Functional annotation and sequence analysis of protein BAH14511.1 for homology modeling to find potential binders by virtual screening and docking analyses

A Major Project dissertation submitted

in partial fulfilment of the requirement for the degree of

Master of Technology

In

Bioinformatics

Submitted by

Manisha

(DTU/12/MTECH/395)

Delhi Technological University, Delhi, India

Under the supervision of

Dr Vimal Kishor Singh

Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road,
Delhi-110042, INDIA

CERTIFICATE

This is to certify that the M. Tech. Dissertation entitled **"Functional annotation and sequence analysis of protein BAH14511.1 for homology modeling to find potential binders by virtual screening and docking analyses."**, submitted by **MANISHA (DTU/12/MTECH/395)** in partial fulfillment of the requirement for the award of the degree of Master of Engineering, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate's own work carried out by her under my guidance.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honoring of any other degree.

Date:

Dr Vimal Kishor Singh
(Project Mentor)
Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering, University of Delhi)

ACKNOWLEDGEMENT

First and foremost, my sincere thanks to Prof B. D. Malhotra, Head, Department of Biotechnology, Delhi Technological University, for giving me an opportunity to study and work in this prestigious institute.

No words are adequate to express my feeling of profound gratitude to my mentor Dr. Vimal Kishor Singh not only for giving me the opportunity to work under him, but also for his guidance, valuable suggestions and persistent encouragement and generosity which inspired me to submit this work in the present form. He has been responsible for smoothing all the rough edges in this investigation by their constructive criticism and deep insight.

MANISHA
2K12/BIO/13

CONTENTS

LIST OF FIGURES

LIST OF TABLES

.

LIST OF ABBREVIATIONS

ADME- Absorption, Distribution, Metabolism, Excretion

ATP- Adenosine triphosphate

CDCP- Centre for Disease Control & Prevention

cGMP- cyclic guanosine-3',5'-monophosphate

CTNNB1-Catenin Beta-1

EGF- Epidermal Growth Factor

EGFR- Epidermal Growth Factor Receptor

ERK- Extracellular signal Regulated Kinase

FOXO4- Factor Forkhead box O

GRAVY- Grand Average of hydropathycity

GTP- Guanosine triphosphate

JNK- c-Jun N- terminal Kinase

MAPK- Mitogen Activated Protein Kinase

NAACCR- North American Association of Central Cancer Registries

NCHS- National Centre for Health Statistics

NCI- National Cancer Institute

NTP- Nucleotide Tri Phosphate

PDB- Protein Data Bank

PDE- Phosphodiesterases

PKG- cGMP-dependent Protein Kinase

TCF- T-Cell Factor

Functional annotation and sequence analysis of protein BAH14511.1 for homology modeling to find potential binders by virtual screening and docking analyses.

MANISHA

Delhi Technological University, Delhi, India

ABSTRACT

A protein with Accession No. BAH14511.1 isolated form tongue tumor tissue of *Homo sapiens* was taken. Non-redundant BLAST revealed this protein to be identical with cGMP-dependent protein kinase II isoform b for which the sequence has been put online at NCBI site recently. PKGII has been reported to be very important for the regulation of different signalling pathways related to cancer and hence, prediction of the structure of PKGIIb would be important for finding potential binders that can further be used for the purpose of designing a drug that might help in the regulation of PKGIIb. For functional annotation, PROTPARAM, SMART, TMHMM, SignalP, SecretomeP, NetChop and NetPhos were used. BLAST against PDB entries was used to find out any homologs for which secondary and tertiary structures were compared with that of the protein by using different tools before homology modelling. The 3D structure predicted by Homology Modelling was validated by Verify3D and WhatIF. The binding site was predicted by using SiteMap which gives 5 sites but top scoring site was used. It was used for virtual screening to find structures that might act as ligands for it. This was done by using two- Schrödinger Fragments and NCI database. Many chemical structures were formed by different combinations of the fragments in Schrödinger Fragments which were then virtually screened. The virtual screening of Schrödinger fragments revealed that many chemical structures docked with the protein had a score of less than even -9.000 and were found to be following Lipinski's Rule. The predicted protein 3D structure was also used for docking to chemical structures of NCI database with docking score as less as -12.547 which correspond to a good docking. The top 5 structures, each from compounds made from Schrödinger fragments and NCI database, having lowest docking scores have been found as potential binders for the protein.

INTRODUCTION

Cancer is condition which is characterized by uncontrolled growth of cells or tissues. This uncontrolled growth of cells leads to tumour formation. This tumor may spread to nearby cells and affect nearby parts of body. Many types of cancer are there depending on the origin of the tumor. Cancer is a severe disease which is the cause of numerous deaths worldwide. The cases of deaths that occur from cancer have been estimated to increase in coming years. Oral cancer refers to the cancer of oral cavity. The protein BAH14511.1 has been isolated from tongue tumor tissue. This protein was found to be identical to PKGII which has role in the regulation of cancerous cells and their apoptosis. For finding potential binders, homology modelling followed by virtual screening and docking analyses was done.

Homology modelling is a method of predicting the 3D structure of a protein on the basis of the structure of its homologs (template). The homology is found out by aligning the sequence of the query protein by using BLAST. BLAST aligns the sequence of query protein with the sequences of the entries included in the database specified before run. For homology modelling, BLAST against PDB entries is done to find out homologs which already have a 3D structure determined. The secondary and tertiary structures are compared for both, query and template before moving onto modelling. Once the secondary and tertiary structures are found similar, modelling is done. Two approaches- knowledge based and energy based homology modelling can be adopted. In knowledge based method, the insertions are constructed and gaps are closed by using segments from the structures already known. In energy based method, the residues are constructed and refined on the basis of energy. The 3D structure of protein serves as the base for finding the chemical structures that would bind to its binding site predicted. For this, if the number of chemical structures to be tested, is very high then, the database, is first screened by virtual screening before docking, and if the database is not so big, direct docking is done.

Virtual screening is the method to screen a large database of compounds against a target to decrease the size of chemical compounds for further experimental screening. It helps to save time and effort that is usually associated with the identification of lead molecule. There are two types of screening- ligand based and receptor based. In ligand based screening, the properties of chemical compounds are compared with that of the ligands and, on the basis of similarity, the chemical structures are screened. In receptor-based screening, the 3D structure of the receptor protein is used to filter the compounds that interact with the active site of the receptor. For this work, receptor-based virtual screening was done. After compounds have been filtered by virtual screening, they are docked to the receptor protein.

Docking is the prediction of binding geometry of two interacting molecules. Docking can be done for a protein-protein, protein- chemical compound, and protein-nucleic acid complexes. Two types on docking approaches are there- local and global. In local docking, the site where the interaction is to be done is known, whereas no site is known in global docking. The site of the receptor is predicted by tools that look for binding sites. In our work, local alignment was carried out.

REVIEW OF LITERATURE

The beginning of cancer occurs in a cell, the building blocks of life. The normal conditions of the body are when old cells are dying and these cells get replaced by the new cells formed, but when this equilibrium gets disturbed, cancer occurs. In cancer, due to different reasons, new cells form but the old cells which must die do not die. Due to this imbalance, mass of cells forms. This mass of cell is called as 'tumor'. Tumors can be of two types: benign (Latin *bene* means 'well' and *–genus* 'born') and malignant (Latin *male* means 'badly' and *-genus* 'born'). Benign tumours are located at a specific location only, and do not spread to nearby cells and hence, are not classified as cancer. But, malignant tumors are those which spread to nearby cells. If malignant cancer keeps on spreading, it invades tissues and other body parts as well. This spreading of cancer from one to other body parts is called as metastasis (Greek *methistanai* means 'to change'). There are many causes of incidence of cancer including lifestyle, environmental factors, eating habits etc. Different factors for the prevalence of cancer in India have been depicted in Fig 1.

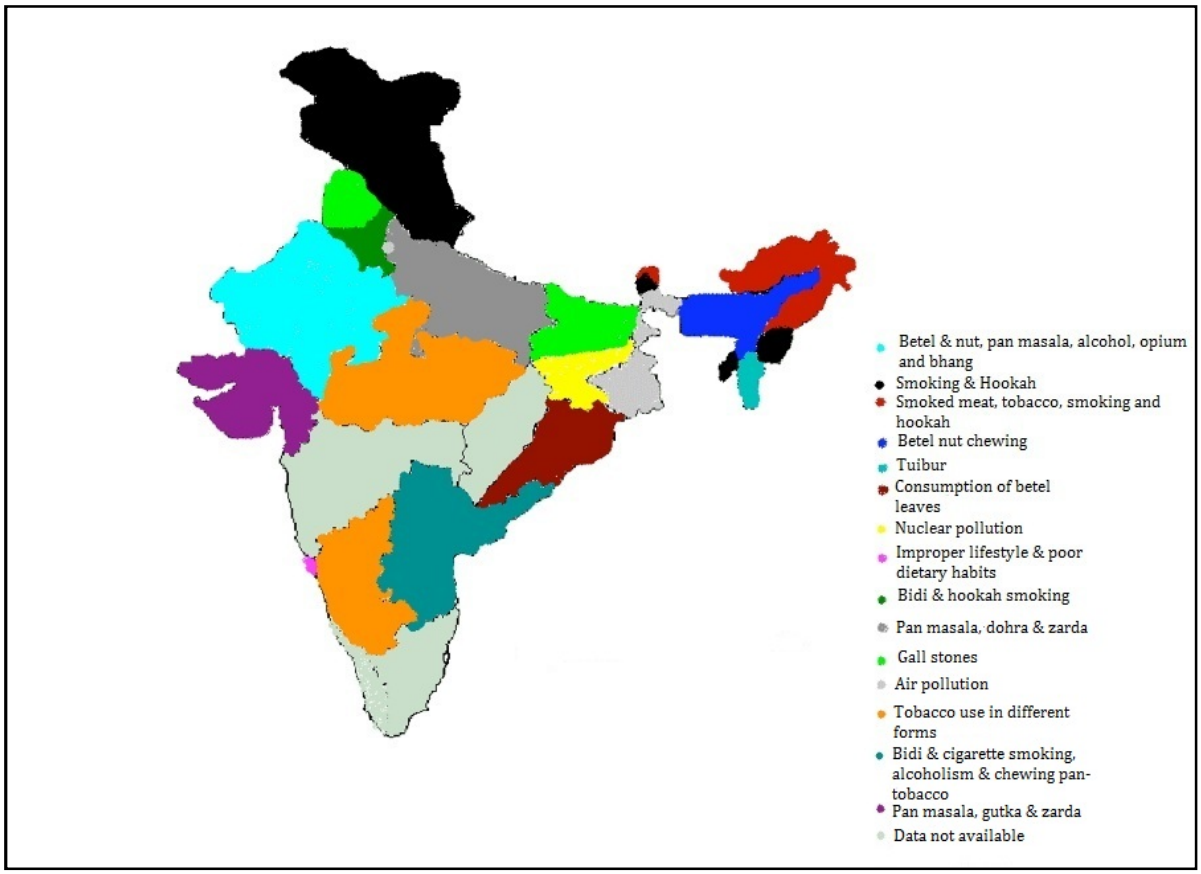


Figure 1: Diagrammatic description of different lifestyle habits & environmental factors leading to different cancer in some parts of the country (Taken from Ali et al 2011).

Cancer is a single word, but it comprises many diseases. More than 100 various types of cancer are there. The cancer is named on the organ from where it started Cancer if not treated leads to death and hence, it is a very serious problem worldwide. Every year, many people all around the world die because of one or the other type of cancer and this number is estimated

to increase in coming years. According to the incidence data from National Cancer Institute (NCI), Centers for Disease Control and Prevention (CDCP), and North American Association of Central Cancer Registries (NAACCR) and mortality data from National Centre for Health Statistics (NCHS), 1,665,540 new cancer cases and 585,720 deaths are estimated to occur in 2014 in United States alone (Siegel et al 2014). In India, it is estimated that cases for different cancers will increase from 979786 to 1,148,757 from 2010 to 2020 (Takiar et al 2010).

Oral cancer includes oral cavity and the back of it (oropharynx) and is majorly squamous cell carcinoma. It makes around 1-2% of all the cancers that arise, and, is the 6th most common type of cancer in the world and 3rd most common type of cancer in India and accounts for over 30% of total cancers in India (van der Waal et al 2013, Coelho et al 2012). The data for the incidence and mortality of Indian men and women as compared with men and women from all over the world, due to oral cancer, has been shown in Table 1.

	Male		**Female**	
	India	**World**	**India**	**World**
Incidence				
Annual new cases	53,842	1,98,975	23,161	1,01,398
Crude incidence rate	8.3	8	3.8	2.9
Mortality				
Annual new cases	36,436	97,919	15,631	47,409
Crude mortality rate	5.6	2.8	2.6	1.4

Table 1: Incidence and mortility data for for oral cancer in Indian men and women compared with men and women from across the world. Modified from Bruni et al 2014.

Kinases are the enzymes that carry out the transfer of a phosphate to a substrate at a specific location. Usually, the molecule from which phosphate are transferred are nucleotide triphosphates. This transfer results in the changes in substrate that are functional in nature. Most of these are protein kinases (Wolfertstetter et al 2013). Human kinome encodes more than 500 protein kinases and much more if splice variants are considered. These protein kinases contribute nearly 2% of the genome and are extremely important for the regulation of different biological events. The two most important type of protein kinases are serine/threonine kinases and tyrosine kinases.

Each protein kinase has two lobes which are functionally and structurally defined- N-lobe and C-lobe. N-lobe helps in the positioning of γ-phosphate of nucleotide triphosphate for catalysis. β-subdomain of C-lobe which contains 4 β sheets (β6-9) contains the machinery required for the phosphate transfer from nucleotide triphosphate to protein. β(8-9) of C-lobe flank the motif important for the recognition of one of the nucleotide-triphosphate bound Mg^{++} ions (Taylor et al 2011). The main NTPs are ATP and GTP which by the action of Adenylyl Cycalse and Guanylyl Cyclase, respectively get converted to cAMP and cGMP.

cGMP is a second messenger that is produced from GTP by the action of guanylyl cyclase (soluble or particulate). Three cGMP receptor proteins are known- cyclic nucleotide-gated (CNG) cation channels, cGMP-regulated PDEs and cGMP-dependent protein kinases (PKG or cGK) (Feil et al 2003) as shown in Fig 2.

