Mutational Study of RAS Subfamily 2CE2 Domain and Fosfestrol as an Ideal Drug Candidate against NSCLC

Jayant Mishra[1], J. K. Srivastava[2]

[1] Himalayan University, Department of Biotechnology, Arunanchal Pradesh, India

[2]Department of Biotechnology, Amity University, Lucknow, India

Abstract: *Lung cancer is a deadly disease that is prevalent as one of the largest contributor to higher mortality rates. Most commonly EGFR is the widely studied factor in almost 90% of the cases that are considered in the epidemiological study of lung cancer. The present work deals primarily in considering KRAS as an important factor for impending cancer deaths, particularly NSCLC and Designing a potent drug Fosfesterol as a potent target for the same. The basis of the study includes the analysis of a few drug candidates and their activity against NSCLC in comparison to Fosfestrol. Fosfestrol proved to be an effective drug candidate which is ascertained by protein blockage due to correct docking. So fosfestrol may be considered as potent drug against NSCLC for further clinical trials.*

Keywords: lung cancer; K-ras; smokers; fosfestrol; NSCLC

1. Introduction

Fosfestrol: fosfestrol chemical name by IUPAC is α, α´-Diethylstilbene-4, 4´-diol bis(dihydrogen phosphate) or (E)-4, 4´-(1, 2-Diethylvinylene)bis(phenyl dihydrogen orthophosphate). It appears as off-white, odourless, crystalline powder which is sparingly soluble in water soluble in alcohol and in dilute alkali. It is basically an synthetic nonsteroidal oestrogen used in treatment of prostate and colorectal cancer. It currently available in market as injection as well as in tablet forms and It requires conversion from dephosphorylation to diethylstilbestrol for being active. So it is stored in airtight containers at a temperature not exceeding 21°C. The structure of fosfesterol is shown in fig 1.The use of fosfestrol has been found to be effective against prostate cancer. [1]
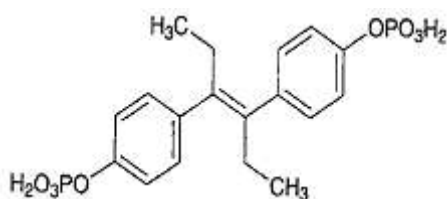

Figure 1: Structure of fosfesterol

Lung cancer is among the most common types of cancer rated second after oral cancer and the number is rising with the increase in no. of smokers, smoking being the chief cause, though passive smoking counts as much as 24% of cases. Lung cancer is characterized by uncontrolled cell division in lung tissues followed by metastasis, which marks its invasion to neighboring tissues. Lung cancer can be classified into two sub categories on the basis of microscopic appearance and their respective treatment as:

1. **Small cell lung cancer:** It responds better against chemotherapy but is more aggressive, yet better as only 20% of majority suffering lung cancer suffers SCLC. The population affected is observed to be of smokers majorly.

2. **Non small cell lung cancer:** NSCLC is relatively insensitive to chemotherapy and is treated surgically. It is the most common type of lung cancer affecting 80% of the total majority.

Therefore NSCLC contributes more than 80% to the frequency of histological types of lung cancer and can be variously be caused by many factors. The chief mutational factors known mostly include EGFR and interleukins. To investigate mutations in NSCLC pathway by both RAS-RAF dependent and RAS-RAF independent mechanism, the gene of choice is KRAS. The impressive use of kras being a marker for nsclc with or without is reported in the work of Aviel-Ronen S et.al [2]

RAS Family
RAS gene: It was discovered in 1960 in mouse leukemia virus isolated from leukemic mouse. In 1967, named the gene was named as Harvey (Ha or H) ras and Kirsten (Ki or K) ras after the name of discoverer. A year later another *ras* family gene, neuroblastoma- or *N-ras*, was identified. Ras family contains 21-kD GTP binding protein that regulate cell growth, differentiation, apoptosis which also includes MAPK (mitogen-activated protein kinase), STAT (signal transducer and activator of transcription), and PI3K (phosphoinositide3-kinase) signaling cascades. Much about the signaling pathway is known for pancreatic cancer through Kras gene shown by Zenyu ji et.al [3]

Affect of RAS gene mutation upon cell cycle: The binding of a growth factor molecule to a receptor on the cell membrane serves as stimulus and cell division begins. Initially RAS protein remains inactive and is associated with GDP, when activated it switches GDP for the more active GTP. However, a mutation in the ras gene results in the

formation of a protein that cannot convert GTP to GDP. Therefore, the ras protein remains in the active state and though there is no stimulus from growth factors, it keeps on sending signals and so the cell cycle runs uncontrollably and results in excessive cell proliferation and cancer.

