International Journal of Engineering Research and Development
ISSN: 2278-067X, Volume 1, Issue 8 (June 2012), PP.08-11
www.ijerd.com

Efficient Protein Interaction Hotspot Identification in Proteins

P.V.S.Lakshmi Jagadamba[1], Prof. M.S.Prasadbabu[2], Prof Allam Apparao[3]
[1]*Principal Investigator, Women Scientist Scheme A ,(WOS-A),DST Project,,JNTUK, Kakinada,AP,India*
[3]*Director, CR Rao Advanced Institute for Mathematics, Statistics & Computer Science (AIMSCS) University of Hyderabad, India*
[2]*Department of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, AP,India*

Abstract—*The pathological functionality of an organism is done by protein-protein interaction. In protein-protein interaction the proteins bind to each other and the binding energy is not uniformly distributed. Much of the binding energy is caused by some critical amino acids comprising of a small fraction of interfaces. An attempt is in this paper to develop basic frame work for Hotspot Identification to identify the hotspots in a protein amino acid sequence. The accuracy of this method is compared with the results of the existing methods.*

Keywords*—Primary Protein Structure, Amino acids, HISPPS, Digital filtering.*

I. INTRODUCTION

Bioinformatics is an amalgamation of computational and biological sciences. In the biological process, proteins undergo interactions. These interactions are mediated by molecular mechanisms. During this interaction, a small set of amino acids (residues) play a vital role. Such amino acids (residues) are called hotspots. The analysis of protein-protein interaction enables one to identify the hotspots from biological sequences accurately and efficiently. This analysis helps to predict the functionality of the residue for drug development.

Analysis of Protein structure and functioning plays a vital role in Bioinformatics. In general, protein sequence (also known as protein primary structure) folds into protein secondary/tertiary structure which binds with other proteins for functioning of a protein. Protein primary structure is its amino acid sequence (protein sequence). The amino acid chain forms in to a few secondary structures, based on the interactions of the peptide bond with nearby amino acids. Each secondary protein structure has one or more chain sequence segments determined by the primary protein structure i.e. there are several secondary protein structures for a protein sequence. Proteins generally interact with other molecules while functioning. During protein interactions, the energies are not uniformly distributed and some critical residues comprising only a small fraction of interactions account for the majority of the binding energy. The amino acids/ residues present in the active site are called hot spots. Hot spots form tightly packed regions in protein interactions [1] and identification of hot spots is helpful in estimating the efficiency of functioning of a protein [2].

II. MULTIPLE SEQUENCE ALIGNMENT USING PARTICLE SWARM OPTIMIZATION

Multiple sequence alignment algorithms are used to detect conserved regions in genetic sequences and to identify evolutionary relationships among organisms. MSAs are used to predict the functional segments of a sequence (genes) and the areas of a protein under selective pressures [3]. MSAs are also the primary input for reconstructing phylogenetic trees, or phylogenies.

Multiple sequence alignments are often used in identifying conserved (similar or identical regions) regions across a group of sequences assumed to be evolutionarily related. Such conserved regions can be used in conjunction with structural information to locate the active sites of protein-protein interaction. Hence multiple sequence alignment is used to identify hotspots in protein sequence.

III. METHODOLOGY

Nature of biological research is complex and widely distributed and the biological data is spread over many redundant databases and each group maintains a different type of database independently. One gene/protein could have different identifiers within one, or many databases. Similarly different gene identifiers for the same gene could be collected in different levels across different databases. Thus most gene functional annotation databases are in a gene-associated format with corresponding gene or protein identifiers. Such a format provides a good opportunity for researchers to integrate heterogeneous annotation resources through their common gene or protein identifiers.

There are dozens of types of gene or protein sequence identifiers that are redundant across several independent groups, such as GenBank Accession; GenBank ID; RefSeq Accession; PIR ID; PIR Accession; UniProt ID; UniProt Accession; SwissProt id; Pfam id; pdb is etc. The major challenge of integration comes from the weak cross-reference of different types of gene identifiers used by different functional annotation databases.

In this work an attempt is made to retrieve the general features of the protein data instead of simply retrieving the sequence, in a convenient way using with uniprot protein sequence identifier. As the details of a specific protein can be known from different databases, one need to know different identifiers to access the information from these databases. Instead of using different identifiers, in this work the data is retrieved by using only one identifier Uniprot id to retrieve the general features of a given protein from four (4) different databases [4-7].

Sequence alignment is the arrangement of two or more amino acid/ nucleotide sequences from an organism or organisms in such a way as to align areas of the sequences sharing common properties. The degree of relationship between the sequences is predicted computationally or statistically based on weights assigned to the elements aligned between the sequences. Sequence alignment of two sequences is known as pair wise sequence alignment and more than two sequences is known as multiple sequence alignment. The standard algorithm to align the sequences is Needleman and Wunsch algorithm [8], which uses the concept of dynamic programming. MSA is used to solve several biological problems such as Structure Prediction, Protein Family, Pattern Identification and so on.

