International Journal of Current Engineering and Technology
©2017 INPRESSCO®, All Rights Reserved

E-ISSN 2277 – 4106, P-ISSN 2347 – 5161
Available at http://inpressco.com/category/ijcet

Research Article

Protein Active Site Structure Prediction Strategy and Algorithm

Ayan Chatterjee[†*], Uttam Kumar Roy[†] and Dinesh Halder[†]

[†]Department of IT, Jadavpur University, Kolkata, India
[†]Department of Botany & BioInformatics, Kalyani University, Nadia, India

Accepted 25 May 2017, Available online 02 June 2017, **Vol.7, No.3 (June 2017)**

Abstract

Computational Drug Design is important as it reduces conventional research deadline and cost. Proper Drug Design is a challenge till date for complex diseases in reduced time and cost. Few days back we were completely dependent on wet chemistry lab but now with the help of Bio-Informatics, a perfect blending of computer science & biological science with chemistry, we are able to do complex drug design analysis in dry lab. Different structure prediction tools are discovered with Bio-Informatics to determine 3D structure of protein that helps in classification, prediction of functions for uncharacterized proteins, interaction with other macromolecules, interactions with small ligands like metal ions, nucleotides, substrates, cofactors and inhibitors, enzyme mechanism, structure based drug development, understanding of different bonding, predicting active site for targeted therapy like cancer, experimental evidence for transmembrane domains. In this paper, a new algorithm has been proposed that determines the active site of a protein and its implementation has been shown in Java Programming. Once the geometry of the active site is obtained, its corresponding drug structure is predicted.

Keywords: *Polar, Hydrophobic, Hydrophilic, Residue, Gene, Java3D, JFreeChart, Java, Functional Group, Protein, Bonding, Active Site, Ligand etc.*

1. Introduction

Computer Aided Drug Design is taking focus in latest Bio-Informatics research as it reduces deadline and cost. Here we have considered structure based drug design identifies a protein target (active site) and going ahead we have to find a molecule or ligand that nicely fits there. The main aim at drug designing is to develop or search a stable drug that cures disease with minimal cost and time and produces minimum interaction energy. Different docking strategies are taking place to make it more efficient with evolutionary algorithms. The problems can be classified as optimization problems with NP Completeness. Here, we have used Java language to find out the probable orientation or geometry of an active site to produce a stable drug. Active sites are some area on the surface of a protein molecule and that area becomes active when it comes in contact with an appropriate drug. As a result, the normal function of the protein is hampered.

In this paper, we have used 5a35.pdb and developed algorithm to predict protein active site. At the output, we have prepared the geometry of active site both in 2D & 3D in reduced complexity as we have used binary tree algorithm (Goh G *et al*, 2000; Luis Rueda *et al*; Piyali Chatterjee *et al*, 2007).

2. Bacteria Protein and Ligand Structure

We have used www.pdb.org repository to find our required protein structure file or pdb.

Name: 5A35.pdb

Protein Details: Crystal structure of Glycine Cleavage Protein H-Like (GcvH-L) from Streptococcus pyogenes. Source: This is a family of glycine cleavage H-proteins, part of the glycine cleavage multienzyme complex (GCV) found in bacteria and the mitochondria of eukaryotes.

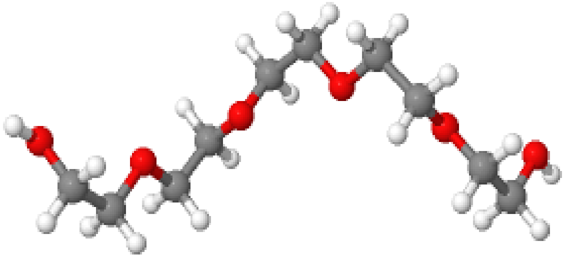


Fig. 1 1PE.pdb structure in 3D

Streptococcus pyogenes, a Group A gram positive bacteria that shows streptococcal Group A antigen on

*Corresponding author: **Ayan Chatterjee**
DOI: http://Dx.Doi.Org/10.14741/Ijcet/22774106/7.3.2017.53

its cell wall and responsible for pharyngitis, cellulitis, tonsillitis, rheumatic fever, scarlet fever etc. Based on the statistics, 700 million infections occur worldwide each year with a mortality rate of 25 Percent. The Ligand Structure (1PE.pdb or PENTAETHYLENE GLYCOL) as shown in fig.1

2.1 Active Site

They are basically hydrophobic pockets inside of a protein or enzyme that involves side chain atoms. They are just like a pocket where ligands can bind and do reaction in order to cure disease. Protein 3D structure prediction is important in order to predict active site. An active site consists of both binding site and catalytic side (to enhance the reaction).

