THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

PRO-COMP-IDS TOOL (Protein's Component- Interaction Determing Server Tool)

Neeti Kulkarni
M.Tech 4th Sem (Bioinformatics), Department of Biotechnology
KLE DR M.S Sheshgiri College of Engineering and Technology, Udyambag, Belgaum, India
Rakesh N. R.
Assistant Professor, Department of Biotechnology
KLE DR M.S Sheshgiri College of Engineering and Technology, Udyambag, Belgaum, India

Abstract:
Many of the current protein active site prediction tools finally fail to develop an active Site in target. The advancement in structure elucidation of proteins will open up a route for structure based target assessment especially during in-silico drug design. The current developing tool is based on Gaussian function and statistical analysis of the components involved in the protein active site formation and which could be a beneficial data for identifying a potential drug target and to overcome the gaps of other prediction tools.

Key words: *DoG(Difference of Gaussian), pocket atom, JVM(Java Virtual Machine), protein interaction,PDB(Protein Data Bank)*

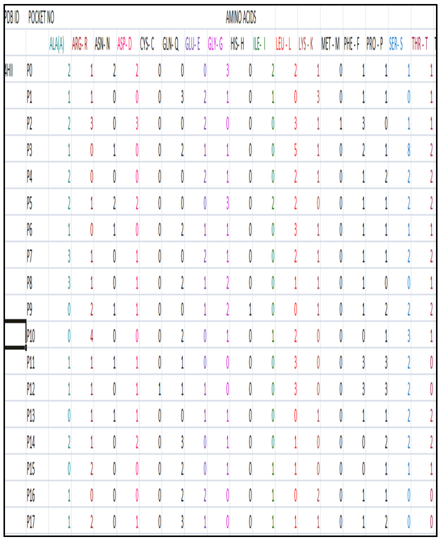
1. Introduction

The number of available bioinformatics tools has grown dramatically over the past two decades. However, as a result of increasing complexity of the tools and a lack of rigorous software engineering practices, using, installing and configuring a bioinformatics application, along with ensuring that it is communicating well with other programs, can be a challenging task. Increasing structural genomics projects have led to the exponential growth of the number of available protein structures. Rating the attractiveness of a drug target is one of the major challenges in the early stages of drug discovery. Besides attractivity assessment based on medical rationale and commercial viability, the properties of the target and its ability to be modulated by small drug-like compounds (further referred to as druggability) have to be analyzed. Due to the large amount of available crystal structures, the automatic collection of target information gains importance. In a first step, binding pockets have to be detected on the protein surface. Some methods fulfilling this task are available via web- services, e.g., QSite-Finder, CASTp, SCREEN, PocketDepth, MetaPocket and Fpocket . The next step on the path towards target classification or druggability prediction is the annotation and comparison of target specific pocket properties. Some servers exist that allow besides binding site prediction - for their analysis and functional classification, e.g., FINDSITE (Brylinski and Skolnick, 2008), SplitPocket (Tseng *et al*., 2009), fPOP (Tseng *et al*., 2010), ProBis (Konc and Janezic, 2010), and SiteComp (Lin *et al*., 2012). Many of these approaches search for structural similarities, which can help to predict side effects of known drugs or to identify the role of yet uncharacterized proteins**.**
While methods for fully automatic structure-based druggability predictions like SiteMap (Halgren, 2009), Fpocket (Schmidtke and Barril, 2010) and DLID (Sheridan *et al.*, 2010) exist, none of these methods is available online for predictions on new targets. Fpocket allocates a web-service where druggability scores and information can be requested (Schmidtke and Barril, 2010) but only for precalculated data points. The human genome project and developments in functional genomics are promising to present researchers with a number of clinically important targets. Attempts to generate three-dimensional structures of the target proteins are moving equally fast. We have been focusing on providing freely accessible computational tools for developing reliable in silico suggestions of candidate molecules against biomolecular targets (www.scfbio-iitd.res.in). Here, we introduce an automated version of active site (potential ligand binding site) detection, docking, and scoring methodology for any target protein. PRO-COMP IDS Tool provides the functionality to detect potential binding pockets and sub pockets of a protein of interest. PRO-COMP IDS Tool is been evaluated on a large data set containing around 1000 structures and shows prediction accuracies of 88%. Thus, the method provides valuable information for target assessment and can be accessed via a web-server.

