinities (the standard free energies of binding between target macromolecule and the candidate drug). The computational pathway paves way expressly towards lead molecule design starting with framing innumerable number of known or new candidate molecules out of a non-redundant small but versatile set of building blocks called templates. These candidates are then screened for oral bioavailability, followed by geometry optimization, determination of partial atomic charges and assignment of other force field parameters. The candidates thus prepared, are then docked in the active site of a given biological target using a Monte Carlo method and the interaction/binding energy is estimated using an all-atom energy based function. This binding affinity information is employed to judge whether the candidate molecule is lead-like. In propitious cases, molecular dynamics simulations can be performed with explicit solvent and salt on the biomolecular target, the candidate and the complex followed by a rigorous analysis of the binding free energy for further optimization. The methodology integrates tools of computational chemistry and molecular biophysics, some known and some new, to generate candidate molecules to any given target and to screen them based on binding affinities. The protocols developed draw strength from first principles and assure rigor, transferability and potential for automation. A proof of concept of the methodology developed obtained with the estrogen receptor target is represented in Fig. 3.

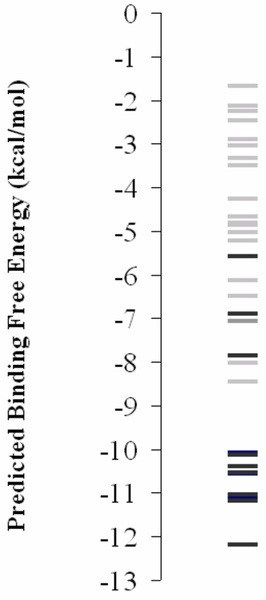


Figure 3. A typical separation achieved by *Sanjeevini* methodology of known marketed drugs from non-drugs for a biological target (Estrogen receptor in this case).

As a part of Sanjeevini pathway, to enable a rapid scanning of candidate molecules, empirical energy based scoring functions have been developed for both, protein-drug [9] as well as DNA-drug [10-11] complexes. The scoring function for protein-drug complexes shows a high correlation (coefficient ~ 0.92) with experimental binding affinity data on 161 protein-drug complexes comprising 55 unique drug targets without the need for system dependent parameterization. The energy function for DNA-drug complexes shows correlation coefficients of 0.95 and 0.97 against experimental data on binding free energies and changes in melting temperatures upon DNA-drug binding for 50 DNA-drug complexes. These scoring function protocols have been web enabled for free usage by the scientific community.

3. Products Developed

3.1 ChemGenome

ChemGenome can be accessed at www.scfbio-iitd.res.in/chemgenome2. The web server offers platform-independent advantages to the user, requires no installation and presents a friendly browser interface for data entry and display. Users can input the whole genome sequence or part of genome of an organism. The sequence can be uploaded or alternatively pasted or typed into the query window of the browser. A tabular output displays the strand name, and the predicted gene boundaries.

3.2 Bhageerath

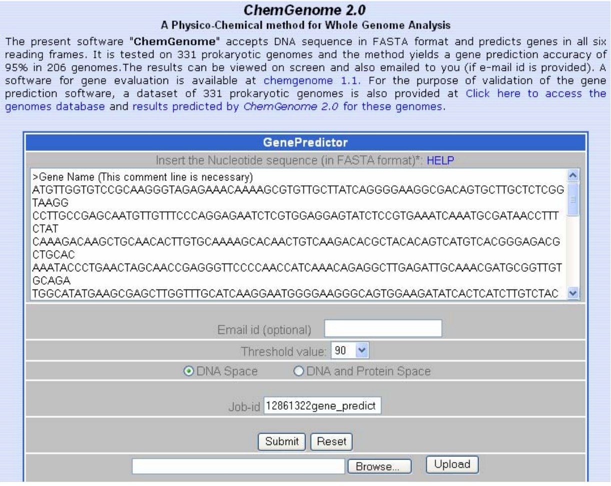
Bhageerath is a fully automated web enabled (http://www.scfbio-iitd.res.in/bhageerath/index.jsp) *ab initio* protein structure prediction software suite that is made available through a convenient user interface which returns 10 candidate structures for a given protein query sequence. A click on the *Bhageerath* server opens into a window wherein a user can paste a query protein sequence in FASTA format. The current version supports continuous sequences up to 100 amino acids. The user is prompted for amino acid range as secondary structural input. Upon submission the user receives a unique job id for his/her sequence. User has an option to provide an email ID to receive an output link, which contains the results. The expected prediction time with *Bhageerath* web server depends on the length of the sequence, number of secondary structure elements and the number of structures accepted by biophysical filters for processing the energetics of each trial structure at the atomic level. It is currently able to process around 4-5 normally sized jobs per day on a 32 processor cluster.

