Contents lists available at ScienceDirect

Asian Pacific Journal of Tropical Disease

journal homepage:www.elsevier.com/locate/apjtd

Document heading

In silico modelling and validation of differential expressed proteins in lung cancer

Bhagavathi S[1], Gulshan Wadhwa[2], Anil Prakash[1]

1 -Research Scholar, Department of Biotechnology & Bio Informatics Centre, Barkatullah University, Bhopal-462026, India
1-Head & Coordinator, Department of Biotechnology & Bio Informatics Centre, Barkatullah University, Bhopal-462026, India
2 - Joint Director, Department of Biotechnology, Ministry of Science & Technology, Govt. of India, New Delhi-110003, India

ARTICLE INFO

Article history:
Received 25 June 2012
Received in revised from 5 July 2012
Accepted 7 Octoberr 2012
Available online 28 October 2012

Keywords:
Lung cancer
Matrix metallo proteinase
Trophinin
Thrombomodulin
Polo like kinase

ABSTRACT

Objective: The present study aims predict the three dimensional structure of three major proteins responsible for causing Lung cancer. **Methods:** These are the differentially expressed proteins in lung cancer dataset. Initially, the structural template for these proteins is identified from structural database using homology search and perform homology modelling approach to predict its native 3D structure. Three-dimensional model obtained was validated using Ramachandran plot analysis to find the reliability of the model. **Results:** Four proteins were differentially expressed and were significant proteins in causing lung cancer. Among the four proteins, Matrixmetallo proteinase (P39900) had a known 3D structure and hence was not considered for modelling. The remaining proteins Polo like kinase I Q58A51, Trophinin B1AKF1, Thrombomodulin P07204 were modelled and validated. **Conclusions:** The three dimensional structure of proteins provides insights about the functional aspect and regulatory aspect of the protein. Thus, this study will be a breakthrough for further lung cancer related studies.

1. Introduction

Cancer is associated with multiple genetic and regulatory aberrations in the cell. It is a highly heterogeneous disease, both morphologically and genetically [1]. Analysis of cancer pathways shows a number of interrelated markers responsible for oncogenesis. Lung cancer constitutes one of the leading causes of death in industrialized countries, and its incidence is rapidly growing in developing nations worldwide. Although tobacco smoke and other environmental pollutants are responsible for more than 80-90% of the cases in men [2] , it is well established that less than 10-15% of smokers develop lung cancer, indicating that other factors might contribute to the development of this disease [3][4]. Preclinical studies have provided evidence that matrix metalloproteinases (MMPs), a family of zinc-containing proteolytic enzymes, facilitate tumor invasion, the establishment of metastases, and the promotion of tumor-related angiogenesis. Matrix metalloproteinase inhibitors (MMPIs) have been shown to inhibit tumor growth and dissemination in preclinical models [5]. Human lung cancer cells have been found to express varying degrees of several kinds of onco developmental antigens, such as carcino embryonic antigen and stage-specific embryonic antigen related antigens, which are found expressed in stage-specific lung buds of human embryos and may play some role in the cell-to-cell interactions. Preclinical studies have also provided evidence that Thrombomodulin is not only a thrombin receptor but also an onco developmental antigen, found to be expressed in lung cancer cells as thrombomodulin is expressed in the lung bud epithelium. Extensive studies have shown that Plk1 expression is elevated in non-small-cell lung cancer, head and neck cancer, esophageal cancer, gastric cancer, melanomas, breast cancer, ovarian cancer, endometrial cancer, colorectal cancer, gliomas , and thyroid cancer. Plk1 gene and protein expression has been proposed as a new prognostic marker for many types of malignancies, and Plk1 is a potential target for cancer therapy [6]. Selection of a potential target for therapy is a daunting task. In-silico modeling is a multidisciplinary method integrating mathematical models with experimental (in vitro and in vivo) and clinical data [7]. Homology or evolutionary

*Corresponding author: Bhagavathi S. A-22, Arihant Heirloom, Navalur, Via thalambur, Chennai-603103; Department of Biotechnology, Bioinformatics Centre Barkatullah University, Bhopal-462026, INDIA.
Email:bhagavathikanagaraj@gmail.com

relatedness represents a key concept in studying protein sequence, structure, and function. Homologs can be inferred by sequence similarity search tools such as the popular sequence-profile comparison method PSI-BLAST [8]. Basic Local Alignment Search Tool (BLAST) provides an "expect" value, statistical information about the significance of each alignment [9]. MACS (multiple alignments of complete sequences) are typically used to perform comparative analysis at the genome level, to define the phylogentic relationships between organisms in evolutionary studies, to identify conserved functional residues, motifs or domains and to predict protein [10]. Comparative, or homology, modelling structures is the most widely used prediction method when the target protein has homologues of known structure [11]. This study is aimed at modeling and evaluating the structure of four major proteins actively involved in lung cancer.