Structure of Ras protein: RAS protien comprises 189 amino acids of which 165 amino acids show homology while 25 are heterologous at carboxyl end. It comprises 5 discontinuous domains which are conserved and contribute to function namely L1(5-64), L2(77-92), L3(109-123), L4(139 - 169), L5 (186-189).The paper by Macaluso et.al[4] shows the role of ras family genes, cancer and checkpoints of the cell cycle.

2. Materials and Methods

Several insilico tools shown in Table 1 have been used to ascertain the drug binding activity of Fosfesterol against mutated Domain of KRAS protein. The strategy used has been depicted in the Figure 2.

Ligand Screening
In the ligand screening procedure, a list of compounds was selected depending on the disease from pubchem, Chebi. Each of their IUPAC name, molecular weight, logp values and canonical smile values were noted. Then using prodrug server bond changes and ambiguity values were determined. Following this ADME/tox values were calculated for the above selected compounds & also for like compounds. From the above analysis absorptivity, distribution, ionization, solubility, p-gp specificity of the compounds were conveniently deciphered. Thus a complete profiling and screening of ligand was done considering several parameters. Initially the energy minimization was done using Argus lab, followed by sybyl. Then the ligand was ready for docking with the receptor binding site of protein.

Protein Analysis:
Protein analysis forms an imperial part of the drug binding strategy. It is necessary to analyse a protein in order that catalytic binding site available for ligand binding are detected. Moreover binding of the ligand to various receptor sites gives various poses with different energy calculations. In the first step sequence analysis is done followed by blast and phylogenetic analysis. Sequences retrieved from protein databases are subjected to multiple alignments to determine their relative closeness and evolutionary distances. Subsequently, primary, secondary and tertiary structure analysis is done by using several tools like *protpram, cph, phyre, sopma and HHpred*. Apart from this a domain analysis using *SMART and PRODOM. Tools* like Q-sitefinder, pocket finder and ibm pattern discovery show the amino acid sites, repeatations and point mutation. Finally a cast-p prediction is done to systematically point out pockets having large surface area, volume accessible surfaces. After The above analysis was done, the protein was minimized using Argus Lab, SYBYL and finally docked to the screened & stable ligand found above using hex and Molegro.

Earlier studies of mutations in kras sequences were reported by E.M.NEVILLEet.al[6] In our strategy, for the mutational analysis of KRAS gene and its role in NSCLC, the sequence of the gene can be taken from the database of NCBI. Further by the use of advanced tools of bioinformatics the biological attributes of the protein can be studied and the functional annotations and various domains can be prioritized. The sequence analysis of the protein also gives broad knowledge about the family profiling, which includes dendogram generation, domain functionality and nature of the surface pockets accessible for drug delivery. The following tools of bioinformatics were used in the study:

Table 1: list of softwares and tools used for Analysis

Sequence analysis	BLAST
Phylogenetic analysis	Clustal W, Clustal distance matrix, BOX SAHDE, TEX SHADE.
Structural analysis (a)Primary structure (b)Secondary structure (c)Tertiary structure	PROTPARAM. SOPMA & HNN. CPH, HHPRED.
Target Identification and validation	Q-site finder Pocket finder CASTp IBM pattern discovery
Drug identification and validation	Pubchem Drugbank Dunde Prodrug server
Visualization	RASMOL
Database for 3D structure	PDB
Geometry and energy optimization Docking	AURGUS LAB HEX

GENE CARDS: HUMAN GENE DATABASE
Gene Cards is a searchable, integrated data base of human genes that provides concise genomic, proteomic, transcriptomatics, genetics and functional information on all known and predicted human genes.

3. Analysis of Protein

A) Sequence Analysis
NCBI: The National Centre for Biotechnology Information advances science and by providing access to by medical and genomic information.
BLAST (Basic Local Alignment Search Tool): The Basic Local Alignment Search Tool (BLAST) finds reason of local similarity between sequences. The program compares nucleotide or protein sequences to databases and calculates the statistical significance of matches.

B)Phylogenetic Analysis
CLUSTAL W/ TeXshade / BOXSHADE: Clustalw, texshade and box shade are multiple alignment tools which produces meaningful biological divergence to give evolutionary relationship and identical, non identical, similar and consensus sequences.

Structure Prediction
Expasy Proteomic Server: The EXPASY (Expert Protein Analysis System) proteomics server of the Swiss Institute of Bioinformatics (SIB) is dedicated to the analysis of protein sequences and structures as well as 2-D PAGE. It allows you to browse through a number of databases for identification of proteins, their analysis and prediction of their tertiary structure.

