International Journal of Computer Science and Engineering (IJCSE)
ISSN 2278-9960
Vol. 2, Issue 1, Feb 2013, 19-28
© IASET

International Academy of Science, Engineering and Technology
Connecting Researchers; Nurturing Innovations

MODELING AND DOCKING STUDIES OF BIOLOGICAL DATA STAT4 INVOLVED IN CANCER

CHUKKA SANTHAIAH[1] & A.RAMA MOHAN REDDY[2]

[1]Department of Computer Science and Engineering, S. V. University, Tirupati, Andhra Pradesh, India

[2]Department of Computer Science and Eng, Head & Prof of C.S.E Dept, S.V.University, Tirupati, Andhra Pradesh, India

ABSTRACT

The amount and variety of data in natural sciences increases rapidly. Data abstraction, Data manipulation and Pattern discovery techniques are of great need in order to deal with such large quantities. Integration between different sources of data is also of major interest, as complex relations may arise. Biology is a good example of a field that provides extensive, highly variable and multi-sources data. The technical advances achieved by the genomics, metabolomics, transcriptomics and proteomics technologies in recent years have significantly increased the amount of data that are available for biologists to analyze different aspects of an organism. Bioinformatics is an interdisciplinary research area at the interface between computer science and biological science.

In order to identify a better drug for cancer, STAT4 (signal transducers and activators of transcription) protein was choosed as target. A three dimensional (3D) model of the STAT4 is generated based on the crystal structure of 1Y1U template by using Modeller software. With the aid of the molecular mechanics and molecular dynamics methods, the final model is obtained and is further assessed by Procheck and Verify 3D graph programs, which showed that the final refined model is reliable. With this model, a flexible docking study is performed with different drugs. From the docking studies, we also suggest that MET3, ARG4, THR5 in STAT4 domain are three important residues in binding. The hydrogen bonding interactions play an important role for stability of the complex. Our results may be helpful for further experimental investigations.

KEYWORDS: Biological Data, Cancer, Docking Studies, Naphazoline, Modeling and STAT4

INTRODUCTION

Bioinformatics differs from a related field known as computational biology. Bioinformatics is limited to sequence, structural, and functional analysis of genes and genomes and their corresponding products and is often considered computational molecular biology. However, computational biology encompasses all biological areas that involve computation. For example, mathematical modeling of ecosystems, population dynamics, application of the game theory in behavioral studies, and phylogenetic construction using fossil records all employ computational tools, but do not necessarily involve biological macromolecules.

There are a variety of other advanced skill sets that can add value to this background: molecular evolution and systematics, physical chemistry kinetics, thermodynamics and statistical mechanics, statistics and probabilistic methods, database design and implementation, algorithm development, molecular biology laboratory methods and others. A understanding of biological processes has grown and deepened, it isn't surprising, then, that the disciplines of computational biology and, more recently, bioinformatics, have evolved from the intersection of classical biology, mathematics, and computer science.

There is a wide range of topics that are useful if you are interested in pursuing bioinformatics, and it's not possible to learn / mastering them all. However, in our daily lab procedure protocols and conversations with researchers working at various levels, we have picked up on the following **"core requirements"** for bioinformaticians:

- Deep background in some aspect of molecular biology, biochemistry, molecular biophysics, or even molecular modeling.
- You must absolutely understand the central dogma of molecular biology. Understanding how and why DNA sequence is transcribed into RNA and translated into protein is vital [meaning is same].
- You should have substantial experience with at least one or two major molecular biology software packages, either for sequence analysis or molecular modeling. The experience of learning one of these packages makes it much easier to learn to use other software quickly. You should be comfortable working in a command-line computing environment. Working in Linux or Unix will provide this experience. You should have experience with programming in a computer language such as C/C++, as well as in a scripting language such as Perl or Python.

Cancer is a class of diseases in which a group of cells display uncontrolled growth (division beyond the normal limits), invasion (intrusion on and destruction of adjacent tissues), and sometimes metastasis (spread to other locations in the body via lymph or blood). These three malignant properties of cancers differentiate them from benign tumors, which are self-limited, and do not invade or metastasize. Most cancers form a tumor but some, like leukemia, do not. The branch of medicine concerned with the study, diagnosis, treatment, and prevention of cancer is oncology.