2. Materials and methods

2.1 Sequence Retrieval from Swissprot

Amino acid sequences retrieved from swissprot/uniprot (www.uniprot.org) provides descriptions of a non redundant set of proteins including their function, domain structure, posttranslational modifications and variants [12] [13]. This database merges all proteins in single entry coded by one gene so as to minimize redundancy and improve reliability with fully featured information. Cross-references with others databases modemize swissprot entries to hold detailed expertise [14].

2.2 Template selection and Target Structure Modeling

Structural homologous entries were obtained for proteins through local alignment search using BlastP (Basic Local Alignment Search Tool) [15], against Protein Data Bank (PDB) [16]. Comparison of homology models with known structure (Template) may also reveal similarities which allow biochemical and biological functions to be inferred . The alignment was used for comparative modeling to build 3D model by satisfaction of spatial restraints using Modeller9v7 [17].The core modeling procedure begins with an alignment of the sequence to be modelled (Target) with related known 3D structures (templates). This alignment is usually input to the program. The output is a 3D model for the target sequence containing all main chain and side chain non hydrogen atoms. Ramachandran Analysis was performed to determine the stability of the modelled structure. Subsequently the model structure was validated using PROCHECK, which determine stereo chemical aspects along with main chain and side chain parameters with comprehensive analysis.

3. Results

Matrix metallo proteinase with accession number P39900 was retrieved from swissprot.

And it has an already available 3D structure which was retrieved from PDB (1ROS) and visualised using Rasmol (Fig1.)

Fig 1: Structure of Matrix mettalo proteinase (1ROS) visualised using Rasmol.

Polo like kinase I (Q58A51) was subjected to homology search against PDB database using BlastP to identify significant structural homolog's to be used as template for homology modelling. The results indicated the presence of Pkc like super family domain and the best homolog was 3KB7 with 99 % identity with the query protein and thus served as a template for modelling and the modelled protein obtained is shown in Fig2 and Validation was done using Ramachandran map (Fig.3.) after loop refinement.