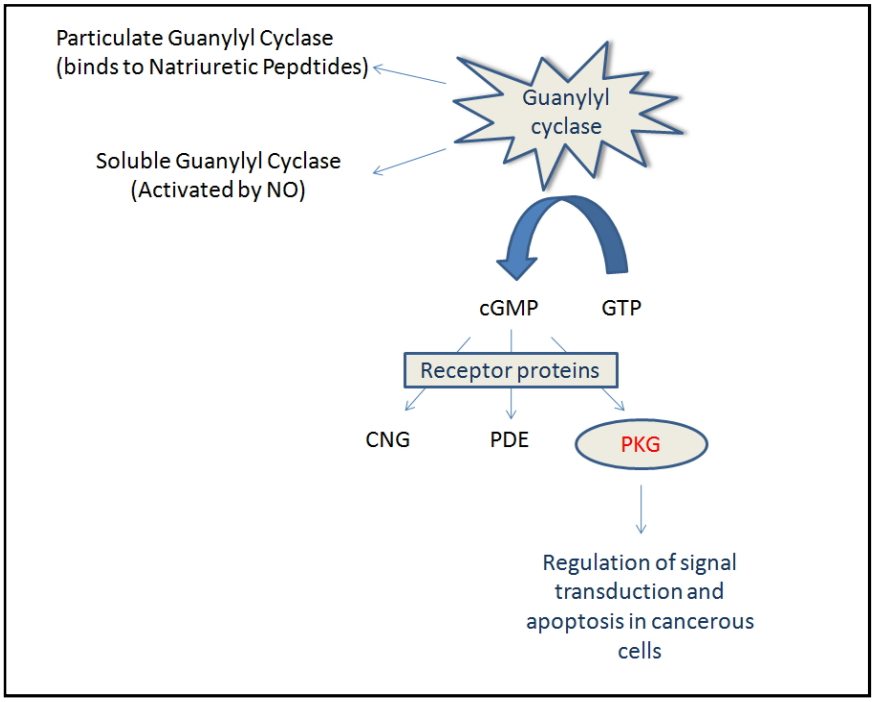


Figure 2: Production of cGMP by guanylyl cyclase and receptor proteins of cGMP.

cGMP-dependent protein kinases (PKG) belong to the family of serine/threonine protein kinases. Two types of PKG are there- PKGI and PKGII, with different isoforms of each.

PKGII inhibits EGF mediated MAPK/ERK transduction by blocking the phosphorylation of EGFR by EGF. EGF phopshorylates EGFR which mediates MAPK mediated signal transductions. Higher levels of EGF increase the activation of EGFR but decreases apoptosis. An increase in the activity of PKGII and stimulation with cGMP has been shown to do the reverse, that is, inhibit the phosphorylation and hence activation of EGFR and increase apoptosis in breast cancer cells. PKGII has also been reported to inhibit EGF-induced MAPK/JNK signal transduction in breast carcinoma (Lan et al 2012).

PKG has been found to lower the levels of β-catenin by causing inhibition of transcription of CNNTB1 gene in colon cancer cells. It also does not stimulate protein degradation. PKG has also been reported to activate FOXO4 which inhibits TCF dependent transcription in colon cancer cells. This proves cGMP signalling to be anti-tumor in colon carcinoma (Kwon et al 2010). Constitutive active expression of PKG has been reported to induce cell cycle arrest, decrease angiogenesis, inhibit migration of tumor cell, and promote apoptosis (Fajardo, A. 2014).

PKG signalling has been demonstrated to be very important in cancers. Many drugs which are having anti tumor activity work with the involvement of cGMP signalling. cGMP PDE is

an enzyme that hydrolyses 3'-5'-phosphodiester bond in cGMP and cAMP due to which PKG and PKA does not get activated. Due to this inactivation cGMP, cAMP signalling stops. A drug Sulindac sulfone can cause inhibition of cGMP PDE and hence, inhibition of hydrolysis of cGMP. It also induces apoptosis and elevates the level of intracellular cGMP and activation of PKG and thus, inhibits breast tumor growth. Indomethacin, meclofenamic acid, exisulind and celecoxib have also been reported to inhibit cGMP PDE (Tinsley et al 2009, Piazza et al 2001). Exisulind has been reported to induce apoptosis in bladder cancer in humans and mouse (Piazza et al 2001). Exisulind also induces apoptosis in colon cancer cells (Kwon et al 2010).

cGMP has important role in other diseases as well. PKG plays an important role in the regulation of thrombosis and vascular remodelling. For more than 100 years, the use of drug that cause an elevation in the level of cGMP, Glyceral trinitrate, is being used to treat cardiovascular diseases (Feil et al 2003).

METHODOLOGY

The structure of PKGII isoform b is not available, so prediction of 3D structure of PKGIIb was thought of. PKGII has an important role in the regulation of cancer cells in different cancers and hence finding chemical structures that might prove as good ligands for PKGIIb was considered for work in this report. The following methodology has been adopted:

1. Sequence analysis
    - Sequence retrieval from NCBI
    - Alignment using BLAST
2. Secondary structure analysis
3. Tertiary structure analysis
4. Homology modelling
    - PDB template selection by BLAST
    - Homology modelling
    - Structure validation
5. Binding site prediction
    - Protein Preparation
    - Active site prediction
6. Receptor preparation steps
    - Generation of Grid on the active site
7. Virtual Screening
    - Schrödinger fragments
        - Screening the database formed by joining various fragments to find potential ligands that bind to the modelled protein.
        - Computation of different properties of the chemical structures found to be having a good docking score.
    - NCI database
        - ADME to filter structures
        - Docking of filtered structures with the 3D structure predicted for PKG.

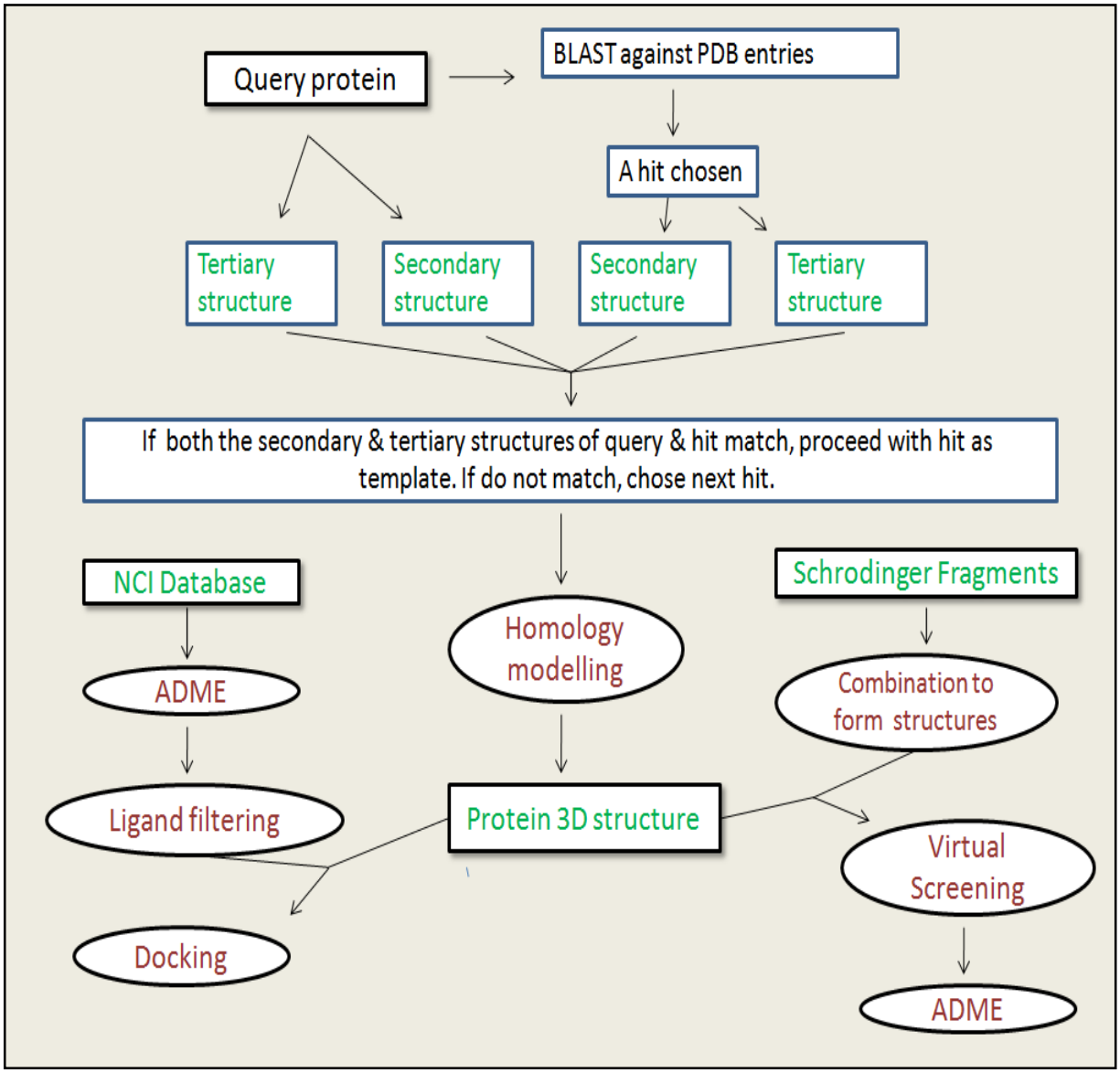


Figure 3: A flowchart describing the methodology.

1. Sequence analysis

The protein product BAH11451.1 which was isolated from tumour tissue of tongue from Homo sapiens was run on BLAST. The BLAST results showed that this protein which was released by NCBI in 2008 has been predicted to be cGMP dependent protein kinase 2 isoform b, the sequence for which was released in 2014.

On the basis of results, best scoring hits that could be used as template were chosen and searched for their PDB structure. If there existed both, high similarity and PDB structure, the hit was further considered for comparison of secondary and tertiary structures with query.

2. Secondary structure analysis

Before proceeding to the step of building the structure by homology modelling, the secondary and tertiary structures of both, the query and the template were predicted and compared.

The tools used for secondary structure comparison were

- SOPMA (http://npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=/NPSA/npsa_sopma.html),
- SABLE (http://sable.cchmc.org/),
- JPRED (http://www.compbio.dundee.ac.uk/www-jpred/) and
- PSIPRED (http://bioinf.cs.ucl.ac.uk/psipred/).

3. Tertiary structure analysis

For tertiary structure comparison, Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2) was used.

Once the structures are found to be having similarity, template can be further used for modelling purpose, and hence, the Chain A of Pkab3 protein was finalised as template for Homology Modelling.

4. Homology Modelling

On the basis of analysis of alignment by BLAST, and the comparison of secondary and tertiary structures of query with the BLAST hit, Chain A of Pkab3 protein (PDB ID: 3L9M) was taken as template for the prediction of structure by using Homology modelling.

After once the template was chosen, its PDB structure (3L9M) was downloaded from the PDB site (www.rcsb.org). For modelling of the query protein, the “Prime” tool of “Maestro server” from “Schrödinger Suite” was used.

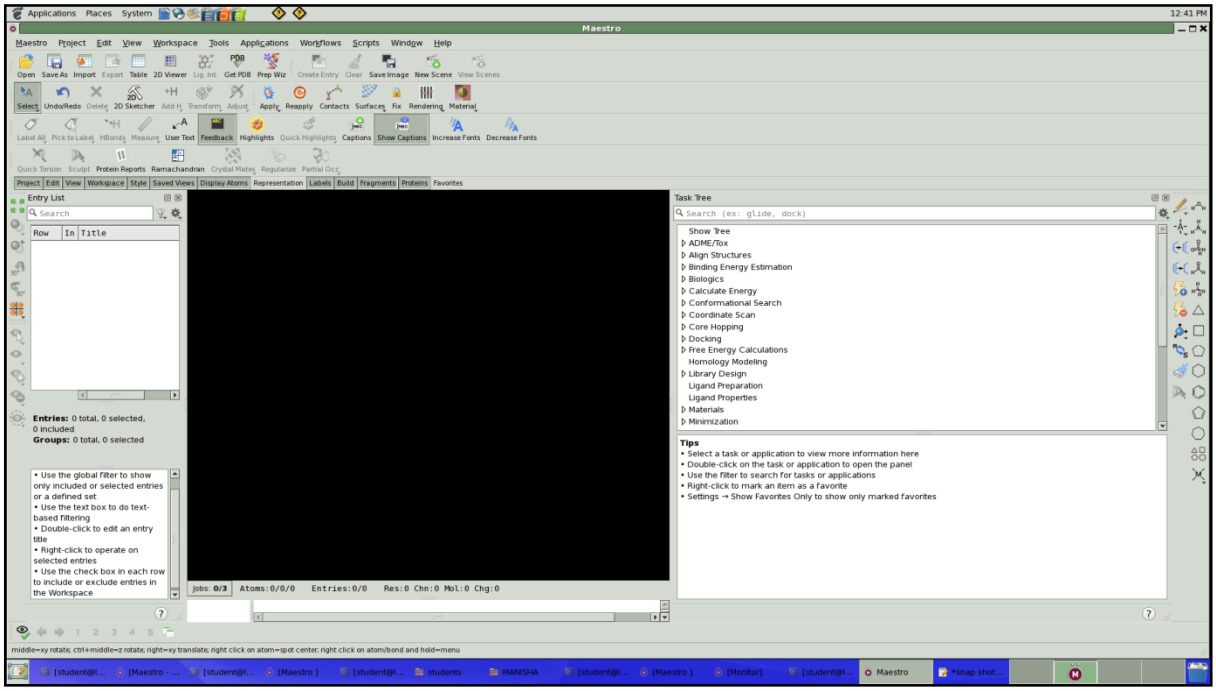


Figure 4: Graphical interface of Maestro

After launching Maestro, Directory, the location where project gets saved (Project > Change Directory) was changed, and a name was given to the project (Project > Save As).