Cancer affects people at all ages with the risk for most types increasing with age. Cancer caused about 13% of all human deaths in 2007 (7.6 million).

STATs (Signal Transducers and Activators of Transcription)

STATs (signal transducers and activators of transcription) are members of a recently identified family of transcription factors that activate gene transcription in response to a number of different cytokines. The STATs are latent cytoplasmic proteins that are promptly activated by tyrosine phosphorylation by the cytokine receptor associated JAK (Janus) kinases after cytokine exposure. STAT phosphorylation allows the dimerization of individual STAT proteins via their SH2 (src homology 2) domains.

The resulting functional STAT dimer is then capable of migrating directly to the nucleus where it can bind DNA and directly activate cytokine responsive gene transcription. To date seven different STAT proteins have been described each activated by specific cytokine/cytokine receptor combinations.

MATERIALS AND METHODS

Software's and Internet Servers

Blast Database

BLAST searches for high scoring sequence alignments between the query sequence and sequences in the database using a heuristic approach that approximates the Smith-Waterman algorithm. The exhaustive Smith-Waterman approach is too slow for searching large genomic databases such as GenBank. Therefore, the BLAST algorithm uses a heuristic approach that is slightly less accurate than Smith-Waterman but over 50 times faster. The speed and relatively good accuracy of BLAST are the key technical innovation of the BLAST programs and arguably why the tool is the most popular bioinformatics search tool.

BLAST is actually a family of programs (all included in the blastall executable). The following are some of the programs, ranked mostly in order of importance:

- Nucleotide-nucleotide BLAST (blastn): This program, given a DNA query, returns the most similar DNA sequences from the DNA database that the user specifies.
- Protein-protein BLAST (blastp): This program, given a protein query, returns the most similar protein sequences from the protein database that the user specifies.
- Position-Specific Iterative BLAST (PSI-BLAST): One of the more recent BLAST programs, this program is used for finding distant relatives of a protein. First, a list of all closely related proteins is created. Then these proteins are combined into a "profile" that is a sort of average sequence. A query against the protein database is then run using this profile, and a larger group of proteins found. This larger group is used to construct another profile, and the process is repeated. By including related proteins in the search PSI-BLAST is much more sensitive in picking up distant evolutionary relationships than the standard protein-protein BLAST.
- Nucleotide 6-frame translation-protein (blastx): This program compares the six-frame conceptual translation products of a nucleotide query sequence (both strands) against a protein sequence database.
- Nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx): This program is the slowest of the BLAST family. It translates the query nucleotide sequence in all six possible frames and compares it against the six-frame translations of a nucleotide sequence database. The purpose of tblastx is to find very distant relationships between nucleotide sequences.
- Protein-nucleotide 6-frame translation (tblastn): This program compares a protein query against the six-frame translations of a nucleotide sequence database.
- Large numbers of query sequences (megablast): When comparing large numbers of input sequences via the command-line BLAST, "megablast" is much faster than running BLAST multiple times. It basically concatenates many input sequences together to form a large sequence before searching the BLAST database, and then post-analyze the search results to glean individual alignments and statistical values".

Homology Modeling

All homology-modeling methods consist of the following four steps:

- Template selection
- Target template alignment
- Model building
- Evaluation

These steps can be iteratively repeated, until a satisfying model structure is achieved. Several different techniques for model building have been developed. The SWISS-MODEL server approach can be described as rigid fragment assembly is outlined briefly.

Alignment

Up to five template structures per batch are superposed using an iterative least squares algorithm. A structural alignment is generated after removing incompatible templates, i.e. omitting structures with high C_α root mean square deviations to the first template. A local pair-wise alignment of the target sequence to the main template structures is calculated, followed by a heuristic step to improve the alignment for modeling purposes. The placement of insertions and deletions is optimized considering the template structure context. In particular, isolated residues in the alignment ('islands') are moved to the flanks to facilitate the loop building process.

Docking

Docking studies are computational techniques for the exploration of the possible binding modes of a substrate to a given receptor, enzyme or other binding site. Docking studies have become nearly indispensable for study of macromolecular structures and interactions. Mechanical model construction requires heroic patience and endurance to complete a structure which may contain several thousand atoms while computer graphics can build and display in seconds. Macromolecular modeling by docking studies provides most detailed possible view of drug-receptor interaction and has created a new rational approach to drug design where the structure of drug is designed based on its fit to three dimensional structures of receptor site, rather than by analogy to other active structures of random leads.

