In Silico Screening of Putative Drug Molecules to Target MSI Pathway for Colorectal Cancer and HNPCC.

Submitted in partial fulfillment of the award of Degree

Bioinformatics B.Tech

DEPARTMENT OF BIOTECHNOLOGY AND BIOINFORMATICS

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY

WAKNAGHAT

Submitted by:

Rahul Dabra (131508)

Name of supervisor

Dr. Tiratha Raj Singh

Assistant Professor (Senior Grade)
Jaypee University of Information Technology
Waknaghat, Solan

Dr. Raghu M. Yennamalli

Assistant Professor (Grade II)
Jaypee University of Information Technology
Waknaghat, Solan

CERTIFICATE

This is to certify that the work entitled **In Silico Screening of Putative Drug Molecules Target MSI pathway for Colorectal Cancer and HNPCC** pursued by **Rahul Dabra (131508**) in partial fulfillment for the award of degree Masters of Technology in Biotechnology from Jaypee University of Information Technology, Waknaghat has been carried out under my supervision. This part of work has not been submitted partially or wholly to any other University or Institute for the award of any degree or diploma.

Dr. Tiratha Raj Singh
Assistant Professor (Senior Grade)

Jaypee University of Information Technology
Waknaghat, Solan

Dr. Raghu M. Yennamalli

Assistant Professor (Grade II)

Jaypee University of Information Technology
Waknaghat, Solan

DATE :

ACKNOWLEDGEMENT

All praise belongs to the almighty lord to whom I thank for the strength, courage and perseverance bestowed upon to me to undertake the course of the study.

I hereby acknowledge with deep gratitude the cooperation and help given by all members of Jaypee University in helping with my project.

With proud privilege and profound sense of gratitude, I acknowledge my indebtedness to my guide **Dr. Tiratha Raj Singh** and **Dr. Raghu M. Yennamalli**, Assistant professor, Jaypee University of Information and technology for their valuable guidance, suggestions, constant encouragement and cooperation.

I express my thanks to Prof. Rajinder Singh Chauhan, Dean, Department of Biotechnology and Bioinformatics, Jaypee University of Information and Technology.

I would also like to extend my gratitude towards Mr. Ashwani Kumar, Ms. Ankita Shukla and other staff members for their constant help and motivation for successfully carrying my research work.

Date:

Place:

Rahul Dabra (131508)

SUMMARY

The findings that different molecular pathways are involved in colorectal cancer development have helped researchers build different models and understand how colorectal cancer initiates and progresses. However, the application of molecular markers on large scale populations is now facilitating the understanding of the peculiar role of these alterations on disease behavior, prognosis and response to treatments.

- We predicted damage for 457 different CRC associated SNP's out of which total 108 are found to be highly damaged.

- Better understanding of mutations at various positions in genome.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATION

ABBREVIATED WORD	WORD
HNPCC	Hereditary Nonpolyposis Colorectal Cancer
FAP	Familial Adenomatous Polyposis
SNP	Single Nucleotide Polymorphism
SIFT	Scale Invariant Feature Transform
MLH	MutL Homolog
CNS	Central Nervous System
MSI	Micro Satellite Instability
MMR	Mis-Match Repair
5FU	5 Fluorouracil
FU-LV	Fluorouracil-Leucovorin

CHAPTER 1: INTRODUCTION

1.1 Overview:

Colorectal cancer is cancer that starts in the colon or rectum. Most colorectal cancers are adenocarcinomas. Colorectal cancer begins as a growth called a polyp thus later on may form on inner wall of the colon and rectum. Some polyps become cancer over time. Finding and removing polyps can prevent colorectal cancer. Colorectal cancer is the third most common type of cancer in men and women in USA. Deaths from colorectal cancer decreased with use of colonoscopies and fecal occult blood tests. In India, the annual incident rates for colon cancer and rectal cancer in men and women are 4.4 and 4.1 per 100000, respectively. For better understanding of CRC we have to understand concept of MSI-pathways and history of HNPCC [1]. For further screening of putative drug molecule targeting MSI-pathways we have to extract 4P7A-protein from PDB database and dock it with possible drug chemicals available to find out the best ligand-protein complex. SNP prediction will allow us the identification of deleterious mutations generally harmful for humans.

1.2 Description of the project:

Inherited gene mutations that increase the risk of colon cancer:

It can be passed through families, but these inherited genes are linked to only a small percentage of colon cancers.The most common forms of inherited colon cancer syndromes are:

- **Hereditary nonpolyposis colorectal cancer (HNPCC)**

HNPCC, also called Lynch syndrome, increases the risk of colon cancer and other cancers. People with HNPCC tend to develop colon cancer before age 50.

- **Familial adenomatous polyposis (FAP)**

FAP is a rare disorder that causes you to develop thousands of polyps in the lining of your colon and rectum. People with untreated FAP have a greatly increased risk of developing colon cancer.FAP, HNPCC colon cancer syndromes can be detected through genetic testing.

1.3 MSI- pathways:

This pathway describes form of genomic instability that is involved in genesis of approx. 15% of sporadic CRC cancer and >95% of HNPCC syndrome [2]. MSI is caused due to inactivity of the DNA Mismatch Repair. The MMR system is a multi-protein system that acts like proofing machine that increases fidelity of DNA replications by identification and direct repair of mismatched nucleotides [2]. In human cells MMR system is comprised of multiple interaction of proteins including the human MutS homologue (MSH) 2, and human MutL homologue (MLH) 1. MSI is very good diagnostic marker in determination of lynch syndrome and to determine prognosis for cancer treatments. The NCI has agreed on five microsatellite markers necessary to determine MSI presence: two mononucleotides, BAT25 and BAT26, and three dinucleotide repeats, D2S123, D5S346, and D17S250. MSI-H tumors are derived by MSI of greater than 30% of unstable MSI biomarkers. MSI-L tumors result from less than 30% of unstable MSI biomarkers. MSI-L tumors are termed as tumor of alternative etiology. Several studies illustrates that MSI-H patients responds good to surgery alone, rather than chemotherapy, therefore preventing patients from needlessly experiencing chemotherapy [12].