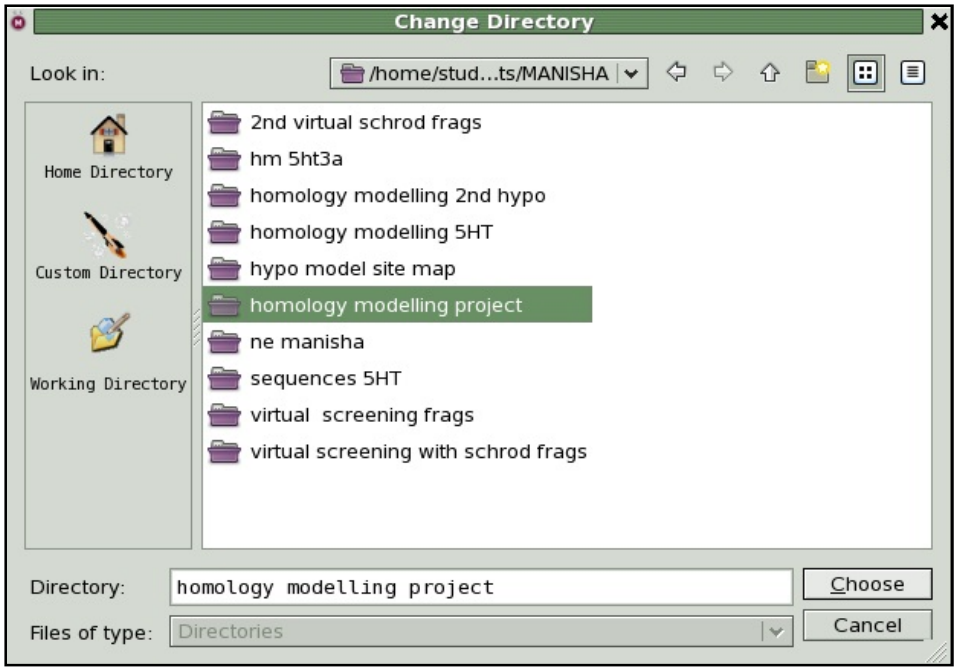


Figure 5: Change Directory Panel

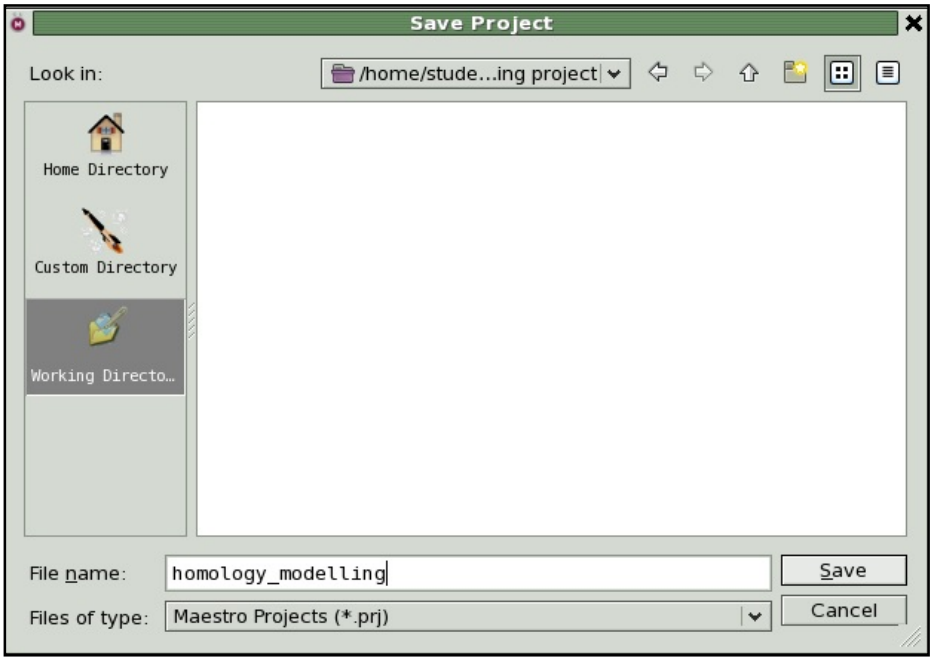


Figure 6: Save Project Panel

The PDB file of the template was imported. Before working on any protein, it needs to be prepared by using PrepWiz. Under PrepWiz, the template was preprocessed, reviewed (for removal of chains, water molecules, or heteromolecules), energy-minimized and optimized.

From the PDB of 3L9M, only Chain A was to be used as a template, hence, other chains (Chain B, C, & D), water molecules, and heteromolecules were deleted.

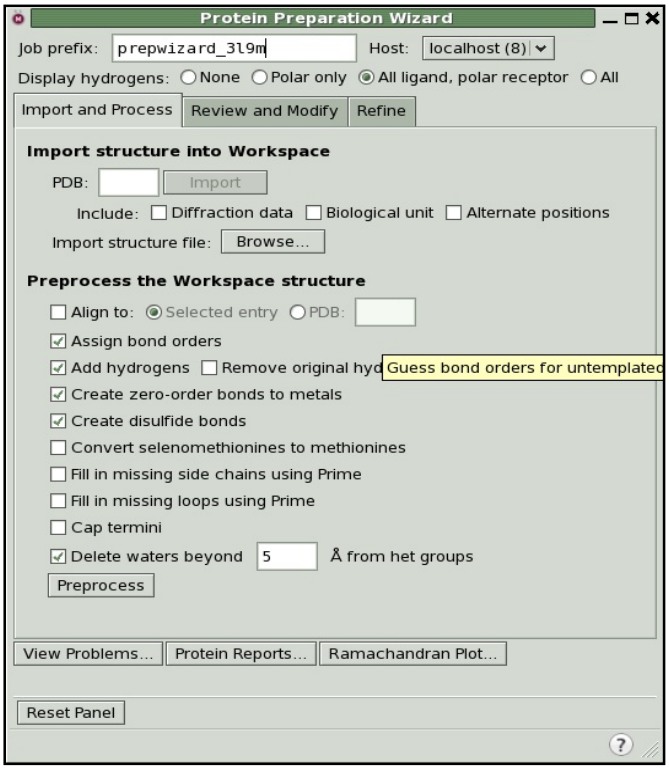


Figure 7: Protein Preparation Wizard PrepWiz.

After preparation of protein, Prime (Applications > Prime > Structure Prediction) was used for Homology Modelling.

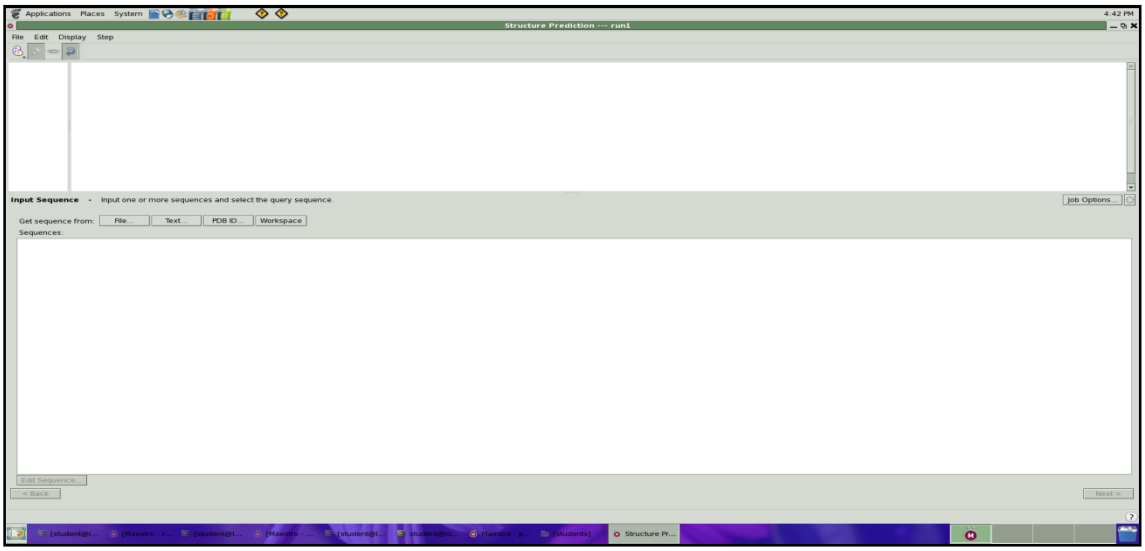


Figure 8: Homepage of Prime.

The sequence of the query protein was given as input to Prime by browsing the FASTA file or directly pasting the sequence in the text box. After clicking Next, Blast was done and the results were analysed. From results, entry for template was selected.

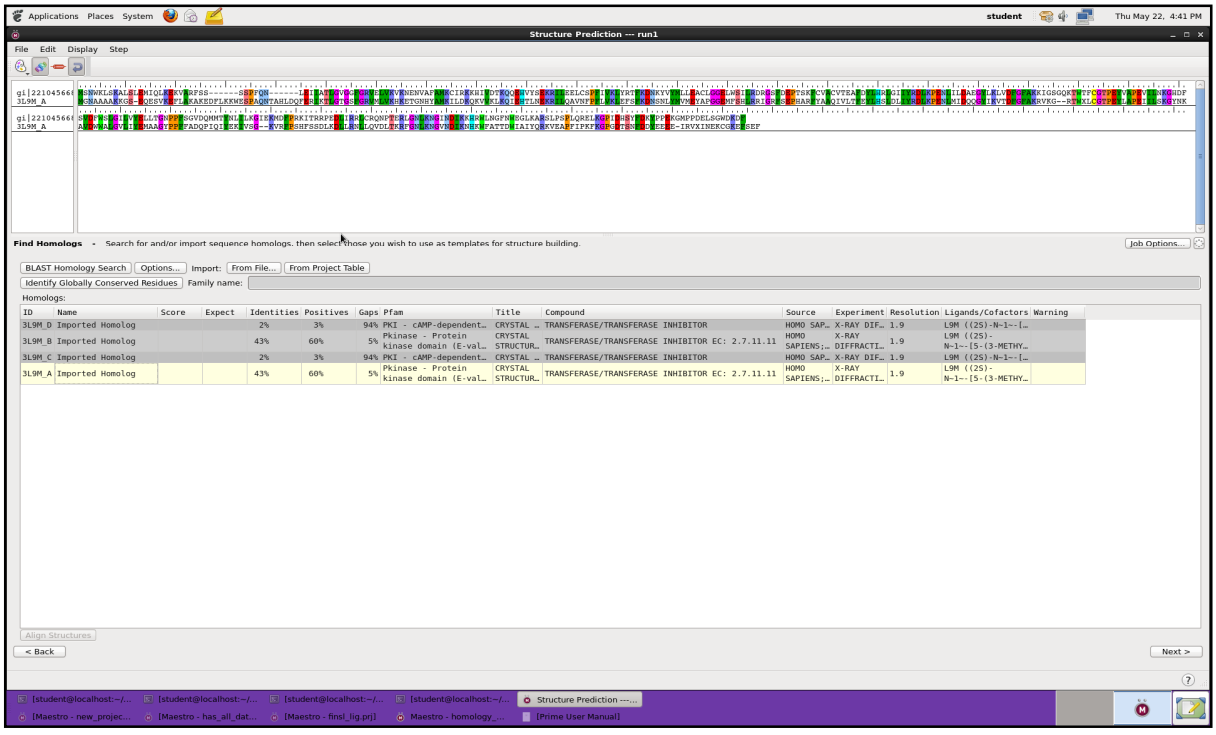


Figure 9: Import Homolog for use as template

Homolog which was to be used as template for building the structure was imported. BLAST homology search can also be done to find out different homologs and their score. The interested homolog was chosen and was further proceeded to building the structure.

In the next step, the secondary structure of protein was determined and alignment is predicted by ClustalW.

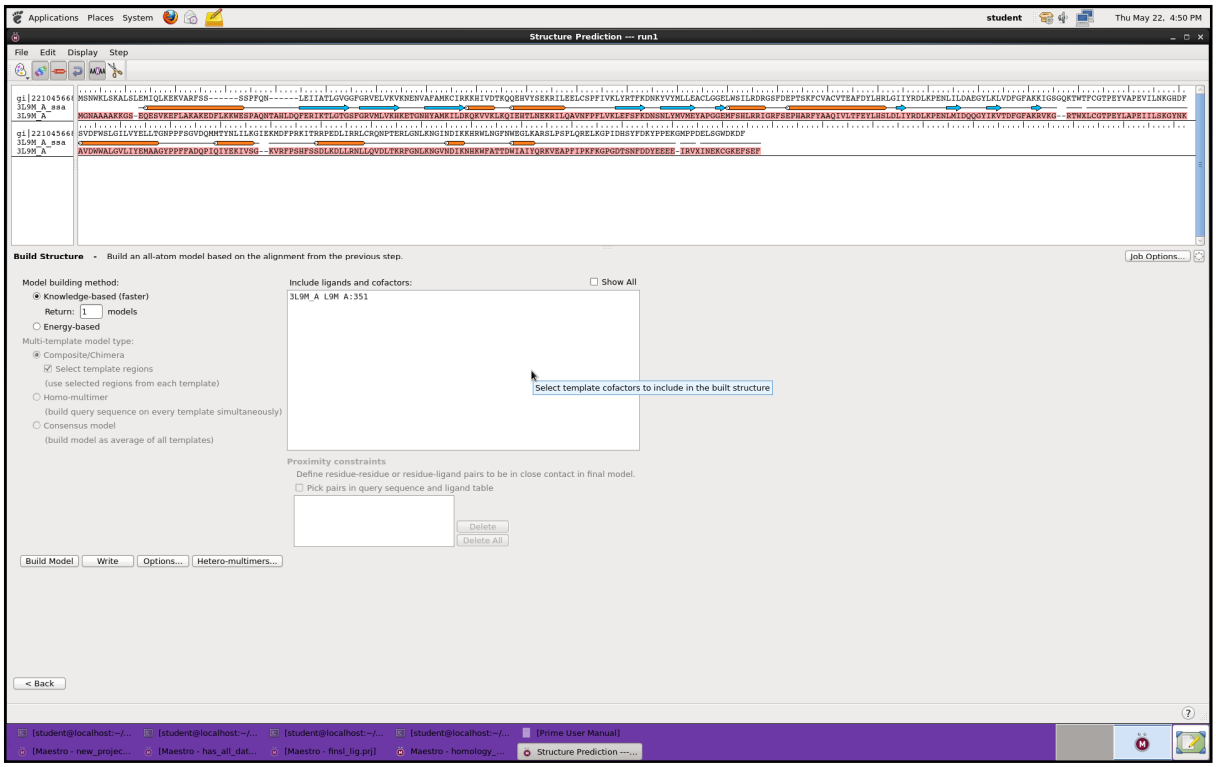


Figure 10: Two different approaches (Knowledge-based and Energy-based) are available to built protein structure in Prime.

Next, the structure of the query protein (cGMP dependent protein kinase 2 isoform b) was built on the basis of the template (Chain A of Pkab3 protein). Before building the structure, it was confirmed that there is no gap in the template that comes within loop, sheet or helix. If by chance, there is gap, it is filled by enabling “Fill missing chains using Prime” and “Fill missing loops using Prime” during Preprocessing by PrepWiz. The built structure was exported as PDB file and Ramachandran Plot (Tools > Ramachandran Plot) was checked.

This PDB structure was further validated by Verify3D (*nihserver.mbi.ucla.edu/**Verify_3D/***) and WhatIF (http://swift.cmbi.ru.nl/whatif/).

5. Binding Site Prediction

For any operation to be done on a protein, it first needs to be prepared by PrepWiz as described earlier.

For Virtual Screening, there is need to specify the site where the ligand is to be docked. As the protein structure that was used for screening was determined by using Homology Modelling and no previous data was available for its 3-D structure and the site for its interaction with any ligand, different possible sites were determined by using the tool “SiteMap” from Maestro. SiteMap returned 5 sites in the protein. The site with the best score was selected.