METHODOLOGY

3D Model Building

The initial model of Signal transducer and activator of transcription 4 (STAT4) was built by using homology-modeling methods and the MODELLER software; a program for comparative protein structure modeling optimally satisfying spatial restraints derived from the alignment and expressed as probability density functions (pdfs) for the features restrained. The pdfs restrain C^{α}-C^{α} distances, main-chain N-O distances, and main-chain and side-chain dihedral angles. The 3D model of a protein is obtained by optimization of the molecular pdf such that the model violates the input restraints as little as possible. The molecular pdf is derived as a combination of pdfs restraining individual spatial features of the whole molecule.

The optimization procedure is a variable target function method that applies the conjugate gradients algorithm to positions of all non-hydrogen atoms. The query sequence from Homo sapiens was submitted to domain fishing server Signal transducer and activator of transcription 4 prediction. The predicted domain was searched to find out the related protein structure to be used as a template by the BLAST (Basic Local Alignment Search Tool) program against PDB (Protein Databank). Sequence that showed maximum identity with high score and less e-value were aligned and was used as a reference structure to build a 3D model for Signal transducer and activator of transcription 4. The sequence of Signal transducer and activator of transcription 4 (Q14765) was obtained from NCBI.

The co-ordinates for the structurally conserved regions (SCRs) for Signal transducer and activator of transcription 4 were assigned from the template using multiple sequence alignment, based on the Needleman-Wunsch algorithm. The structure having the least modeller objective function, obtained from the modeller was improved by molecular dynamics and equilibration methods.

Finally, the structure having the least energy with low RMSD (Root Mean Square Deviation) was used for further studies. In this step, the quality of the initial model was improved. The final structure obtained was analyzed by Ramachandran's map using PROCHECK (Programs to check the Stereo chemical Quality of Protein Structures) and

environment profile using ERRAT graph (Structure Evaluation server). This model was used for the identification of active site and for docking of the substrate with the enzyme.

Docking Method

Docking was carried out using GOLD (Genetic Optimization of Ligand Docking) software which is based on genetic algorithm (GA). This method allows as partial flexibility of protein and full flexibility of ligand. The compounds are docked to the active site of the STAT4. The interaction of these compounds with the active site residues are thoroughly studied using molecular mechanics calculations. The parameters used for GA were population size (100), selection pressure (1.1), number of operations (10,000), number of island (1) and niche size (2). Operator parameters for crossover, mutation and migration were set to 100, 100 and 10 respectively. Default cutoff values of 3.0 A° (dH-X) for hydrogen bonds and 6.0 A° for vanderwaals were employed.

During docking, the default algorithm speed was selected and the ligand binding site in the alpha glucosidase was defined within a 10 A° radius with the centroid as CE atom of ALA410. The number of poses for each inhibitor was set 100, and early termination was allowed if the top three bound conformations of a ligand were within 1.5A° RMSD. After docking, the individual binding poses of each ligand were observed and their interactions with the protein were studied. The best and most energetically favorable conformation of each ligand was selected.

RESULTS AND DISCUSSIONS

Homology Modeling of STAT4 Protein (Q14765)

A high level of sequence identity should guarantee more accurate alignment between the target sequence and template structure. In the results of BLAST search against PDB, only two-reference proteins, including 1Y1U A (Chain A, Structure Of Unphosphorylated Stat5a) has a high level of sequence identity and the identity of the reference protein with the Q14765_STAT4 domain are 31%. Structurally conserved regions (SCRs) for the model and the template were determined by superimposition of the two structures and multiple sequence alignment.