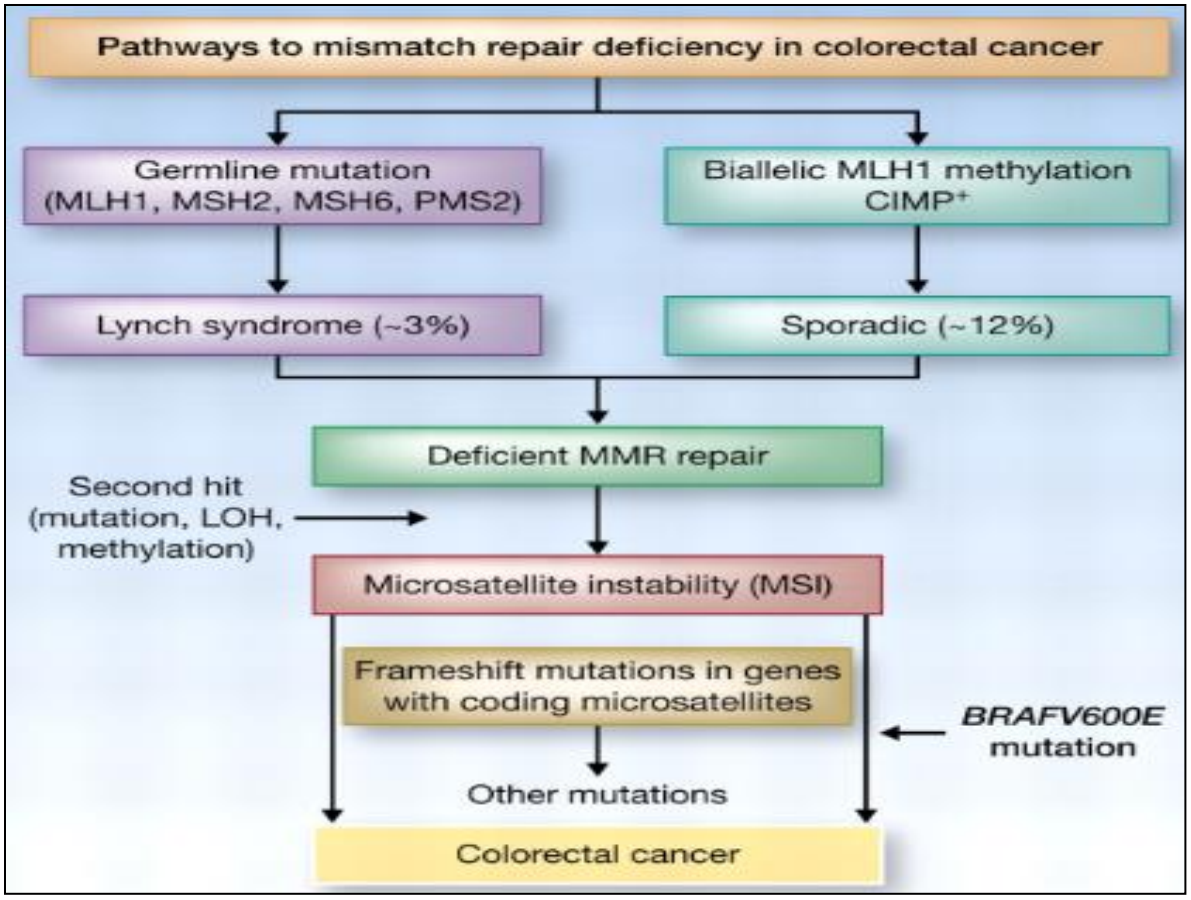


Figure 1: MSI pathway colorectal cancer

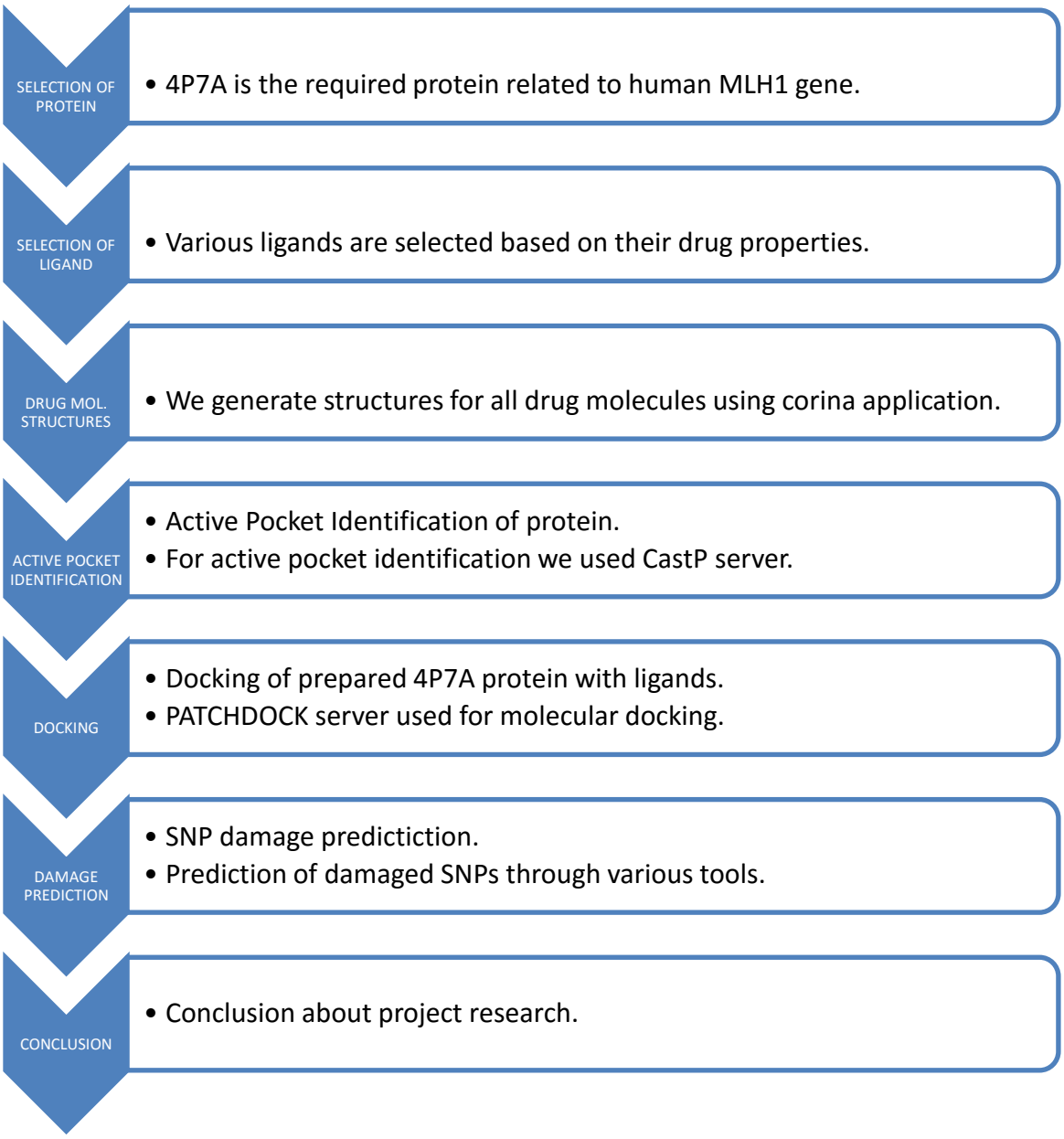
1.4 Study of HNPCC:

Autosomal inherited disorder of cancer with the higher penetrance rate (80– 85%), 31 and many up to date mutations are described by 5 mismatch repair genes i.e. hMSH2, hMLH1, hPMS1and hPMS2[3] along with direct contrast with majority of colorectal cancers, which are aneuploid in chromosomal constitution. Reasons for lacking of detectable mutations in hMLH1 and hMSH2 in sporadic cancers with micro-satellite instability lead to hypothesis that there may be other genetic loci for encoding proteins responsible for DNA mismatch repair [3].

1.5 Colorectal cancer risk factors:

AGE	• Risk of colorectal cancer increases as people get older 90% of CRC occures in people over age of 50.
GENDER	• Men have slightly more risk of developing CRC than women.
FAMILY HISTORY	• If a person has a family history of CRC his or her risk of developing disease is nearly doublethe average risk of CRC.
RARE INHERITED CONDITIONS	• Members of families with uncommon inherited conditions, also have a significantly increased risk of CRC include: FAP, ATTENUATED FAP, LYNCH SYNDROME etc.

CHAPTER 2: MATERIAL AND METHODS



2.1 Protein structure:

Here is the crystal structure of the mlh1 gene, obtained from PDB database. N- termination is denoted by blue ribbons and C- termination is shown by red ribbon. Metal ion represented by spheres.

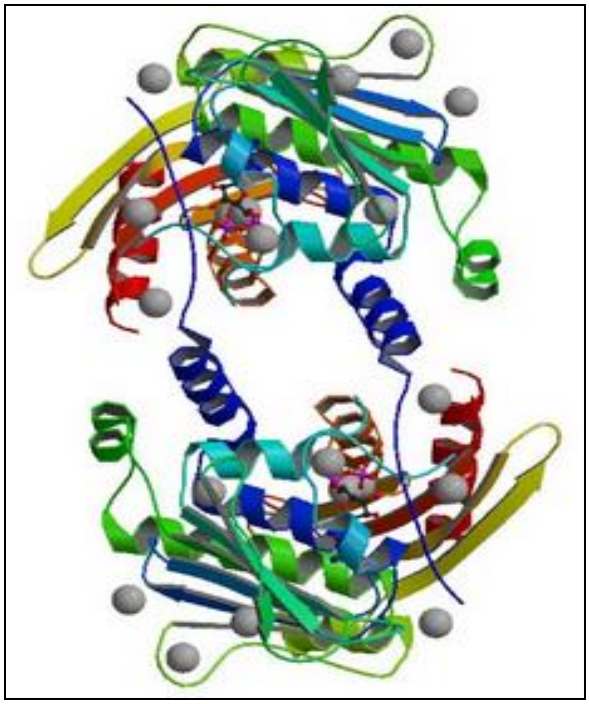


Figure2:4P7A Crystal Structure of human MLH1.

Experimental Data of protein 4P7A:

- **Method**: X-RAY DIFFRACTION
- **Resolution**: 2.3 Å
- **R-Value Free**: 0.254
- **R-Value Work**: 0.203

4P7A is a 1 chain structure with sequence from human. This structure shares a high degree of similarity with previously determined prokaryotes MLH 1 homologs, however this structure affords a more accurate platform for MLH 1 variants classification.

2.2 Drug molecules:

PubChem is a database for chemicals and their molecules and provide their activities against biological assays. The system is run by the National Center for Biotechnology Information (NCBI) [5].

TABLE 1. Drug molecules and their properties : We collected molecules information from various sources include: NCBI Pubchem, National cancer institute and WIKIPEDIA and on the basis of their bioavailability, half-life, and trade name we classified them as below:

DRUGS	TRADE NAME	BIOAVAILABLE	BIO. HALF-LIFE	FORMULA
Avastin	Avastin	100% (IV only)	20 days	$C_{6638}H_{10160}N_{1720}O_{2108}S_{44}$
Bevacizumab	Avastin	100% (IV only)	20 days	$C_{6638}H_{10160}N_{1720}O_{2108}S_{44}$
Camptosar				$C_{33}H_{38}N_4O_6$
Capecitabine	Xeloda	Extensive	38—45 minutes	$C_{15}H_{22}FN_3O_6$
Cetuximab	Erbitux		114 hrs	$C_{6484}H_{10042}N_{1732}O_{2023}S_{36}$
Cyramza	Cyramza			$C_{6374}H_{9864}N_{1692}O_{1996}S_{46}$
Eloxatin	Eloxatin	Complete	10 - 25 minutes	$C_8H_{14}N_2O_4Pt$
5-FU	Adrucil, Carac	28 to 100%	16 minutes	$C_4H_3FN_2O_2$
Fluorouracil Injection	Adrucil, Carac	28 to 100%	16 minutes	$C_4H_3FN_2O_2$
Oxaliplatin	Eloxatin	Complete	10 - 25 minutes	$C_8H_{14}N_2O_4Pt$
Panitumumab	Vectibix		9.4 days	$C_{6398}H_{9878}N_{1694}O_{2016}S_{48}$
Leucovorin Calcium	Many	Dose dependent	6.2 hours	$C_{20}H_{23}N_7O_7$
Lonsurf	Lonsurf			
Ramucirumab	Cyramza			$C_{6374}H_{9864}N_{1692}O_{1996}S_{46}$
Regorafenib	Stivarga	69-83%	20-30 hours	$C_{21}H_{17}ClF_4N_4O_4$
Stivarga	Stivarga	69-83%	20-30 hours	$C_{21}H_{17}ClF_4N_4O_4$
Vectibix	Vectibix		9.4 days	$C_{6398}H_{9878}N_{1694}O_{2016}S_{48}$
Wellcovorin	Many	Dose dependent	6.2 hours	$C_{20}H_{23}N_7O_7$
Xeloda	Xeloda	Extensive	38—45 minutes	$C_{15}H_{22}FN_3O_6$
Zaltrap	Eylea, Zaltrap			$C_{4318}H_{6788}N_{1164}O_{1304}S_{32}$

2.3 Drug combinations :

We get information about these drug combinations from NCBI PubChem and National cancer institute where we get the detail how these combinations are formed by combinations of 2 or more drug molecules (Table 2).

Table 2: Drug combinations and their constituents:

DRUG NAME	**CONSTITUENT COMBINATIONS**
Capox	Capecitabine, Oxaliplatin
Folfiri-Cetuximab	Leucovorin Calcium, Fluorouracil, Irinotecan Hydrochloride + Cetuximab
Folfiri-Bevacizumab	Leucovorin Calcium, Fluorouracil, Irinotecan Hydrochloride + Bevacizumab
Folfox	Fluorouracil, Oxaliplatin
Fu-Lv	Fluorouracil, Leaucovorin Calcium
Xelox	Capecitabine, Oxaliplatin
Xeliri	Capecitabine, Irinotecan Hydrochloride

2. 4Corina:

Open source and freely available fast and powerful 3D structure generator for drug-like molecules. Corina provide robust data, and therefore provides best application to convert large chemical databases. Generates a single high quality and low- energy conformation. Input is SMILES string then press enter button.