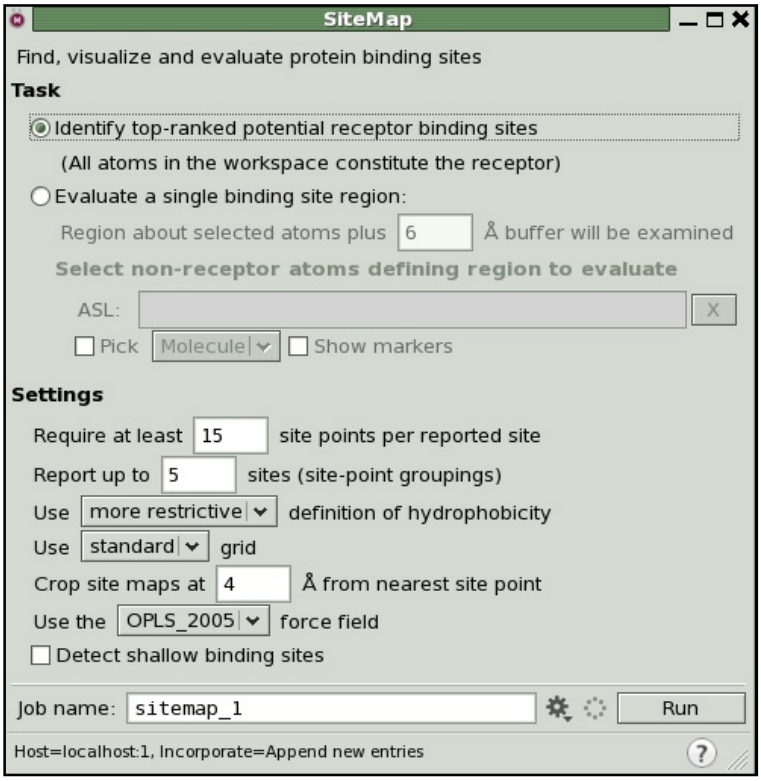


Figure 11: SiteMap panel

6. Receptor protein preparation

Once the site was selected, a grid was formed by using Receptor Grid Generation (Applications > Glide > Receptor Grid Generation). This grid was further used for the specification of the site where screening was to be done.

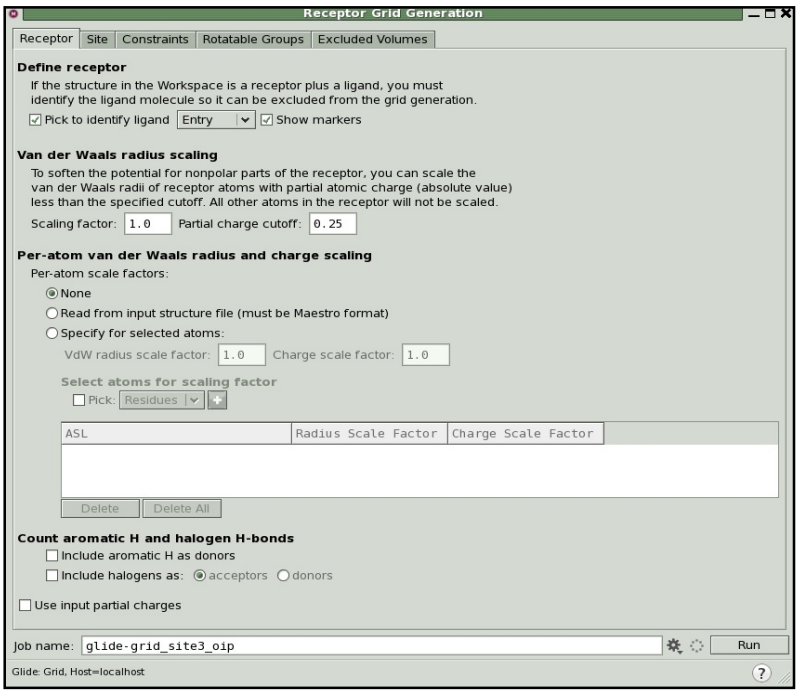


Figure 11: Receptor-grid generation panel.

7. Virtual Screening

For Virtual Screening, Virtual Screening Workflow (Applications > Glide > Virtual screen Workflow) was used.

Virtual screening was done on two collections.

- Combined Schrödinger fragments
- National Cancer Institute (NCI) Database (release 2012)

In Schrödinger Suite, there is a collection of different small fragments. This collection was used for screening. These small fragments were first joined together to form different combinations and hence, form different chemical compounds for the screening of the ones that could bind to the modelled protein structure. The fragments were combined by using the "Combine" tool in Maestro. This tool combines fragments and gives back the number of fragments that was specified before running the tool. There were 667 fragments before combining. The combine tool returned 1152 chemical compounds that were then screened for their affinity for the modelled protein structure. The first 5 results with highest Glide scores have been reported here.

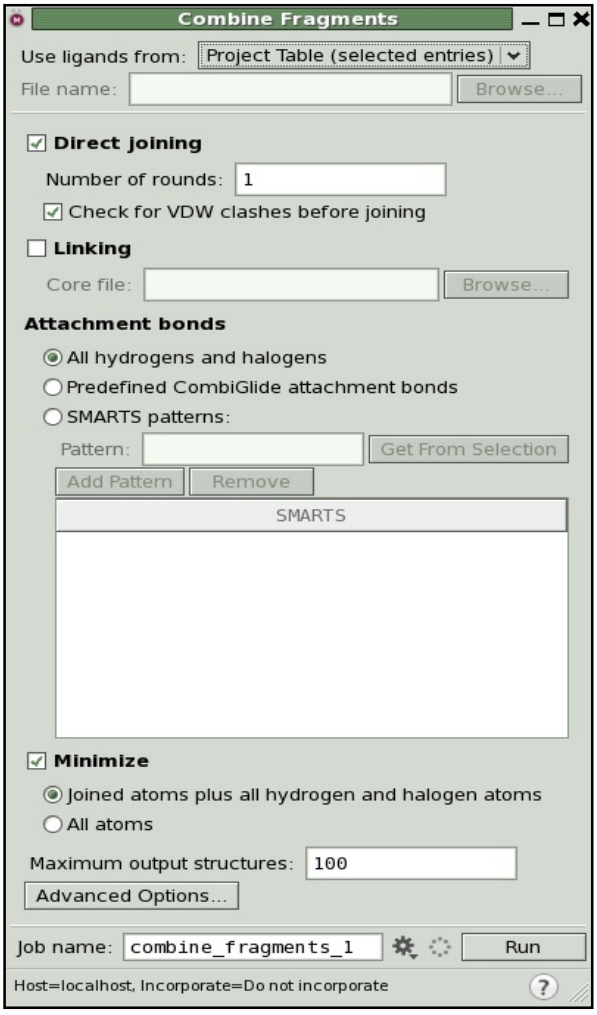


Figure 13: Combine Fragments Panel.

For screening the NCI database, firstly, ADME was done for all the chemical structures by QikProp. After ADME, properties of only 55400 chemical structures were given as result. The values of different properties predicted by QikProp were used to filter ligands by the criteria “Rule of Five” by using “Ligand Filtering”. The characteristics that are included in Rule of Five or Lipinski’s Rule were specified in Ligand Filtering Panel.

After filtration, 24905 structures were returned. These structures were docked to the 3D structure of the protein using “Glide”. The 5 structures having top Glide scores have been reported here.

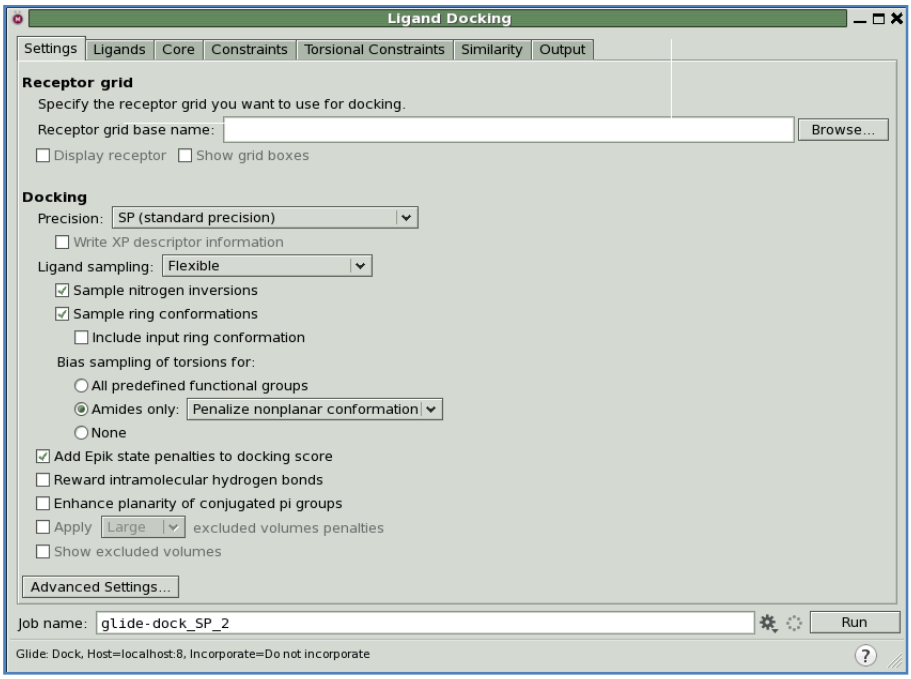


Figure 14: Glide panel

RESULTS

The BLAST results showed 47 % identity with 88% query coverage and an E- value of 2e-98 to Chain A of Pkab3 protein (PDB: 3L9M).

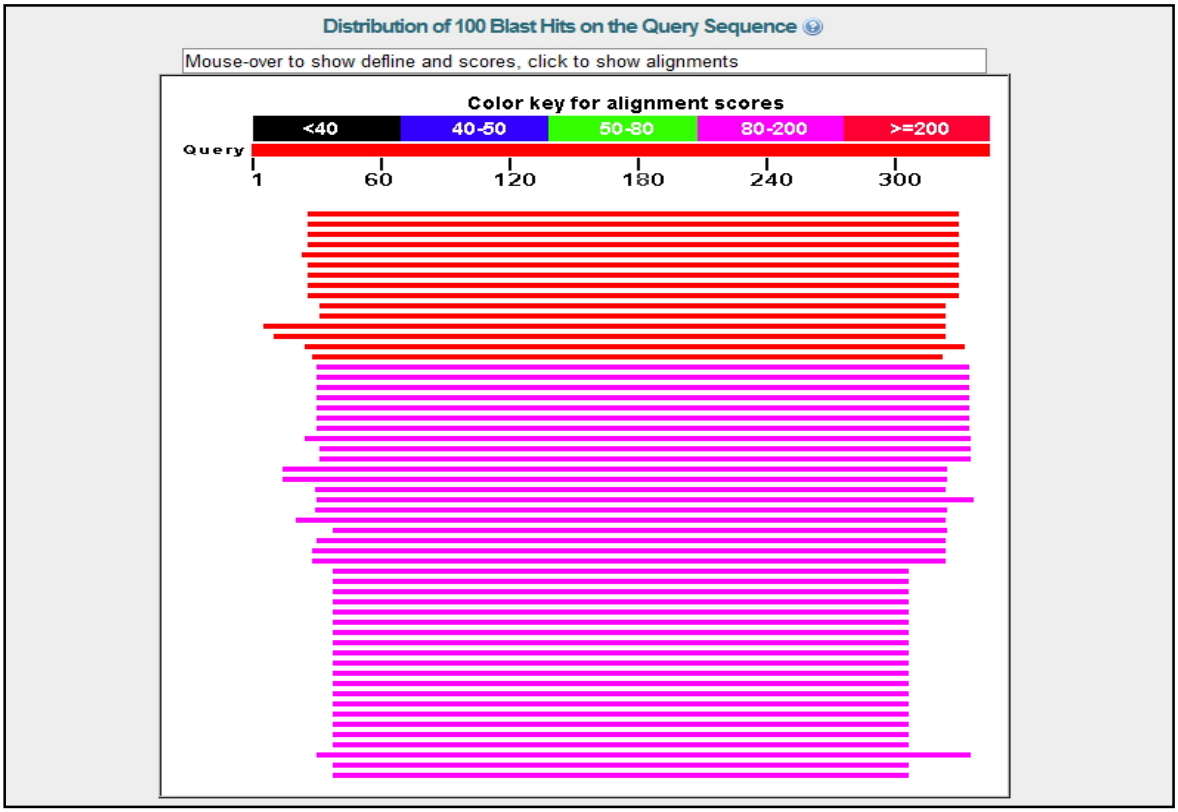


Figure 15: Colour coded graphical version of BLAST results

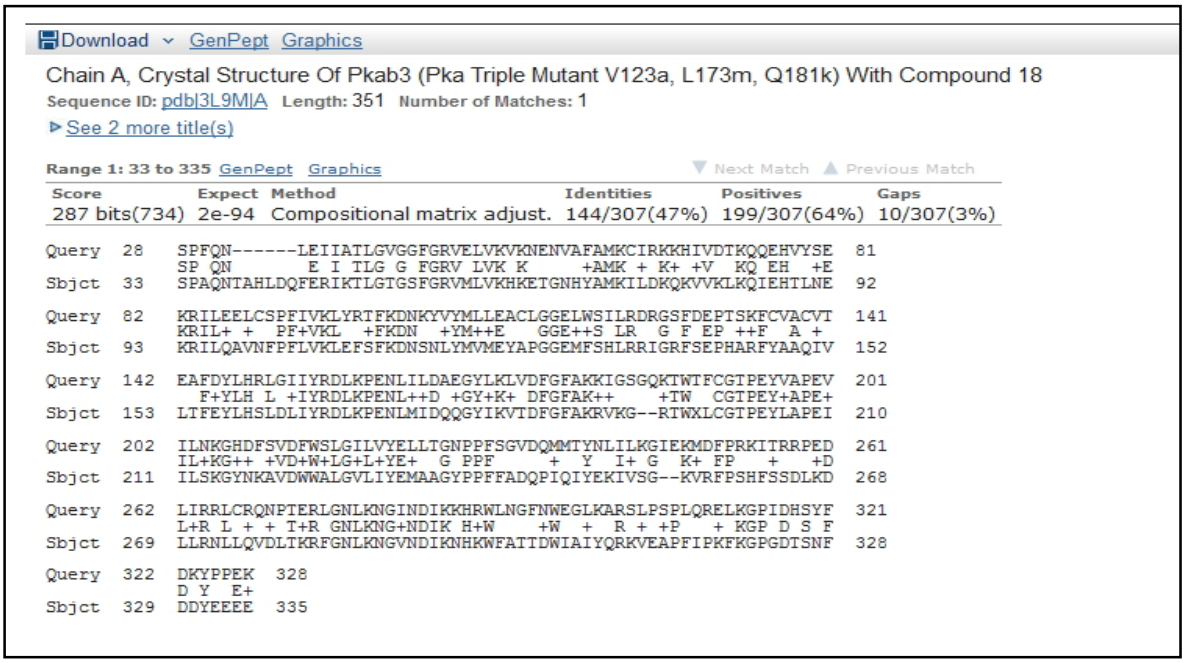


Figure 16: Alignment of the query protein with Crystal Structure of Pkab3 protein (PDB: 3L9M) from BLAST results.