2. Methods

The first step in the PRO-COMP IDS Tool procedure is the collection of the data from the DoGSite Scorer. A excel sheet is prepared which consists of the protein PDB ID and their respective subsequent pockets and the occurrence of the amino acid in the pockets with their score. Subsequently, a Difference of Gaussian (DoG) filter is applied. Each atom within 4 ˚A of any pocket point is considered a pocket atom. Pocket atom counts or functional groups and amino acid compositions describe the physicochemical features of the pocket. Furthermore, the lipophilic character of the pockets is addressed by the lipophilic surface and the overall hydrophobicity ratio. A java programming is used to code for this particular tool. Java is a computer programming language that is concurrent, class-based, object-oriented, and specifically designed to have as few implementation dependencies as possible. It is intended to let application developers "write once, run anywhere" (WORA), meaning that code that runs on one platform does not need to be recompiled to run on another. Java applications are typically compiled to bytecode (class file) that can run on any Java virtual machine (JVM) regardless of computer architecture. The language derives much of its syntax from C and C++, but it has fewer low-level facilities than either of them. Swing is a graphical user interface library for the Java SE platform. It is possible to specify a different look and feel through the pluggable look and feel system of Swing. Swing is the primary Java GUI widget toolkit. It is part of Oracle's Java Foundation Classes (JFC) — an API for providing a graphical user interface (GUI) for Java programs.

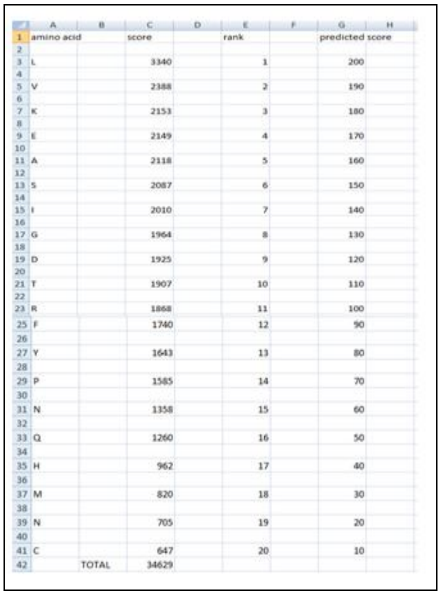
Swing was developed to provide a more sophisticated set of GUI components than the earlier Abstract Window Toolkit (AWT). The model has been trained and tested on the non-redundant version of the druggable data set (Schmidtke and Barril, 2010). External cross validation, randomly taking one half of the data as training and the other half as test set, showed a mean accuracy of 90%. (Volkamer et al., 2012) For each input structure, the method predicts the interaction probability of the protein in the particular active site and gives the score between zero and one . The higher the score the more interaction probability is estimated to be. The excel sheet is been shown in fig(1).



PDB ID	POCKET NO	AMINO ACIDS																
		ALA(A)	ARG- R	ASN- N	ASP- D	CYS- C	GLN- Q	GLU- E	GLY- G	HIS- H	ILE- I	LEU - L	LYS - K	MET - M	PHE - F	PRO - P	SER- S	THR - T
4HII	P0	2	1	2	2	0	0	0	3	0	2	2	1	0	1	1	1	1
	P1	1	1	0	0	0	3	2	1	0	1	0	3	0	1	1	0	1
	P2	2	3	0	3	0	0	2	0	0	0	3	1	1	3	0	1	1
	P3	1	0	1	0	0	2	1	1	0	0	5	1	0	2	1	8	2
	P4	2	0	0	0	0	0	2	1	0	0	2	1	0	1	2	2	2
	P5	2	1	2	2	0	0	0	3	0	2	2	0	0	1	1	2	2
	P6	1	0	1	0	0	2	1	1	0	0	3	1	0	1	1	1	1
	P7	3	1	0	1	0	0	2	1	0	0	2	1	0	1	1	2	2
	P8	3	1	0	1	0	2	1	2	0	0	1	1	0	1	0	0	1
	P9	0	2	1	1	0	0	1	2	1	0	0	1	0	1	2	2	2
	P10	0	4	0	0	0	2	0	1	0	1	2	0	0	0	1	3	1
	P11	1	1	1	1	0	1	0	0	0	0	3	0	0	3	3	2	0
	P12	1	1	0	1	1	1	1	0	0	0	3	0	0	3	3	2	0
	P13	0	1	1	1	0	0	1	1	0	0	0	1	0	1	1	2	2
	P14	2	1	0	2	0	3	0	1	0	0	1	0	0	0	2	2	2
	P15	0	2	0	0	0	2	0	1	0	1	1	0	0	0	1	1	1
	P16	1	0	0	0	0	2	2	0	0	1	0	2	0	1	1	0	0
	P17	1	2	0	1	0	3	1	0	0	1	1	1	0	1	2	0	0

Fig ure 1: The collection of the data from the DoGSite Scorer

In the excel sheet with the respective PDB ID and the subsequent pockets and their amino acid score.