LINK: https://www.mn-am.com/products/corina

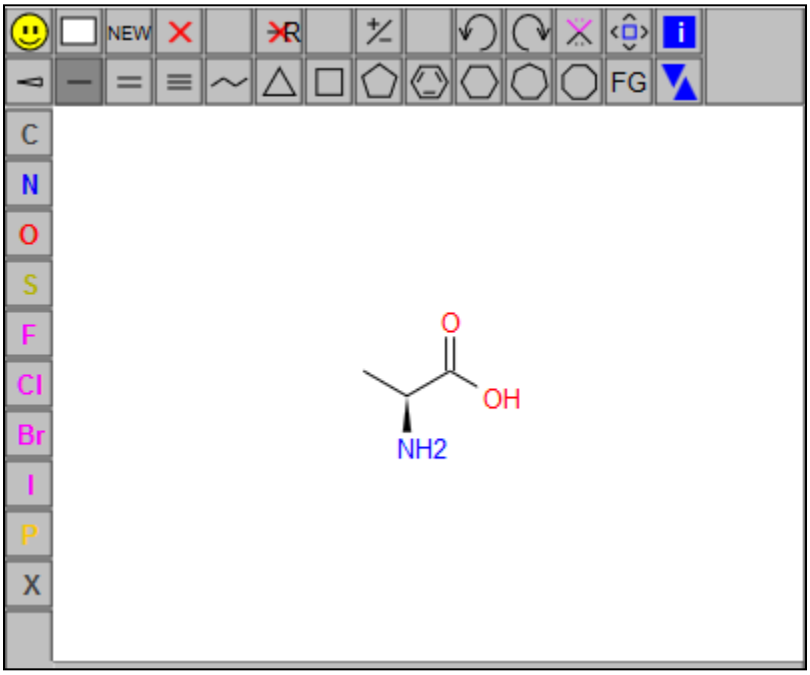


Figure 3. Corina Application

2.5 CastP:

CastP server use weighted Delaunay triangulation and alpha complex for shape measurements. Used of identification and measurement of active binding pockets and also for measuring interior inaccessible cavities, for proteins and other small molecules. Allow us to measure number of mouth openings, area of the openings, circumference of mouth lips, in both SA and MS surfaces for each binding pocket [7].

```
ATOM    106  CA  ASN A  17      30.826 -37.725  14.075  1.00 61.16  45  POC
ATOM    136  CB  ALA A  20      29.234 -38.488   9.679  1.00 62.00  45  POC
ATOM    137  N   ALA A  21      30.218 -35.497   9.548  1.00 55.65  45  POC
ATOM    160  CG1 VAL A  24      25.779 -32.400   5.863  1.00 67.47  45  POC
ATOM    169  CD1 ILE A  25      27.897 -28.813   6.825  1.00 59.05  45  POC
ATOM    231  O   GLU A  34      21.221 -22.973   6.196  1.00 36.41  45  POC
ATOM    238  CA  MET A  35      23.118 -21.046   6.706  1.00 33.53  45  POC
ATOM    257  CB  GLU A  37      18.516 -22.791   3.735  1.00 38.89  45  POC
ATOM    264  C   ASN A  38      17.842 -19.362   7.970  1.00 35.18  45  POC
ATOM    270  N   CYS A  39      18.594 -18.457   7.359  1.00 34.63  45  POC
ATOM    288  CB  ASP A  41      13.881 -21.127   7.849  1.00 40.85  45  POC
ATOM    440  O   ILE A  61      25.353 -14.257   9.848  1.00 37.19  45  POC
ATOM    461  OD2 ASP A  63      20.018 -14.934  10.060  1.00 32.55  45  POC
ATOM    471  CA  GLY A  65      15.483 -12.424  10.943  1.00 43.95  45  POC
ATOM    479  OG1 THR A  66      13.019 -15.907  13.571  1.00 49.87  45  POC
ATOM    481  N   GLY A  67      15.771 -14.478  14.868  1.00 42.02  45  POC
ATOM    486  CA  ILE A  68      17.383 -17.612  18.402  1.00 41.60  45  POC
ATOM    526  CB  ASP A  72      16.866 -22.049  21.950  1.00 45.06  45  POC
ATOM    553  CD1 ILE A  75      17.374 -27.577  22.725  1.00 61.72  45  POC
ATOM    555  CA  VAL A  76      20.975 -25.119  19.059  1.00 47.20  45  POC
ATOM    570  O   GLU A  78      21.607 -30.407  17.032  1.00 47.53  45  POC
ATOM    586  NH2 ARG A  79      23.228 -38.953  13.904  1.00 95.09  45  POC
ATOM    600  C   THR A  81      17.095 -28.977  16.568  1.00 46.84  45  POC
ATOM    610  OG1 THR A  82      14.168 -27.950  13.303  1.00 39.06  45  POC
ATOM    612  N   SER A  83      13.635 -25.945  15.599  1.00 41.56  45  POC
ATOM    625  CE  LYS A  84      14.182 -23.641  10.579  1.00 48.63  45  POC
ATOM    637  C   GLY A  98      16.706 -35.389   6.673  1.00 88.55  45  POC
ATOM    642  O   PHE A  99      17.685 -32.940   9.173  1.00 52.11  45  POC
ATOM    650  N   ARG A 100      15.574 -32.892  10.045  1.00 55.25  45  POC
ATOM    670  O   GLY A 101      20.416 -31.036  13.809  1.00 46.01  45  POC
ATOM    671  N   GLU A 102      20.292 -31.035  11.560  1.00 42.77  45  POC
ATOM    684  CB  ALA A 103      22.432 -27.645   8.731  1.00 42.84  45  POC
ATOM    685  N   LEU A 104      23.624 -27.217  11.580  1.00 40.56  45  POC
ATOM    697  CB  ALA A 105      24.556 -30.884  14.729  1.00 41.17  45  POC
ATOM    698  N   SER A 106      26.402 -30.420  12.341  1.00 40.95  45  POC
ATOM    724  ND1 HIS A 109      29.188 -32.231  15.568  1.00 47.96  45  POC
ATOM    990  O   THR A 148      22.602 -13.211  11.762  1.00 36.63  45  POC
ATOM   1010  CD1 ILE A 150      24.147 -18.888  13.144  1.00 47.54  45  POC
ATOM   1843  O   ASN A 263      12.451 -25.702  -1.172  1.00 66.94  45  POC
ATOM   1868  NH2 ARG A 265      18.305 -29.665   0.784  1.00100.74  45  POC
ATOM   2106  CA  SER A 299       4.774 -32.249  -2.827  1.00116.87  45  POC
ATOM   2110  N   PRO A 300       5.525 -34.425  -1.868  1.00135.24  45  POC
```

Figure 4. CASTp extracted data of 4P7A

2.6 Pymol Viewer:

It is an Open-source, molecular visualization system. Currently it is commercialized by Schrödinger Inc. PyMOL, can produce high-quality 3D images of small molecules and biological macromolecules, like proteins.

2.7 PatchDock:

PatchDock is an algorithm for molecular docking. We give input of two molecules of any type: like, proteins, DNA, peptides, drugs etc. The output is a list of potential complexes sorted by shape complementarity criteria [6].

2.8 Ligplot:

It is used for the generation of 2-D representations of protein-ligand complexes from standard Protein Data Bank file input.

2.9 Damage prediction Tools:

- **PROVEAN PREDICTION:**

Provean is used to identify if amino acid substitution has an effect on protein function. Filter sequence variations and identify functionally important nonsynonymous variations [8].79.5% accuracy for human UniProt protein variations. Low score than -2.5 indicated the variants to be damaged.

- **PANTHER PREDICTION:**

Protein Analysis through Evolutionary Relationships classify proteins in order to process high throughput analysis[10]. It is used worldwide for protein evolutionary and functional classification. It estimates likelihood that SNPs will affect protein function.

- **SNP & GO:**

Predict single point protein mutations, that cause diseases in humans. Classify mutations as neutral or disease related.

- **MUTATION ACCESSOR:**

Use multiple sequence alignment. Conversation score is combined with specificity score to give functional impact score. Variants are labelled as ‘neutral’, ‘low’, ‘medium’, ‘high’. High FI (Functional Impact) score > 3.5.

- **SIFT (Scale Invariant Feature Transform):**

Predicts if an amino acid substitution affecting the function of protein or not based on sequence homology method and using physical properties of amino acids [9]. It can be applied to nonsynonymous polymorphisms and missense mutations.

- **POLYPHEN 2:**

Use to predict amino acid substitution impact on structure and function of human protein. It uses physical and comparative considerations. SNP is predicted to be highly deleterious if PSIC score is 1.

- **PHD SNP:**

Predicts human deleterious single nucleotide polymorphisms. Based on decision tree with SVM sequence coupled to SVM profile. It classifies mutations as neutral or disease-related.

- **MutPRED:**

Classify amino acid substitution as disease associated or neutral in human. Also predicts molecular cause of disease. Deleterious if g score is high than 0.5.

- **PREDICT SNP:**
  Predicts disease related mutations. For the effect of amino acid substitutions and nucleotide substitutions. Predict SNP consists of six prediction tools (MAPP, SNAP, POLYPHEN 1, POLYPHEN 2, SIFT AND PHD-SNP).

- **SNAP2:**

Classifier based on machine learning tool “neural network”. Distinguish between effect and nonsynonymous SNPs by taking sequence and variant featured in to account. It has accuracy of 82%.

2.10 Selection of active sites in 4P7A Protein:

We get this detailed structural information from PDB database about active sites in 4P7A-proteinby clicking in Ligands section. In the following structure black dashed lines shows hydrogen bonds, salt bridges, and metal interactions. Green solid line indicate hydrophobic interactions and green dashed lines show π-π and π-cation interactions.

