International Journal of Bioinformatics Research
ISSN: 0975–3087 , E-ISSN: 0975–9115, Vol. 3, Issue 1, 2011, pp-103-113
Available online at http://www.bioinfo.in/contents.php?id=21

IN SILICO RETRIEVAL OF NOVEL GENES USING SAGE GENIE AND COMPARATIVE MODELING OF IMMUNOGLOBULIN LAMBDA LOCUS (IGL@) GENE: A POTENT TARGET OF LUNG CANCER

SUBHALAXMI NAYAK.[1], UMADEVI K.[1]* AND REDDY M.N.[2]
[1]DBT – Bioinformatics Programme, Department of Marine Living Resources, Andhra University, Visakhapatnam –530 003, AP India, snnayak.subhalaxmi@gmail.com; andhrauniv.btisnet@nic.in
[2]Coordinator, Bioinformatics Centre, Department of Applied Microbiology, Sri Padmavati Mahila Visvavidyalayam, Tirupati – 517 502, AP, India, mopuri_nr@yahoo.com
*Corresponding author. E-mail: andhrauniv.btisnet@nic.in

Received: September 18, 2010; Accepted: October 13, 2010

Abstract- Lung cancer is a disease of uncontrolled cell growth in tissues of the lung. This growth may lead to metastasis, which is the invasion of adjacent tissue and infiltration beyond the lungs. The vast majority of primary lung cancers are carcinomas of the lung, derived from epithelial cells. Therefore, the present study is aimed at detecting the highly expressed genes, responsible for lung cancer through Serial Analysis of Genome Expression (SAGE) Genie at Cancer Genome Anatomy Project (CGAP). IGL@ (Immunoglobulin lambda locus) gene is predominantly highly expressed in the lung cancer and is considered to be a new target for the Cancerous diseases. The protein was retrieved from the Swissprot/Uniprot KB with accession number Q6GMX3 and was modelled using MODELLER9v7 for predicting the 3D structure of the IGL @ protein which provides an accurate and efficient module to build loops and side chains, found to be identical in sequence. Modelled structure revealed appreciable measures when subjected to structure verification and evaluation using PROCHECK. Ramachandran plot signified the present work undertaken through conformational parameters $^{\Phi}$ (phi) and ψ (psi) angles calculated from model with 93.4% residues in most favoured region. Further, PROCHECK results confirmed acceptance of model through main and side-chain values. Root mean square distance of planarity was found below 0.02. Hence the model was revealed to have good stereochemistry. Structure of IGL@ Protein can be important tool for future endeavours towards structure based drug designing techniques to impel the search of new efficient inhibitors.

Keywords- SAGE, Lung Cancer, Highly Expressed Genes, Comparative Modeling, PROCHECK

Introduction

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. Cancer cells often have patterns of gene expression that differ from their normal cell counterparts. Some genes are expressed at higher levels (Over-expression), and others are expressed at lower levels or not at all. Genes that are over-expressed in the cancerous tissue are of particular interest because over-expression is a trait that we would expect of a gene that is causing the cancer to grow. For example if a gene codes for a protein that usually functions to cause a cell to divide, then making more if this protein may lead to uncontrolled growth. On the other hand, other gene products that normally function to control or inhibit growth may be lost the normal cell functions and becomes a cancer cell. If we compare the cell to a car, over-expression of a growth-stimulating gene is analogous to jamming the accelerator down, whereas loss of an inhibitory gene is analogous to losing the ability to apply the brakes. The Cancer Genome Anatomy Project (CGAP) sought to determine the gene expression profiles of normal, precancer, and cancer cells, leading eventually to improved detection, diagnosis, and treatment for the patient. A gene's expression pattern provides clues to its role in normal physiology and disease. To provide quantitative expression levels on a genome-wide scale, the CGAP uses Serial Analysis of Gene Expression (SAGE) [1]. SAGE is a powerful tool that allows the analysis of overall gene expression patterns with digital analysis. Because SAGE does not require a pre-existing clone, it can be used to identify and quantitate new genes as well as known genes. SAGE works on three principles: 1. A short sequence tag (10-14bp) contains sufficient information to uniquely identify a transcript provided that the tag [2] is obtained from a unique position within each transcript; 2. Sequence tags can be linked together from long serial molecules that can be cloned and sequenced; and 3. Quantitation of the number of times a particular tag is observed provides the expression level of the corresponding transcript. Over 5 million transcript tags from more than 100 human cell types have been assembled. To enhance the utility of this data, the CGAP SAGE project created SAGE Genie, a set of tools for the analysis and presentation of SAGE data within a biological context [3]. The SAGE Genie provides highly intuitive, visual displays of human and mouse gene expression, based on a unique analytical process that reliably matches SAGE tags, 10 or 17 nucleotides in length, of known genes. SAGE Genie provides an automatic link between gene names and SAGE transcript levels, accounting for alternative transcription and many potential errors. These advances informatics provide a rapid and intuitive view of transcript expression in the human body or brain, displayed on the SAGE Anatomic Viewer. The lungs are a vital organ in our body, located in our chests. These pairs of cone-shaped breathing organs bring oxygen into our body and releases carbon dioxide. The Human body is made up of millions of cells. These cells normally divide and multiply in an orderly fashion. New cells replace older cells and shape our growth patterns by applying the specific genes and living habits of the body. Cell birth and renewal is the processes occur constantly in each and every living body. When an action of uncontrolled growth of abnormal cells occur, the population of cells become overflowing, and instead of creating thin, uniformly arranged cells they create lumps in the tissue and form a tumour. Lung cancer is a disease of uncontrolled cell growth in tissues of the lung. This growth may lead to metastasis, which is the invasion of adjacent tissue and infiltration beyond the lungs. Lung cancers can arise in any part of the lung. The vast majority of primary lung cancers are carcinomas of the lung, derived from epithelial cells. Lung cancer, the most common cause of cancer-related death in men and women, is responsible for 1.3 million deaths worldwide annually, as of 2004 [4]. The most common symptoms are shortness of breath, coughing (including coughing up blood), and weight loss [5]. An estimated 219,440 new cases of lung cancer are expected in 2009, accounting for about 15% of cancer diagnoses. The incidence rate is declining significantly in men, from a high of 102.1 cases per 100,000 in 1984 to 73.2 in 2005. In women, the rate is approaching a plateau after a long period of increase. Lung cancer is classified clinically as small cell (14%) or non-small cell (85%) [6] for the purposes of treatment. The most common cause of lung cancer is long-term exposure to tobacco smoke. Lung cancer is a highly aggressive malignancy presenting as metastatic disease with extremely poor prognosis. A comprehensive understanding of molecular genetics of lung cancer is required in order to develop new approaches for early diagnosis and targeted therapy. To identify genes involved in lung cancer, a method i.e. SAGE, for global analysis of gene expression patterns is employed. Several genes were identified as being differentially expressed. There are many cancerous genes which are highly expressed in cancerous diseases but not having tertiary structure. So these require structure validation for further drug development to inhibit the cancerous diseases. IGL@ (Immunoglobulin Lambda Locus) gene is one of the key factor gene which is over-expressed in lung cancer, may be helping the development and growth of the tumour. IGL@ is a region on human chromosome 22 that contains genes for the lambda light chains of antibodies (or immunoglobulins). The total number of human IGL genes per haploid genome is 84-93 (90-99 genes, if the orphons are included) of which 37-42 genes are functional. Proteins encoded by the IGL locus are the immunoglobulin lambda chains. The human IGL locus is located on chromosome 22 on the long arm, at band 22q11.2. The orientation of the locus has been determined by the analysis of translocations, involving the IGL locus, in leukemia and lymphoma. Immunoglobulins recognize foreign antigens and initiate immune responses such as phagocytosis and the complement system. Each immunoglobulin molecule consists of two identical heavy chains and two identical light chains. There are two classes of light chains, kappa and lambda. This region represents the germline organization of the lambda light chain locus. The locus also includes several non-immunoglobulin genes, many of which are pseudogenes or are predicted by automated computational analysis or homology to other species. Our goal is to identify molecular-level differences between normal human tissue and cancerous tissue. Specifically we have identified genes that are over-expressed in cancerous tissue derived from the lung through SAGE Genie, available at the CGAP website. The output is a list of short sequence tags and the number of times it is observed. Statistical methods can be applied to tag and count lists from different samples in order to determine which genes are more highly expressed. An attempt was to predict the 3D model structure for the new protein drug target using the spatial restraints obtained from the template structure and evaluate the structure using the validation server. The IGL@ protein 3D model was generated using MODELLER9v7. Computational modeling has grown with a faster rate to overcome the difference between protein sequences available and structures determined using crystallography and spectroscopic techniques. These techniques have added high levels of understandings in structural studies to functional role of protein involved in efforts to improve human health. At the end of structure validation brings final appraisal in modeling efforts.