FUNCTIONAL ANNOTATION

TOOL	LINK	PURPOSE	RESULTS
PROTPARAM	http://web.expasy.org/protparam	Physiochemical characterization	Molecular weight, half-life, stability and aliphatic index predicted
SMART (Simple Modular Architecture Research Tool)	http://smart.embl.de/	analysis of domain architechture	2 domains annotated
TMHMM	http://www.cbs.dtu.dk/services/TMHMM/	prediction of transmembrane helices	No transmembrane helix predicted
SignalP 4.1	http://www.cbs.dtu.dk/services/SignalP/	prediction of cleavage sites in signal peptide	No signal peptide cleavage site
SecretomeP 2.0	http://www.cbs.dtu.dk/services/SecretomeP/	prediction of possibility of non-classical secretion	score of 0.390 indicates no possible secretion
NetChop 3.1	http://www.cbs.dtu.dk/services/NetChop/	prediction of cleavage sites	125 cleavage sites were predicted
NetPhos 2.0	www.cbs.dtu.dk/services/NetPhos/	prediction of Serine, Thereonine and tyrosine phosphorylation sites	phosphorylation sites predicted
CELLO: Subcellular Localization Predictive System	http://cello.life.nctu.edu.tw/	determination of cellular localization of proteins.	protein was predicted to be cytoplasmic

Table 2: Predictions from different tools used for the functional annotation of the protein.

Property	Value
Number of negative residues	44
Number of positive residues	51
Molecular weight	39446.7
Isoelectric point	8.87
Instability Index	40.3
Aliphatic Index	87.19
Grand averagy of hydropathycity	-0.361

Table 3: Predicted physiological properties of the protein from PROTPARAM tool.

SECONDARY STRUCTURE ANALYSIS

Tool	Helix	Strand	Turn	Coil
SOPMA	43.02%	14.25%	7.41%	35.33%
	38.30%	14.62%	9.06%	38.01%

Table 4: Predicted percentage of different secondary structures in 3L9M and PKGIIb by SOPMA.

TERTIARY STRUCTURE ANALYSIS

The tertiary structure of 3L9M and PKGIIb were compared by the visual analysis of their PDB structures

The secondary structure and tertiary structure of cGMP dependent protein kinase 2 isoform b and Chain A of Pkab3 protein (PDB ID: 3L9M) were compared and found to be similar. And hence homology modelling of PKGII was done on the basis of structure of 3L9M.

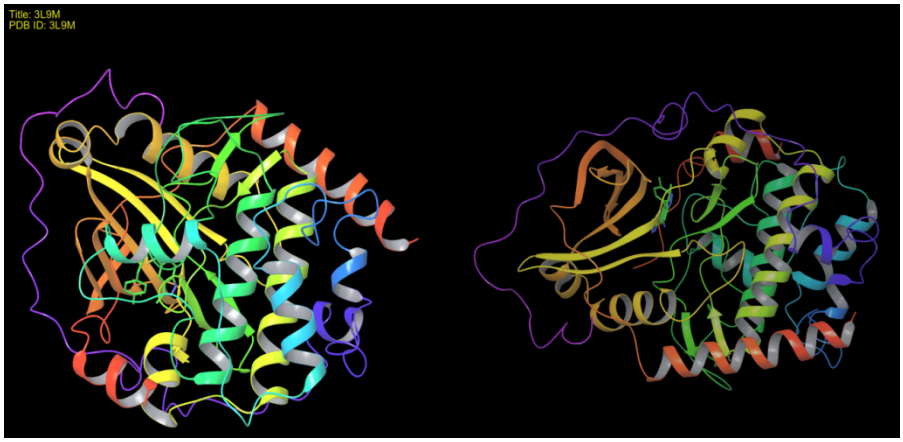
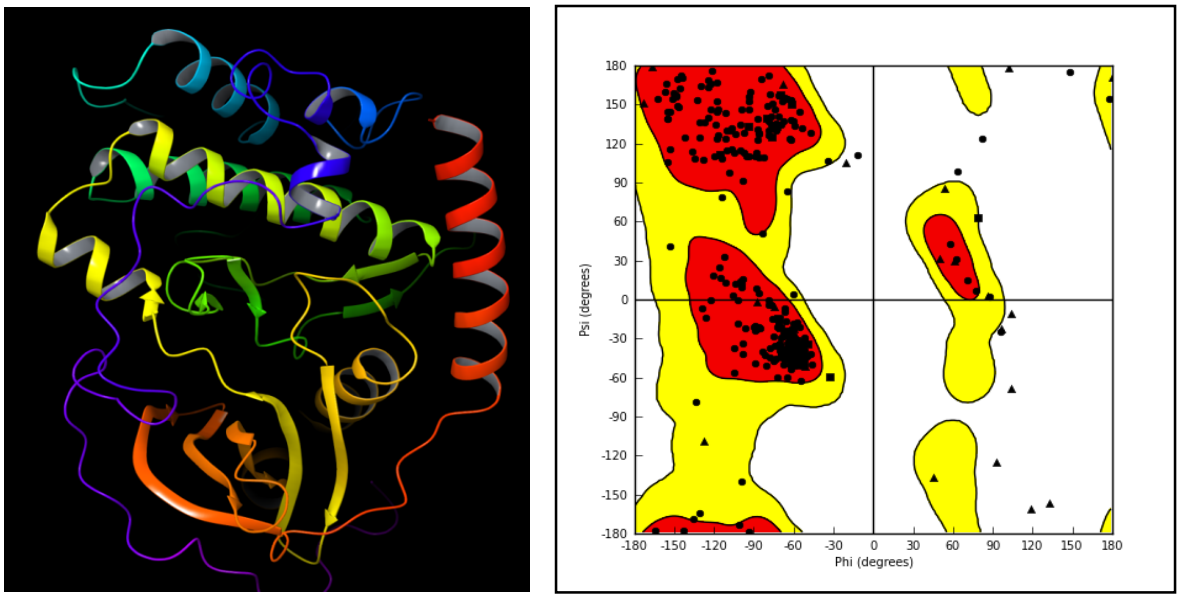


Figure 17: The PDB structure of 3L9M contains two subunits- catalytic subunit alpha, and inhibitor alpha, each containing two chains, Chain A & B and Chain C & D respectively.



A B

Figure 18: A. Homology model for PKGII built on 3L9M, and B. Ramachandran Plot of the model built by using Homology.

Site	Score	Size	Volume
1	1.114	132	348.145
2	1.024	131	353.23
3	1.008	73	111.475
4	0.96	98	222.607
5	0.907	53	105.644

Table 5: Summary of different sites found on the protein

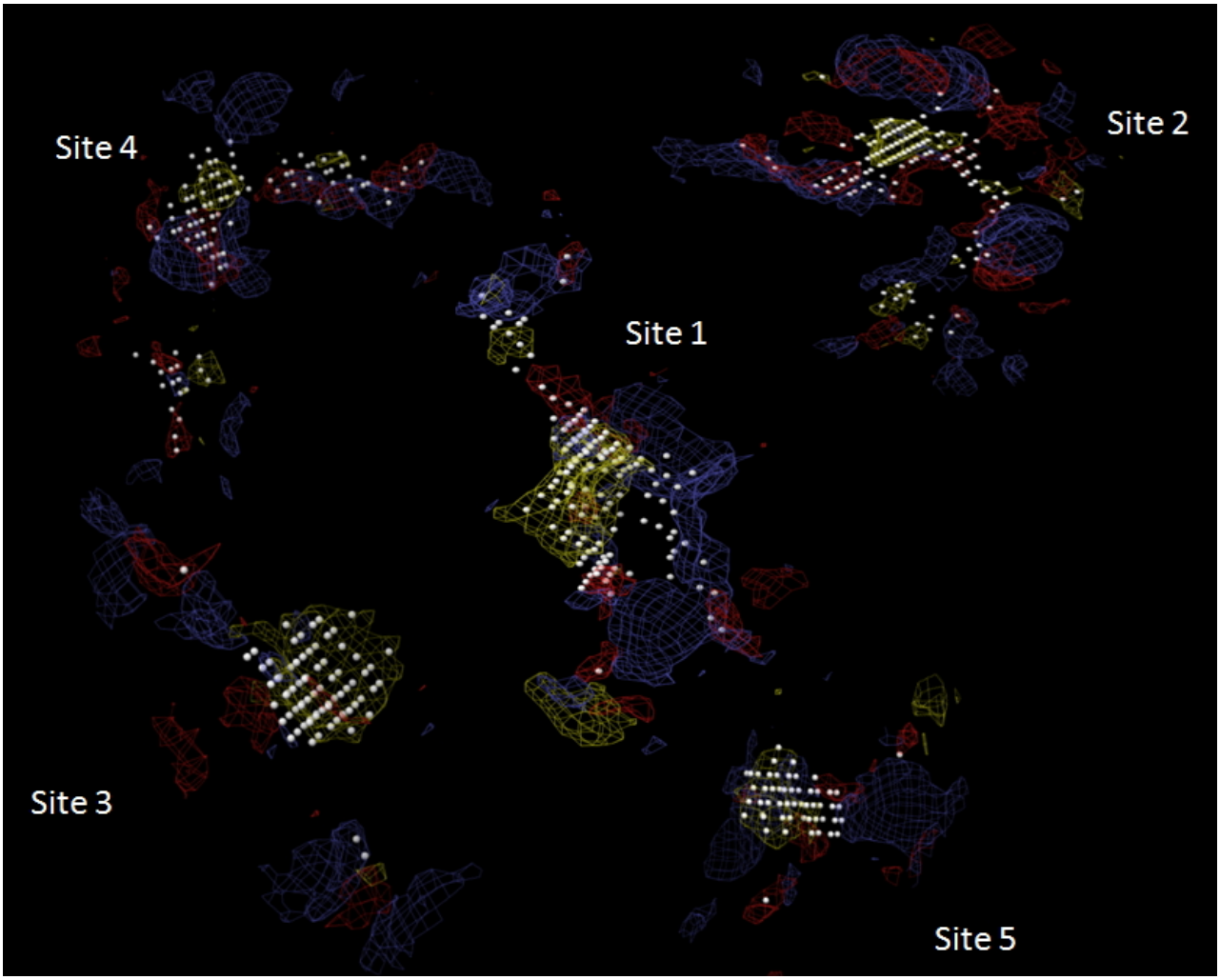


Figure 19: Postioning of all 5 sites on the receptor protein

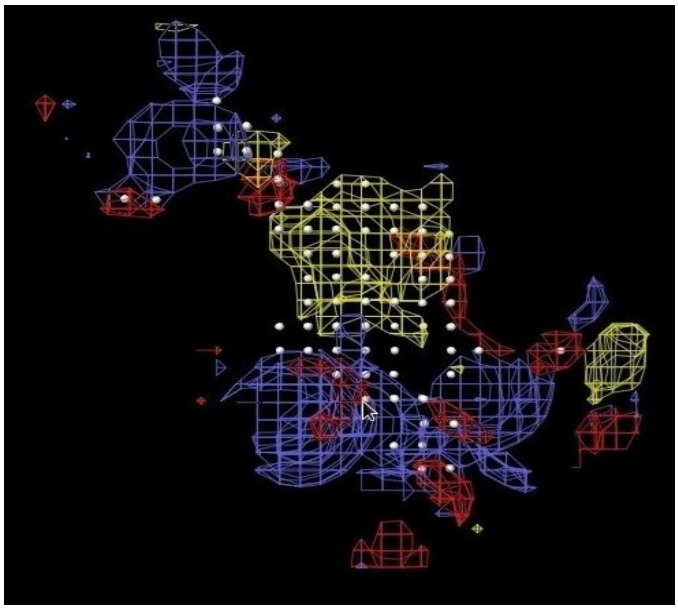


Figure 20: The site with the best score (Site 1).

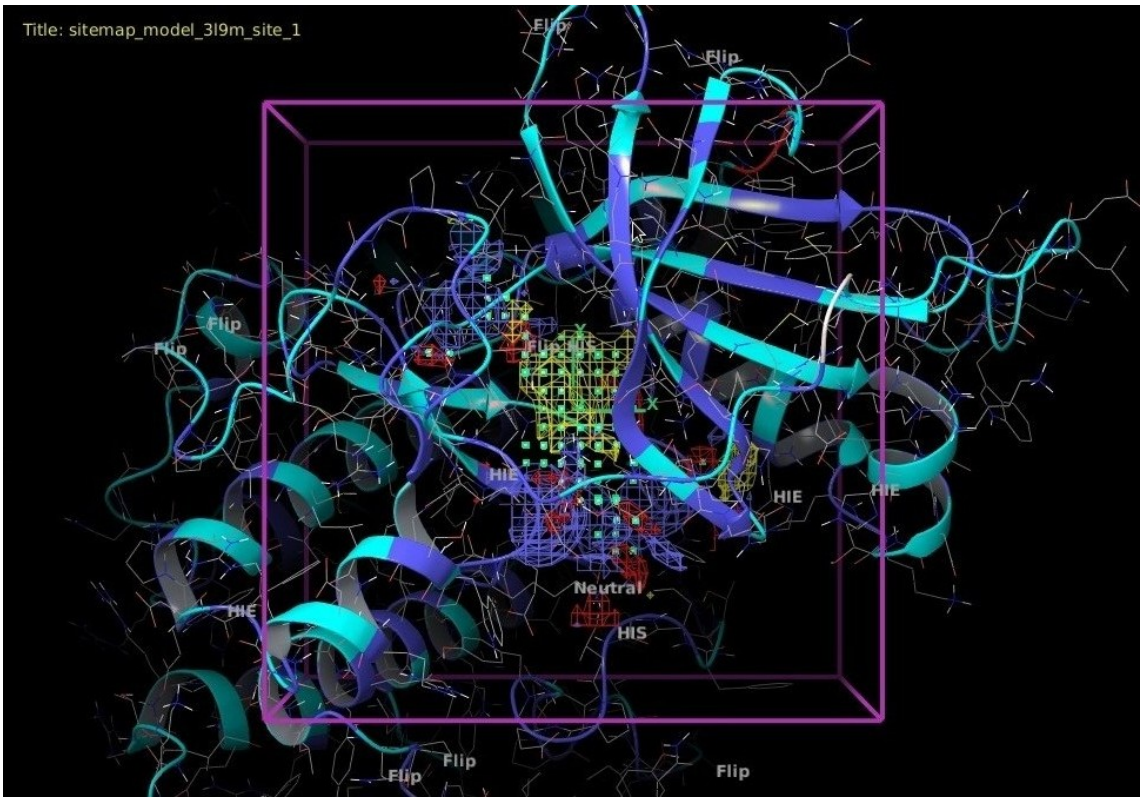


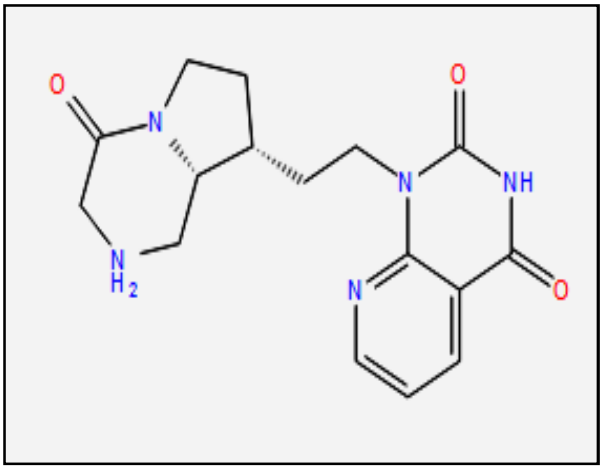
Figure 21: The grid generated on the site predicted by SiteMap.