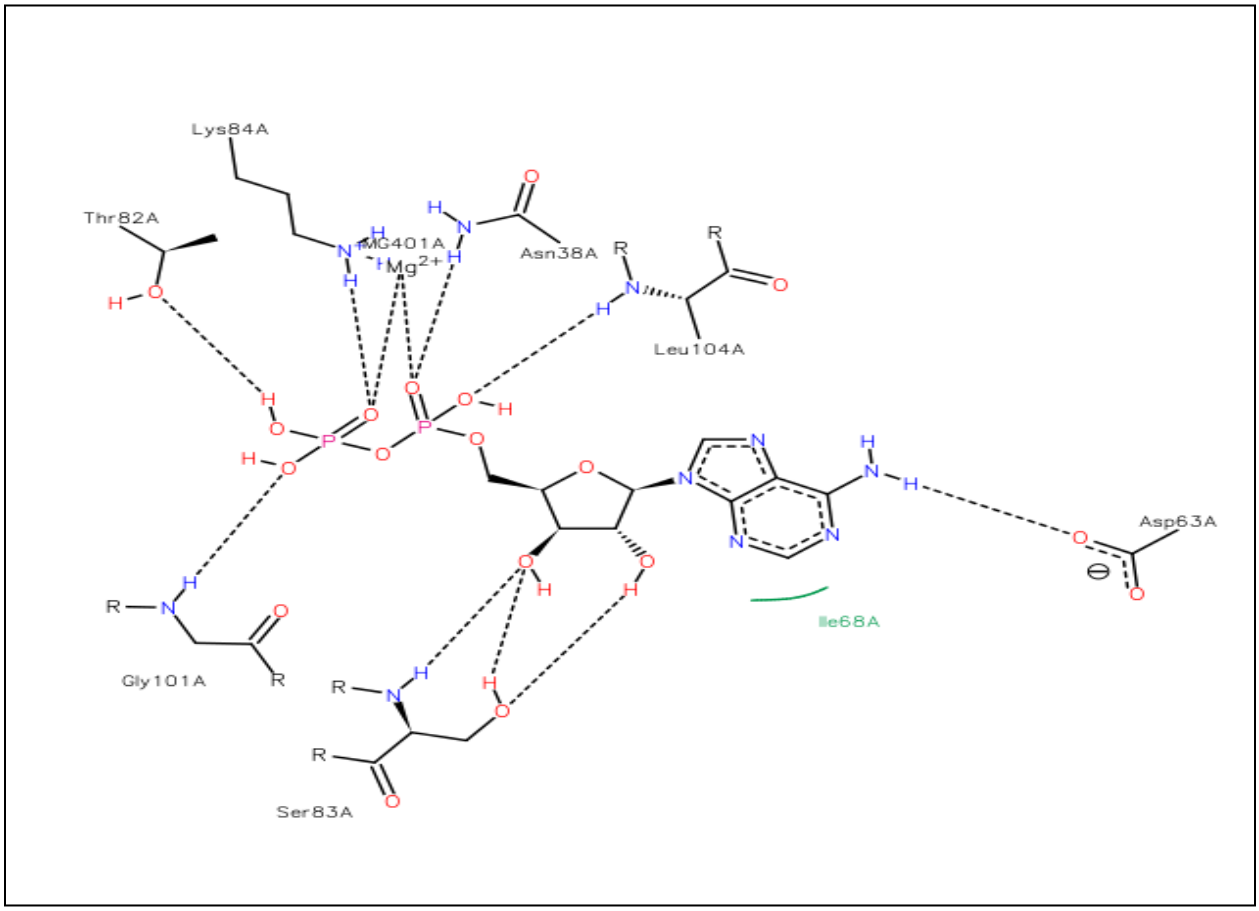


Figure 5. ACTIVE SITES IN 4P7A

While performing predictions of the active sites we came to know following active sites :

Asn38A, Asp63A, Se68A, Thr82A, Ser83A, Lys84A, Gly101A, Leu104A.

2.11 Extraction of prepared4P7A from wild type:

We opened wild 4P7A-Protein in notepad and edit to extract all the active sites, in the 4P7A-protein. Then we saved the new prepared 4P7A-Protein, later which we used for docking with ligand molecules.

2.12 Docking with Ligands:

Using PatchDock server we will do Docking of prepared 4p7a-protein with following ligand and the following ligands are selected on the basis if their drug properties mentioned in Table 1:

- **5-FLUOROURACIL**
- **Avastin**
- **Bevacizumab**
- **Folinic Acid**
- **Campostar**
- **Capecitabine**
- **Cetuximab**
- **Eloxatin**
- **Stivarga**
- **Leucovorin Calcium**
- **Welcovorin**

Before submitting each of the ligands with 4P7A protein, we have to fix Clustering RMSD value to 1.5, and select complex type "protein-small ligand" also in Advanced Options we have to submit ACTIVE SITE information in receptor binding site option. Then put your mail in the column, and after proceeding the run process PatchDock will send all the results to your respective e-mail. Using PyMOL viewer we can study the docking structure and represent final dock result and save it.

2.13 Damage prediction and Extraction of damaged snp:

- **METHODOLOGY:**

Firstly, we have to collect SNP data of CRC from NCBI SNP database. Then we have to select those SNP about which we have complete information and has known biological significance. Therefore, we extracted 457 CRC associated SNPs. Now damage prediction is to be done to get highly damaged SNP.

Figure 6. Methodologies for SNP prediction

- **HIGHLY DAMAGED SNPs:**

Here is the table of highly damaged SNPs predicted by various prediction tools. Below are the scores (PROVEAN, SNP & GO, POLYPHEN 2 and PHD-SNP) for respective highly damaged SNPs. Red highlighted SNP are most highly damaged based on their prediction scores.

Damaged SNP's	PROVEAN	SN & G0 RI	SNP & GO PROBABILITY	POLYPHEN 2	PHD SNP
R18C	-7.221	7	0.826	1	3
		1	0.563		
E23K	-3.63	8	0.911	0.999	0
		6	0.789		
R27W	-7.105	9	0.928	1	9
		6	0.803		
D63N	-4.854	8	0.914	1	6
		5	0.762		
N64S	-4.587	6	0.789	1	6
		0	0.51		
G65V	-8.737	8	0.908	1	4
		7	0.86		
G65D	-6.795	8	0.919	1	4
		7	0.863		
G67R	-7.766	9	0.933	1	6
		8	0.889		
G67E	-7.766	9	0.935	1	2
		8	0.895		
I68S	-5.824	9	0.928	1	5
		6	0.784		
I68N	-6.795	9	0.925	1	2
		5	0.77		
L73P	-5.35	8	0.883	1	7
		9	0.929		
		8	0.881		

R79S	**-5.828**	**6**	**0.823**	**1**	**3**
		8	**0.876**		
		6	**0.784**		
T82I	**-5.828**	**7**	**0.84**	**1**	**2**
		10	**0.999**		
		4	**0.72**		
S83I	**-5.828**	**7**	**0.868**	**1**	**2**
		9	**0.968**		
		5	**0.737**		
K84E	**-3.886**	**7**	**0.835**	**1**	**4**
		10	**0.997**		
		7	**0.841**		
K84N	**-4.857**	**6**	**0.81**	**1**	**2**
		10	**0.998**		
		5	**0.729**		
D154V	**-7.089**	**7**	**0.867**	**0.999**	**1**
		7	**0.867**		
		7	**0.836**		
F156L	**-5.619**	**7**	**0.853**	**0.862**	**3**
		8	**0.893**		
		4	**0.71**		
G244D	**-5.455**	**8**	**0.901**	**0.966**	**5**
		6	**0.824**		
		6	**0.813**		
N263S	**-4.8**	**7**	**0.852**	**0.994**	**0**
		9	**0.932**		
		5	**0.763**		
V303E	**-5.938**	**7**	**0.826**	**0.999**	**3**
		8	**0.895**		
		6	**0.802**		
N306K	**-6**				**3**
H315P	**-7.567**	**4**	**0.688**	**0.999**	**7**
		2	**0.603**		
		3	**0.648**		

Table 3: Highly damaged SNPs

2.14 PATCHDOCK METHODOLOGY

- **Patchdock methodology:**

This server runs PatchDock algorithm with following default values:

Receptor molecule	• Upload file in PDB format.
Ligand molecule	• upload file in pdb format same as receptor molecule.
E-Mail Address	• Link with the result of your job request will send to your specified mail. There we can see docking results.
Clustering RMSD	• Higher the values smaller the number of results we will get. Best value for protein-small molecule docking is 1.5 .
Complex Type	• For protein-small molecule docking the algorithm use parameter set optimization.
Receptor Binding Site	• Upload a file containing data of potential binding sites in the receptor in the format : • 88 L, 89L, 102 H, 105 H.
Ligand Binding site	• List residues of ligand potential binding site.
Distance constraints	• We have to specify distance constraints between pairs of atoms, one in receptor and one in ligands.

- **PatchDock Table Format:**

1. **Solution No.:** Contain number of solutions.
2. **Score:** Shows Geometric shape complementarity.
3. **Area:** Provide Approximate interface area of complex.
4. **ACE :** Atomic Contact Energy.
5. **Transformation :** 3D transformation, should be applied on ligand molecule.
6. **PDB file download :** Download the predicted complex structure in PDB format.

CHAPTER 3: RESULTS

Cartoon representation where helices are shown and Beta-strands are shown as arrows. The receptor is colored in rainbow representation with the N-ter in blue and the C-ter in red. The transition colors indicates the sequence in between the N and C terminals. The ligand shown is the top ranked confirmation obtained from PatchDock, where the ligand is shown in stick confirmation. The metal ion (mg2+) is shown in sphere and colored green. Below PatchDock protein-small molecule complex structures are visualized by using Pymol.