Materials and Methods

Identification of Highly Expressed Genes through SAGE Genie

The SAGE technique has been extensively used for the genetic analyses of various types of cancers consistent with its conception in an oncology laboratory. It has been used to create a Tumour Gene Index, an archived database of SAGE tags from many different types of cancers or tissues, on the CGAP Web site. There are many tools and techniques available for studying and analyze gene expression. Here an attempt had been made to use SAGE genie as one of the main method for carrying out SAGE analysis. SAGE Genie provides a computational platform on which not only more than two horizontal comparisons (e.g., normal brain versus brain tumors) but also a nearly infinite number of vertical comparisons (e.g., different tissue or organ types) in gene expression at a global scale can be conducted. The data output can be presented with interfaces such as SAGE Digital Gene Expression Display (SDGED), SAGE Anatomic Viewer(SAV), SAGE Experimental Viewer(SEV), SAGE Absolute Level Lister(SALL), SAGE Library Finders(SLF) for any given SAGE tag or gene transcript of interest, thus providing a quick glance at, when and where a gene may be expressed. The SDGED distinguishes significant differences in gene expression profiles between two pools of SAGE libraries. It evaluates the statistical significance of the differences using the sequence odds ratio and a statistics analysis . The user has the option of comparing two human SAGE library pools, one human SAGE library pool against a user-defined file, or two user-defined files. User-defined files consist of lines of tab-separated SAGE tags and their frequencies. The SAGE libraries of Lung were used for studying differential expression. There are four libraries available in SDGED. Out of them three are made up of cancerous tissue while the first one is made up of normal tissue. These SAGE libraries named for SDGED have been labelled by using a consistent naming convention that contains organ site of tissue origin, tissue histology or pathology, a code for type of tissue purification (or culturing), and a unique identifier. Name of the four lung libraries were given. Library 1: SAGE_Lung_normal_B_1, Library 2: SAGE_Lung_adenocarcinoma_MD_L9, Library 3: SAGE_Lung_adenocarcinoma_B_1, Library 4: SAGE_Lung_adenocarcinoma_MD_L10. SDGED identified genes that are differentially expressed after comparing the gene expression in normal lung tissue to gene expression in cancerous lung tissue. All the statistical parameters were needed to be set i.e. expression factor (F) = 2, false discovery rate (Q) = 0.1. Increasing F, or decreasing Q, will increase the stringency of the search; the search will yield fewer genes, but their differential expression will be more statistically significant. F is set by default to "2" but this number may be set to any number greater than or equal to 1. As F increases, fewer results will be reported. Range of Q should be within 0 (show only most significant results) to 1 (show all results). Q is computed using the Benjiamini Hochberg algorithm. There are so many cancer-associated genes that are over-expressed in lung cancer tissue in a comparative way. The genes that have been identified to date have been categorized into two broad categories, depending on their normal functions in the cell. 1. Genes whose protein products stimulate or enhance the division and viability of cells. This first category also includes genes that contribute to tumour growth by inhibiting cell death. 2. Genes whose protein products can directly or indirectly prevent cell division or lead to cell death. A major goal of structural biology is to predict the three-dimensional structure from the amino acid sequence of a protein which would lead to smart drug discovery, a pursuit that has not yet been realized for many proteins. Hence there are many important proteins where the sequence is available but the three-dimensional structure is not known. Alternative computational strategies are being applied to develop models of protein structure when the data from X-ray diffraction or NMR spectroscopy are not available.