	A	B	C	D	E	F	G	H
1	amino acid		score		rank		predicted score	
2								
3	L		3340		1		200	
4								
5	V		2388		2		190	
6								
7	K		2153		3		180	
8								
9	E		2149		4		170	
10								
11	A		2118		5		160	
12								
13	S		2087		6		150	
14								
15	I		2010		7		140	
16								
17	G		1964		8		130	
18								
19	D		1925		9		120	
20								
21	T		1907		10		110	
22								
23	R		1868		11		100	
25	F		1740		12		90	
26								
27	Y		1643		13		80	
28								
29	P		1585		14		70	
30								
31	N		1358		15		60	
32								
33	Q		1260		16		50	
34								
35	H		962		17		40	
36								
37	M		820		18		30	
38								
39	N		705		19		20	
40								
41	C		647		20		10	
42		TOTAL	34629					

Figure 2: The altered excel sheet which consists of the amino acid with their respective summed up score and the given rank and the predicted score

3. Implementation (Present and Future Work)

The PRO-COMP IDS Tool requires a PDB code or a user-specified PDB sequence. The algorithm of the flow of the tool works as follows.

- Step1. The user will specify the PDB ID or PID in the first textbox and clicks on the submit button.
- Step2. The application downloads the particular fasta format of the sequence of the specified ID and calculates the result i.e the interaction property of the particular protein.
- Step3. The user can also browse the sequence instead of the specifying the PDB ID in the next textbox.
- Step4. First we create a statistics table or a probability table of the amino acids of the protein that are been participating in the active site based on the output given by DoGSite scorer.
- Step5. The table is been created for around 500 proteins.
- Step6. The total occurrence of all amino acids is been summed up to get a score or to specify a rank for the amino acids.
- Step7. The highest total of an amino acid is been given the first rank and the following rank is been given based on the decreasing order of the total of amino acids.
- Step8. We assign an simple score for each of the amino acid based on the rank specified.
- Step9. When the user specifies the sequence or browses a sequence the application first calculates the length of the sequence.
- Step10. Next it calculates the occurrences of the amino acids in the sequence i.e it calculates how many times does a particular amino acid occurs in a sequence.
- Step11. The probability is been calculated by the following formula
- Total occurrence of amino acid/ (length of the sequence * highest score given to the first rank).
- Step12. The probability is been displayed in between 0.1 and 1.
- Step13. Based on this probability the interaction of a particular protein is been determined.

4. Conclusion

PRO-COMP IDS web-server provides an easy to use interface to predict protein interaction of a protein structure of interest. Furthermore, key properties characterizing the protein and druggability estimations are supplied.

5. Acknowledgments

Our thanks to respected senior lecturers and experts who guided in this development of the template. The research is supported in our KLE college biotechnology department.

6. References

1. Andrea Volkamer 1, Daniel Kuhn 2, Friedrich Rippmann 2, Matthias Rarey, “DoGSiteScorer: A web-server for automatic binding site prediction, analysis, and druggability assessment,” Bioinformatics Advance Access published May 23, 2012
2. Halgren, T. (2009).” Identifying and characterizing binding sites and assessing druggability”. J. Chem. Inf. Model., (49), 377–389.
3. Volkamer, A. et al. (2010). “Analyzing the topology of active sites: on the prediction of pockets and subpockets.” J. Chem. Inf. Model., 50(11), 2041–2052.
4. Lin, Y. et al. (2012).” Sitecomp: a server for ligand binding site analysis in protein structures”. Bioinformatics (Oxford, England)..
5. K. Sun and F. Bai, “Mining Weighted Association Rules without Preassigned Weights,” IEEE Trans. Knowledge and Data Eng., vol. 20, no. 4, pp. 489-495, Apr. 2008.
6. http:// en.wikipedia.org/wiki/Protein_Data_Bank
7. http://en.wikipedia.org/wiki/Java_(programming_language)
8. Tseng, Y. et al. (2009). Splitpocket: identification of protein functional surfaces and characterization of their spatial patterns. Nucleic Acids Research, 37(Web Server issue), W384–9