1. 4P7A_5FU

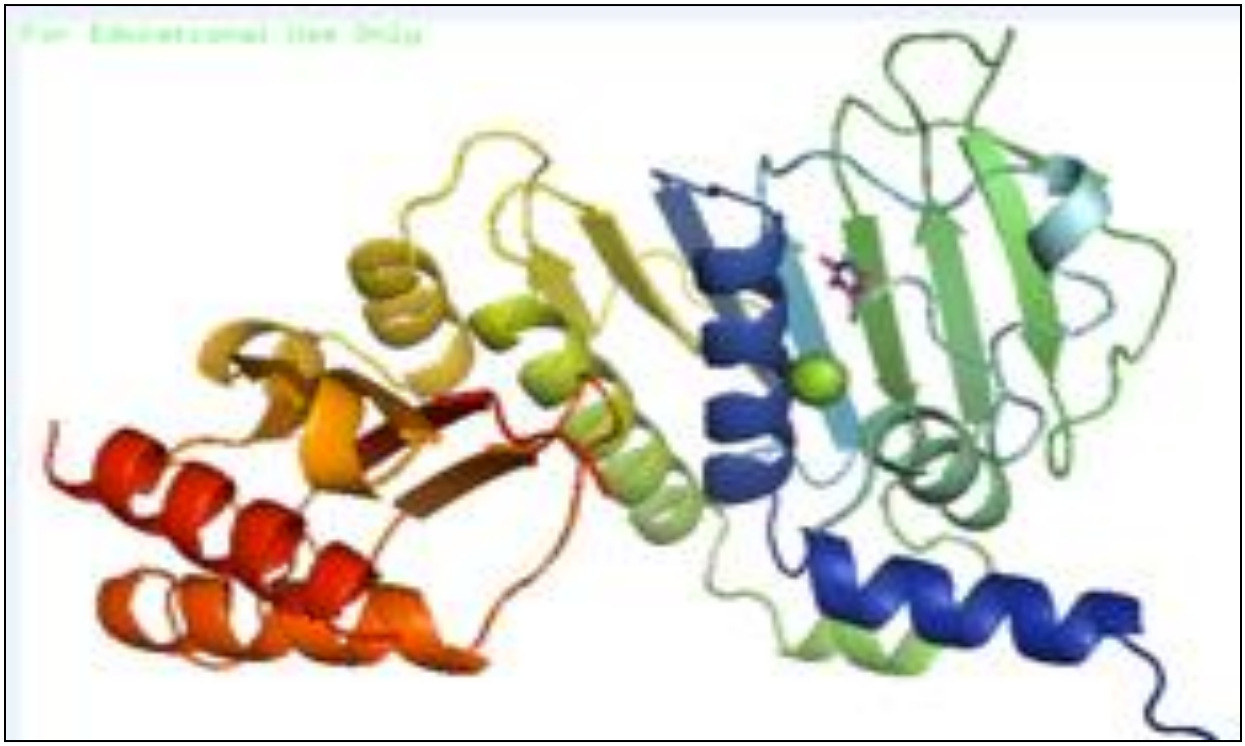


Figure 7. 4P7A_5FU Docking result

Solution No	Score	Area	ACE
1	1962	201.60	-58.75
2	1902	218.80	-51.82
3	1814	205.10	-33.61
4	1812	208.80	-47.83
5	1798	203.50	-27.63
6	1794	210.90	-40.03
7	1792	202.80	-44.09
8	1788	208.20	-31.13
9	1768	201.00	-2.64
10	1766	203.70	-31.38

TABLE 4

Above table contains top 10 results for the docking of 4P7A with ligand 5FU.Solution 1 with highest score 1962, area 201.60, ACE -58.75 is the optimal site for protein-small molecule docking complex.

2. 4P7A_Avastin

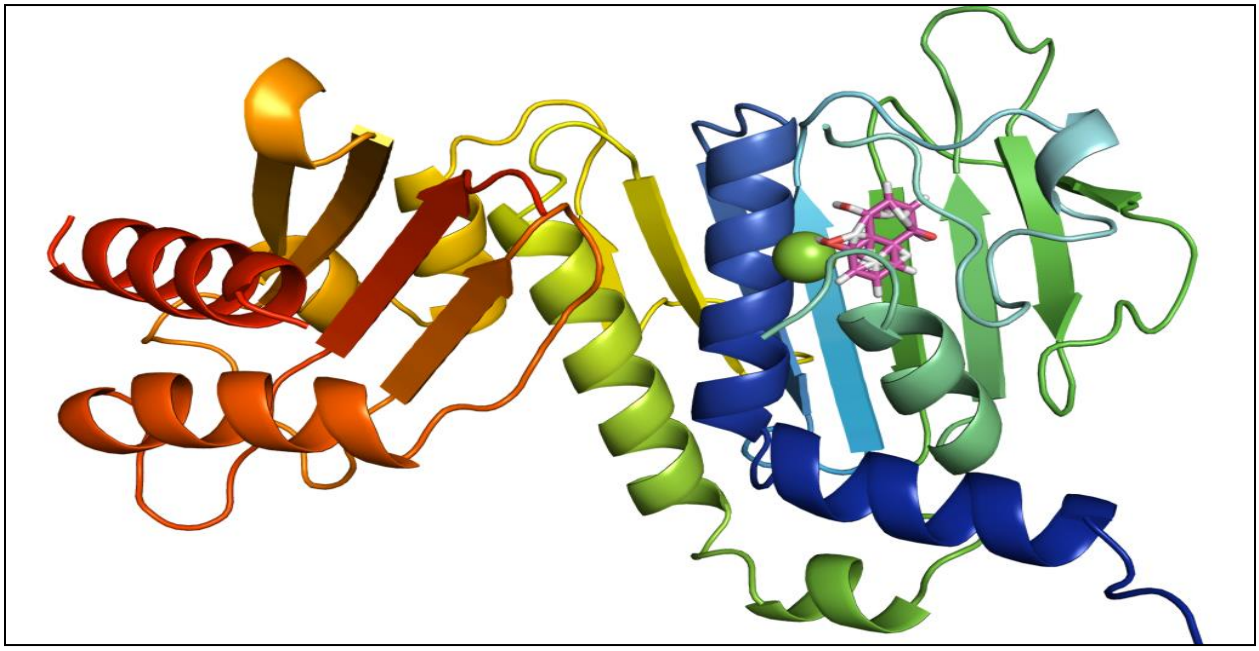


Figure 8. Docking result for 4P7A_AVASTIN

Solution No	Score	Area	ACE
1	2966	321.60	-95.49
2	2876	324.60	-95.05
3	2840	332.90	-72.83
4	2744	318.50	-68.50
5	2730	321.80	-63.98
6	2700	333.20	-53.74
7	2688	297.60	45.68
8	2678	318.80	-75.82
9	2652	298.20	48.88
10	2616	287.30	-118.72

Table 5

Above table contains top 10 results for the docking of 4P7A with ligand AVASTIN. Solution 1 with highest score 2966, area 321.60, ACE -95.49 is the optimal site for protein-small molecule docking complex.

3. 4P7A_Bevacizumab

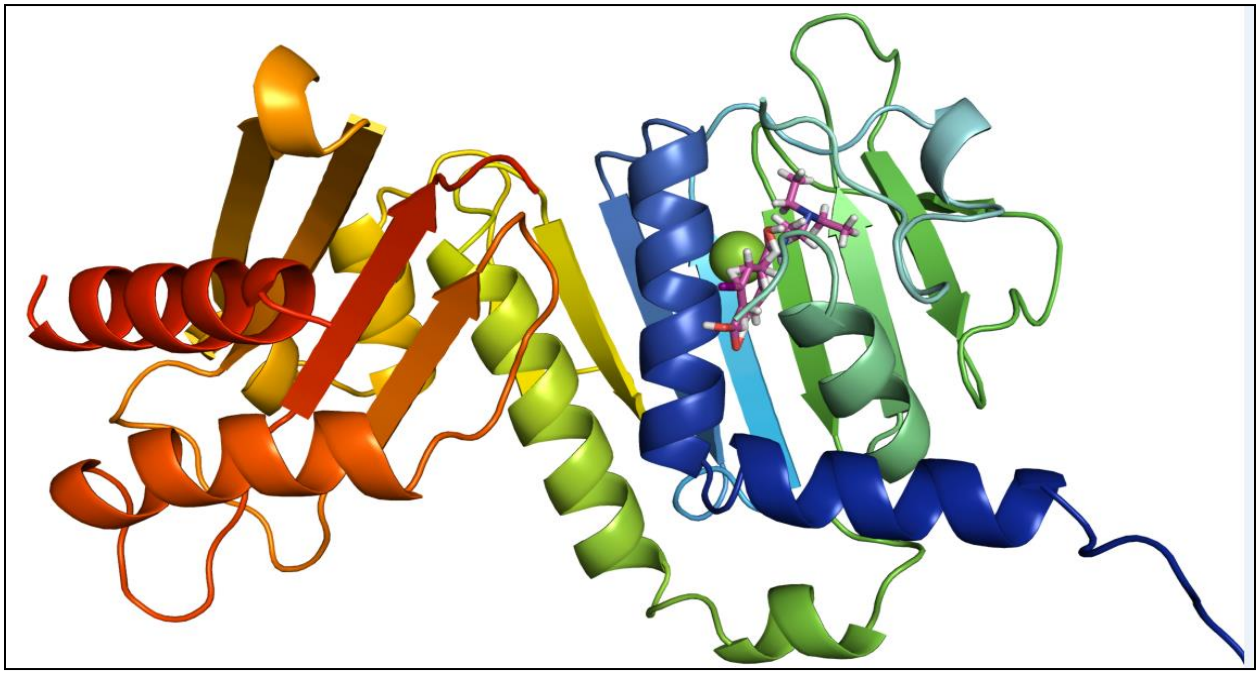


Figure 9. 4P7A_BEVACIZUMAB Docking result

Solution No	Score	Area	ACE
1	4558	516.50	-133.89
2	4456	528.40	-117.52
3	4442	516.20	-113.87
4	4418	543.10	-155.48
5	4398	516.80	-147.19
6	4352	481.60	-91.08
7	4320	490.30	-98.81
8	4294	502.80	-90.52
9	4254	474.40	-112.98
10	4236	520.70	-121.67

TABLE 6

Above table contains top 10 results for the docking of 4P7A with ligand BEVACIZUMAB. Solution 1 with highest score 4558, area 516.50, ACE -133.89 is the optimal site for protein-small molecule docking complex.