1) Primary Structure
PROTPARAM: Protparam is a tool which allows the computation of various physical and chemical parameters for a given protein stored in SWISS-PROT for a user entered sequence.

2) Secondary Structure
A) **SOPMA:** SOPMA is a secondary structure prediction method. Sopma (self-optimized prediction method with alignment) is an improvement of sopma method; these methods are based on the homologue method of Levin et al.
B) **HHPRED:** HHPred's sensitivity is competitive with the most powerful servers for structure prediction currently available. It is the first server that is based on the pair wise comparison of profile markov models (HMM).

3) Tertiary Structure
A) **CPH model**: CPH model is a protein homology modeling server. The template recognition is based on the profile–profile alignment using secondary structure and exposure predictions.
B)**Phyre**: Phyre provides a guide to interpreting the output of structure prediction servers in general and one such tool in particular, the protein homology/analogy recognition engine (Phyre). The Phyre server uses a Structural Classification of Proteins (SCOP) database as library for entered sequence and is then augmented with newer depositions in the Protein Data Bank (PDB).

Sdsc biology workbench: It is a server which helps us in analysis of protein in number of ways using several tools and algorithms phylogenetic analysis is one of its prime objectives.

4) Structure Analysis : Structural Classification
a) Protein Data Bank (PDB): The Protein Data Bank (PDB) is a repository for the 3D structural data of large biomolecules, such as proteins and nucleic acids. The data, typically obtained by X-ray crystallographic or NMR spectroscopy and submitted by biologists and biochemists from around the world, can be accessed at no charge on the internet.

b) The Structural Classification Of Protiens (SCOP): SCOP database is a largely manual classification of protein structural domains based on similarities of their amino acid sequences and 3D structures. SCOP utilizes 4 levels hierarchic structural classification namely Class, Fold, Superfamily, Family

c) CATH: The CATH protein structure classification is a semi automatic, hierarchical classification of protein domains. The four main levels of CATH hierarchy are Class, Arhitecture, Topology, and Homology.

d) PFAM: The Pfam database contains information about protein domains and family.

1. Identification of Drug Molecule:
DRUGBANK: The drugbank database available at the University of Alberta is a bioinformatics and chem-informatics resource that combines detailed drug (i.e. sequence, structure, pathway) information. The database contains nearly 4800 drug entries

Pubchem: Pubchem is a database of chemical molecules .the system is maintained by the national center for biotechnology information (NCBI), a component of the national library of medicine, which is the part of the National Of Health (NIH).

RASMOL: Rasmol is a computer program written for molecular graphics visualization intended and used primarily for the depiction and exploration of biological macromolecule structures, such as those found in the PDB. These have been uploaded by researchers who have characterized the structure of molecules usually by X-Ray crystallography or NMR spectroscopy.

2. Finding domain and motifs

a) PROSITE: It is database of protein families and domains. It comprises entries describing the domains, families, functional sites and patterns of amino acids their signatures and their profiles.
b) Prodom: It is another database of protein families which tells about the active domains of the protein. Thus it helps in localizing action plan to a smaller and well understood region.

c) SMART
d) TOPDOM
3. Receptor identification and designing
a) Receptor identification
IBM Pattern discovery: It helps user in locating the specific amino acid residue whose mutation gives rise to particular abnormalities. Hence it helps in choosing the best geometry for docking of the drug in order to make study more precise and specific.

b) Receptor design:
ARGUS LAB: Argus Lab is a program to built graphic representation of molecular models. Using this program, user is able see molecular models or even design matters by combing different elements. Every component can be edited to meet user's needs and he can join those atoms using any kind of bond possible. This way he can build simple or complete molecules.

IV) Toxicity and Tolerability Test
ADME-TOX WEB
ADME: ACD/ADME Suite is a collection of software modules that provide predictions relating to the pharmacokinetics profiling compounds, specifically their absorption, distribution, metabolism and excretion properties.