Chemical Structure No.	1st Schrödinger fragment	2nd Schrödinger Fragment	Join Score	Glide Score
ChemStr1	274	411	10.256	-9.038
ChemStr2	274	387	9.471	-9.551
ChemStr3	220	277	9.250	-9.549
ChemStr4	210	277	9.383	-9.091
ChemStr5	277	385	11.421	-9.557

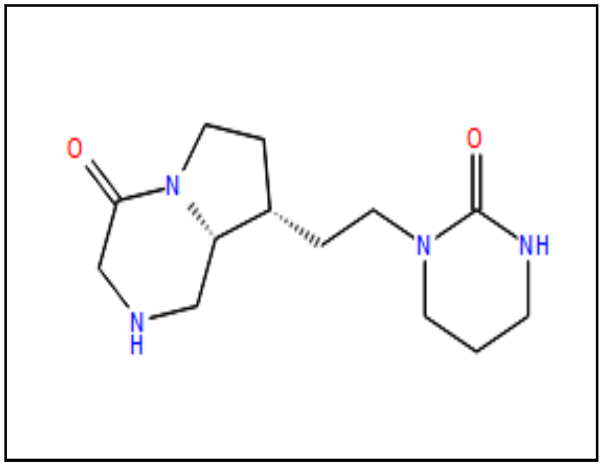
Table 6: Chemical structures formed by Schrödinger fragments with their Join score and Glide score.

ChemStr no.	Mol Mw	Log Po/w	Log s	Log BB	PMDCK	Human oral absorption	Rule of 5
ChemStr1	329.358	-0.754	-0.146	-0.809	16.931	46.018	0
ChemStr2	266.342	0.194	-0.767	-0.346	74.583	62.518	0
ChemStr3	278.31	-17.84	0.336	-1.086	9.758	35.997	0
ChemStr4	313.358	-0.878	0.363	-0.547	39.491	51.384	0
ChemStr5	315.374	-0.807	0.178	-0.518	44.677	52.695	0

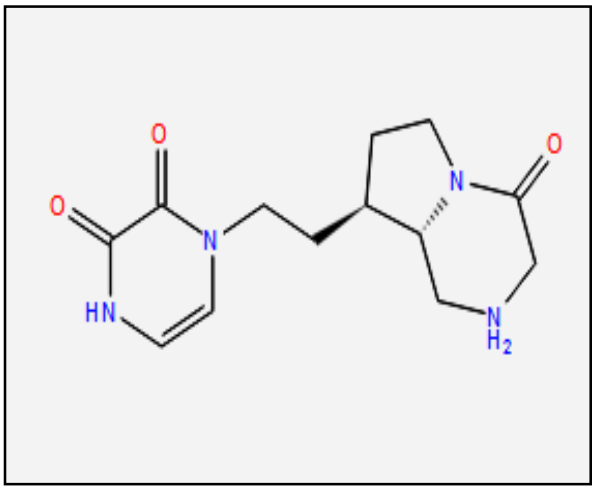
Table 7: ADME properties of ChemStr1-5.



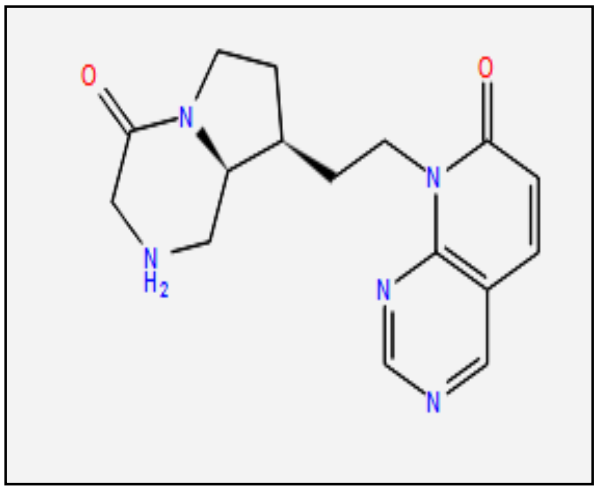
ChemStr1



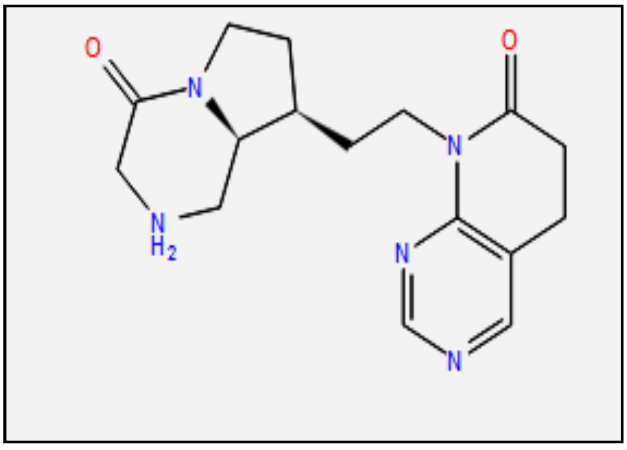
ChemStr2



ChemStr3

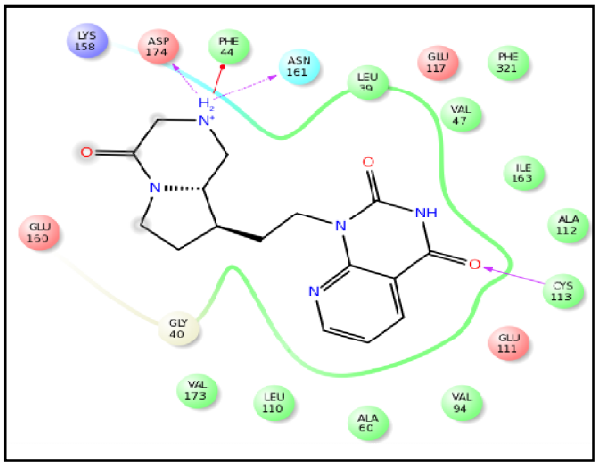


ChemStr4

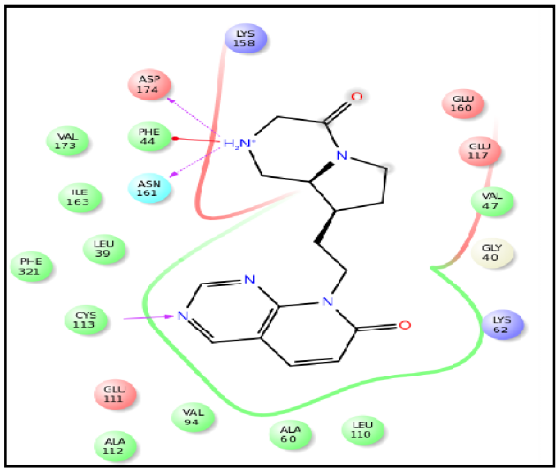


ChemStr5

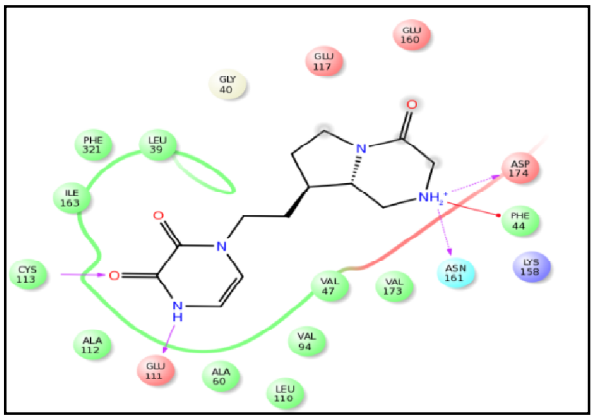
Figure 22: Chemical Structures of ChemStr1-5



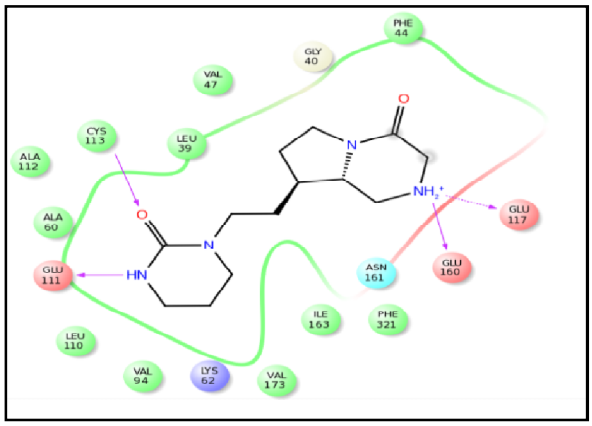
A



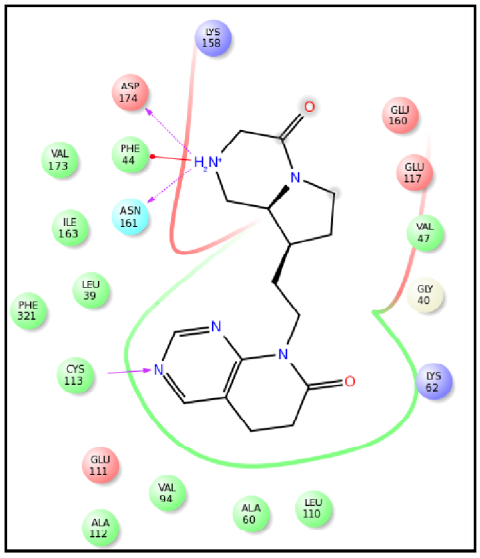
B



C



D



E

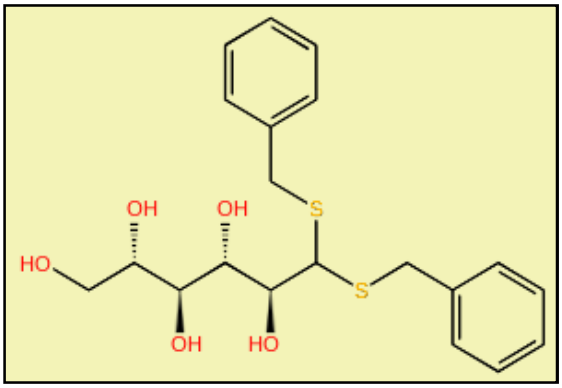
Figure 23: (A-E) Ligplots (Ligand Interaction Plots) of ChemStr(1-5)

NSC no	mol MW	Glide score	Glide energy
1972	410.542	-12.547	-49.872
12102	273.29	-11.117	-37.143
26850	416.387	-10.712	-60.272
37721	260.356	-10.502	-31.015
14778	288.256	-10.461	-45.866

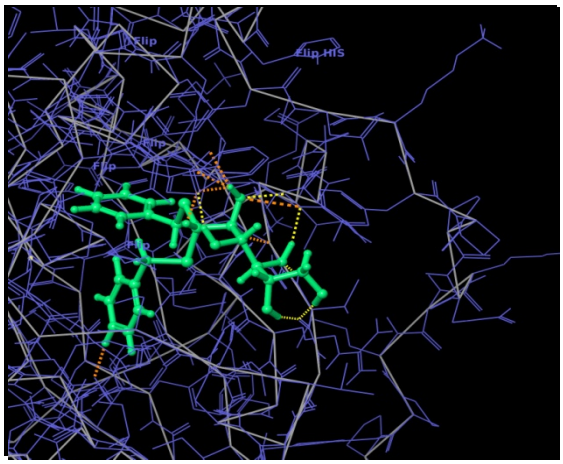
Table 8: Chemical Structures from NCI database having top Glide Scores.

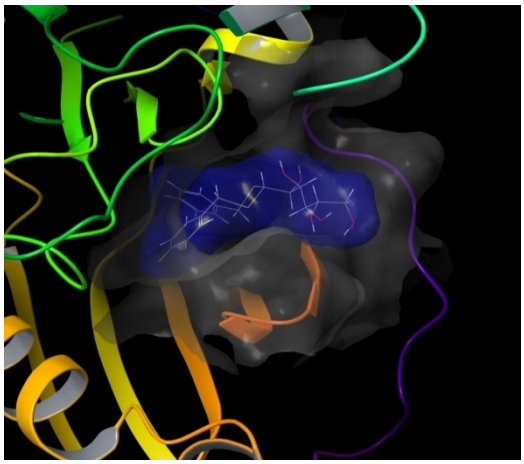
NSC no.	mol MW	QPlogPo/w	QPlogS	QPlogBB	QPPMDCK	Percent Human Oral Absorption	Rule Of Five
1972	410.542	2.234	-3.526	-2.18	155.664	82.901	0
12102	273.29	2.792	-3.501	-0.375	446.858	96.255	0
26850	416.387	2.176	-5.548	-4.242	0.134	30.838	0
37721	260.356	2.893	-3.912	-0.752	742.086	100	0
14778	288.256	2.32	-3.228	-2.359	0.618	42.689	0

Table 9: ADME properties of 5 NCI chemical compounds with highest Glide score.