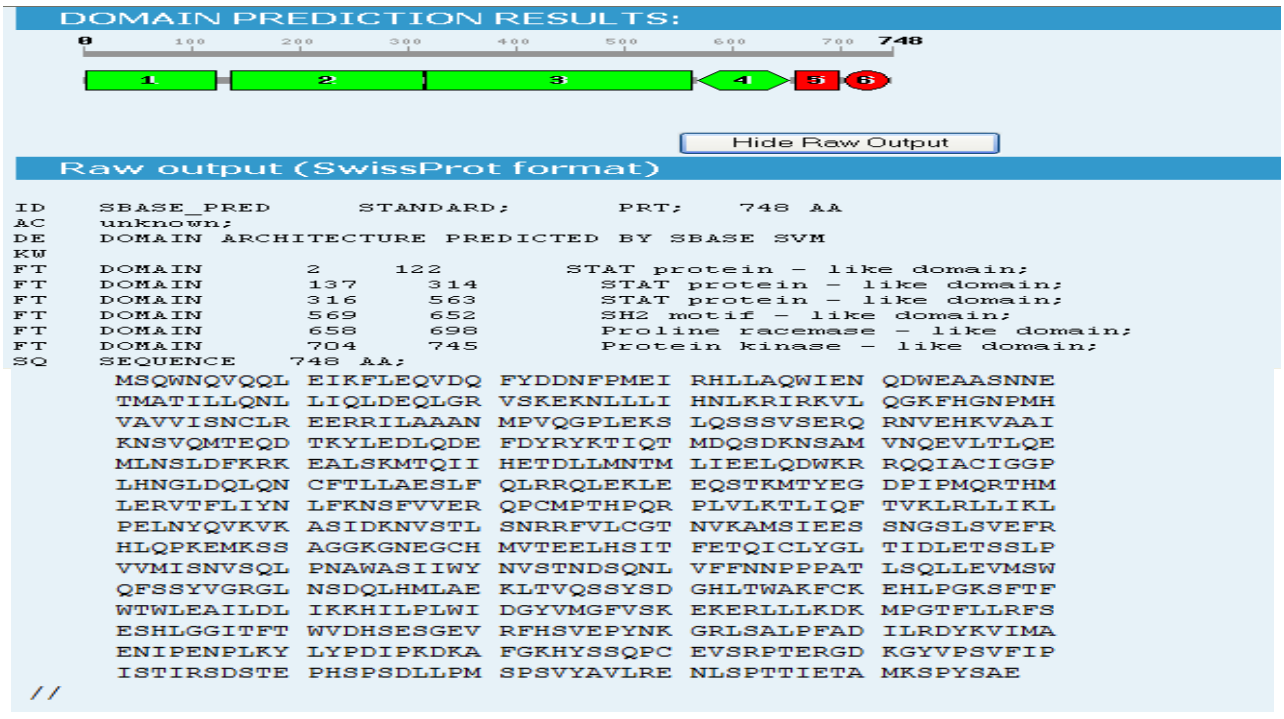
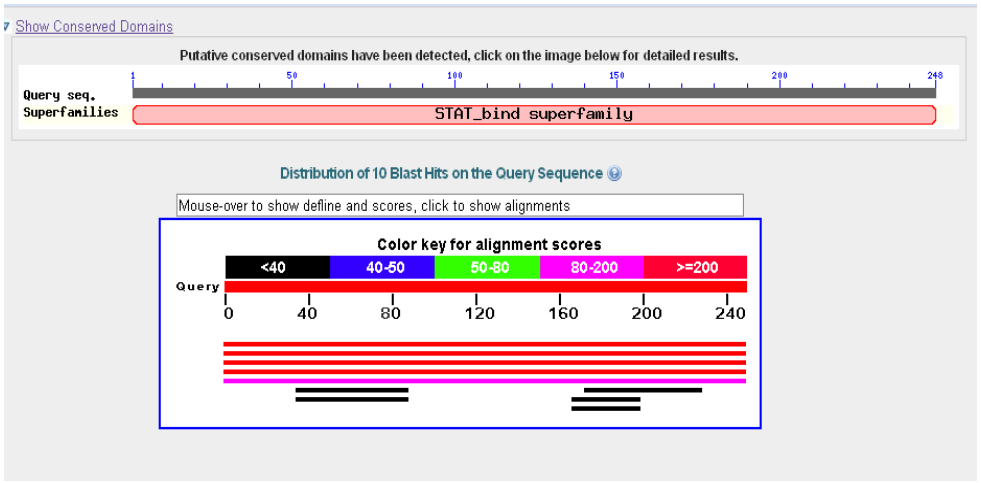