Sequence Retrieval of IGL@ gene

Amino acid sequences retrieved from swissprot/uniprot provides descriptions of a nonredundant set of proteins including their function, domain structure, posttranslational modifications and variants [7, 8]. Immunoglobulin Lambda Locus protein was retrieved as query sequence with total length of 236 amino acids. 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 [9].

Template for Modeling

Structural homologous entries were obtained for IGL protein through local alignment search using BlastP (Basic Local Alignment Search Tool) [10] against Protein Data Bank (PDB) [11, 12]. Results from BlastP show 95% identity [13, 14], 96% positives, 417 score bits and 2e-117 E value with above query from which is a structure of Proprotein convertase subtilisin/kexin type 9 (PCSK9) from Cabbage looper and E.coli Sp. 3H42 is PDB entry ID of template which is a X-ray crystallized structure at 2.30 Ang. and selected for backbone alignment with two domains identified in secondary structure studies. Comparison of homology models with known structure (Template) may also reveal similarities which allow biochemical and biological functions to be inferred [15].

Comparative Modeling of IGL@ protein and structure validation using PROCHECK

The BlastP alignment was further refined using sequence alignments in ClustalX [16, 17]. This alignment was used for comparative modeling to build 3D model by satisfaction of spatial restraints using Modeller9v7 [18, 19]. Modeller implements an automated approach to comparative protein structure modeling by satisfaction of spatial restraints. Briefly 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. A protein 3D structure was predicted based on crystal structure of 3H42. Ramachandran Analysis was performed to determine the stability of the modeled structure. Subsequently the model structure was validated using PROCHECK [20, 21] which determine stereo chemical aspects along with main chain and side chain parameters with comprehensive analysis [22]. Optimization and analysis of bond length and bond angles is reffered from Cambridge Structural Database, CSD after studying 100, 00 Structures [23].