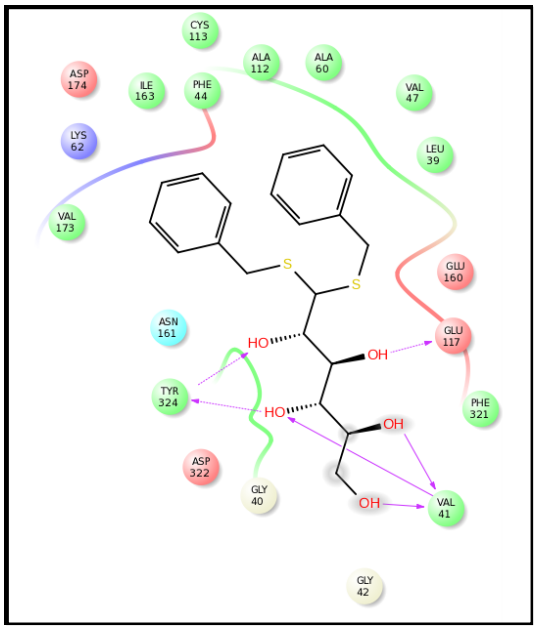


A



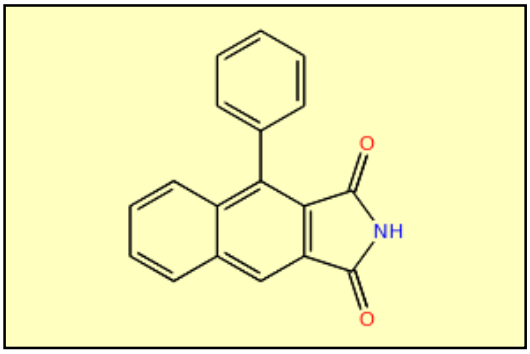


B C

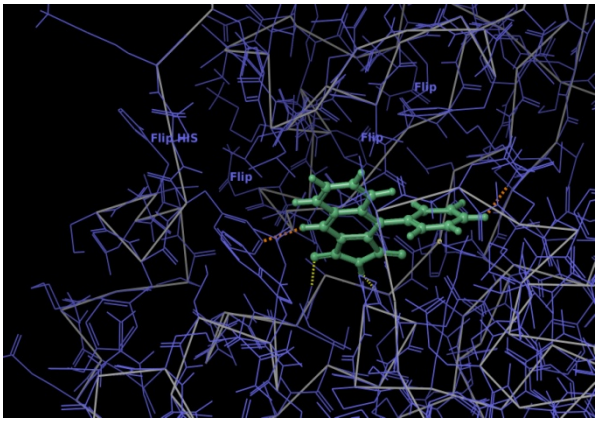


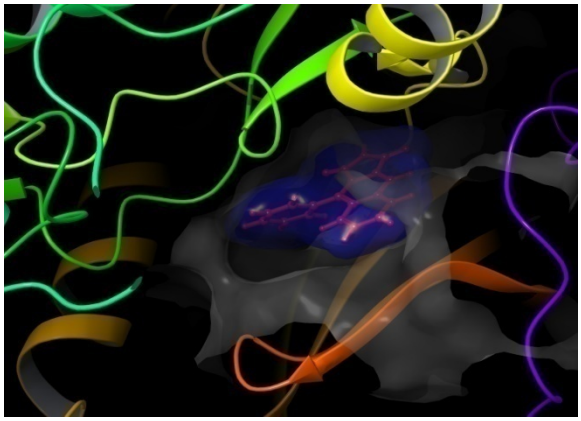
D

Figure 24: A. Chemical structure, B. H-bonds & contacts with protein, C. Electrostatic surfaces of protein & ligand, and D. LigpProt of NSC1972

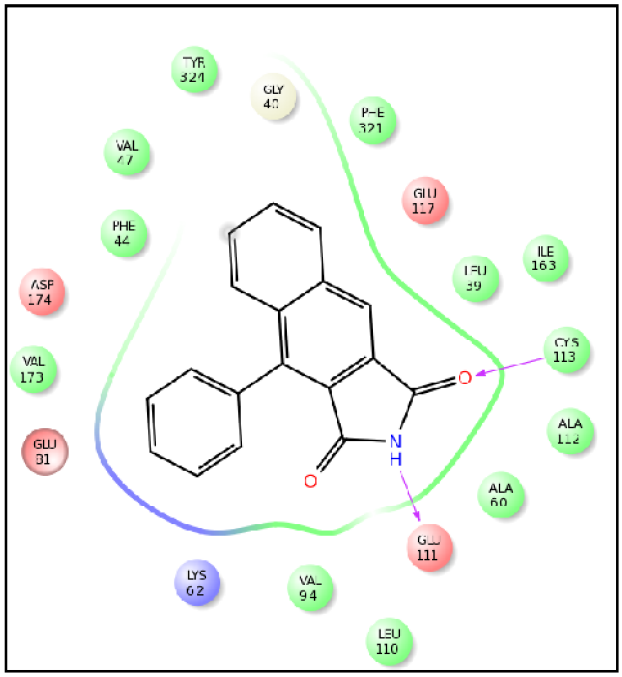


A



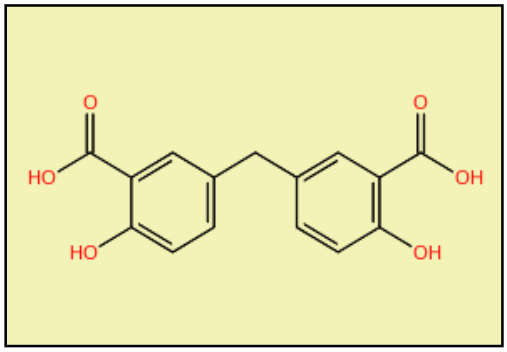


B C

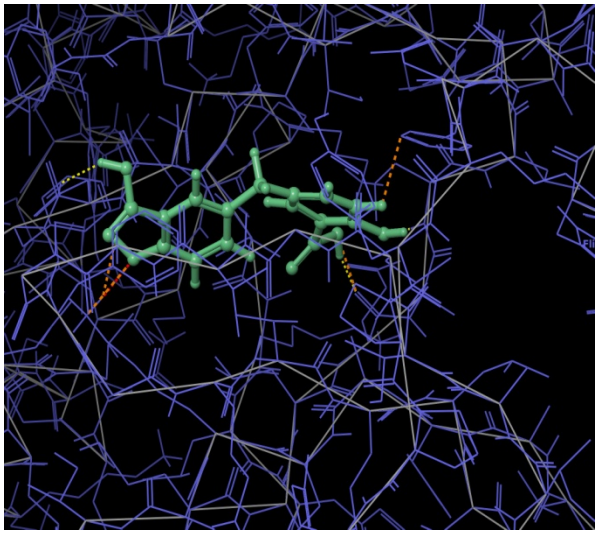


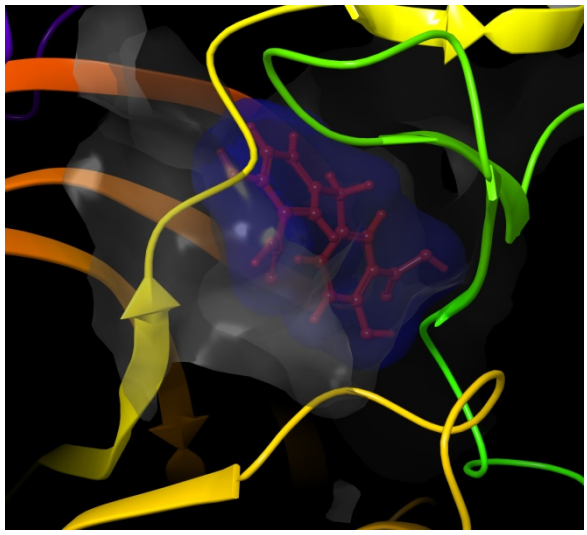
D

Figure 25: A. Chemical structure, B. H-bonds & contacts with protein, C. Electrostatic surfaces of protein & ligand, and D. LigpProt of NSC12102

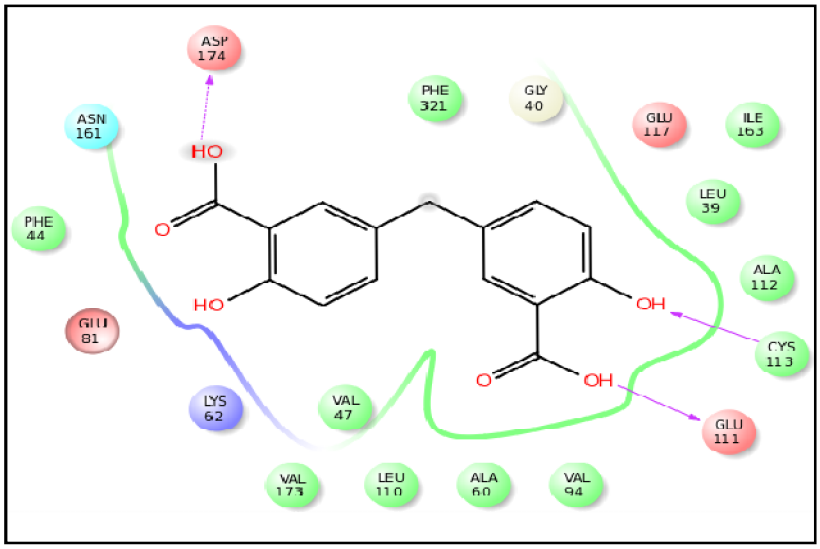


A



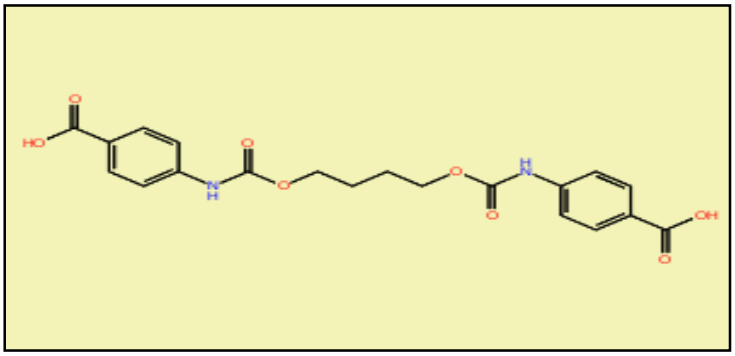


B C

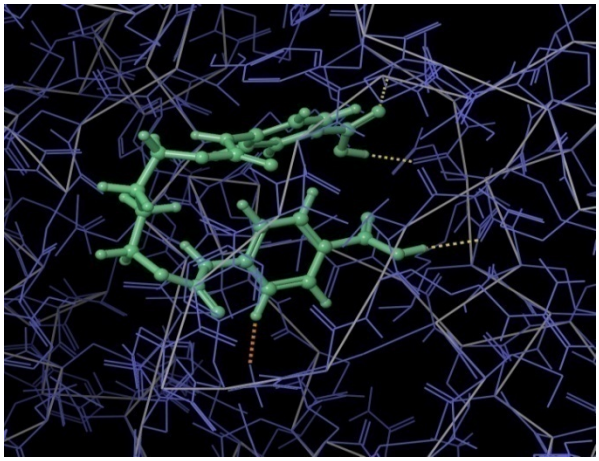


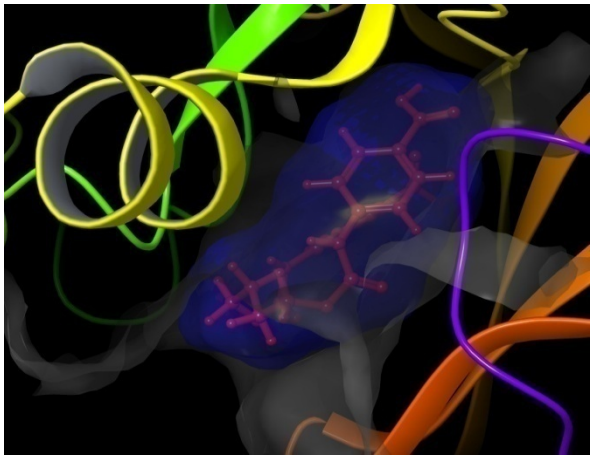
D

Figure 26: A. Chemical structure, B. H-bonds & contacts with protein, C. Electrostatic surfaces of protein & ligand, and D. LigpProt of NSC14778

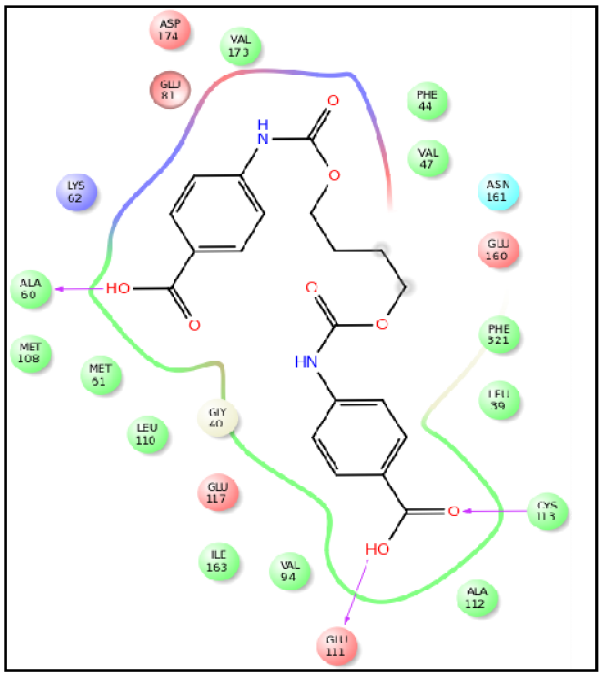


A



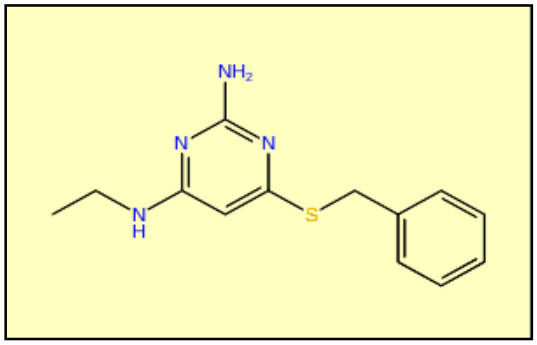


B C

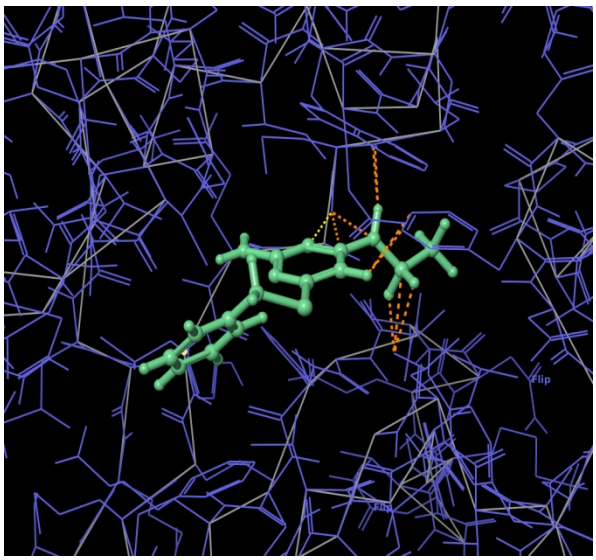


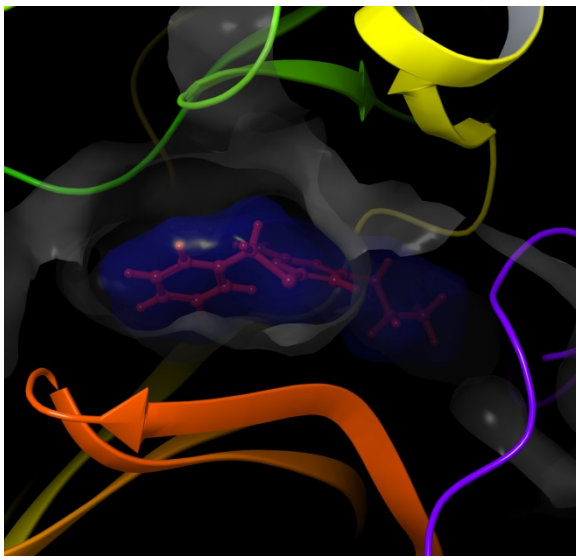
D

Figure 27: A. Chemical structure, B. H-bonds & contacts with protein, C. Electrostatic surfaces of protein & ligand, and D. LigpProt of NSC26850