Figure 1: Identification of Domain Region Using SBASE Server




```
>pdb|1Y1U|A  Chain A, Structure Of Unphosphorylated Stat5a
 pdb|1Y1U|B  Chain B, Structure Of Unphosphorylated Stat5a
 pdb|1Y1U|C  Chain C, Structure Of Unphosphorylated Stat5a
Length=585

 Score =  106 bits (264),  Expect = 1e-23, Method: Compositional matrix adjust.
 Identities = 85/267 (31%), Positives = 139/267 (52%), Gaps = 34/267 (12%)

Query  1    FVVERQPCMPTHPQRPLVLKTLIQFTVKLRLLIKLPELNYQV---KVKASI--DKNVSTL  55
            F++E+QP          P VLKT  +F   +RLL+   +LN   +    +VKA+I   ++    +L
Sbjct  205  FIIEKQP---------PQVLKTQTKFAATVRLLVG-GKLNVHMNPPQVKATIISEQQAKSL  255

Query  56   ---SNRRFVLCGTNVKAMSIEE--SSNGSLSVEFRHLQPKEMKSSAGGKGNEGCHMVTEE  110
               N R    G  +    + E    + G+LS   FR++    K +K +     G    VTEE
Sbjct  256  LKNENTRNECSGEILNNCCVMEYHQATGTLSAHFRNMSLKRIKRA----DRRGAESVTEE  311

Query  111  LHSITFETQICLYG--LTIDLETSSLPVVMISNVSQLPNAWASIIWYNVSTNDSQNLVFF  168
             ++ FE+Q  +    L   ++T SLPVV+I + SQ  NA A+++W N      +  V F
Sbjct  312  KFTVLFESQFSVGSNELVFQVKTLSLPVVVIVHGSQDHNATATVLWDNAFAEPGR--VPF  369

Query  169  NNPPPATLSQLLEVMSWQFSSYV--GRGLNSDQLHMLAEKLTVQSS-----YSDGHLTWA  221
               P       QL E ++ +F + V    RGL   + L   LA+KL    SS      Y+    ++W+

Sbjct  370  AVPDKVLWPQLCEALNMKFKAEVQSNRGLTKENLVFLAQKLFNISSNHLEDYNSMSVSWS  429

Query  222  KFCKEHLPGKSFTFWTWLEAILDLIKK  248
            +F +E+LPG ++TFW W + +++++KK
Sbjct  430  QFNRENLPGWNYTFWQWFDGVMEVLKK  456
```

Figure 2: Blast Result with a Similar Template Having 31% Identity with STAT4

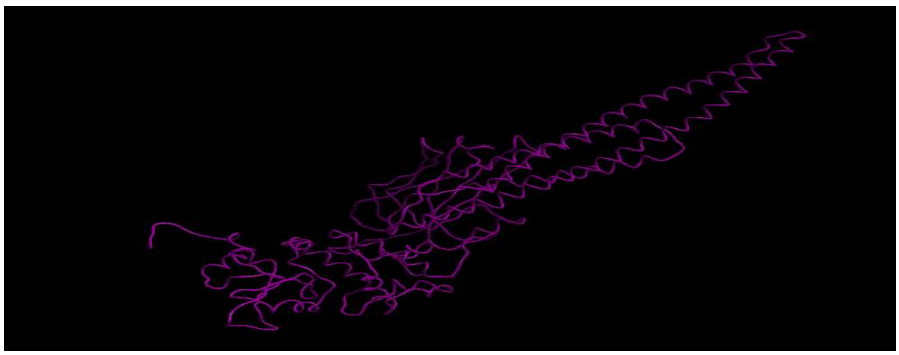


Figure 3: The Stable Structure of the 1Y1U A (Chain A, Structure of Unphosphorylated Stat5a) Protein Obtained

ClustalW2 Results

Results of search	
Number of sequences	2
Alignment score	324
Sequence format	Pearson
Sequence type	aa
JalView	Start Jalview
Output file	clustalw2-20100619-1603543653.output
Alignment file	clustalw2-20100619-1603543653.aln
Guide tree file	clustalw2-20100619-1603543653.dnd
Your input file	clustalw2-20100619-1603543653.input