Residues in the binding site form hydrogen bonds or hydrophobic interactions, or temporary covalent interaction with the substrate in order to form protein-substrate complex. Once the binding is done and oriented in the active site, catalysis can begin. The residues of the catalytic site are naturally very close to the binding site and some residues can also have dual-roles in both binding and catalysis. [Wikipedia]

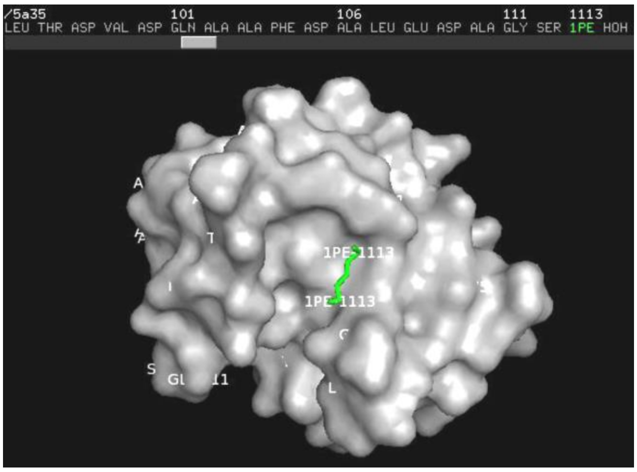
2.2 Active Site Determination with MOE Site Finder

Active sites in protein can be obtained by calculating minimal interaction energy in between receptor and ligands or probes which requires assignment of proton locations and partial charges to the receptors that is not always easy & that is the reason, we have discussed Geometry Method here to locate active site inside a protein:

a. Identify the regions where atoms are tightly packed.
b. Remove those sites that are much more exposed to solvent.
c. Use hydrophobic or hydrophilic classifications
d. Don't use grid based methods as they are not invariant to rotation of the atomic coordinates and can consume large amounts of memory.

2.3 Active Site Determination with PyMol

- Open PyMol
- Do Upload PDB of 5a35.pdb
- Hide all the objects loaded into PyMol by using the command hide. (Hide → Everything)
- Represent entire protein with surface representation, setting with a 50% transparency. (Show Surface)
- set transparency 0.5 (Color → grays → gray50)
- Upload ligand & show as sticks (select ligand, resn 1PE)
- disable the ligand by clicking on the ligand object, pink dots disappear under



```
/5a35                    101                      106                      111       1113
LEU THR ASP VAL ASP GLN ALA ALA PHE ASP ALA LEU GLU ASP ALA GLY SER 1PE HOH
```

Fig.2 5a35.pdb structure Active site

2.4 Active Site Determination with Online Research Portals or Tools

Case#1

We have used IIT Delhi's active site prediction tool http://www.scfbio-iitd.res.in/dock/ActiveSite.jsp and found list of active sites: http://www.scfbio-iitd.res.in/dock/69978516ACTIVE/list with details for 5a35.pdb: http:// www.scfbio-iitd. res.in/dock/ 69978516ACTIVE/

Case#2

We have used http://zhanglab.ccmb.med.umich.edu, a Michigan University Lab portal for Department of Computational Medicine and Bioinformatics to determine binding site residues for *.pdb and below result as follows for 5a35.pdb:

http://zhanglab.ccmb.med.umich.edu/COACH/output/CH036020/
The drug:
http://zhanglab.ccmb.med.umich.edu/BioLiP/sym.cgi?code=OSS
The detail drug information:
http://zhanglab.ccmb.med.umich.edu/BioLiP/getaid.cgi?id=0021298.

3. Proposed Algorithm & Implementation

3.1 Algorithm to determine protein structure from PDB and structure analysis

Step#1: Reading protein co-ordinate file in pdb format & parse it
Step#2: Separate out Protein Chain
Step#3: For every chain do

- Separate out Amino Acids & put them in AVL Tree Structure with all ATOM Information
- Classify amino acids based on color code

- Generate FASTA sequence & Amino Acid sequence
- Generate Nucleotide sequence
- Use ATOM Co-ordinates to generate 3D Architecture
- Use ATOM co-ordinates with HELIX & SHEET information to generate view of secondary structure atoms in 3-dimentional space.
- Calculate phi & psi angle in between connective amino acids for checking bond Rotation (based on the principle of Ramachandran Plot)
- Determine bonds, angles
- Continue for all chains
- Determine the best active site structure from given list obtained from site prediction tool

Step#4: Determine Active sites and for each active site generate a 3D & 2D structure
Step#5: Determine 3D structure of Ligand with atomic coordinates
Step#6: Ends

N.B: Ramachandran plot graphically represents allowed bond angles. Not all Phi & Psi bonds are rotatable due to Steric hindrance of R (side chain) of Alpha Carbon. So, without Ramachandran Plot it would be difficult to obtain graphically which bond angles are allowable and it also help in classifying most allowed region (GLY) and least allowed region (PRO).