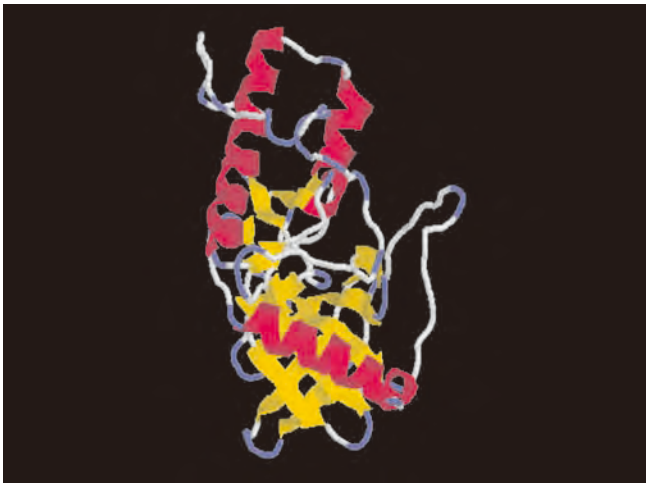

Fig 2: Structure of Polo like kinase I visualised using Rasmol

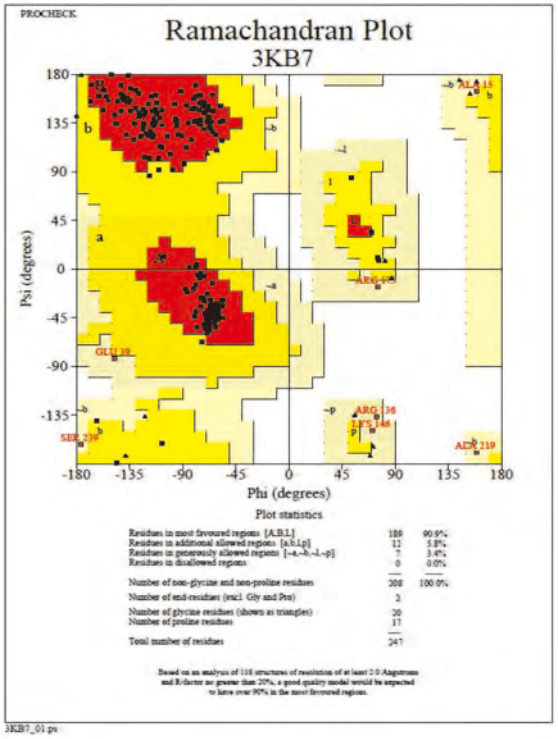

Fig.3: Ramachandran plot of Polo like kinase I

Trophinin (B1AKF1) was subjected to homology search against PDB database using BlastP to identify significant structural homologs to be used as template for homology modelling. The results indicated the presence of MAGE super family domain and the best homolog was 2WA0 with 40 % identity with the query protein and thus served as a template for modelling and the modelled protein obtained is shown in Fig4 and Validation was done using Ramachandran map (Fig.5) after loop refinement.

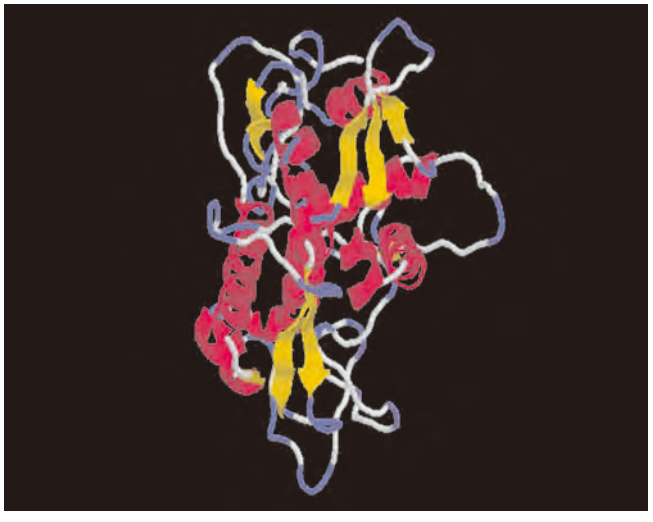

Fig.4: Structure of Trophinin visualised using Rasmol

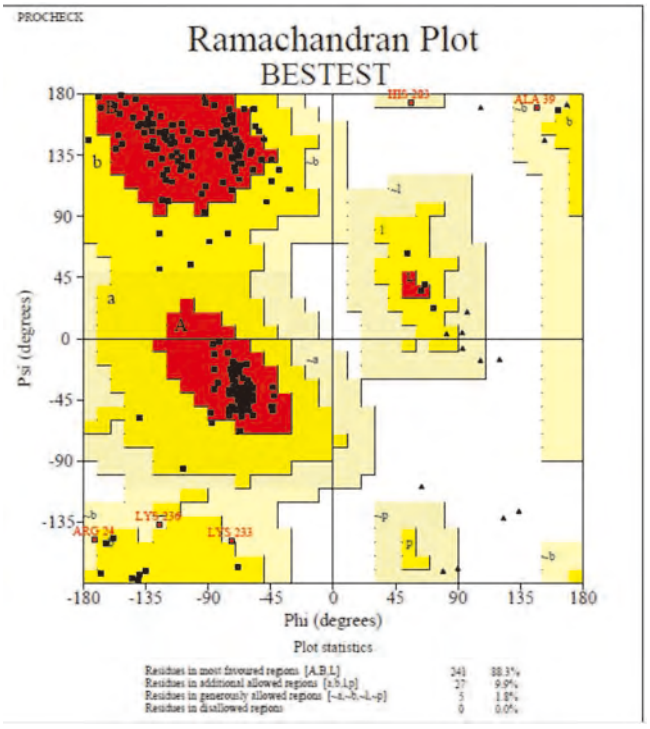

Fig.5: Ramachandran plot of Trophinin

Thrombomodulin (P07204) was subjected to homology search against PDB database using BlastP to identify significant structural homologues to be used as template for homology modelling. The results indicated the presence of CLECT super family domain and the best homolog was 3P5B with 42 % identity with the query protein and thus served as a template for modelling and the modelled protein obtained is shown in Fig.6. and Validation was done using Ramachandran map (Fig.7.) after loop refinement.