TOXBOXES: Genotoxicity describes a deleterious action on a cells genetic material affecting its integrity. Genotoxic substances are known to be potentially carcinogenic specifically those capable of causing genetic mutation and of contributing to the development of tumors.
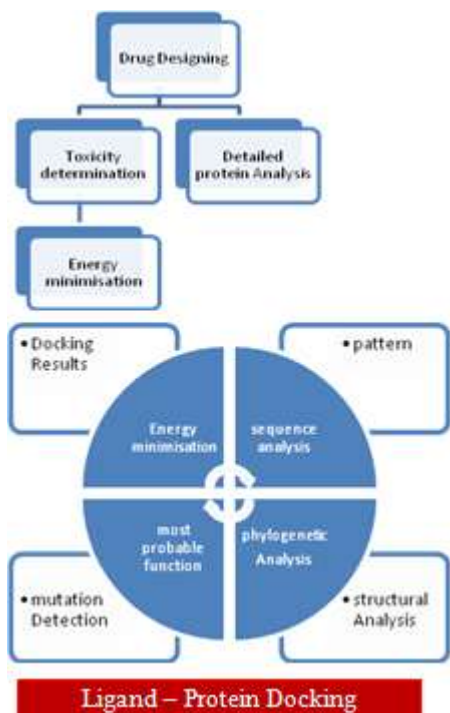

Fig 2: Strategy for Insilico Drug Designing

4. Results

The fasta sequence is obtained from the NCBI protein database and the length of the KRAS protein for humans is 189 amino acids,

Kras Sequence:
>sp|P01116|RASK_HUMAN GTPase KRas OS=Homo sapiens GN=KRAS PE=1 SV=1
MTEYKLVVVGAGGVGKSALTIQLIQNHFVDEYDPT IEDSYRKQVVIDGETCLLDILDTAGQEEYSAMRDQY MRTGEGFLCVFAINNTKSFEDIHHYREQIKRVKDSE DVPMVLVGNKCDLPSRTVDTKQAQDLARSYGIPFI ETSAKTRQRVEDAFYTLVREIRQYRLKKISKEEKTP GCVKIKKCIIM

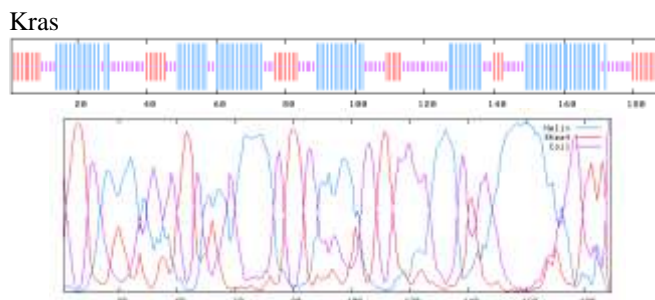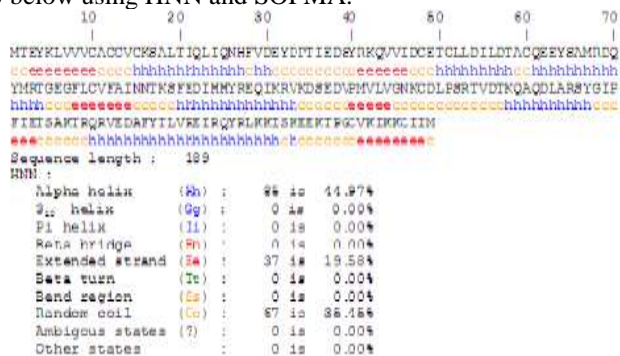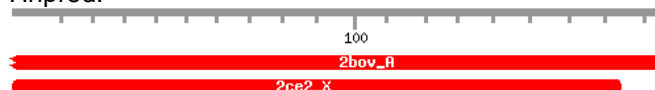The secondary structure analysis results area shown in figure 3 below using HNN and SOPMA:

Kras

Figure 3 : secondary structure Analysis

For Kras:
CPH Model: `entry: 2CE2 chain: X score: 297 E: 4e-81`
Hhpred:


```
No Hit Prob E-value P-value Score SS Cols
Query HMM Template HMM
22ce2 X GTPase HRAS; signaling 100.0 0 0
284.4 24.0 166
```

CPH Model: `entry: 2CE2` RAP2C_HUMAN_Ras-rela

(9) 0.573 0.573 0.573 0.593 0.586 0.612 0.591 0.621

0.000 0.142 0.399 0.383 0.638 0.687 0.701 0.689 0.722

0.713 0.643 `chain: X score: 336 E: 1e-92`

HHpred: `No Hit Prob E-value P-value Score SS Cols Query HMM Template HMM`
`12ce2 X GTPase HRAS; signaling 100.0 0 0 284.1 22.8 166`

2CE2 domain was scanned by PROSITE and PRODOM to reaffirm same domain while PDB gave more information about CATH and PFAM of the domain. Further 'Q-site finder' was used to find out the site of presence of amino acid 12 which was found to be present in 3 sites. The affect of their mutation are as follows shown in figure 4 while binding site prediction is done by using CASTp in figure 5:

IBM pattern discovery result

Figure 4: mutation sites by using IBM pattern discovery tool
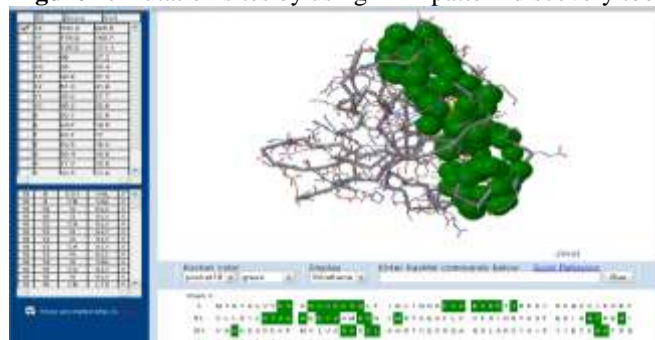

Figure 5: Binding site Prediction using CASTp

Pubchem

Different drug variants can be identified using PubChem and can be verified with the Drug bank followed by structure study and its ambiguity using Dunde Prodrug server. Table 2 shows the Toxicity analysis of the compounds using ADMETOX.

The details of the optimization are shown in the figure 6 and final configurations with results are shown in figure 7. During clustering the final poses shown are 27 configurations followed by a Re-clustering. The best ligand pose energy is -6.60725 kcal/mol.
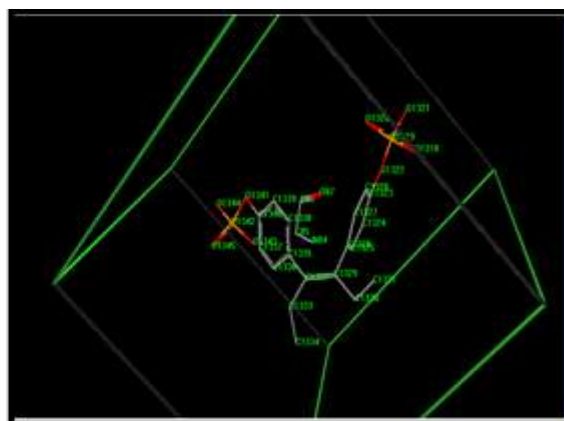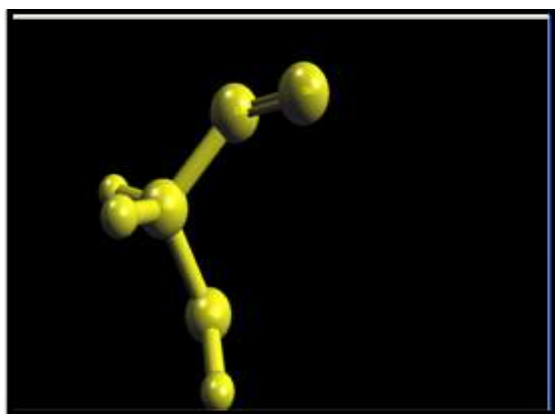
5. Conclusion and Discussion

The KRAS protein or its homologues where found in various organisms using the alignment tool BLASTP. By this we try to find out if there is any ancestor organism so that we can conclude with an idea that whether the protein is only functional in human or any other organism also. In case of most of the heritable diseases we conclude that the protein is transmitted from other organism to human or human to other organisms.

Through BlastP we also try to find out the organism having similar protein like human KRAS, so that we can target those organisms as animal models during drug validation phase of preclinical studies, closer organism are targeted for finding the tolerability and toxicity of the dose, we give it to humans. Using the above FASTA sequence BlastP was performed for various organism like *Homo sapiens, Pan troglodytes, Rattus norvegicus, Bos taurus, Canis lupus, Mus musculus, Danio rerio, Caenorhabditis elegan and E.Coli* and found that gene H-Ras/N-Ras/K-Ras are highly similar and their homologues in other related organisms are involved in the similar function. Proceeding with the structure prediction primary structure analysis for K-ras and its nearest homologue H-ras using PROTPARAM was done and various physico chemical properties of protein were studied. as protein is having the instability index greater than 40, this clarifies the protein is unstable and hydrophilic in nature because the hydropathic (GRAVY) value is -0.417 for Kras and -0.432 for Hras. After taking the drug it will create less toxicity as the aliphatic index is very high (81.96and85.03 respectively).