Results and Discussion

SAGE Genie is a logistically laid out suite of bioinformatics tools that allow automatic and reliable matches of SAGE tags to known gene transcripts. This process was accomplished first by filtering out experimentally obtained SAGE tags [24] that had incorrect linker sequences, appeared only once, or were generated by sequencing errors, from millions of tags collected from over 100 different human cell types as part of the National Institutes of Health Cancer Genome Anatomy Project. The resulting confident SAGE tags (CSTs) then were used to evaluate and match the virtual SAGE tags predicted from known mRNA transcript (cDNA) sequences of different publicly available databases, including full-length cDNAs or 3' ESTs. The virtual tags were divided into different groups based on the origin of the databases from which the tags were generated, the absence and presence of poly adenylation signals and poly(A) tails, and whether the tags represented differentially spliced or internal (non-3) transcript sequences. The match in percentage of virtual tags to CSTs allows ranking of available databases with known transcript sequences. Reciprocal cross-referencing between virtual tags and CSTs provides not only the best match of a CST to a known gene transcript sequence, but also confirmation that experimentally obtained SAGE tags indeed come from mostly 3' ends of mRNA transcripts. The resulting bioinformatics interface allows automatic tag-to-gene identification, measurement of gene expression normalized to the occurrence of a tag per 200,000 tags collected from a SAGE experiment, and the origins from which a tag is counted. Here the normal library is compared against all 3 libraries to find out the tags of differential expression and result obtained contain the information about the genes that are not found in cancerous cell, down expressed in cancer, over expressed in cancer cell and not found in normal but only expressed in cancerous cell respectively. The SDGED results page contains: 1. the UniGene Build number, 2. the total number of sequences or tags in each pool, 3. the total number of libraries in each pool, and 4. a table listing the genes or tags found to be expressed with a statistically significant difference between pools A and B (Table 1). Used together, the seqs odds ratio and the significance test provide a measure of confidence that the difference in the expression of a gene or tag is "real" and not due sampling error. The result is further analyzed and from the result following interesting patterns are emerged. Some of the genes are found in normal case but totally absent in cancerous library. The cause behind it is in confusion but it may be a case that inefficient enzymatic reactions that occur during the generation of a SAGE library can lead to inaccurate data. Few genes are present in both the library but down expressed in cancerous cell. These genes can be predicted as tumour suppressor gene. Some of the genes are over expressed in cancerous library. They may be tumour proliferator gene. By identifying these genes and their reason for over expression can be studied. Through SDGED total 787 tags were retrieved. Among 787 tags we had taken only 15 genes (Table 2) that are over-expressed in lung cancer tissue according to their frequency (B>A, where B is the pool of genes from cancerous tissue and A is the pool of genes from normal tissue). Among them IGL@ gene is the best novel marker for lung cancer carcinogenesis. The IGL@ gene is found in both cancer libraries i.e. SAGE_Lung_adenocarcinoma_MD_L9 and SAGE_Lung_adenocarcinoma_MD_L10, but not in the normal library. It was found 1225 times in the pool of genes from the cancerous tissue. This looks like a gene that is really over-expressed in lung cancer. As IGL@ is the best gene for tag, the link under Tag that has the nucleic acid sequence AAGGGAGCAC showed the SAGE Anatomy Viewer "Fig. (1)". The picture of a body under SAGE Anatomic Viewer displayed the data comparing expression of IGL@ in a variety of normal and cancerous tissues "Fig. (2)". The red was indicating over-expression. Noted that IGL@ over-expression is associated with lung cancer, but not with other types of cancer. The "Monochromatic SAGE/cDNA Virtual Northern" link under "Gene Expression Data" on the IGL@ Gene Info page had given a visual look at the expression level of IGL@ in normal and cancerous tissues. IGL@ is over-expressed in lung cancerous tissue, compared to normal lung tissue. A northern is a type of mRNA blot used to measure gene expression levels. High expression levels will yield a dark band like the one on the Virtual Northern "Fig. (3)". We screened out lots of genes which are highly expressed in Lung cancer. Among them IGL@ gene is the best novel marker for lung cancer carcinogenesis. IGL@ is an important gene where the sequence is available but the three-dimensional structure is not known. Knowledge of the three-dimensional structure is a prerequisite for the rational design of site-directed mutations in a protein and can be of great importance for the design of drugs. Homology modeling method is to build the protein molecule from the primary sequence of desired protein with help of template as 3-D structure. As the 3D structure of IGL@ gene is unknown, for that purpose we had proceeded to predict the three-dimensional structure. The quality of the homology model is dependent on the quality of the sequence alignment and template structure. The approach can be complicated by the presence of alignment gaps (commonly called indels) that indicate a structural region present in the target but not in the template, and by structure gaps in the template that arise from poor resolution in the experimental procedure (usually X-ray crystallography) used to solve the structure. Comparative modeling can produce high-quality structural models when the target and template are closely related, which has inspired the formation of a structural genomics consortium dedicated to the production of representative experimental structures for all classes of protein folds. The general information of IGL@ gene was given in (Table 3). The protein sequence of IGL@ gene was extracted from SwissProt / Uniprot KB with accession number Q6GMX3, which is a curated and annotated database. Swiss-Prot provides descriptions of a nonredundant set of proteins, including their function, domain structure, posttranslational modifications and variants [25]. The protein was queried against PDB using BlastP to find structural homologues. The Fasta format sequence of IGL@ is subjected to NCBI – BLAST which yields the template structure as Proprotein convertase subtilisin/kexin type 9 (PCSK9) with PDB id: 3H42.The crystal structure 3H42, having residue length 217 was identified as structural homologous protein showing 95% identity with IGL@ protein [26]. The PCSK9 structure was used to find out topologically equivalent residues based on structural alignment and the structurally conserved regions (SCRs) were modelled. In the absence of experimentally determined protein structures, homology-based models may serve as working models for the investigation of sequence-structure-function relationships between IGL@ and Proprotein convertase subtilisin/kexin type 9 (PCSK9). Homology modelled structures may be of too resolution to characterise the protein-protein or protein-DNA contacts at the atomic level, but they can suggest which sequence regions are individual amino acids are essential components of the binding surfaces. In particular, identification of the amino acids potentially involved in protein-DNA contacts may guide mutagenesis experiments aimed at the engineering protein variants with novel specificities. However, comparative modeling requires a homologous template structure to be identified and the sequence of the protein of interest (a target) to be correctly aligned to the template. The target and template protein sequences were aligned using ClustalX "Fig. (4)" and generated an alignment file to develop *IGL@* protein model using Modeller9v7. Modeller calculates three-dimensional (3D) model of a protein by optimally satisfying spatial restraints derived from the alignment and expressed as probability density functions (pdfs) for the features restrained. The modelled structure was selected for validation by checking the stereochemistry "Fig. (5)". A good quality model would be expected to have over 90% in the most favoured regions. PROCHECK analysis reveals in Ramachandran plot concluding Phi and Psi angles to contribute in conformations of amino acids excluding glycine and proline with 93.4% in (184) residues in most favoured region, 4.6% (9) [27] in additional allowed region, 1.5% (3) in generously allowed region and 0.5% (1) residues in disallowed region "Fig. (6)". Although standards allow model acceptance in 90% residues in most favoured region [28] less similarity between query and template may account for it and additionally other statistical parameters are found in support of structure modelled. Plot statistics is collected in “Fig. (7)”. This outcome was compared with the validation report of template structure (3H42) which has shown 91.5% residues in most favored region and 8.0% residues in allowed region (Table 4). These results suggest that the model is valid with good stereochemical quality. Glycine and Proline found in regions of acceptance and shown in “Fig. (8)” [29]. Main-chain parameters and side-chain parameters calculated at 2.0 Ang [30 - 33] of resolution signifies modeling of IGL@ Protein. Root Mean Square Distances from planarity is found below 0.02 when plotted against amino acids frequency in sequence. This reveals that the predicted structure could serve as a good target model for the design of the drug. Thus over all it can be said that 98.0% of the residues are in the allowed portions of the plot and can be exploited for drug designing after further energy minimising the model and performing molecular dynamic simulations. Thus, the initial structure of modeling was revised by means of refining loops and rotamers, checking bonds and adding hydrogen atoms and then molecular dynamic simulations optimise the initial modeling structure. The refined model structure thus obtained after energy minimisation and molecular dynamic simulations was used for docking studies.