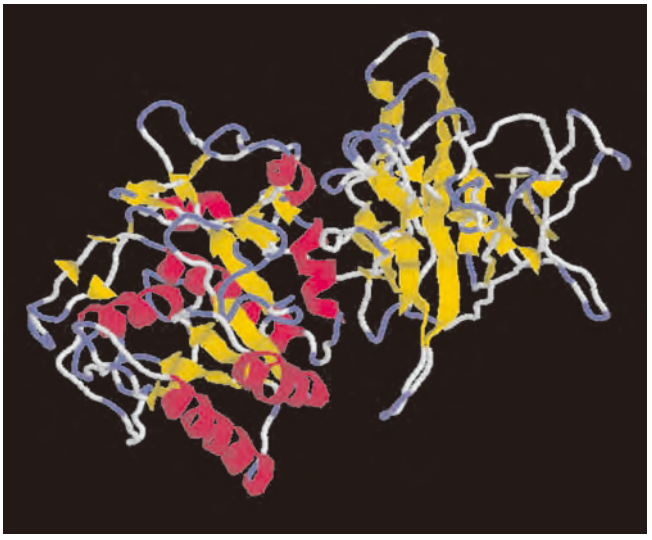

Fig.6: Structure of Thrombomodulin visualised using Rasmol

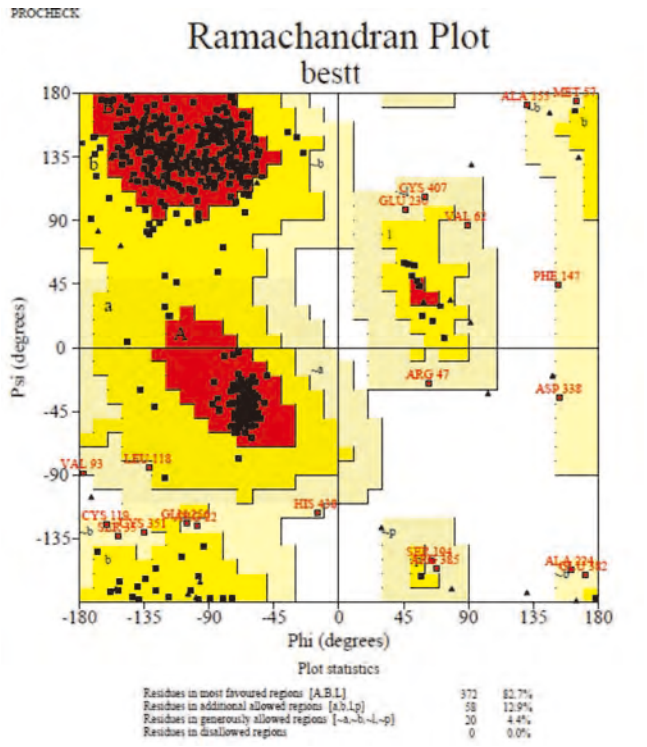

Fig.7: Ramachandran plot of Thrombomodulin

4. Discussion

The results indicate effective modelling of the proteins responsible for lung cancer. The validation using Ramachandran plot confirms the location of most of the significant amino acid in the allowed region; thus confirming its reliability. Analysis of Model was done by the DOPE (Discrete Optimized Protein Energy, a statistical potential used to assess homology models in protein structure prediction. DOPE is based on an improved reference state that corresponds to non-interacting atoms in a homogeneous sphere with the radius dependent on a sample native structure; it thus accounts for the finite and spherical shape of the native structures) method is generally used to assess the quality of a structure model as a whole. It is implemented in the popular homology modeling program MODELLER and used to assess the energy of the protein model generated through many iterations by MODELLER, which produces homology models by the satisfaction of spatial restraints.DOPE is implemented in Python and is run within the MODELLER environment. The biological role of a protein is determined by its function, which is in turn largely determined by its structure. Thus there is enormous benefit in knowing the three dimensional structures of all the proteins. Although more and more structures are determined experimentally at an accelerated rate, it is simply not possible to determine all the protein structures from experiments. As more and more protein sequences are determined, there is pressing need for predicting protein structures computationally. Decades of intense research in this area brought about huge progress in our ability to predict protein structures from sequences only.