Table 2: ADMETOX values for final screened compounds

Properties / drug	Bio-availability	Absorption trans-cellular	Log D (blood)	Log P	Solubility (buffer-blood pH7.4)	Pgp inhibitor specificity	AMES test	LD$_{50}$ mouse (oral) mg/kg	LD$_{50}$ rat (oral) mg/kg
Tibolone	30-70%	100%	3.37	3.37	-4.85	0.326	0	860	530
Liarazole	30-70%	100%	2.65	3.0	-3.46	--------	0.926	180	64
Hexestrol	<30%	100%	5.1	5.1	-4.28	0.079	0.006	1400	1600
Fosfesterol	<30%	100%	-2.19	2.78	0.83	0.058	0.563	130	230
Tamoxifen	30-70%	100%	5.2	6.52	-4.02	Large amphi-phillic base (>360 mol.wt.)	0.012	680	680
Fadrozole	>70%	100%	2.08	2.13	-3.04	Small molecules (<250 mol. Wt.)	0.009	350	540
Azopxy-methane	>70%	98%	0.84	0.84	-0.57	Small molecules (<250 mol. Wt.)	0.436	370	420
Endometrion	>70%	100%	2.29	2.29	-4.22	-------	0.005	900	1100
Solumedrol	<30%	100%	-0.35	2.14	-1.35	-------	0.018	1900	2500

Figure 6: Ligand Receptor and labelled view of drug ligand

Surface view of ligand and receptor:

Active region of drug that adheres to the receptor:

Opaque view of ligand and receptor assembly
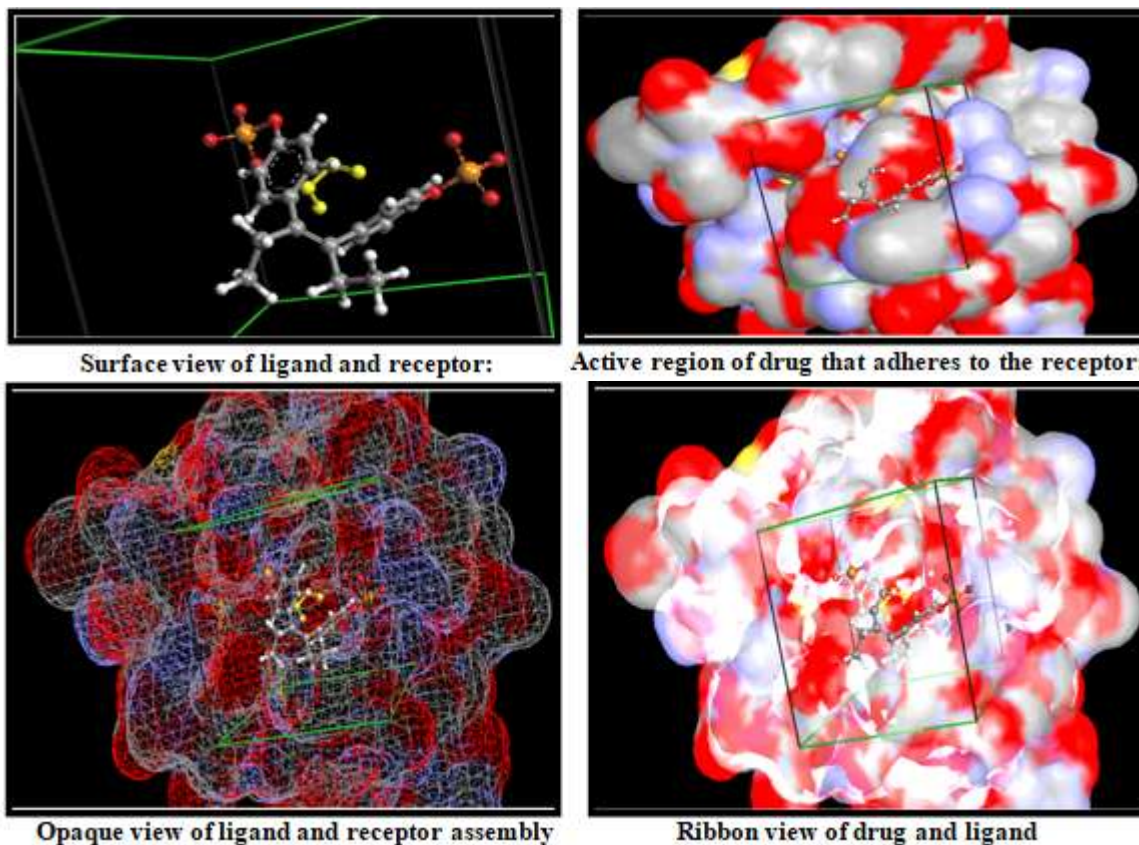
Ribbon view of drug and ligand

Figure 7: Geometry optimization and Docking Results

SOPMA and HNN both gave us the secondary structure we shows alpha rich regions. From secondary structure analysis, compute that both SOPMA and HNN protein have maximum percentage of alpha helix which is more than beta sheet. Thus protein is hydrophilic in nature and unstable. And as PROPARAM result also satisfies it, it is a membrane protein. As the above results were in favorable the tertiary structure analysis using PHYRE, HHPred and CPH model was carried and it was found that 2CE2 domain of the two proteins was the largest similar domain and it also harbors the 12th position amino acid glycine. The same values fall in for H-Ras also. Thus we tried to optimize the geometry of this domain using ARGUS lab in order to study the effect of drug after its docking into the optimized geometry.