Omega (C-N : 180 degree and non rotatable due to partial double bond character), Phi (N - Alpha C : Rotatable) : freedom of rotation, Psi (Alpha C – C : Rotatable): freedom of rotation

Torsion Angle Calculation with Java for Phi & Psi:

```
double torsionAngle(Atom a, Atom b, Atom c, Atom d) {
Atom ab = subtract(a,b); Atom cb = subtract(c,b);
Atom bc = subtract(b,c);
Atom dc = subtract(d,c);
Atom abc = vectorProduct(ab,cb);
Atom bcd = vectorProduct(bc,dc);
double angl = angle(abc,bcd) ;
    /* calc the sign */
    Atom vecprod = vectorProduct(abc,bcd);
    double val = scalarProduct(cb,vecprod);
    if (val<0.0) angl = -angl ;
return angl;
}
```

3.2 Active site determination with Java programming

Based on our Algorithm, the best active site structure of 5a35.pdb.

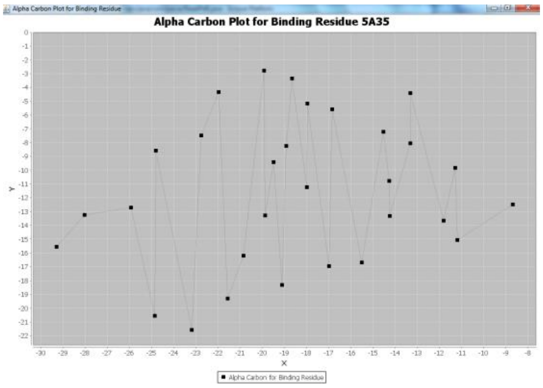


Fig. 3 5a35.pdb structure Active site in 2D in JFreeChart

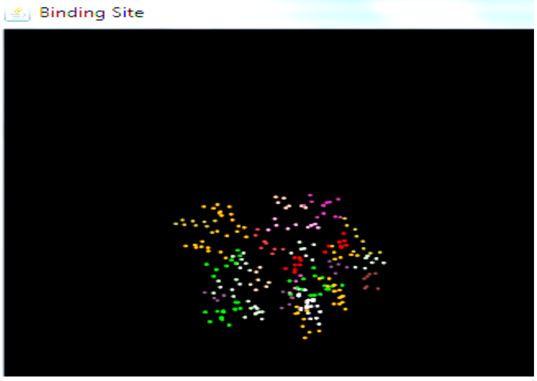


Fig.4 5a35.pdb structure Active site in 3D

Case#2: Based on our Algorithm, the best active site structure of 5a35.pdb

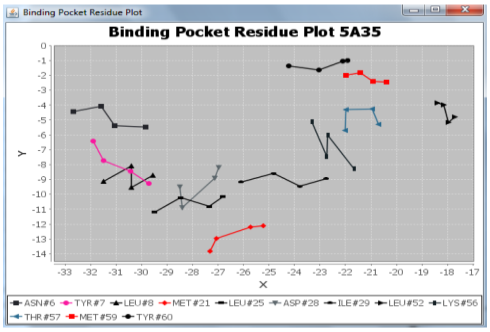


Fig. 5 5a35.pdb structure Active site amino acid in 2D

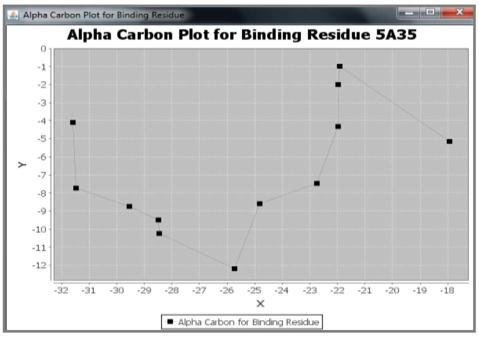


Fig.6 5a35.pdb structure Active site alpha carbon in 2D

The co-ordinate value f(X, Y) of CA/Alpha Carbon:

-31.584,-4.099 | -31.485,-7.72 | -29.547,-8.727 | -25.726,-12.198 | -28.462,-10.22 | -28.49,-9.472 | -24.808,-8.597 | -17.948,-5.15 | -22.741,-7.462 | -21.977,-4.321 | -21.972,-1.988 | -21.909,-0.988