4. 4P7A_Campostar

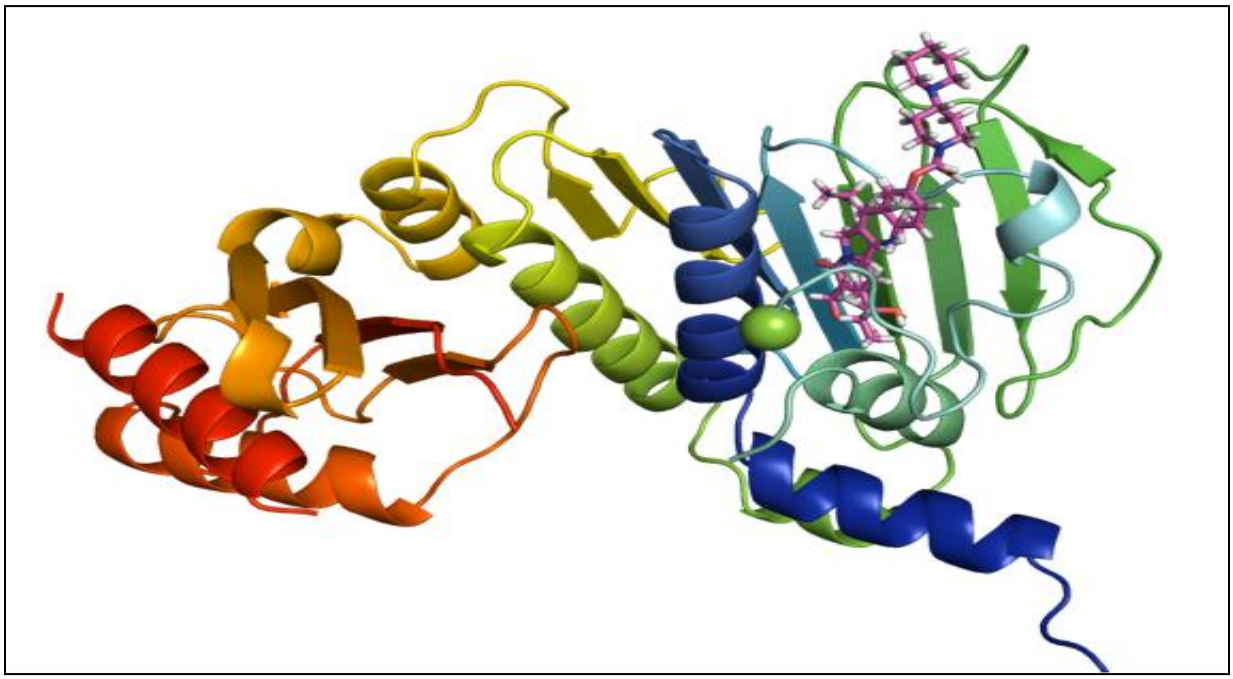


Figure 10. 4P7A_CAMPOSTAR Docking result

Solution No	Score	Area	ACE
1	6772	834.50	-259.27
2	6722	888.60	-253.83
3	6716	841.30	-255.16
4	6650	881.70	-241.85
5	6630	839.10	-246.29
6	6544	847.60	-184.12
7	6400	782.40	-168.98
8	6386	799.90	-135.43
9	6378	794.50	-133.54
10	6366	851.70	-178.43

TABLE 7

Above table contains top 10 results for the docking of 4P7A with ligand CAMPOSTAR. Solution 1 with highest score 6772, area 834.50, ACE -259.27 is the optimal site for protein-small molecule docking complex.

5. 4P7A_Capecitabine

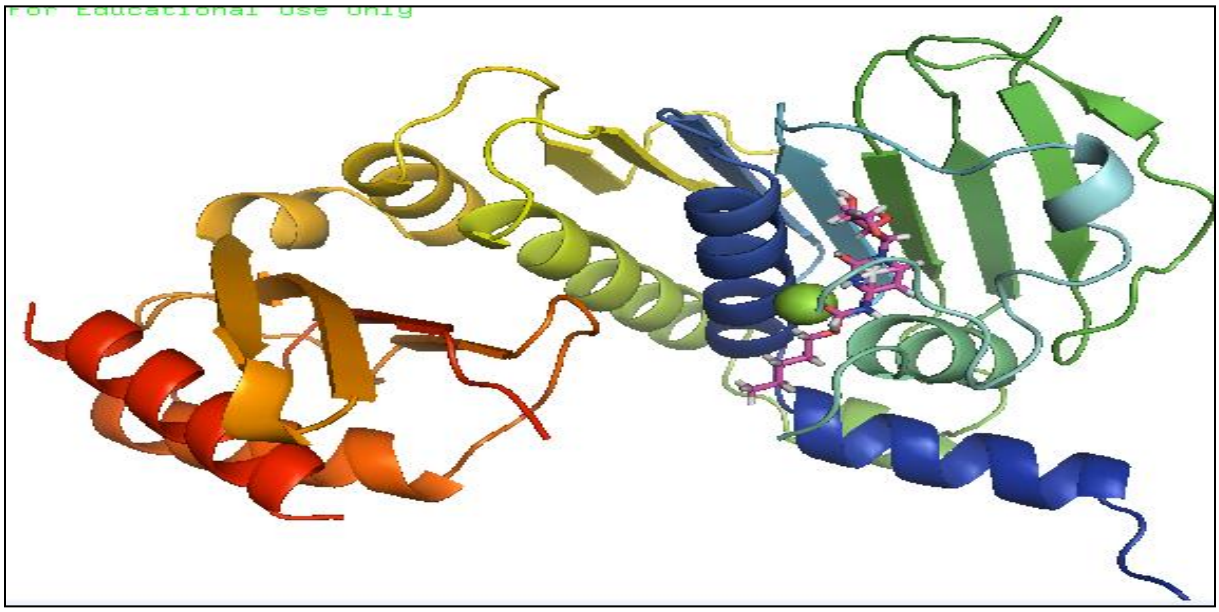


Figure 11. 4P7A_CAPECITABINE Docking result

Solution No	Score	Area	ACE
1	5082	572.60	-64.53
2	5036	560.10	-121.59
3	4890	560.90	-68.33
4	4850	573.80	-48.51
5	4850	586.40	-85.90
6	4844	579.60	-133.49
7	4832	567.80	-68.27
8	4816	559.30	-85.15
9	4800	572.70	-133.16
10	4766	557.30	-56.94

TABLE 8

Above table contains top 10 results for the docking of 4P7A with ligand CAPECITABINE. Solution 1 with highest score 5082, area 572.60, ACE -64.53 is the optimal site for protein-small molecule docking complex.

6. 4P7A_Cetuximab

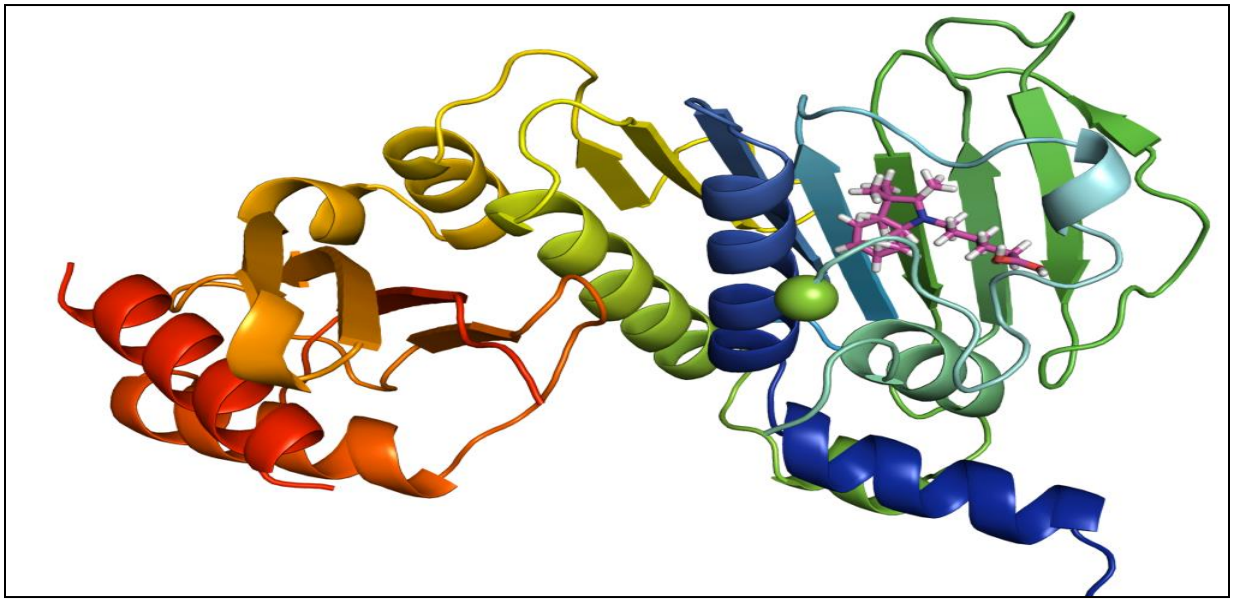


Figure 12. 4P7A_CETUXIMAB Docking result

Solution No	Score	Area	ACE
1	4494	479.50	-141.62
2	4388	511.40	-179.83
3	4348	479.20	-172.31
4	4336	534.50	-112.12
5	4308	499.80	-28.75
6	4260	526.10	-126.04
7	4228	489.40	-37.69
8	4224	516.30	-199.75
9	4218	500.10	-132.31
10	4218	515.30	-113.97

TABLE 9

Above table contains top 10 results for the docking of 4P7A with ligand CETUXIMAB. Solution 1 with highest score 4494, area 479.50, ACE -141.62 is the optimal site for protein-small molecule docking complex.