This study concentrates in the pattern recognition i.e. identifying the residues/amino acids responsible for functional site using conserved region, which can be identified using MSA. Performing MSA for the family of sequences of the given input sequence to identify the hot spot residues in the given input sequence, the families of sequences are retrieved from PFam database. In this context the proposed two algorithms for identification of hot spots in a given protein sequence are Pair wise Sequence Alignment using Particle Swarm Optimisation (PSAPSO) algorithm [9] to align a pair of sequences and Multiple Sequence Alignment using Particle Swarm Optimisation (MSAPSO) algorithm [10] using PSAPSO to align multiple sequences.

IV. PERFORMANCE EVALUATION

There are no tools available to find the hot spots in a protein sequence. But there are some tools/ published results existing for the identification of hot spots in secondary structure of the protein. Comparison become easy for the position of hotspots with other studies for select Protein secondary structures, as all the hot spots are identified by observing all different possible 'Secondary Protein structures ' for a given 'Primary protein structure' in this study.

The existing methods namely ASM, Digital filtering, S- Transform filtering and Hotpoint focused on identification of hot spots in the select 'secondary protein structures' but not on 'primary protein structure' (amino acid sequence). As there are several secondary protein structures from a primary protein structure, hot spot identification using secondary protein structures may not yield efficient results. As all the hotspots in a Primary Protein Structure are important in binding to form three dimensional structure which influences the biological function of a cell, an attempt is made in this work to identify all the hotspots in 'primary protein structure' through Multiple Sequence Alignment using Particle Swarm Optimisation (MSAPSO).

V. HOTSPOT POSITIONS OBTAINED THROUGH DIFFERENT METHODS

Identification of hotspots is essential to understand the function of proteins and their interactions. In all biological processes proteins interact with each other. To understand biological process one need to study the principles of protein interactions. In the process of interaction, a small subset of residues (amino acids) is recognized or bound. These residues are commonly referred as hotspots [11]. These residues are defined as residues that impede protein- protein interactions if mutated.

Steve Buckingham [12] in his study attempted to investigate interactions between proteins using Proteomics. According to him revealing information about protein-protein interactions could provide the targets for a generation of new drugs. Protein-protein interactions also control the localization of proteins and their substrate processing activity NurcanTuncbag et al [13] in another study presented a web server HotPoint that predicts hotspots in protein interactions using an empirical model. The empirical model incorporates a few simple rules consisting of occlusion from solvent and total knowledge based pair potentials of residues.

Computational methods need basic data about protein finding to identify and design better drugs and also to predict potential interaction sites .In contrast to BIND (Biomolecular Interaction Network Database) [14-15] and DIP (Database of Interacting Proteins) [16] the Binding Interface Database (BID) presents data about the residues that make interactive function.

5.1. Efficiency of HISPPS – A Comparative Study : The measure F score is used to judge the efficiency of proposed method. F score is obtained for different available methods generally used for identification of hotspots in secondary structure considering Alanine Scanning Mutageneous (ASM) method as standard.

VI. RESULTS AND DISCUSSION

Sitanshu Sekhar Sahu and Ganapati Panda [17-19] developed an algorithm namely digital filtering to find the hotspots in select protein secondary structures. He compared his results with Stockwell et al [20] results obtained through S-transform time-frequency analysis technique. The results obtained in this work for these select secondary protein structures.

6.3. **F-score-a measure of overall model's accuracy** : The accuracy of these results has been examined using F-Score which is a measure of overall model's accuracy. F-Score value varies between zero and one where zero is worst and one is best value.

F- Score is calculated using two values precision p and recall R . Precision P is the fraction of predicted hot spots that are true hot spots. Recall R is the fraction of true hot spots that are predicted. To compute P and R one has to compare two methods and calculate Tp (True-positives), Fp (False Positives) and Fn (False-Negatives) for results to test [21].
The definitions are as follows.

- True-Positive (Tp): Residue is both a predicted hotspot and also an actual hotspot.
- False-Positive (Fp): Residue is a predicted hotspot but not an actual hotspot.
- False-Negative (Fn): Residue is not a predicted hotspot but an actual hotspot.
- Precision p = Tp/(Tp+Fp)
- Recall R = Tp/(Tp+Fn)
- F-Score = 2*P*R/(P+R)

The F-Score for the proposed method is found to be higher (0.87) than for other methods. The extent of efficiency reveals that HISPPS the proposed method/ algorithm is efficient by more than 20% over a) Digital filtering and b) S-transform filtering and more than 50% over c) Hotpoint method and hence concluded that the results obtained in this work for identification of hotspots through proposed method are relatively better than the existing methods.