The results of docking the drug were very good as energy is coming under the free energy difference (5 to 15 kcal/mol). And thus the current experiment ends with the docking the final pose of energy -6.60725 kcal /mol, which is a standard difference of energy and comes under free energy difference.
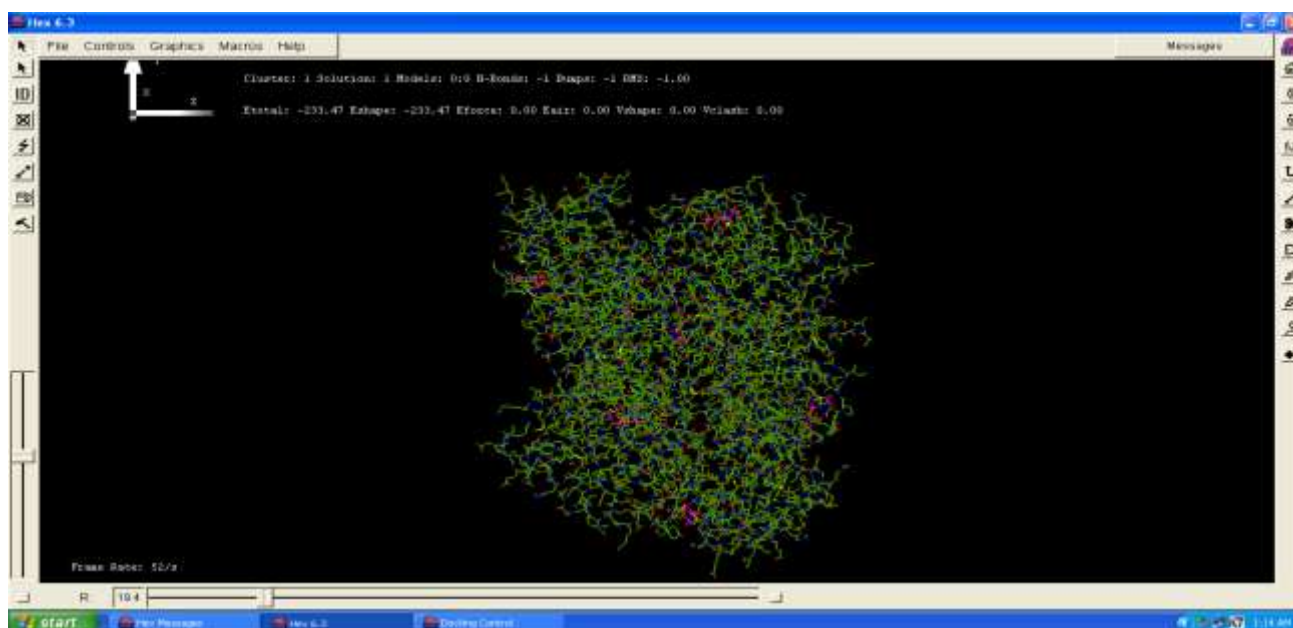