7.4P7A_ Eloxatin

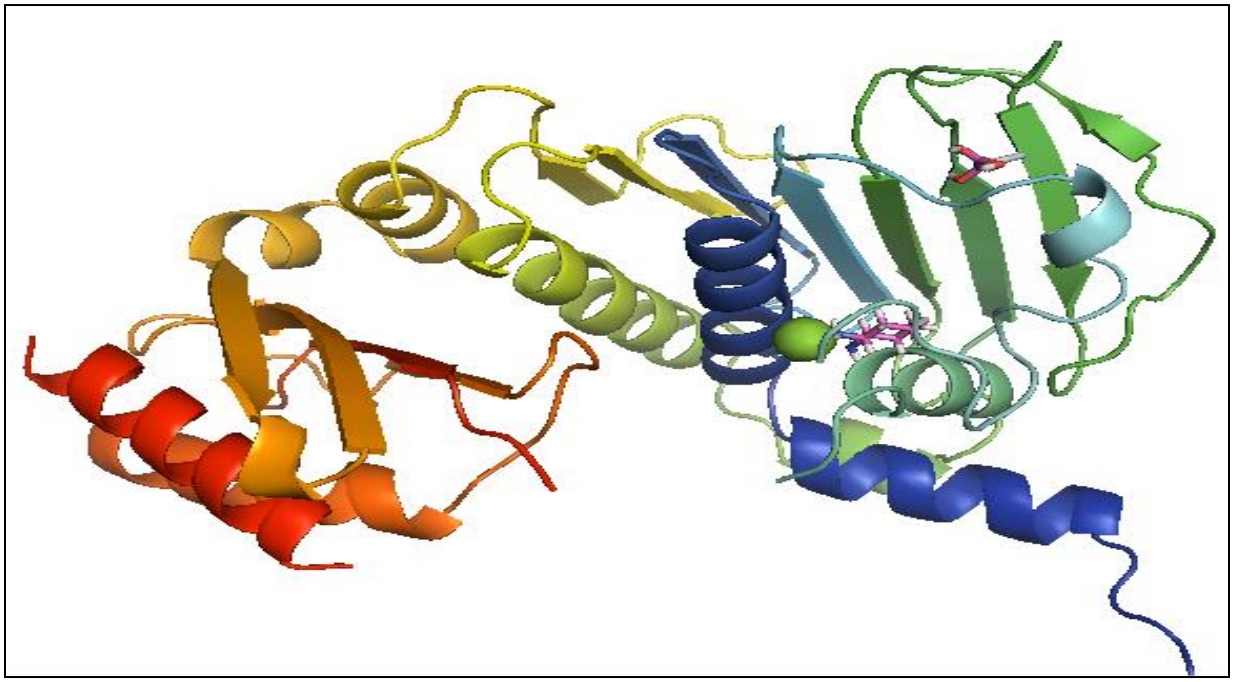


Figure 13. 4P7A_ELOXATIN Docking result

Solution No	Score	Area	ACE
1	3972	460.90	-61.40
2	3536	446.20	-63.56
3	3468	394.00	-45.41
4	3428	453.60	-47.29
5	3412	397.50	-117.01
6	3358	386.50	-49.44
7	3328	385.00	-64.94
8	3240	428.00	-70.54
9	3236	395.30	-45.55
10	3218	399.70	-67.46

TABLE 10

Above table contains top 10 results for the docking of 4P7A with ligand ELOXATIN. Solution 1 with highest score 3972, area 460.90, ACE -61.40 is the optimal site for protein-small molecule docking complex.

8. 4P7A_Folinic Acid

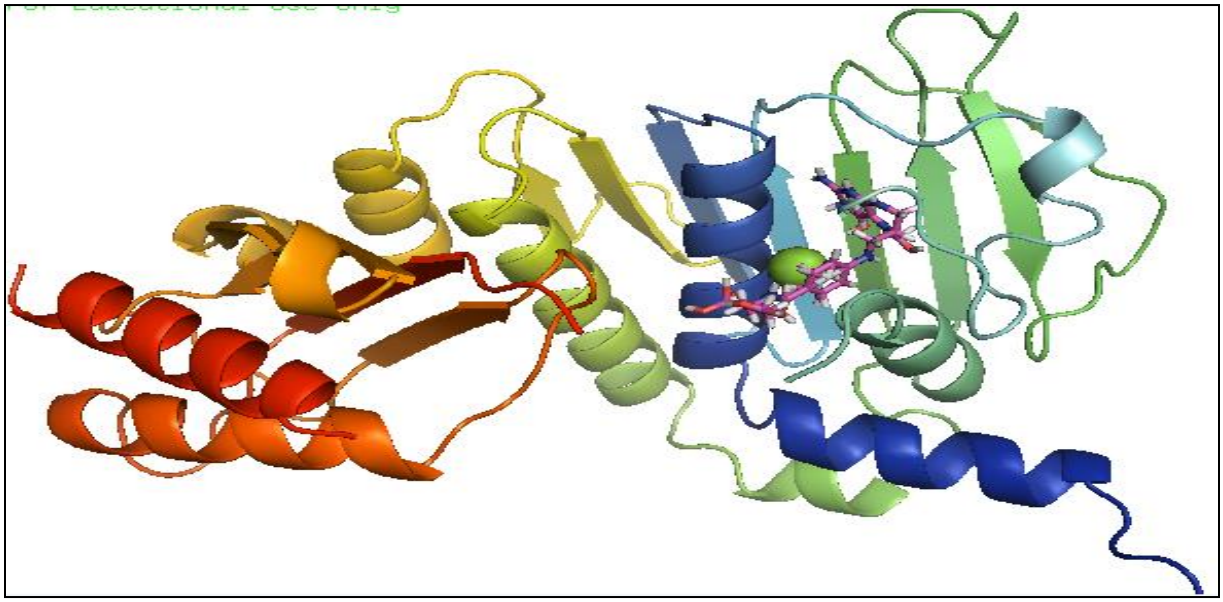


Figure 14. 4P7A_FOLINIC ACID Docking result

Solution No	Score	Area	ACE
1	5910	656.90	-102.05
2	5708	667.30	-140.72
3	5676	714.60	-234.12
4	5638	672.60	-188.34
5	5576	639.60	-149.87
6	5458	632.50	-93.10
7	5432	617.70	-135.89
8	5432	634.20	-95.07
9	5420	657.60	-35.27
10	5416	653.50	-182.04

TABLE 11

Above table contains top 10 results for the docking of 4P7A with ligand FOLINIC ACID. Solution 1 with highest score 5910, area 656.90, ACE -102.05 is the optimal site for protein-small molecule docking complex.

9. 4P7A_Leucovorin Calcium

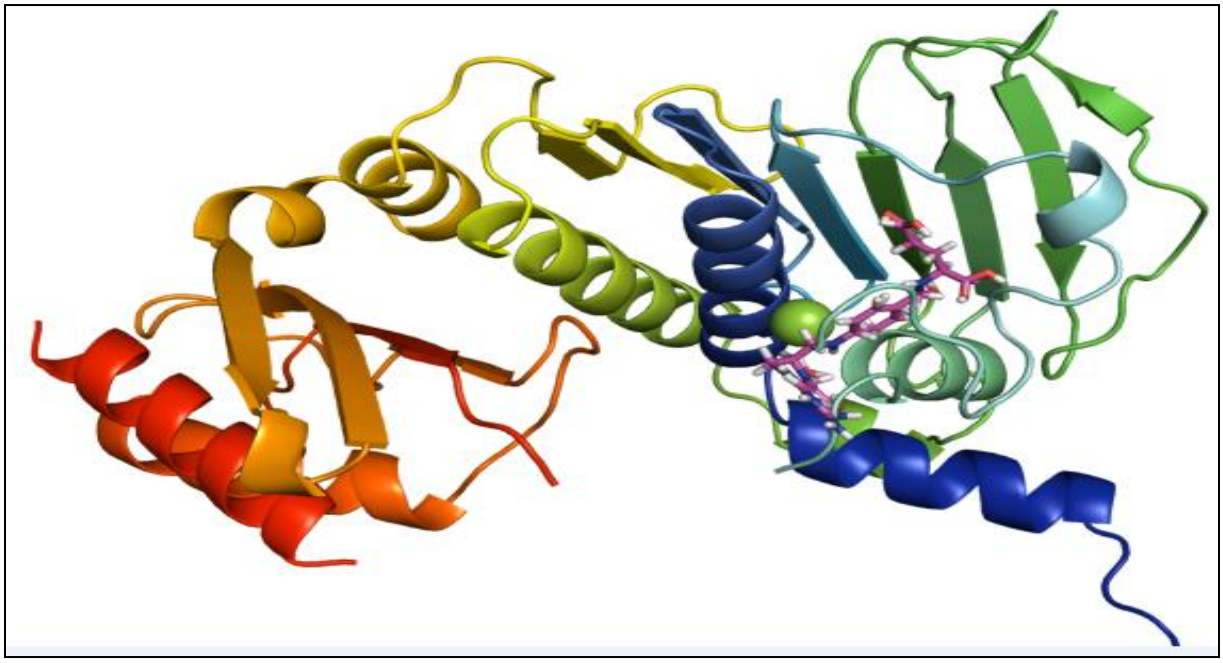


Figure 15. 4P7A_LEUCOVORIN CALCIUM Docking result

Solution No	Score	Area	ACE
1	6104	792.20	-285.62
2	5936	781.50	-184.87
3	5870	725.20	-107.74
4	5832	672.00	-106.43
5	5820	701.20	-212.19
6	5796	716.20	-97.49
7	5792	726.70	-159.10
8	5780	665.80	-82.45
9	5768	709.40	-9.92
10	5764	708.40	-208.67

TABLE 12

Above table contains top 10 results for the docking of 4P7A with ligand LEUCOVORIN CALCIUM. Solution 1 with highest score 6104, area 792.20, ACE -285.62 is the optimal site for protein-small molecule docking complex.

10. 4P7A_Stivarga

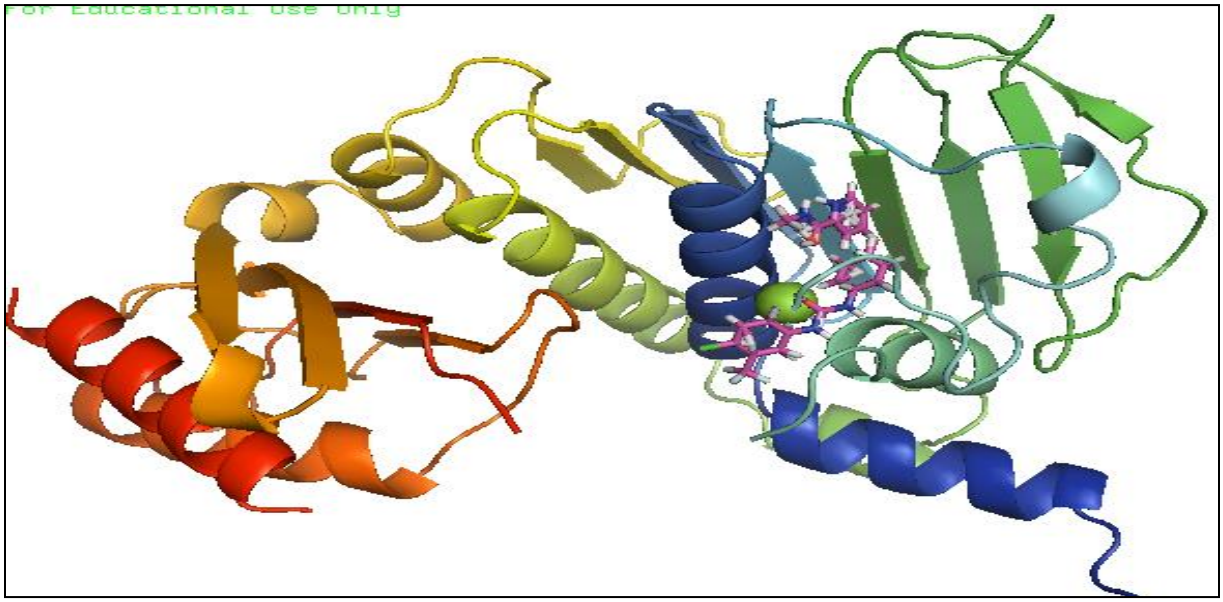


Figure 16. 4P7A_STIVARGA Docking result

Solution No	Score	Area	ACE
1	6296	731.50	-28.45
2	6092	731.00	-76.23
3	6080	714.10	-72.87
4	6076	710.50	-85.77
5	5914	707.70	-85.01
6	5870	702.90	-47.24
7	5844	705.80	-79.29
8	5786	710.10	-82.19
9	5756	686.30	-50.73
10	5724	738.80	-105.05

TABLE 13

Above table contains top 10 results for the docking of 4P7A with ligand STIVARGA. Solution 1 with highest score 6296, area 731.50, ACE -28.45 is the optimal site for protein-small molecule docking complex.