3.3 Sanjeevini

Sanjeevini (http://www.scfbio-iitd.res.in/research/drugdesign.htm) is a comprehensive active site-directed lead molecule design protocol. The scoring protocols associated with Sanjeevini for ranking candidate molecules for any biomolecular target are web-enabled for free access by the scientific community. These include BAPPL sever (http://www.scfbio-iitd.res.in/software/drugdesign/bappl.jsp) that computes the binding free energy of a protein-ligand complex, PreDDICTA (http://www.scfbio-iitd.res.in/software/drugdesign/preddicta.jsp) which calculates the Drug-DNA interaction energy and Lipinski Filter (http://www.scfbio-iitd.res.in/utility/LipinskiFilters.jsp) that checks whether a candidate satisfies Lipinski's rules.

4. Conclusions and Perspectives

"Gene to Drug" is the culmination of a functional prototype software suite completely based on *ab initio* methodologies, developed in-house. *ChemGenome* is a novel gene identification model based on physicochemical properties of the double helix. The relatively high sensitivity of the method, coupled with database independent and hypothesis driven nature of the algorithm should make it a useful tool for studies in structural genomics. *Bhageerath* is energy based computational web server for tertiary structure predictions of proteins. The validation of the computational protocol on a few globular proteins shows that that the web server predicts a structure within an RMSD of 3-5 Å with respect to the native in the 10 lowest energy structures. *Sanjeevini* is a comprehensive drug design software suite, which could sort out drugs from non-drugs in an automated mode based on binding affinity calculations. The key driving forces for current



ChemGenome 2.0

A Physico-Chemical method for Whole Genome Analysis

The present software "**ChemGenome**" accepts DNA sequence in FASTA format and predicts genes in all six reading frames. It is tested on 331 prokaryotic genomes and the method yields a gene prediction accuracy of 95% in 206 genomes.The results can be viewed on screen and also emailed to you (if e-mail id is provided). A software for gene evaluation is available at chemgenome 1.1. For the purpose of validation of the gene prediction software, a dataset of 331 prokaryotic genomes is also provided at Click here to access the genomes database and results predicted by *ChemGenome 2.0* for these genomes.

GenePredictor

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

```
>Gene Name (This comment line is necessary)
ATGTTGGTGTCCGCAAGGGTAGAGAAACAAAAGCGTGTTGCTTATCAGGGGAAGGCGACAGTGCTTGCTCTCGG
TAAGG
CCTTGCCGAGCAATGTTGTTTCCCAGGAGAATCTCGTGGAGGAGTATCTCCGTGAAATCAAATGCGATAACCTTT
CTAT
CAAAGACAAGCTGCAACACTTGTGCAAAAGCACAACTGTCAAGACACGCTACACAGTCATGTCACGGGAGACG
CTGCAC
AAATACCCTGAACTAGCAACCGAGGGTTCCCCAACCATCAAACAGAGGCTTGAGATTGCAAACGATGCGGTTGT
GCAGA
TGGCATATGAAGCGAGCTTGGTTTGCATCAAGGAATGGGGAAGGGCAGTGGAAGATATCACTCATCTTGTCTAC
```

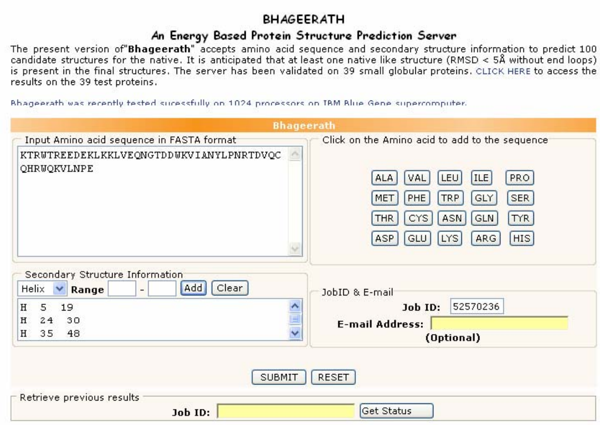
Email id (optional)

Threshold value: 90

● DNA Space ○ DNA and Protein Space

Job-id 12861322gene_predict

Figure 4. Snap shot of the front end of the *ChemGenome* software server (http://www.scfbio-iitd.res.in/chemgenome/chemgenomenew.jsp)



BHAGEERATH

An Energy Based Protein Structure Prediction Server

The present version of"**Bhageerath**" accepts amino acid sequence and secondary structure information to predict 100 candidate structures for the native. It is anticipated that at least one native like structure (RMSD < 5Å without end loops) is present in the final structures. The server has been validated on 39 small globular proteins. CLICK HERE to access the results on the 39 test proteins.