So far protein prediction methods based on homology have been the most successful. Homology modeling is based on the notion that new proteins evolve gradually from existing ones by amino acid substitution, addition, and/or deletion and that the 3D structures and functions are often strongly conserved during this process. Many proteins thus share similar functions and structures and there are usually strong sequence similarities among the structurally similar proteins. Strong sequence similarity often indicates strong structure similarity, although the opposite is not necessarily true. Homology modeling tries to identify structures similar to the target protein through sequence comparison. The quality of homology modeling depends on whether these exists one or more protein structures in the protein structure databases that show significant sequence similarity to the target sequence.

One major progress in Homology modeling is the very sensitive profile based sequence comparison method such as PSI-BLAST and profile sequence comparison. Profile-profile based sequence comparison methods are usually superior in that such methods can pick up possible homologous structure templates even when the sequence identity is very low and that profile-profile comparison can align the sequence to the structure template more accurately, producing more accurate structure models. As more and more novel sequences are produced from the genome projects, the profile-based methods can be expected to become even more sensitive. The motive of homology modeling is based on the observation that protein tertiary structure is better conserved than amino acid sequence [18].

Thus, even proteins that have diverged appreciably in sequence but still share detectable similarity will also share common structural properties, particularly the overall fold. Because it is difficult and time-consuming to obtain experimental structures from methods such as X-ray crystallography and protein NMR for every protein of interest, homology modeling can provide useful structural models for generating hypotheses about a protein's function and directing further experimental work .There are exceptions to the general rule that proteins sharing significant sequence identity will share a fold. For example, a judiciously chosen set of mutations of less than 50% of a protein can cause the protein to adopt a completely different fold [19] [20].

However, such a massive structural rearrangement is unlikely to occur in evolution, especially since the protein is usually under the constraint that it must fold properly and carry out its function in the cell. Consequently, the roughly folded structure of a protein (its "topology") is conserved longer than its amino-acid sequence and much longer than the corresponding DNA sequence; in other words, two proteins may share a similar fold even if their evolutionary

relationship is so distant that it cannot be discerned reliably. For comparison, the function of a protein is conserved much less than the protein sequence, since relatively few changes in amino-acid sequence are required to take on a related function however, more effective chemotherapy is needed to control cancer , which is the desired effect for successful cancer treatment [21]. Programmed cell death or apoptosis also plays an important role for balancing cell proliferation and cell death and contributes to an effective cancer therapy [22-30].

Over the past few years, there has been a gradual increase in both the accuracy of comparative models and the fraction of protein sequences that can be modelled with useful accuracy. The magnitude of errors in fold assignment, alignment, and the modeling of side chains, loops, distortions, and rigid body shifts have decreased measurably. This is a consequence of both better techniques and a larger number of known protein sequences and structures. Nevertheless, all the errors remain significant and demand future methodological improvements. In addition, there is a great need for more accurate detection of errors in a given protein structure model. Error detection is useful both for refinement and interpretation of the models .The biological role of a protein is determined by its function, which is in turn largely determined by its structure. Thus there is enormous benefit in knowing the three dimensional structures of all the proteins. Although more and more structures are determined experimentally at an accelerated rate, it is simply not possible to determine all the protein structures from experiments. As more and more protein sequences are determined, there is pressing need for predicting protein structures computationally. Decades of intense research in this area brought about huge progress in our ability to predict protein structures from sequences only. This process is an efficient way for enriching potential target genes, and for identifying those that are critical for normal cell functions [31].

Protein structure prediction aims to model the three-dimensional (3D) structure of so far structurally uncharacterised proteins from their amino acid sequence. Motivated by the observation that homologous proteins with related amino acid sequences have similar 3D structures, protein homology modelling uses comparative methods to generate models for a target protein based on one or more related proteins with known 3D structure. The coordinates of the model are generated based on alignments between the target′s and template′s amino acid sequences, which define the correspondence between residues in both proteins. Ultimately, the quality of a computational model determines its usefulness for specific biomedical applications. Therefore, model quality estimation methods are used to identify unreliable or erroneous regions in the resulting models, and to estimate the overall accuracy of a model. Homology modelling (or comparative modelling) is currently the most accurate computational method available to routinely generate models of sufficient quality for various applications in life science research. Comparative protein modelling methods have been completely automated in recent years, and several Internet servers offer protein modelling services which are reliable and easy to use – also for the non expert in computational biology[32]. The stereochemical quality of the predicted structures was measured employing PROCHECK which yielded Ramachandran Plots displaying favourable conformations which plays an important role in validating the predicted structures [33.]