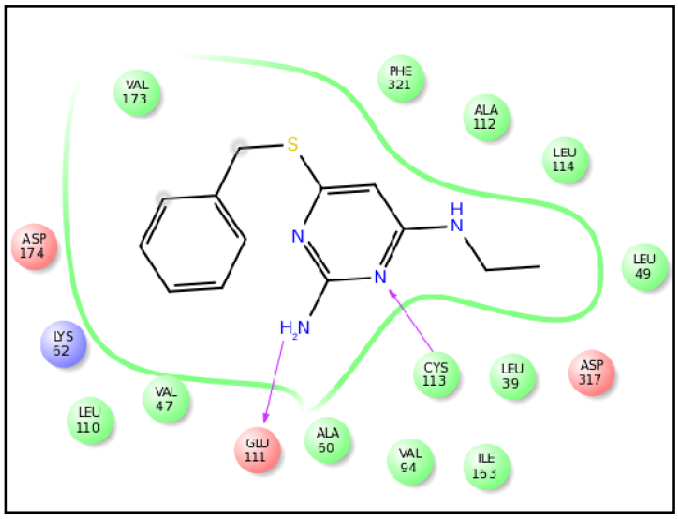


A





B C



D

Figure 28: A. Chemical structure, B. H-bonds & contacts with protein, C. Electrostatic surfaces of protein & ligand, and D. LigpProt of NSC37721

DISCUSSION

The protein with accession no. BAH14511.1was isolated from tongue tumor. This protein when run on BLAST, revealed that it is identical to cGMP-dependent Protein Kinase II isoform b (PKGIIb). Seeing that protein sequence of BAH14511.1 was released on NCBI site in 2008 and the protein sequence for PKGIIbwas released on NCBI site in 2014, clears that it has been named to cGMP-dependent Protein Kinase II isoform b in 2014. Much information about isoform b of PKG II is not known but two forms of PKG- PKGI and PKII are well known.

PKG signalling is very important for the regulation of cancerous cells. In cancerous cells, the levels of cGMP and PKG are very low. If these cancerous cells, PKG is activated and somehow induced, it leads to suppression of cancer and apoptosis. PKG also blocks different signal transductions in cancerous cells which help in the regulation of growth. This ultimately results in decrease in growth of cancer and increase in apoptosis of cancerous cells. It has been reported in many cancers. It has also been reported to decrease invasive activity of tumour.

There are drugs which help in treatment of cancer by activation of PKG and by increasing the levels of cGMP. These drugs are used to treat different forms of cancer. But PKGIIb is not reported and hence, it was hypothesized that PKGIIb, like already known PKGs, it might also possess some activity in the regulation of cancer. Hence, it was thought to find the chemical structures that might bind to it with good score to have chemical structures that might be used for drug designing of PKGIIb in near future.

Functional annotation and prediction of physiological properties revealed that PKGIIb is a cytoplasmic protein with molecular weight 39466.7 and pI 8.87. The instability constant was found to be 40.3 which signify protein to be unstable. An aliphatic index of 87.19 was predicted. The protein had no transmembrane helix and no signal peptide cleavage sites. Phosphorylation sites and 125 different cleavage sites were predicted. GRAVY was found out to be -0.361. Negative value of GRAVY signifies the protein to be hydrophillic.

For the purpose of finding potential binders, structure had to be known for PKGIIb and hence, homology modelling with 3L9M was done. The PDB structure contained two structural units which contained 4 chains in total (A & C and B & D). A & B are catalytic sub-unit alpha, and C & D are inhibitor alpha. For homology modelling, the chains B, C, and D were removed while preparation of 3L9M protein structure and Chain C was taken as template. In homology modelling, first the sequence of the query protein was given as input and BLAST was done to find the homologs of the input sequence or a homolog can also be given as input. After choosing homolog, the gaps in the query and template were filled by PrepWiz. The knowledge based approach for building model was used.

The Binding site prediction for the protein was done. The results were having 5 sites with different site scores and volume & location on the modelled protein structure, from which, the top scoring site having a site score of 1.114 was proceeded with, for virtual screening. For finding potential binders, Schrödinger fragments and NCI database were used.

The chemical structures formed from Schrödinger fragments were virtual screened and top 5 chemical structures which gave good docking score have been reported in this work. The most negative score was -9.557. ADME properties were also calculated for these chemical structures. Violations of Lipinski's Rule of Five (MW< 500 Da; LogP< 5; H-bonds< 5, H-bond donor <=5, H-bond acceptor <=10) were zero for all the structures which shows that all these structures follow drug-like properties and hence can be used for further work of designing drug for PKGIIb.

NCI database contained ~275,000 chemical structures. This is a huge database and hence, it was better to filter out structures on some basis. Firstly, the ADME properties were predicted for all the compounds, by using QikProp. In results, ADME properties of only 55,400 structures were returned. Theses ADME properties were used to further filter out the structures. For this, Ligand Filtering was used. The criteria for filtering were specified as per Lipinski's Rule of Five. Finally, after filtering, 24,905 compounds of NCI database were left, which were then docked to the binding site of the receptor protein and the results showed even better results than that of chemical structures from Schrödinger Fragments. The most negative docking score was -12.547. The more negative the docking score is, the better the docking is. Around 25 NCI entries showed scores lesser than -10 but only top 5 has been reported here. The violations of Rule of Five were zero for all the 5 compounds.

CONCLUSION

ChemStr1 to ChemStr5, formed by Schrödinger fragments, and NSC1972, NSC12102, NSC26850, NSC37721, and NSC14778 from NCI database were found to have good docking scores. All these compounds followed Lipinski's Rule. Good docking scores show good binding affinity towards Site 1 (with highest site score) of PKGIIb. These ligands might prove to be good drugs for PKGIIb but for validation, pharmacological studies will be needed to be done in humans.

REFERENCES

Adamczak, R; Porollo, A; and Meller,J. (2005). Combining Prediction of Secondary Structure and Solvent Accessibility in Proteins. Proteins: Structure, Function and Bioinformatics. 59:467-75.

Bendtsen, JD; Jensen, LJ; Blom, N; von Heijne, G; and Brunak, S. (2004). Feature based prediction of non-classical and leaderless protein secretion. Protein Eng. Des. Sel., 17(4):349-356.

Blom, N; Gammeltoft, S; and Brunak, S. (1999). Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. Journal of Molecular Biology. 294(5): 1351-1362.

Bruni, L; Barrionuevo-Rosas, L; Serrano, B; Brotons, M; Cosano, R; Muñoz, J; Bosch, FX; de Sanjosé, S; Castellsagué, X. (2014). ICO Information Centre on HPV and Cancer (HPV Information Centre). Human Papillomavirus and Related Diseases in India. Summary Report 2014-03-17.

Buchan, DWA; Minneci, F; Nugent, TCO; Bryson, K; and Jones, DT. (2013). Scalable web services for the PSIPRED Protein Analysis Workbench . Nucleic Acids Research . 41 (W1): W340-W348.

Coelho, KR; (2012). Challenges of the Oral Cancer Burden in India. Journal of Cancer Epidemiology Volume 2012, Article ID 701932, 17 pages

Cole, C; Barber, JD; and Barton, GJ;(2008). The Jpred 3 secondary structure prediction server. Nucleic Acids Res. 36(Web Server issue):W197-201

Fajardo, AM; Piazza, GA; and Tinsley, HN. (2014). The Role of Cyclic Nucleotide Signaling Pathways in Cancer: Targets for Prevention and Treatment. Cancers. 6, 436-458;

Feil, R; Lohmann, SM; de Jonge, H; Walter, U; and Hofmann, F. (2003). Cyclic GMP-Dependent Protein Kinases and the Cardiovascular System: Insights From Genetically Modified Mice. Circ Res. 93:907-916

Friesner, RA; Murphy, RB; Repasky, MP; Frye, LL; Greenwood, JR; Halgren, TA; Sanschagrin, PC; and Mainz, DT. (2006). Extra Precision Glide: Docking and Scoring Incorporating a Model of Hydrophobic Enclosure for Protein-Ligand Complexes. J. Med. Chem. 49, 6177–6196.

Gasteiger, E; Hoogland, C; Gattiker, A; Duvaud, S; Wilkins, MR; Appel, RD; and Bairoch, A. (2005). Protein Identification and Analysis Tools on the ExPASy Server. The Proteomics Protocols Handbook, Humana Press. 571-607

Halgren, T. (2009). Identifying and characterizing binding sites and assessing druggability. J. Chem. Inf. Model. 49, 377–389a.

Jacobson, MP; Pincus, DL; Rapp, CS; Day, T JF; Honig, B; Shaw, DE; and Friesner, RA. (2004). A Hierarchical Approach to All-Atom Protein Loop Prediction," Proteins: Structure, Function and Bioinformatics. 55, 351-367

Kelley, LA; and Sternberg, MJE. (2009). Protein structure prediction on the web: a case study using the Phyre server. Nature Protocols 4, 363 – 371.

Krogh, A; Larsson, B; von Heijne, G; and Sonnhammer, EL. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305(3):567-80.

Kwon, I; Thangaraju, M; Shuang, H; Liu, K; Dashwood, R; Dulin, N; Ganapathy, V; and Browning, DD. (2010). PKG inhibits TCF signaling in colon cancer cells by blocking β-catenin expression and activating FOXO4. Oncogene. 29(23): 3423–3434.

Lan, T; Chen, Y; Sang, J; Wu, Y; Wang, Y; Jiang, L; and Tao, Y. (2012). Type II cGMP-dependent protein kinase inhibits EGF-induced MAPK/JNK signal transduction in breast cancer cells. ONCOLOGY REPORTS. 27: 2039-2044

Letunic, I; Doerks, T; and Bork, P. (2012). SMART 7: recent updates to the protein domain annotation resource. Nucleic Acids Res. 40(D1): D302–D305

Luthy, R; Bowie, JU; and Eisenberg, D. (1992). Assessment of protein models with three-dimensional profiles. Nature, 356, 83-85.

Nielsen, M; Lundegaard, C; Lund, O; and Kesmir, C. (2005). The role of the proteasome in generating cytotoxic T cell epitopes: Insights obtained from improved predictions of proteasomal cleavage. Immunogenetics., 57(1-2):33-41.

Petersen, TN; Brunak, S; von Heijne, G; and Nielsen, H. (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. Nature Methods, 8:785-786.

Piazza, GA; Thompson, WJ; Pamukcu, R; Alila, HW; Whitehead, CM; Liu, L; Fetter, JR; Gresh, WE Jr; Klein-Szanto, AJ; Farnell, DR; Eto, I; Grubbs, CJ. (2001). Exisulind, a Novel Proapoptotic Drug, Inhibits Rat Urinary Bladder Tumorigenesis. Cancer Res. 61:3961-3968.

Sastry, GM; Adzhigirey, M; Day, T; Annabhimoju, R; and Sherman, W. (2013). Protein and ligand preparation: Parameters, protocols, and influence on virtual screening enrichments. J. Comput. Aid. Mol. Des. 27(3), 221-234.

Siegel, R; Ma, J; Zou, Z; and Jemal, A. (2014). Cancer Statistics, 2014. CA Cancer J Clin 2014;64:9-29.

Takiar, R; Nadayil, D; and Nandakumar, A. (2010) Projections of number of cancer cases in India (2010-2020) by cancer groups. Asian Pac J Cancer Prev. 11(4):1045-9.

Taylor, SS; and Kornev, AP. (2011). Protein Kinases: Evolution of Dynamic Regulatory Proteins. Trends Biochem Sci. 36(2): 65–77.

Tinsley, HN; Gary, BD; Keeton, AB; Zhang, W; Abadi, AH; Reynolds, RC; and Piazza, GA. (2009). Sulindac sulfide selectively inhibits growth and induces apoptosis of human breast tumor cells by PDE5 inhibition,elevation of cGMP, and activation of PKG. Mol Cancer Ther. 8(12): 3331–3340.

van der Waal, I. (2013). Are we able to reduce the mortality and morbidity of oral cancer; Some considerations. Med Oral Patol Oral Cir Bucal. 18 (1):e33-7.

Vriend, G. (1990). WHAT IF: A molecular modeling and drug design program. J. Mol. Graph. 8, 52-56.

Wolfertstetter, S; Huettner, JP; and Schlossmann, J. (2013). cGMP-Dependent Protein Kinase Inhibitors in Health and Disease. Pharmaceuticals. 6, 269-286.

Yu, CS; Chen, YC; Lu, CH; and Hwang, JK. (2006). Prediction of protein subcellular localization. Proteins: Structure, Function and Bioinformatics. 64:643-651.

APPENDIX

Tool and applications used:

1. Protein structure analysis
    a. Secondary structure analysis
        - SOPMA
        - Sable
        - Jpred
        - Psipred
    b. Tertiary structure analysis
        - Phyre

2. Functional analysis
    - Protparam
    - SMART
    - TMHMM
    - SignalP
    - SecretomeP
    - NetChop
    - NetPhos
    - Cello

3. Homology Modelling
    - Prime

4. Validation of built model
    - Verify3D
    - WhatIF

5. Preparation of protein
    - PrepWiz

6. Binding Site Prediction
    - SiteMap

7. Generation of grid
    - Glide

8. Virtual Screening
    - Glide

9. Docking
    - Glide

10. Prediction of ADME properties
    - QikProp

11. Combination of fragments to form structures
    - Combine

12. Filteration of ligands
    - Ligand filtering