Conclusion

Although there is no doubt that SAGE Genie has greatly enhanced the utility of SAGE in global analysis of gene expression, challenges remain for the method with regard to the comprehensiveness in gene coverage as a function of the number of tags needed to be counted for each SAGE screen [34] and SAGE tags that either failed to match any known gene transcript sequences or matched more than one known transcript. But for biomedical and agricultural research, there seem to be an infinite number of comparisons in gene expression with different biological systems, disease states, developmental stages, drug treatment, and stress conditions, etc., which need to be conducted. Such efforts will still require the use of technologies such as arrays and differential display as well as SAGE for custom gene-expression analysis. With an intuitive web-site-based interface, SAGE Genie offers one of the most comprehensive collections of gene expression data across many different cancer and tissue types, making it a valuable tool for a quick glimpse of expression patterns of any known human gene sequences with the need of only a few strokes on a computer keyboard. SAGE Genie enhanced tag predictions need to be continually improved. SAGE Genie could prove to be a very powerful tool for archiving and analyzing the expression profile for any given gene under any biological context. Thus, the protein structure for the novel gene IGL@ found in lung cancer has been predicted. Significant research has to be carried out to understand the structure, function and variants of the IGL@ gene. Modeling studies manifested good stereochemical placement of main chain parameters. Bond angles and bond lengths are under confined limits although side chain modeling introduced some levels of displacement of residues beyond most favoured regions. More efforts in structural analysis in concern with mutational studies can provide better insight towards development of drug resistance profiles of this gene. Thus the present study of modeling of IGL @ gene has brought future prospective to an early diagnosis and treatment against lung and provide better health standards for community. From Ramachandran plot of modeled protein, the number of amino acids in allowed region is found to be 93.3%. This shows the predicted structure of protein is reliable and can be used to find its functional regions. The structure modeled was found to be stereochemically stable from the Ramachandran analysis. The refined model structure can be exploited for drug designing after further energy minimizing the model and performing molecular dynamics simulations. An attempt has been made in the present *in silico* study to provide structure of IGL@ protein which would definitely assist structural based drug design community to accelerate the search of suitable inhibitors for it. The principal objective is to identify, exploit and analysis of new molecular drug targets at structural level. This computational approach will lead to the discovery and structural development of novel drug targets. Computational community can further explore active site of IGL@ for binding of drug and apply docking studies to indentify amino acids involved in electrostatic, hydrophobic and hydrogen bond formation with inhibitors of this protein.

Acknowledgements

The authors are grateful to Department of Biotechnology (DBT), Ministry of Science and Technology, Govt. of India for their financial assistance.

References

[1] http://cgap.nci.nih.gov/SAGE.

[2] Velculescu V. E., Zhang L., Vogelstein B. and Kinzler K. W. (1995) *Science,*270(5235), 484–487.

[3] El-Deiry W. S. (1998) *Semin Cancer Biol.,* 8,345–357.

[4] W. H. O. (February2006) World Health Organization. Retrieved 2007-06-25.

[5] Minna J. D., Schiller J. H. (2008) *Harrison's Principles of Internal Medicine* (17th ed.), McGraw-Hill. pp, 551–562.

[6] Travis W.D., Travis L.B. and Devesa S.S. (1995) *Cancer*, **75** (Suppl.1), 191–202.

[7] Bairoch A., Boeckmann B., Ferro S. and Gasteiger E. (2004) *Swiss-Prot. Brief.Bioinform*., 5, 39-55.

[8] The Uniprot Consortium, The universal protein resource (Uniprot) (2008) *Nucleic Acids Res*, 36, 190-195.

[9] Boeckmann B., Blatter M. C., Famiglietti L., Hinz U., Lane L., Roechert B. and Bairoch A. (2005) *Comptes Rendus Biologies*, 328,882-899.

[10] Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W. and Lipman D.J. (1997) *Nucleic Acids Research*, 25(17), 3389–3402.

[11] Berman H.L., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., ShindyalovI.N. and Bourne P. E. (2000) *Nucleic Acids Research*, 28(1), 235-242.

[12] Dowlathabad M. R., Anuraj N., Mukesh Y., Showmy K.S. and Disha P. **(**2010) *International Journal of Bioinformatics Research*, 2(1), 05-09.

[13] Burley S. K. (2000) *Nat. Struct. Biol.*, 7 (Suppl.), 932-934.

[14] Marti-Renom M. A., Stuart A. C., Fiser A., Sanchez R., Melo F. and Sali A. (2000)*Annu. Rev. Biophys. Biomol. Struct*., 29, 291-325.

[15] Thornton J. M., Todd A. E., Milburn D., Borkakoti N. and Orengo C. A. (2000)Nat. Struct. Biol., 7 (Suppl.), 991-994.

[16] Thompson J.D., Gibson T.J., Plewniak F., Jeanmougin F. and Higgins D.G. (1997) *Nucleic Acids Research*, 25(24), 4876–4882.

[17] Higgins D. G., Thompson J. D. and Gibson T. J. (1996) *Methods Enzymol., 266*, 383-402.

[18] Sali A. and Blundell T.L. (1993) *Journal of Molecular Biology*, 234(3), 779-815.

[19] Rakesh S., Pradhan D. and Umamaheswari A. (2009) *International Journal of Bioinformatics Research*, 1(2), 85-92.

[20] Laskowski R.A., MacArthur M.W., Moss D.S. and Thornton J. M. (1993) *Journal of Applied Crystallography*, 26(2), 283-291.

[21] Laskowski R. A., Rullmannn J. A., MacArthur M. W., Kaptein R. and Thornton J. M. (1996) *J. Biomol*., 8, 477-486.

[22] Morris A. L., MacArthur M. W., Hutchinson E. G. and Thornton J. M. (1992) *Proteins*, 12, 345-364.

[23] Allen F. H., Bellard S., Brice M. D., Cartwright B. A., Doubleday A., Higgs H., Hummelink T., Hukkelink-Peters B. G., Kennard O., Motherwell W. D. S., Rodgers J. R. and Watson D. G. (1979) *Acta. Cryst*., B35, 2331-2339.

[24] Saha S., Sparks A. B., Rago C., Akmaev V., Wang C. J., Vogelstein B., Kinzler K. W. and Velculescu V. E. *(2002) Nat Biotechnol,* 20,508–512*.*

[25] http://www.genome.gov.

[26] Hirak J. C., Sayak G., Protip B., Paushali R. and Abhijit D. (2010) *International Journal of Bioinformatics Research*, 2(2), 01-06.