5. Conclusion

Over the past few years, there has been a gradual increase in both the accuracy of comparative models and the fraction of protein sequences that can be modelled with useful accuracy. The magnitude of errors in fold assignment, alignment, and the modelling of side chains, loops, distortions, and rigid body shifts has decreased measurably. This is a consequence of both better techniques and a larger number of known protein sequences and structures. Nevertheless, all the errors remain significant and demand future methodological improvements. In addition, there is a great need for more accurate detection of errors in a given protein structure model. Error detection is useful both for refinement and interpretation of the models.Homology modelling play a important bridging role for modelling the 3D -structure of a protein. The above work is an in-silico work; this work can serve as a predicted model and can be useful to develop new inhibitor against Lung cancer. The in-silico approach helps researchers by giving them an in-hand idea so that they can happily advance towards the treatment of the disease

Conflict of interest statement

We declare that we have no conflict of interest.

References

[1] Lu Y, Yi Y, Liu P, Wen W, James M, Wang D, You M. Common Human Cancer Genes Discovered by Integrated Gene-Expression Analysis. *PLoS ONE* 2007; **2**: 114-9.

[2] Levi F. Cancer prevention: epidemiology and perspectives. *Eur J cancer.*1999; **35**(14):1912-1924.

[3] Tardon A, Lee J, Rodriguez D, Dosemeci M, Albanes D, Hoover R, Blair A. Leisure-time physical activity and lung cancer: a meta-analysis. *Cancer Causes Control* 2005; **16**(4):389-397.

[4] Rodriguez V, Tardon A, Kogevinas M, Prieto S, Cueto A, Garcia M, Menendez A, Zaplana J: Lung cancer risk in iron and steel foundry workers: a nested case control study in Asturias, *Spain.Am J Ind Med* 2000; **38**(6):644-650.

[5] Philip Bonomi. Matrix metalloproteinYases and matrix metalloproteinase inhibitors in lung cancer. *Seminars in Oncology*.2009; **29**(1): 78-86.

[6] Takai N, Hamanaka R, Yoshimatsu J, Miyakawa I. Polo-like kinases (Plks) and cancer.*Oncogene*.2005; **24**(2):287-91.

[8] Sanga S, Frieboes H, Zheng X, Gatenby R, Bearer E, Cristini V. Predictive oncology:multidisciplinary, multi-scale in-silicomodeling linking phenotype, morphology and growth. *Neuroimage* 2007; **37**: 120-134.

[9] Kim B,Cheng H, Grishin N, Hor A. Web server to infer homology between proteins using sequence and structural similarity. *Nucleic Acids Res*. 2009; **37**: 532-538.

[10]Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden T. BLAST+: Architecture and applications. *BMC Bioinformatics* 2009; **10**: 421.

[11]Friedrich A, Ripp R, Garnier N, Bettler E, Deléage G, Poch O, Moulinier L. Blast sampling for structural and functional analyses. *BMC Bioinformatics* 2007; **8**: 8-62.

[12]Piedra D, Lois S, Cruz X. Preservation of protein clefts in comparative models. *BMC Struct Biol*. 2008; **8**: 2.

[13]Bairoch A, Boeckmann B, Ferro S. and Gasteiger E. Swiss-Prot. Brief.Bioinform., 2004 ;5: 39-55.

[14]The UniProt Consortium The Universal Protein Resource (UniProt). *Nucleic Acids Res* 2007; **36**: 190-195.

[15]Boeckmann B, Blatter C, Famiglietti L, Hinz U, Lane L, Roechert B and Bairoch A. *ComptesRendusBiologies*. 2005; **328**:882-899.

[16]Altschul S, Madden L, Schaffer A, Zhang J, Zhang Z, Miller W and Lipman D . Gapped BLAST and PSI-BLAST:a new generation of protein database search programs. *Nucleic Acids Research*. 1997; **25**(17): 3389-3402.