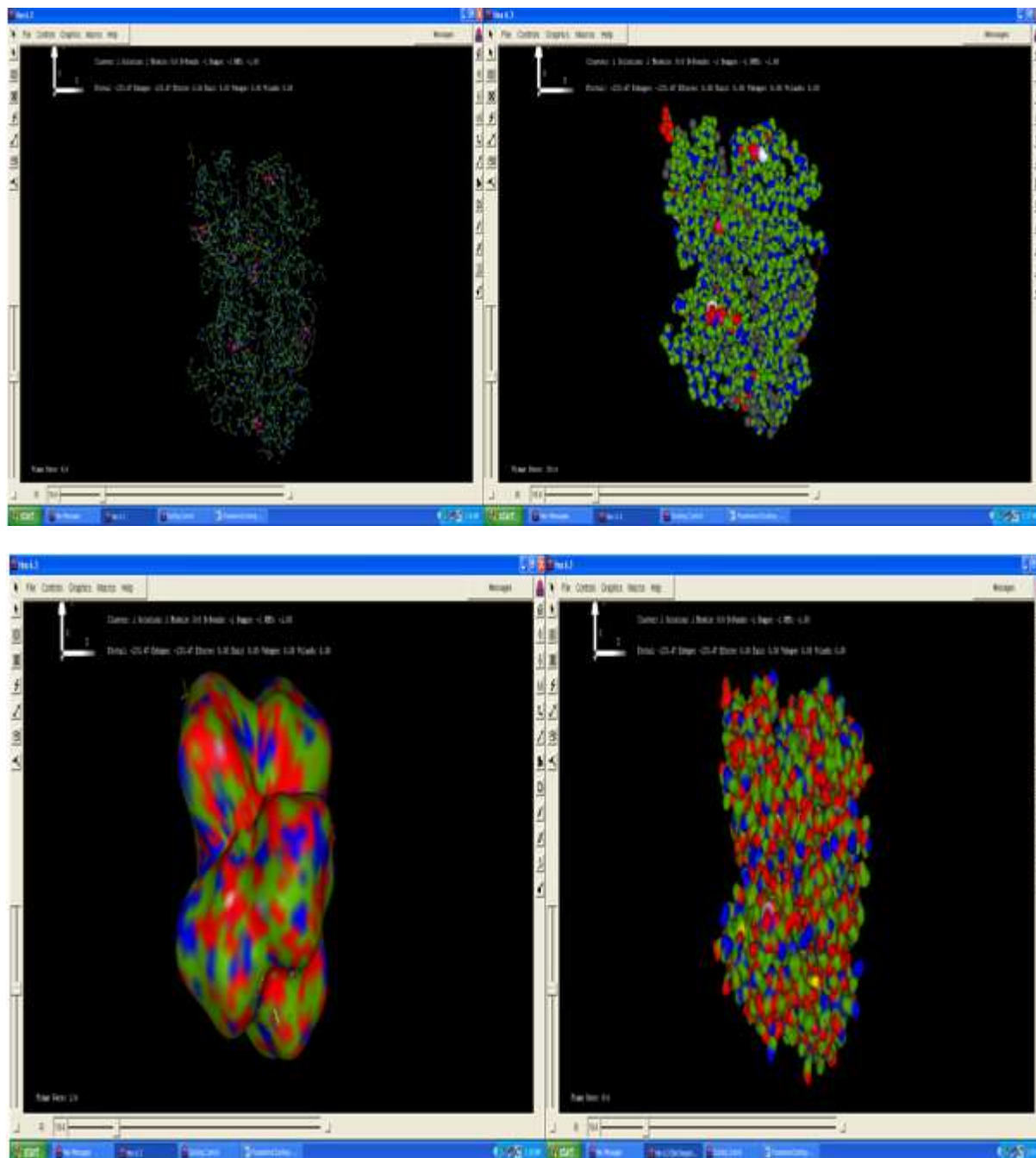**Figure 8:** Fosfesterol Docking By using HEX tool.

Figure 9: Hex results for docking just after energy minimization

References

[1] Low-dose continuous oral fosfestrol is highly active in 'hormone-refractory' prostate cancer**,** M. Orlando, M. Chacon, G. Salum & D. R. Chacon, *Annals of Oncology* 11 177-181. 2000

[2] K-ras mutations in non-small cell lung carcinoma: a review, Aviel Ronen S, blackhall FH, Shepherd FA, Tsao MS, clin Lung cancer.2006 jul;8(1):11-2.

[3] Oncogenic KRAS Activates Hedgehog Signaling Pathway in Pancreatic Cancer Cells, The Journal Of Biological Chemistry Vol. 282, NO. 19, pp. 14048–14055, May 11, 2007.

[4] Ras Family Genes: An Interesting Link Between Cell Cycle and Cancer**,** M. MACALUSO, G. RUSSO, C. CINTI, V. BAZAN, 2 N. GEBBIA, AND A. RUSSO, journal of cellular Physiology 192:125-130(2002).

[5] **Detection of *K-ras* mutations in non-small cell lung carcinoma,** E.M.NEVILLE, G.ELLISON, H. KIARIS, M.STEWART, D.A. SPANDIDOS, J.C.FOX and J. K. FIELD, INTERNATIONAL JOURNAL OF ONCOLOGY 7: 511-514, 1995.

Author Profile

Jayant Mishra received the M.S. degree from BITS pilani in 2009. He is currently pursuing his PhD in Biotechnology from Department of Biotechnology, Himalayan University, Arunanchal Pradesh.