[27] Frosst P., Blom H.J., Milos R., Goyette P., Sheppard C.A., Matthews R.G., Boers G.J., Heijer M., Kluijtmans L.A., Heuvel L.P. and Rozen R.A. (1995) *Nature Genet*., 10,111-113.

[28] http://www.biochem.ucl.ac.uk.

[29] Mukesh Y., Anuraj N., Girish S. R., Aditya J., Ankit V. and Priyanka G. (2010) *International Journal of Bioinformatics Research*, 2(1), 01-04.

[30] http://www.pdb.org.

[31] Schapira M., Raaka B.M., Samuels H.H. and Abagyan R. (2001) *BMC Structural Biology,* 1, 1.

[32] Schapira M., Abagyan R. A. and Totrov M. M. (2002) *BMC Struct. Biol.,* 2(1), 1.

[33] Schapira M., Abagyan R. A. and Totrov M. M. (2003) *J. Med. Chem.,* 46(14), 3045-3059.

[34] Stollberg J., Urschitz J., Urban Z. and Boyd C. D. *(2000) Genome Res.,* 10, 1241–1248*.*

Table 1- Description of SDGED Result page

Field	Description
Tag	The tag is hyperlinked to complete mapping information for the tag.
Gene	The gene symbol (or the UniGene cluster number, in the case of an anonymous gene) that has been mapped to the tag, using the best gene for the tag. If multiple genes are the best choices for the tag, then the gene symbol (or cluster number) is followed by "...". In some cases, a tag has been associated with an accession number for a transcript or EST but not with a UniGene cluster; in other cases, the tag has been associated with neither a UniGene cluster nor an accession.
Libraries A (or B)	The number of libraries which contain this tag
Tags A (or B)	Tag frequency in either Pool A or B.
Tag Odds (A:B)	The odds ratio uses a simple mathematical formula to provide a measure of the relative amount of a tag in pool A to Pool B.
Q	False discovery rate: the smaller the number, the more likely the result is not due merely to sampling error

Table 2- List of Differentially Expressed Genes

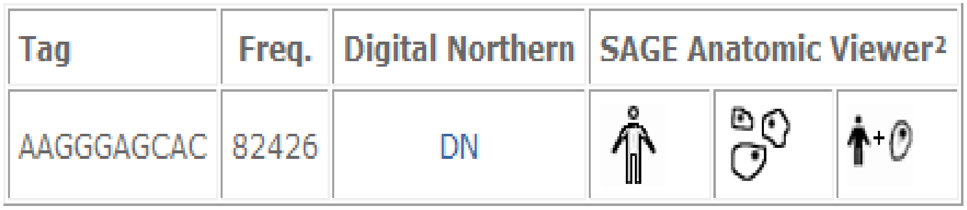
Tags	**Gene or Accession**	**Libraries**		**Tags**		**Tags Odds A:B**	**Q**
		A	**B**	**A**	**B**		
AAGGGAGCAC	IGL@	1	2	42	1225	0.05	3.32e-198
GAAATAAAGC	IGHG1	1	2	94	876	0.17	1.36e-90
CCCGTCCGGA	RPL13	1	3	86	630	0.21	1.38e-54
ATGGGATGGC	SFTPB	1	3	139	531	0.41	5.89e-21
CCTGCTGCAG	MUC5B	0	2	0	513	0	5.55e-107
CCTGTAATCC	RASA4	1	3	113	500	0.35	5.96e-25
CTCCCCCAAG	IGHA1	1	2	47	442	0.17	3.07e-45
ACGCAGGGAG	MALAT1	1	3	1	432	0.00	1.80e.87
CCCTGGGTTC	FTL	1	3	104	394	0.41	3.98e-15
GCCTGTATGA	RPS24	1	3	55	366	0.23	9.54.29
GCCGAGGAAG	RPS12	1	3	38	301	0.20	7.08e-27
AAGCTCGCCG	SCGB3A1	1	3	5	270	0.03	5.45e-47
ATGGCTGGTA	RPS2	1	3	50	259	0.30	1.99e-15
CTCCACCCGA	TFF3	1	3	5	257	0.03	2.00e-44
TCAGACGCAG	PTMA	1	3	6	253	0.04	1.83e-42

Table 3- General Information of IGL@ gene

Immunoglobulin Lambda Locus(IGL@) Gene	
Official symbol	IGL@
Official Full Name	immunoglobulin lambda locus
Other names or Designations	IGL; IGLC6; MGC88804; IGL@
Organism source	Homo sapiens
Gene ID	3535
RefSeq (mRNA)	NG_000002
Chromosome	22
Location	22q11.1-q11

Table 4- Comparison of validation reports of IGL@ protein model and crystal structure 3H42 template

Amino Acid Position	Modeled Protein (%)	3H42 Template Structure (%)
Most favoured region	93.4%	91.5%
Additional allowed region	4.6%	8.0%
Generously allowed region	1.5%	0.1%
Disallowed region	0.5%	0.3%