In 3D, binding site with amino acid color code plot, based on our designed Algorithm:

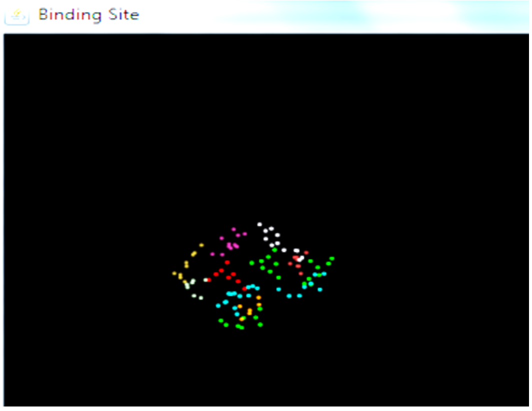


Fig.7 5a35.pdb structure Active site in 3D

3.3 Docking Energy

When a drug interacts with active site of the protein then energy gets produced. Lower is the energy, the stable the drug is and stable drug will be good water soluble. In our study, we have focused on two main forces in order to calculate interaction energy as shown below (Van Der waal & Electrostatic):

$$V_{non-bond} = \sum_{non-bond}\left[\varepsilon_{i,j}\left[\left(\frac{R_{min,i,j}}{r}\right)^{12} - 2\left(\frac{R_{min,i,j}}{r}\right)^{6}\right] + \frac{q_i q_j}{4\pi\varepsilon r}\right]$$

N.B: where ε,i,j is the Lennard-Jones well depth, r is the distance between atoms i and j, Rmin, i,j is the minimum interaction radius & qi and qj are the charges of the two atoms, r is the separation, and ε is the dielectric constant of the surrounding medium.

But in general total energy calculation can be expressed:

Energy = Stretching Energy + Bending Energy + Torsion Energy + Non-Bonded Interaction Energy.

3.4 Experiment and Result

We have used the active site structure as obtained in case#2 as it gives close match to real active site of 5a35.pdb as found in JMol after integrating to our code implementation:

In JMOL, the actual docking diagram & drug (1PE) is at the binding site:

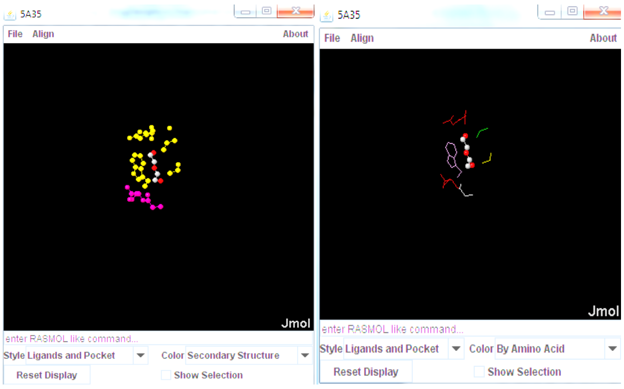


Fig. 8 5a35.pdb structure Active site docking in 3D

In 2D diagram, we have found active site that resembles the active site as shown below with 1PE ligand binding and populated by RASMOL using X-RAY Diffraction:

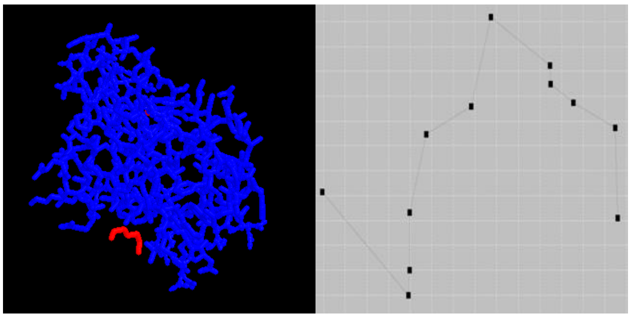


Fig.9 Docking structure comparison

Probable docking structure as shown below:

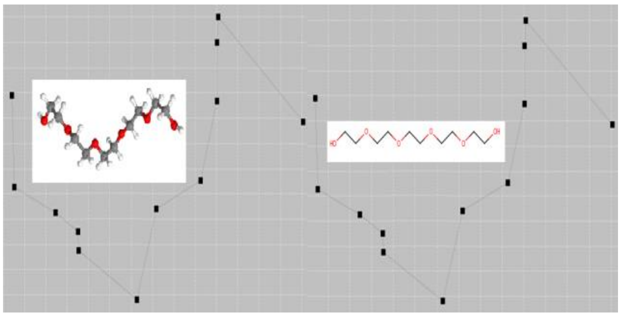


Fig. 10 Ligand fit in active site

Based on our algorithm, the structure of the residue of Rank#1 in Case#2:

The binding site has following residue numbers: 6, 7, 8, 21, 25, 28, 29, 52, 56, 57, 59, and 60.