SUBMIT ANOTHER JOB

Figure 4: Alignment of STAT4 with Template 1Y1U

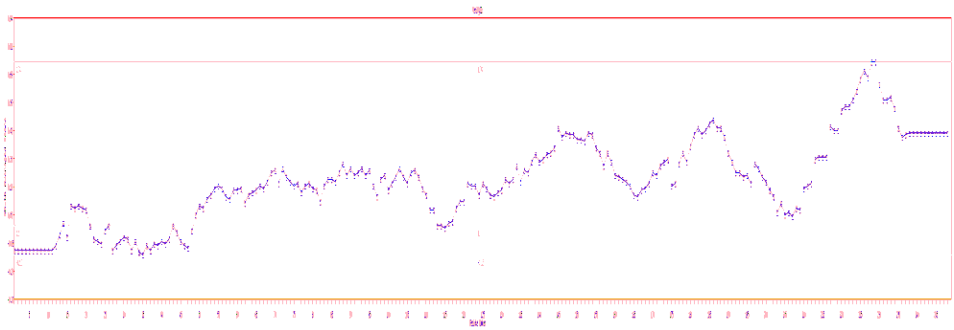


Figure 5: The 3D Profiles Verified Results of Signal Transducer and Activator of Transcription 4 Model; Overall Quality Score Indicates Residues are Reasonably Folded

NAPHAZOLINE

Figure 6: 2-(Naphthalen-1-Ylmethyl)-4, 5-Dihydro-1H-Imidazole Hydrochloride

Docking Results of Naphazoline

Table 1: Fitness List for Ligand File Naphazoline. Sdf, Molecule 1

Mol No	Fitness	S (hb-ext)	S (vdw-ext)	S (hb-int)	S(int)
10	37.62	0.00	30.72	0.00	-4.62
8	37.44	5.80	25.19	0.00	-3.00
3	36.61	5.33	24.89	0.00	-2.94
1	36.31	0.00	28.57	0.00	-2.97
2	36.03	0.00	28.34	0.00	-2.95
7	35.99	6.00	24.29	0.00	-3.40
9	35.81	0.00	28.12	0.00	-2.85
6	35.72	0.00	28.11	0.00	-2.93
4	35.39	0.00	27.80	0.00	-2.83
5	35.17	2.59	25.79	0.00	-2.88
Average Value	**36.21**	**1.97**	**27.18**	**0.00**	**-3.14**

Docking of Inhibitors with the Active Site of Signal Transducer and Activator of Transcription 4

Docking of the inhibitors with Signal transducer and activator of transcription 4 was performed using GOLD 3.0.1, which is based on Genetic algorithimn. This program generates an ensemble of different rigid body orientations (poses) for each compound conformer within the binding pocket and then passes each molecule against a negative image of the binding site. Poses clashing with this 'bump map' are eliminated. Poses surviving the bump test are then scored and ranked with a Gaussian shape function. We defined the binding pocket using the ligand-free protein structure and a box enclosing the binding site. This box was defined by extending the size of a cocrystalized ligand by 4 Å. This dimension was considered here appropriate to allow, for instance, compounds larger than the cocrystallized ones to fit into the binding site. One unique pose for each of the best-scored compounds was saved for the subsequent steps. The compounds used for docking was converted in 3D with SILVER. To this set, the substrate corresponding to the modeled protein were added.

The Chemical Properties these Structures are Tabulated as Follows

S.No	Molecular Formula	Formula Weight	Molar Refractivitycm^3	Index of Refraction	Densityg/Cm^3	Polarisability $10^{-24}cm^3$
1	$C_{26}H_{27}NO_9$	497.49388	123.61 ± 0.4	1.705 ± 0.03	1.56 ± 0.1	49.00 ± 0.5
2	$C_{20}H_{28}O_2$	300.43512	95.52 ± 0.3	1.556 ± 0.02	1.011 ± 0.06	37.87 ± 0.5
3	$C_{46}H_{56}N_4O_{10}$	824.95764	221.08 ± 0.4	1.677 ± 0.03	1.40 ± 0.1	87.64 ± 0.5
4	$C_{16}H_{20}N_2$	240.3434	75.91 ± 0.3	1.556 ± 0.02	1.018 ± 0.06	30.09 ± 0.5
5	$C_{22}H_{28}N_4O_6$	444.48092	119.70 ± 0.3	1.709 ± 0.02	1.450 ± 0.06	47.45 ± 0.5