[17]Dowlathabad M, Anuraj N, Mukesh Y, Showmy S. and Disha P. Comparative modeling of methylentetrahydrofolate reductase (MTHFR) enzyme and its mutational assessment: in silico approach. *International Journal of Bioinformatics Research*.2010; **2**(1): 05-09.

[18]Rakesh S, Pradhan D. and Umamaheswari A. In silico approach for future development of subunit vaccines against Leptospira interrogans serovar Lai . *International Journal of Bioinformatics Research*.2009; **1**(2): 85-92.

[19]Marti-Renom, MA; Stuart, AC; Fiser, A; Sanchez, R; Melo, F; Sali, A. (2000). "Comparative protein structure modeling of genes and genomes". *Annu Rev Biophys Biomol Struct* **29**: 291-325

[20]Dalal S, Balasubramanian S, Regan L. (1997). Transmuting alpha helices and beta sheets. *Fold Des* **2**(5):R71-9.

[21]Dalal, S; Balasubramanian, S; Regan, L. (1997). "Protein alchemy: changing beta-sheet into alpha-helix". *Nat Struct Biol* **4** (7): 548-52.

[22]Locher C, Conforti R, Aymeric L, Ma Y, Yamazaki T, Rusakiewicz9+S, et al. Desirable cell death during anticancer chemotherapy. *Ann N Y Acad Sci* 2010; **1209**: 99-108.

[23]Cummings J, Ward TH, Ranson M Dive C. Apoptosis pathway targeted drugs-from the bench to the clinic. *Biochimica et Biophysica Acta* 2004; **1705**: 53-66

[24]Gulecha V, Sivakuma T. Anticancer activity of Tephrosia purpurea and Ficus religiosa using MCF 7 cell lines. *Asian Pac J Trop Med* 2011; **4**(7): 526-529.

[25]Kumar RS, Rajkapoor B, Perumal P. In vitro and in vivo anticancer activity of Indigofera cassioides Rottl. Ex. DC. *Asian Pac J Trop Med* 2011; **4**(5): 379-385.

[26]Krishnananda P, Gummadi MR, Anjali R. Can antioxidants predispose to cancer recurrence? *Asian Pac J Trop Med* 2011; **3**(6):494-495.

[27]Malik A, Afaq S, Shahid M, Akhtar K, Assiri A. Influence of ellagic acid on prostate cancer cell proliferation: A caspase-dependent pathway. *Asian Pac J Trop Med* 2011; **4**(7): 550-555.

[28]Chanda S, Baravalia Y, Kaneria M. Protective effect of Polyalthia longifolia var. pendula leaves on ethanol and ethanol/HCl induced ulcer in rats and its antimicrobial potency. *Asian Pac J Trop Med* 2011; **4**(10): 673-679.

[29]Nwaehujor CO, Udeh NE. Screening of ethyl acetate extract of Bridelia micrantha for hepatoprotective and anti-oxidant activities on Wistar rats. *Asian Pac J Trop Med* 2011; **4**(10): 796-798.

[30]Arijit M, Tapan KM, Dilipkumar P, Santanu S, Jagadish S. Isolation and in vivo hepatoprotective activity of Melothria heterophylla (Lour.) Cogn. against chemically induced liver injuries in rats. *Asian Pac J Trop Med* 2011; **4**(8): 619-623.

[31]Budhayash Gautam, Gurmit Singh, Gulshan Wadhwa, Rohit Farmer, Satendra Singh, Atul Kumar Singh, Prashant Ankur Jain, Pramod Kumar Yadav, Metabolic pathway analysis and molecular docking, analysis for identification of putative drug targets in Toxoplasma gondii: novel approach, Bioinformation, *Volume* **8**(3), 134-141

[32]Manuel C Peitsch, Torsten Schwede,University of Basel, Switzerland Protein Homology Modelling, Wiley, Els article, Nov,2011.

[33]Kamalika Banerjee, Utkarsh Gupta, Sanjay Gupta, Gulshan Wadhwa, Reema Gabrani, Sanjeev Kumar Sharma1, Chakresh Kumar Jain, Molecular docking of glucosamine-6-phosphate synthase in Rhizopus oryzae, *Bioinformation*. 2011; **7**(6):285-290.