Tag	Freq.	Digital Northern	SAGE Anatomic Viewer2		
AAGGGAGCAC	82426	DN			

Fig. 1 - IGL@SAGE data

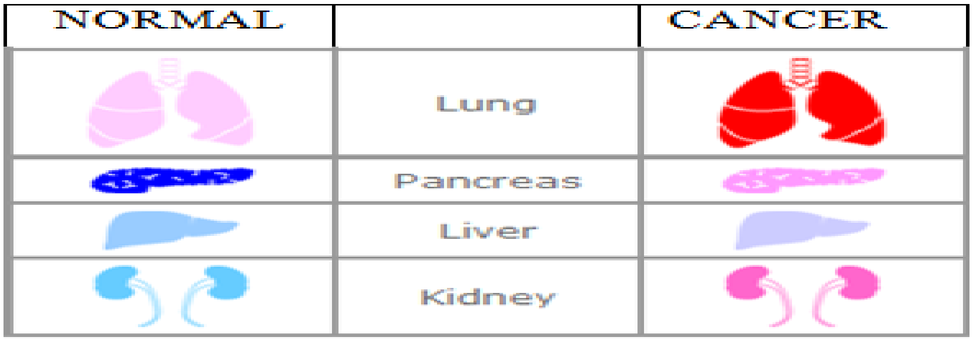
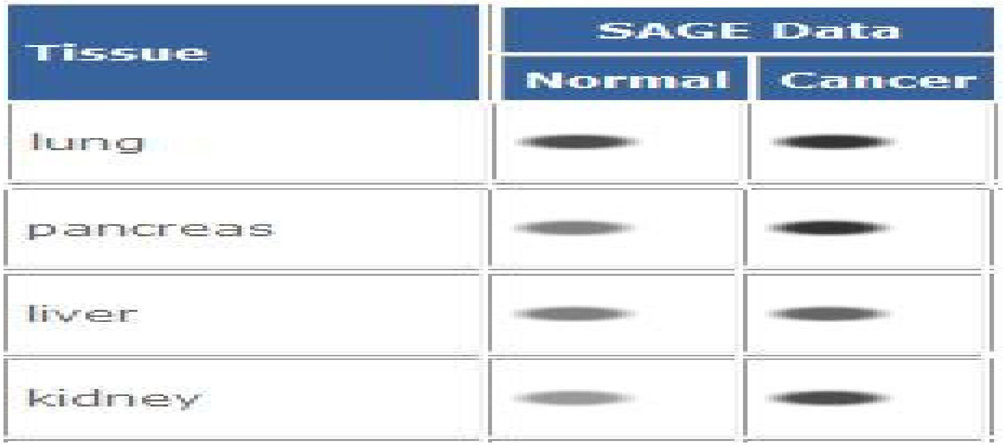


Fig. 2 - IGL@ is over-expressed in lung cancer. Blue indicates average expression level, red indicates high expression level.



Tissue	SAGE Data	
	Normal	Cancer
lung	▬	▬
pancreas	▬	▬
liver	▬	▬
kidney	▬	▬

Fig. 3 - Graphical image of IGL@ gene expression. Note that IGL@ expression is very high in lung cancerous tissue.