VII. CONCLUSION

The main contribution of this paper is development of an efficient algorithm for identification of hotspots in Protein sequence over the existing standard models of identifying hotspots in Protein structure.

REFERENCES

[1]. Bogan and K.S. Thorn, "Anatomy of Hot Spots in Protein Interfaces," J. Molecular Biology, vol. 280, pp. 1-9, 1998.

[2]. Keskin,O., Ma,B. and Nussinov,R. (2005) Hot regions in protein–protein interactions: the organization and contribution of structurally conserved hot spot residues. J. Mol. Biol., 345, 1281–1294.

[3]. Pedro F. Rodriguez, Luis F. Nino and Oscar M. Alonso ' Multiple Sequence Alignment using Swarm Intelligence', International Journal of Computational Intelligence Research, ISSN 0973-1873- Vol 3, No 2(2007),pp 123-130.

[4]. UNIPROT, www.uniprot.org/

[5]. NCBI, www.ncbi.nlm.nih.gov/

[6]. PFAM, http://pfam.sanger.ac.uk/

[7]. PDB, http://www.pdb.org/pdb/home/home.do

[8]. Needleman, Saul B.; and Wunsch, Christian D. (1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". Journal of Molecular Biology 48 (3): 443–53. doi:10.1016/0022-2836(70)90057-4. PMID 5420325.

[9]. P.V.S. Lakshmi Jagadamba, Prof. M.Surendra Prasad Babu, Prof. Allam Appa Rao, P.Krishna Subba Rao, Mr. T. M. N. Vamsi "A Novel Approach for Pair Wise Alignment Using Particle Swarm Optimization", International Journal of Computational Intelligence Research & Applications , July-Dec 2010, Volume 4 , Number 2, ISSN 0973 – 6794, PP 157 - 160.

[10]. P.V.S. Lakshmi Jagadamba, Prof. M.Surendra Prasad Babu , Prof. Allam Appa Rao, .Krishna Subba Rao "An Improved Algorithm for Multiple Sequence Alignment Using Particle Swarm Optimization" IEEE 2nd International Conference on Software Engineering and Service Science (ICSESS2011), July 15th – 17th 2011, ISBN 978-1-4244-9696-9 Page 133- E135 Beijing, China.

[11]. Chao-Yie Yng and Shaomeng Wang, "Computational Analysis of Protein Hotspots",ACS Medicinal Chemistry Letters, 2010,1 (3) pp 125-129.

[12]. Steve Buckingham, "Picking the pockets of Protein-Protein Interactions", Horizon Symposia Charting Chemical Space, April 2004.

[13]. Nurcan Tuncbag, Ozlem Keskin and Attila Gursoy "HotPoint: hot spot prediction server for protein interfaces", W402–W406 Nucleic Acids Research, 2010, Vol. 38, Web Server issue, Published online 5 May 2010.

[14]. Bader GD, Betel D, Hogue CW BIND: the Biomolecular Interaction Network Database. Nucleic Acids Res. 2003;31:248-250.

[15]. Bader G.D., Donaldson,I., Wolting,C., Ouellette,B.F., Pawson,T. and Hogue,C.W. (2001) BIND—the biomolecular interaction network database. Nucleic Acids Res., 29, 242–245

[16]. Xenarios I., Fernandez,E., Salwinski,L., Duan,X.J., Thompson,M.J., Marcotte,E.M. and Eisenberg,D. (2001) DIP: the database of interacting proteins: 2001 update. Nucleic Acids Res., 29, 239–241.

[17]. Sitanshu Sekhar Sahu and Ganapati Panda, "Efficient Localization of Hot Spots in Proteins Using a Novel S-Transform Based Filtering Approach", IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 8, NO. 5, SEPTEMBER/OCTOBER 2011, PP(1235-1246)

[18]. J.A. Wells, "Systematic Mutational Analyses of Protein-Protein Interfaces," Methods Enzymology, vol. 202, pp. 390-411, 1991.

[19]. B.C. Cunningham, P. Jhurani, P. Ng, and J.A. Wells, "Receptor and Antibody Epitopes in Human Growth Hormone dentified by Homologscanning Mutagenesis," Science, vol. 243, no. 4896, pp. 1330-1336, Mar. 1989.

[20]. R.G. Stockwell, L. Mansinha, and R.P. Lowe, "Localisation of the Complex Spectrum: The S-Transform," IEEE Trans. Signal Processing, vol. 44, no. 4, pp. 998-1001, Apr. 1996.
[21]. Qin,L. et al. (2003) ,"Cysteine-scanning analysis of the dimerization domain of EnvZ, an osmosensing histidine kinase", J. Bacteriol., 185, 3429–3435.