Bhageerath was recently tested sucessfully on 1024 processors on IBM Blue Gene supercomputer.

Bhageerath

Input Amino acid sequence in FASTA format

```
KTRWTREEDEKLKKLVEQNGTDDWKVIANYLPNRTDVQC
QHRWQKVLNPE
```

Click on the Amino acid to add to the sequence

Secondary Structure Information

Helix **Range** -

```
H  5   19
H  24  30
H  35  48
```

JobID & E-mail

Job ID: 52570236

E-mail Address:

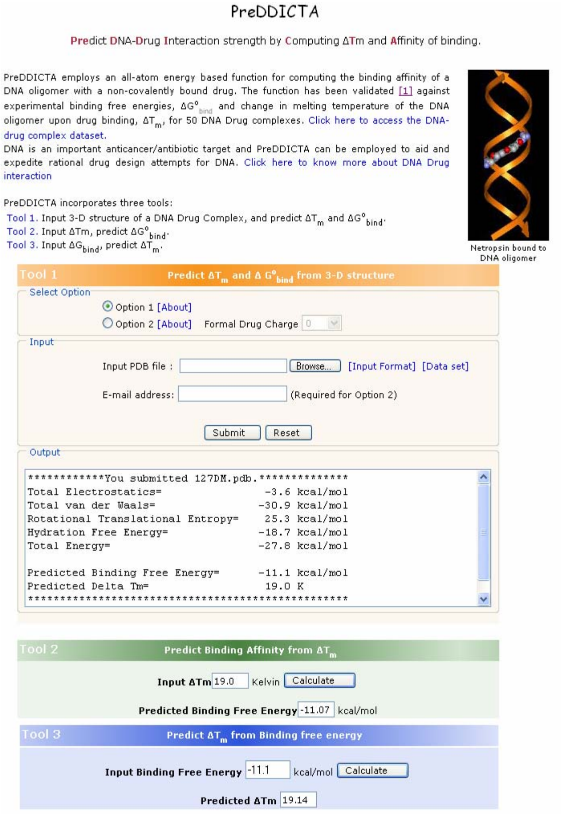
(Optional)

Retrieve previous results

Job ID:

Figure 5. Snap shot of the front end of the *Bhageerath* software server (http://www.scfbio-iitd.res.in/bhageerath/index.jsp)

day *in silico* drug design endeavors are the availability of structural information of the targets, emergence of reliable energy functions and accessibility of high speed computers [12-14]. *Sanjeevini* is a modest demonstration of these developments as an integrated computational pathway targeted to lead molecule discovery.



PreDDICTA

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

PreDDICTA employs an all-atom energy based function for computing the binding affinity of a DNA oligomer with a non-covalently bound drug. The function has been validated [1] against experimental binding free energies, ΔG^{o}_{bind} and change in melting temperature of the DNA oligomer upon drug binding, ΔT_m, for 50 DNA Drug complexes. Click here to access the DNA-drug complex dataset.

DNA is an important anticancer/antibiotic target and PreDDICTA can be employed to aid and expedite rational drug design attempts for DNA. Click here to know more about DNA Drug interaction

PreDDICTA incorporates three tools:

Tool 1. Input 3-D structure of a DNA Drug Complex, and predict ΔT_m and ΔG^{o}_{bind}.
Tool 2. Input ΔTm, predict ΔG^{o}_{bind}.
Tool 3. Input ΔG_{bind}, predict ΔT_m.