Input: the chain name as String & residue numbers as an array

Output: get active site structure in case#2 in O (n*log n) time instead of traditional brute force O (n*n) as we have used BST search (O (log n)) and residue arrangement (from pdb) in an AVL tree.

The corresponding amino acid series: ASN TYR LEU MET LEU ASP ILE LEU LYS THR MET SER

The Protein Structure Analysis Report:

Total Chain from ATOM:[A]

Total Chain from SEQRES:[A]

The chain: A --> MET LYS LYS ILE ALA ASN TYR LEU LEU ILE GLU LYS THR ASP ASP ARG TYR THR ILE SER MET THR PRO GLU LEU GLN ASP ASP ILE GLY THR ILE GLY TYR ALA GLU PHE THR ASP ASN ASP HIS LEU ALA VAL ASP ASP ILE ILE LEU ASN LEU GLU ALA SER LYS THR VAL MET SER VAL LEU SER PRO LEU ALA GLY ALA VAL VAL GLU ARG ASN GLU ALA ALA THR LEU THR PRO THR LEU LEU ASN SER GLU LYS ALA GLU GLU ASN TRP ILE VAL VAL LEU THR ASP VAL ASP GLN ALA ALA PHE ASP ALA LEU GLU ASP ALA GLY SER

The FASTA Sequence:

```
>5A35:A|PDBID|CHAIN|SEQUENCE
MKKIANYLLIEKTDDRYTISMTPELQDDIGTIGYAEFTDN
DHLAVDDIILNLEASKTVMSVLSPLAGAVVERNEAATLTP
TLLNSEKAEENWIVVLTDVDQAAFDALEDAGSGHHHHH
HH
```

No of Atoms in Alpha Helix: 70

AminoAcids in Alpha Helix: [THR, GLY, GLU, THR, PRO, SER, LYS, ASN, ASP, ALA]

No of Atoms in BetaSheet: 319

AminoAcids in BetaSheet: [HIS, LEU, GLY, ARG, GLY, LEU, TRP, THR, VAL, VAL, ARG, MET, TYR, LEU, LEU, THR, LEU, SER, LYS, ALA, LYS, ILE, GLU, ASP, GLU, ILE, ILE, PHE, ILE, ALA, ASN, GLU, VAL, LEU, MET, LEU]

We have obtained the analysis in reduced time due to B-Tree algorithm incorporation.

Conclusions

It has been shown how we have used different online research tools to find active site of a protein and thereby displaying and operating on it in reduced complexity and efficiently with java based implementation.

CADD methods are dependent on bioinformatics tools, knowledge of biology, domain knowledge and DB. Here we have considered structure based drug design where we have identified a protein target and we have to find a molecule or ligand that suitably fits there. The main aim of drug design is to develop or search a stable drug that cures disease with minimal cost and time.

In the upcoming papers, we will be showing how to fit drug in the determined active site in more efficient way and interaction energy optimization with Evolutionary algorithms. So, this study is the footstep of the next phase to design efficient drug with necessary optimization.

References

Filho JLR, Treleavan PC (1994), Genetic Algorithm Programming Environments, IEEE, Comput, pp 28-43.

Goh G, Foster JA (2000), Evolving Molecules for Drug Design Using Genetic Algorithm, Proc. Int. Conf. on Genetic & Evol. Computing, Morgan Kaufmann, pp 27 – 33

J.Kennedy (1999), Small worlds and mega-minds: effects of neighborhood topology on particle swarm performance. Proc. of IEEE Congress on Evolutionary Computation (CEC 1999), Piscataway, NJ. pp. 1931-1938

Luis Rueda, Sridip Banerjee, Md. Mominul Aziz, Mohammad Raza, Protein-protein. Interaction Prediction using Desolvation Energies and Interface Properties, Proceedings of IEEE International Conference on Bioinformatics and Biomedicine, pp: 17-22

Piyali Chatterjee, Subhadip Basu, Mahantapas Kundu, Mita Nasipuri and Dariusz Plewczynski (2007), PPI_SVM: Prediction of Protein-protein Interactions using Machine Learning, Domain-domain Affinities and Frequency Tables, Cellular & Molecular Biology Letters, volume 16 (2011) pp 264-278

Jos´e A. Reyes and David Gilbert, Prediction of protein-protein interactions using one-class classification methods and integrating diverse biological data, Journal of Integrative Bioinformatics 4(3):77