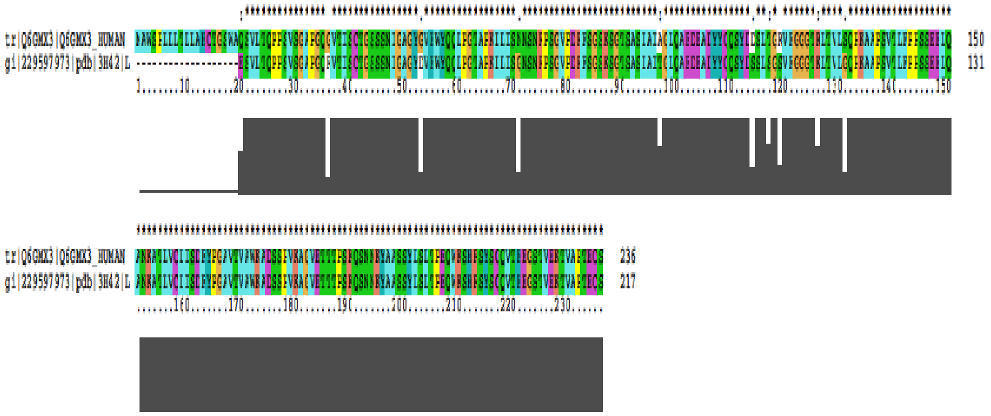


Fig. 4 - ClustalX [16, 17] alignment of target and template sequences

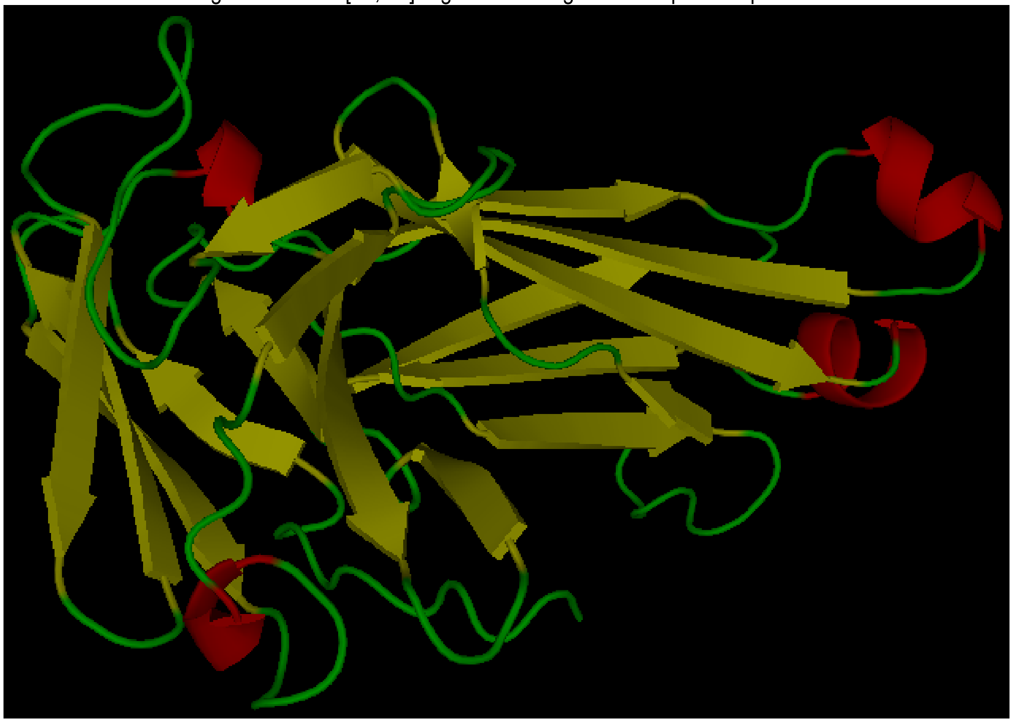


Fig. 5 - Shows the predicted Structure of the protein IGL@

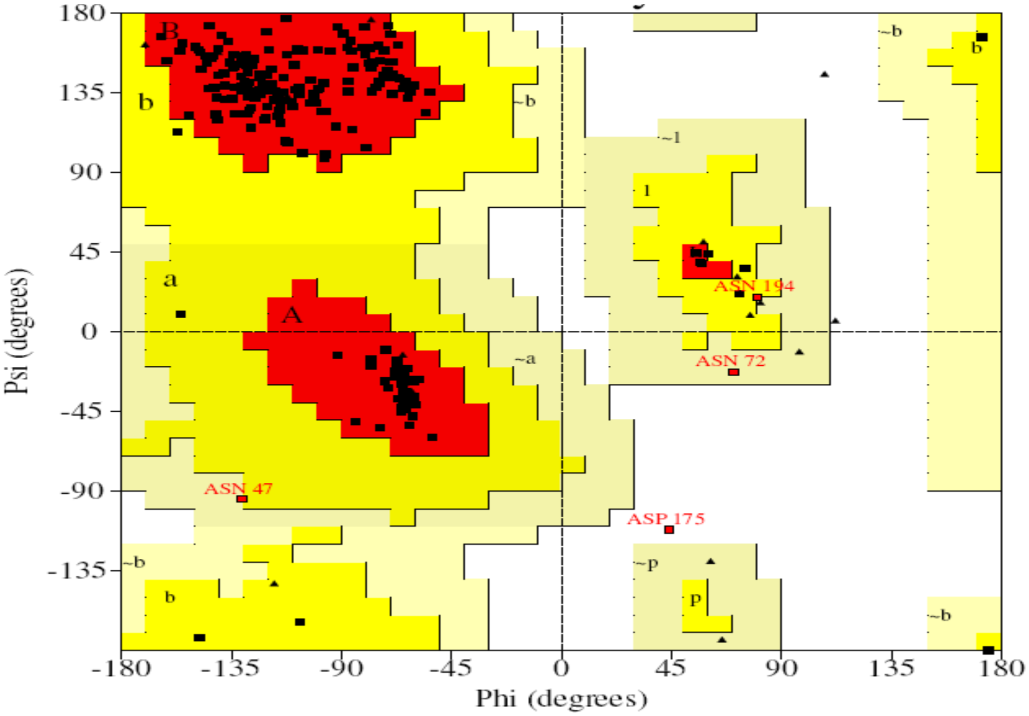
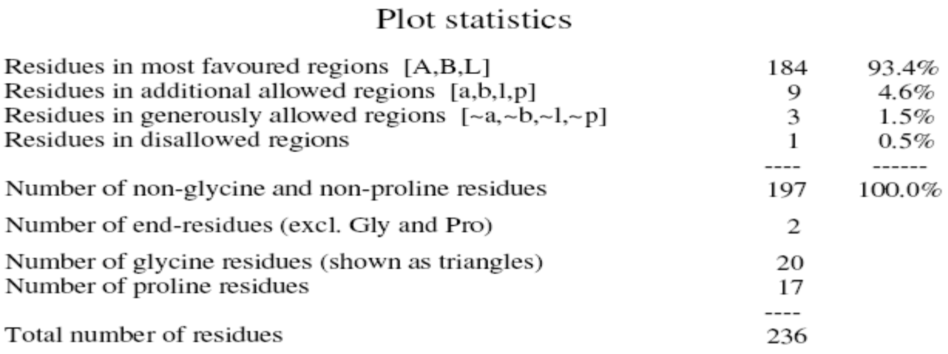


Fig. 6 - Procheck [20, 21] validation report of IGL@ protein model. 93.4% residues in most favoured region, 4.6% in additional allowed region and 1.5% in generously allowed region.



```
                                Plot statistics

Residues in most favoured regions  [A,B,L]               184      93.4%
Residues in additional allowed regions  [a,b,l,p]          9       4.6%
Residues in generously allowed regions  [~a,~b,~l,~p]      3       1.5%
Residues in disallowed regions                              1       0.5%
                                                         ----    ------
Number of non-glycine and non-proline residues           197    100.0%

Number of end-residues (excl. Gly and Pro)                  2

Number of glycine residues (shown as triangles)            20
Number of proline residues                                 17
                                                         ----
Total number of residues                                  236
```

Fig. 7 - Ramachandran plot statistics of IGL@ Protein

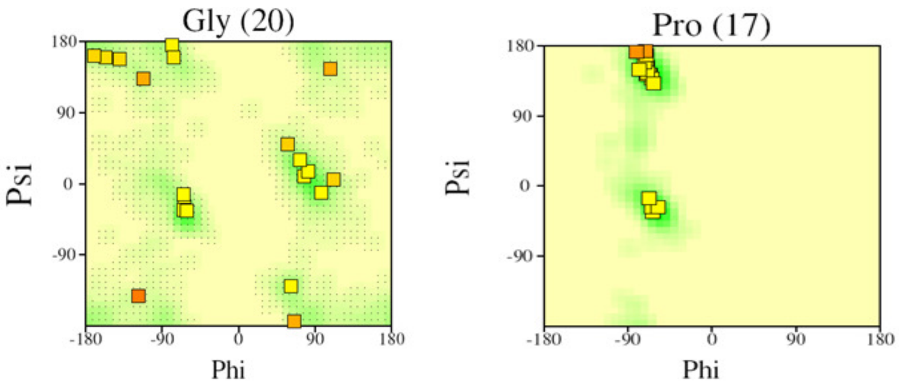


Fig. 8 - Ramachandran plot for Glycine and Proline