Netropsin bound to DNA oligomer

Tool 1 **Predict ΔT_m and Δ G^{o}_{bind} from 3-D structure**

Select Option

● Option 1 [About]
○ Option 2 [About] Formal Drug Charge 0

Input

Input PDB file : [Input Format] [Data set]

E-mail address: (Required for Option 2)

Output

```
************You submitted 127DM.pdb.**************
Total Electrostatics=              -3.6 kcal/mol
Total van der Waals=              -30.9 kcal/mol
Rotational Translational Entropy=  25.3 kcal/mol
Hydration Free Energy=            -18.7 kcal/mol
Total Energy=                     -27.8 kcal/mol

Predicted Binding Free Energy=    -11.1 kcal/mol
Predicted Delta Tm=                19.0 K
**************************************************
```

Tool 2 **Predict Binding Affinity from ΔT_m**

Input ΔTm 19.0 Kelvin

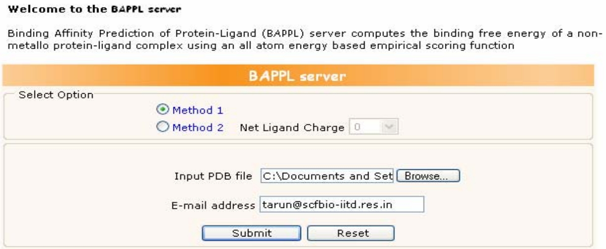
Predicted Binding Free Energy -11.07 kcal/mol

Tool 3 **Predict ΔT_m from Binding free energy**

Input Binding Free Energy -11.1 kcal/mol

Predicted ΔTm 19.14

Figure 6. Snap shot of the front end of PreDDICTA server (http://www.scfbio-iitd.res.in/software/drugdesign/preddicta.jsp)



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

BAPPL server

Select Option

● Method 1
○ Method 2 Net Ligand Charge 0

Input PDB file C:\Documents and Set

E-mail address tarun@scfbio-iitd.res.in

Figure 7. Snap shot of the front end of BAPPL server (http://www.scfbio-iitd.res.in/software/drugdesign/bappl.jsp)

The dream is to realize ‘Gene to Drug’ as a comprehensive bioinformatics application in a completely stand-alone automated pipeline and this appears feasible in the foreseeable future.

References

1. S. Dutta, P. Singhal, P. Agrawal, R. Tomer, Kritee, E. Khurana, B. Jayaram, A Physico-Chemical Model for Analyzing DNA sequences, 2006, *J. Chem. Inf. Mod.*, 46(1), 78-85.
2. P. Singhal, S. Dutta, B. Jayaram, A novel whole genome analysis software suite based on DNA energetics, 2006, Manuscript Communicated (chemgenome2).
3. B. Jayaram, Beyond the wobble: the rule of conjugates, 1997, *J. Mol. Evol.* 45, 704-705.
4. P. Narang, K. Bhushan, S. Bose, B. Jayaram, Protein structure evaluation using all atom energy based empirical scoring function, 2006, J. Biomol. Struct. Dyn. 23, 385-4006.
5. P. Narang, K. Bhushan, S. Bose, B. Jayaram, A computational pathway for bracketing native-like structures for small alpha helical globular proteins, 2005, *Phys. Chem. Chem. Phys*., 7, 2364.
6. B. Jayaram, Kumkum Bhushan, Sandhya R. Shenoy, Surojit Bose, Praveen Agrawal, Debashish Sahu and Vidhu S. Pandey, Bhageerath: An energy based web enabled computer software suite for limiting the search space of tertiary structures of small globular proteins, 2006, Manuscript Communicated.
7. N. Latha, T. Jain, P. Sharma, B. Jayaram, A free energy based computational pathway from chemical templates to lead compounds: a case study of COX-2 inhibitors, 2004, *J. Biomol. Struct. Dyn*, 21, 791-804.
8. N. Latha and B. Jayaram, A Binding Affinity Based Computational Pathway for Active-Site Directed Lead Molecule Design: Some Promises and Perspectives. 2005, *Drug Design Reviews-Online*, 2(2), 145.
9. T. Jain and B. Jayaram, All atom energy based computational protocol for predicting binding affinities of protein-ligand complexes. 2005, *FEBS Letters*, 579, 6659.
10. S. A. Shaikh, S. R. Ahmed and B. Jayaram, A Molecular Thermodynamic view of DNA-drug Interaction: A Case Study of 25 Minor Groove Binders.2004, *Arch. Biochem. Biophys*, 429, 81-99.
11. S. A. Shaikh and B. Jayaram, A Swift All-atom energy based Computational Protocol to Predict DNA-Drug Binding Affinity and delta Tm, 2006, Manuscript Communicated.
12. J. Besemer and M. Borodovsky, GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. 2005, *Nucleic Acids Research*, 33, W451-W454.
13. L-H. Hung, S-C. Ngan, T. Liu and R. Samudrala, PROTINFO: new algorithms for enhanced protein structure predictions. 2005, *Nucleic Acids Research*, 33, W77-W80.
14. W. J. Jorgensen, The many roles of computation in drug discovery. 2004, *Science*, 303, 1813-1818.