CONCLUSIONS

STATs (signal transducers and activators of transcription) is a member of a recently identified family of transcription factors that activate gene transcription in response to a number of different cytokines in Homo sapiens. In

this work, we have constructed a 3D model of Q14765 domain, using the MODELLER software and obtained a refined model after energy minimization. The final refined model was further assessed by ERRAT and PROCHECK program, and the results show that this model is reliable. The stable structure of Q14765 is further used for docking with modified ligand molecules. Docking results indicate that conserved amino-acid residues Signal transducer and activator of transcription 4 main play an important role in maintaining a functional conformation and are directly involved in donor substrate binding. The interaction between the domain and the inhibitors proposed in this study are useful for understanding the potential mechanism of domain and the inhibitor binding.

As is well known, hydrogen bonds play important role for the structure and function of biological molecules. In this study it was found that, ILE 153, LEU 179, VAL 182, MET 183, GLN 186, PHE 187, LEU 195, GLN 199, MET 202, LEU 203, PHE 223, TRY 236, TRP 238, LEU 239, GLU 240, ILE 242, LEU 243, ILE 246 are important for strong hydrogen bonding interaction with the inhibitors. To the best of our knowledge MET1, MET3, ARG4, THR5 are conserved in this domain and may be important for structural integrity or maintaining the hydrophobicity of the inhibitor-binding pocket.

REFERENCES

1. Armitage, James O & Antman, Karen H (2000) High dose cancer therapy: pharmacology, hematopoietins, stem cells. (3rd edition) Philadelphia, Williams & Wilkins. ISBN: 0683306545.
2. Baider, Lea et al. (2001) Cancer and the family. (2nd edition) Chichester, Wiley. ISBN: 0471803006.
3. Bailey, Michael & Sarosdy, Michael (2004) Bladder cancer. (2nd edition) London, Fast facts, Health Press. ISBN: 1903734258.
4. Bleiberg, Harry et al. (2002) Colorectal Cancer. London, Martin Dunitz. ISBN: 185317808X.
5. Bonadonna, Gianni et al. (2001) Textbook of breast cancer: a clinical guide to therapy. (2nd edition) London, Martin Dunitz. ISBN: 1853178241.
6. Canellos, George et al (1998) The lymphomas. WB Saunders. ISBN: 0721650309.
7. Cunningham, David et al. (2003) The effective management of colorectal cancer (3rd edition) London, Aesculapius Medical Press. ISBN: 1903044359.
8. DeVita, Vincent T. et al. (2001) Cancer: principles and practice of oncology. (6th edition) Philadelphia, Lippincott. ISBN: 0781722292.
9. Fentiman, Ian S. (1999) Challenges in breast cancer. Oxford, Blackwell Science. ISBN: 0632052422.
10. Gershenson, David M. & McGuire William P. (1998) Ovarian cancer : controversies in management. New York, Churchill Livingstone. ISBN: 0443078041.
11. Ginsberg, Robert J (2002) Lung cancer. Hamilton, B C Decker. ISBN: 1550090992.
12. Greenberg, Harry S. et al (1999) Brain tumors. Oxford, Oxford University Press. ISBN: 019512958X.
13. Hamilton, Stanley R et al. (2000) Pathology and genetics of tumours of the digestive system. World Health Organisation classification of tumours series. International Agency for Research on Cancer (IARC). ISBN: 9283224108.

14. Hancock, B. W. et al. (2000) Malignant lymphoma. London, Arnold. ISBN: 0340742070.

15. Harris, Jay R. et al (2004) Diseases of the breast. (3rd edition) Philadelphia, Lippincott-Raven. ISBN: 0781746191.

16. Jacobs, I. J. et al. (2002) Ovarian cancer. (2nd edition) Oxford, Oxford University Press. ISBN: 0198508263.

17. Jordan, V. Craig (1996) Tamoxifen: a guide for clinicians and patients New York, PRR. ISBN: 0964182343.

18. Kelsen, David P. et al (2002) Gastrointestinal oncology: principles and practice. Philadelphia, Lippincott, Williams and Wilkins. ISBN: 0781722306.