11. 4P7A_Welcovorine

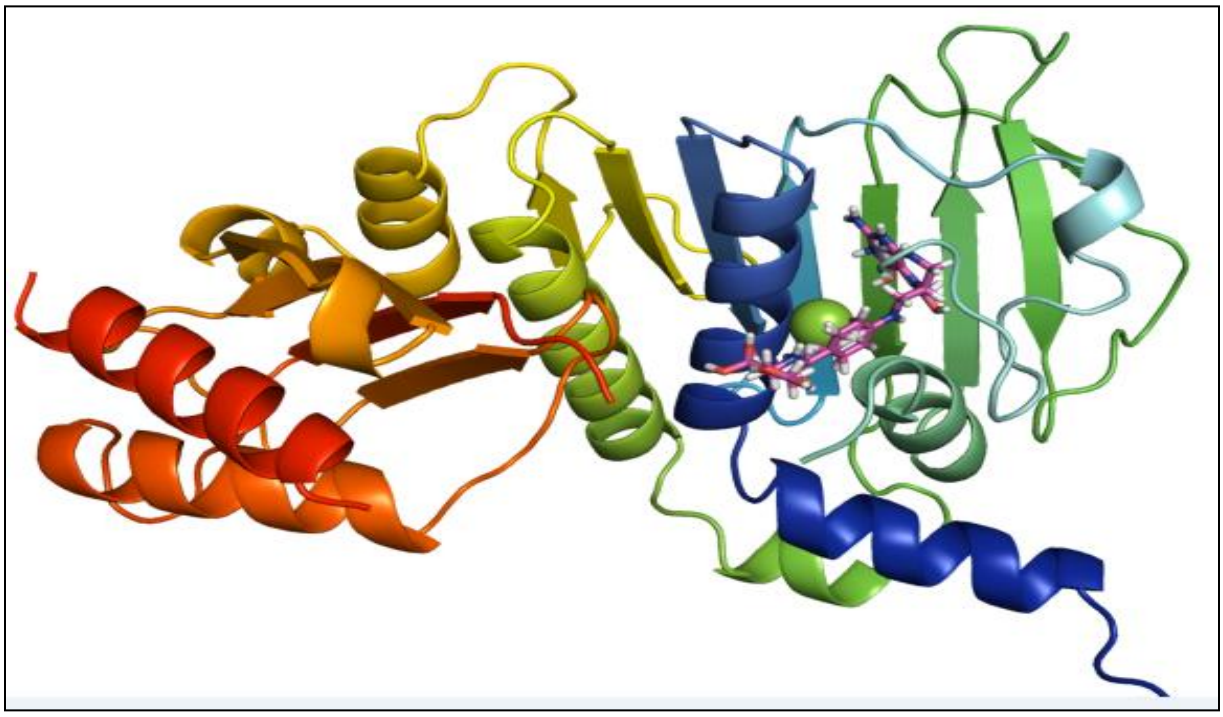


Figure 17. 4P7A_WELCOVORINE Docking result

Solution No	Score	Area	ACE
1	5910	656.90	-102.05
2	5708	667.30	-140.72
3	5676	714.60	-234.12
4	5638	672.60	-188.34
5	5576	639.60	-149.87
6	5458	632.50	-93.10
7	5432	617.70	-135.89
8	5432	634.20	-95.07
9	5420	657.60	-35.27
10	5416	653.50	-182.04

TABLE 14

Above table contains top 10 results for the docking of 4P7A with ligand WELCOVORINE. Solution 1 with highest score 5910, area 656.90, ACE -102.05 is the optimal site for protein-small molecule docking complex.

DISCUSSION:

The output of PatchDock is a big list of candidate complexes between user specified receptor and ligand molecule. The list is in the form of a table comprised of various factors: Solution No., Score, Area, ACE.

Solution with most highest value for score is considered to be the best for docking position. According to our results 4P7A-CAMPOSTAR (6772), 4P7A_STIVARGA (6296), 4P7A_LEUCOVORIN CALCIUM (6104), 4P7A_WELCOVORIN (5910), FOLINIC ACID (5910) are the top best protein-small molecule docking complexes with best scores.

- **Damaged SNPs Results:**

We used 10 different damage prediction tools for prediction of damaged CRC associated SNPs, tools are listed as below:

- **PROVEAN PREDICTION**
- **PANTHER PREDICTION**
- **SNP & GO**
- **MUTATION ACCESSOR**
- **SIFT**
- **POLYPHEN 2**
- **PHD SNP**
- **MutPred**
- **PREDICT SNP**
- **SNAP 2**

Out of total 457 CRC associated SNPs we found total of 24 highly damaged SNPs, and on the basis of these 24 highly damaged nucleotide polymorphisms we can further proceed for mutagenesis (a process used to change the genetic information of an organism, that results in mutation). Criterion for the selection of all SNPs were based on their damaging scores provided by all the prediction tools.

CHAPTER 4: CONCLUSION

In the present study, systematic processes of comparative analysis, subtractive genomic approaches were defined for the identification of novel therapeutic drug targets We compared prediction results of all prediction tools to find out the probability of SNP affecting protein function and be related to a disease. Prediction tools use machine-learning based methods to predict if variants affect functions and lead to related diseases. To identify SNPs to be more 'probably damaged' we combined the results of all 10 prediction tools where SNPs were classifying from most neutral (no damage) to most deleterious (most damaged) [11]. In POLYPHEN 2, MutPred, and MUTATION ACCESSOR high scores relates towards damaged mutations. Whereas in PROVEAN, SIFT and PANTHER low scores relates to damaged SNPs. The non-synonymous polymorphisms found in MLH1 gene were extracted by 10 program tools that use different methods for prediction of SNP. The differences generated in predictions indicates the need for the combined analysis to accurately identify SNPs that are damaging to the MLH1 gene.

CHAPTER 5: REFERENCES

1. Munikumar, Manne. "In Silico Identification Of Common Putative Drug Targets Among The Pathogens Of Bacterial Meningitis". *Biochemistry & Analytical Biochemistry* 01.08 (2012): n. pag. Web. 21 Dec. 2016.
2. Shukla A, Moussa A, Singh TR. DREMECELS: A Curated Database for Base Excision and Mismatch Repair Mechanisms Associated Human Malignancies. PloS one. 2016 Jun 8;11(6):e0157031.
3. Niessen, R. C. et al. "Hereditary Non-Polyposis Colorectal Cancer: Identification Of Mutation Carriers And Assessing Pathogenicity Of Mutations". *Scandinavian Journal of Gastroenterology* 39.241 (2004): 70-77. Web. 21 Dec. 2016.
4. Bank, RCSB. "RCSB Protein Data Bank - RCSB PDB". *Rcsb.org*. N.p., 2016. Web. 21 Dec. 2016.
5. "The Pubchem Project". *Pubchem.ncbi.nlm.nih.gov*. N.p., 2016. Web. 21 Dec. 2016.
6. "Patchdock Server: An Automatic Server For Molecular Docking". *Bioinfo3d.cs.tau.ac.il*. N.p., 2016. Web. 21 Dec. 2016.
7. "Search Castp Database". *Sts.bioe.uic.edu*. N.p., 2016. Web. 21 Dec. 2016.
8. Choi Y., and A. P Chan. "Provean Web Server: A Tool to Predict the Functional Effect of Amino Acid Substitutions and Indels." [In eng]. Bioinformatics 31, no. 16 (Aug 15 2015): 2745-7.
9. Sim, N. L., P. Kumar, J. Hu, S. Henikoff, G. Schneider, and P. C. Ng. "Sift Web Server: Predicting Effects of Amino Acid Substitutions on Proteins." [In eng]. *Nucleic Acids Res* 40, no. Web Server issue (Jul 2012): W452-7.
10. Mi, H., X. Huang, A. Muruganujan, H. Tang, C. Mills, D. Kang, and P. D. Thomas. "Panther Version 11: Expanded Annotation Data from Gene Ontology and Reactome Pathways, and Data Analysis Tool Enhancements." [In eng]. *Nucleic Acids Res* 45, no. D1 (Jan 04 2017): D183-d89.

11. Choi Y., G. E Sims, S. Murphy, J. R. Miller, and A. P. Chan. "Predicting the Functional Effect of Amino Acid Substitutions and Indels." [In eng]. PLoS One 7, no. 10 (2012): e46688.

12. Shukla A, Sehgal M, Singh TR. Hydroxymethylation and its potential implication in DNA repair system: A review and future perspectives. Gene. 2015 Jun 15;564(2):109-18.

13. Sehgal M, Gupta R, Moussa A, Singh TR. An integrative approach for mapping differentially expressed genes and network components using novel parameters to elucidate key regulatory genes in colorectal cancer. PloS one. 2015 Jul 29;10(7):e0133901.