ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 2, August 2013

Statistical Characteristics of Functions of Small Proteins

Jiji N[1], Dr. T. Mahalakshmi[2]

Abstract*— The present paper explores the small proteins from the PDB database to find the existence, if any, of molecular functional characteristics in the chains of these proteins. Such characteristics, if obtained, will help to throw more insight into small proteins as well as for identification of functions of newly discovered proteins. Nine different methods had been used in this paper to find hidden information in the chains and to associate it with its functions. Among them five methods are seen to reveal distinguishable characteristics for the functions and the application of principal component analysis revealed a dendrogram which can be used as signatures to predict molecular functions of small proteins from the amino acids of the chain itself.

***Index Terms*—Dendrogram, Hydrophobicity, Shannon Index, Small Protein.**

INTRODUCTION

Research has revealed that the term "Small Proteins" is being used in two contexts, based on the length of the amino acids in the protein and on the structure classification of proteins. Many of the works seen in literature are based on length of the protein sequence. Each author / authors have their own reason for choosing this length, which varies from 10 to 200 [1]-[3]. These works are based on membrane localization of small proteins [1], identification of short open reading frames [2] and the role of small proteins in cell functions [3].

The second context in which small proteins are used is for the classification of structural proteins having small number of chains as given in SCOP (structural classification of proteins) [4]. Such proteins often lack an extensive hydrophobic core and possess secondary structure elements that are small and irregular [5]. These proteins are generally stabilized either by binding a metal ion or the formation of disulphide bonds; the latter has the ability to stabilize the protein structure by reducing the conformational freedom of protein in the unfolded state [6], [7]. In the present paper the meaning of small proteins are considered as that of a class of structural proteins.

Among the various classifications of structural proteins SCOP classification occupies an important position because it is significantly based on human expertise than seen in its counterpart CATH or FSSP classification [8],[9]. The decision to assign a protein to the same super family or to the same fold is made by human expertise and this approach is believed to produce more accurate and useful results [10]. As per SCOP, small proteins are proteins having one or more of the following properties: (1) Dominated by metal ligand (2) Heme (3) Disulfide bond

The first property refers to the fact that the protein may contain more metals and small number of chains. The second property heme (or haem) stands for a group of organic enzymes that assist in biochemical transformations. They have an iron atom at the centre of their organic ring. The third property disulfide bond (SS-bond) indicates a covalent bond usually derived by the coupling of two thiol groups. They are much easy to rearrange so that a protein can do its designated function that has been assigned to it. Such rearrangement does not change the number of bonds and is faster when compared with similar process like oxidation / reduction reactions [6], [7].

Protein classification on the basis of structural similarity and evolutionary relatedness is a common means of organizing biological data for the purpose of studying various aspects of sequence / structure / functional relationships in proteins, such as structure prediction or identification of functionally important residues [5]. Residues that have similar functions in different proteins are likely to possess similar physicochemical characteristics [11]. Also it is a well known fact that many of the signatures of biological data had been revealed from its sequence.

This led to the thought of exploring the existence of any hidden statistical signatures in the sequence for identifying functional properties of small proteins. If such a signature exists, it will be able to throw more insight to the protein functions from its sequence itself and will help to identify functions of numerous proteins that are being added to the data base for which the functional properties are still unknown. The present paper proposes a combination of signatures for identifying the functions of small proteins based on the statistical features of the chain sequence.

The remaining part of the paper is organized as follows. In section 2 the preparation of the data used in this paper is explained in detail followed by materials and methods in section 3. Discussions is given section 4 followed by conclusion and reference.

DATA SET

From the Protein Data Bank (PDB) [12] small proteins of homosapiens were selected on January 16th 2012 with 100% dissimilar proteins consisting of only protein chains. This gave rise to 136 proteins with 417 chains. Of these chains there were 210 unique chains varying in length from 8 to 434 amino acids. The present paper is based on these 210 unique chains.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 2, August 2013

The first analysis done on these 210 chains is to find the frequency of absence of amino acids in these chains revealed that 46 of the 210 chains are made up using 19 distinct amino acids. It is seen that approximately 54% of small proteins is made using all the 20 amino acids, 22% is made using 19 distinct amino acids, 9% is made using 18 distinct amino acids and the remaining 15% is made using 17 or less distinct amino acids. The second analysis was done on the length of the chains. It was found that 23 of them have sequence length less than 50, 58 of them have same length exactly 192 and 33 have length greater than 200.

The third type of analysis on the data set was to find the number of chains not containing a particular amino acid. Among the 210 chains amino acid 'W' is not occurring in 48 of them , 'M' is not occurring in 43 of the chains , 'H' is not occurring in 32 of the chains. The final analysis that was done on this data set was to group the function domain of the protein obtained from the data base. It was found that this set contained a wide variety of protein function domains ranging from BLOOD CLOTTING to SIGNALLING PROTEIN. There were 70 function domains for these 210 chains. It is seen that many of the function domain names seem to be given in a synonymous manner. Hence some of the function domains were grouped together which yielded 11 classes that is listed in column 2 of Table I. In column 3 the frequency of chains in each function group is given, Column 4 indicates the number of chains in the functional groups that has length less than 50 amino acids, column 5 gives the difference between column 3and column 4, column 6 gives the group number of 7 classes on which further analysis being done in this paper.

MATERIALS AND METHODS

The frequencies of chains in each functional group are 10, 25, 33, 35, 19, 26 and 24. On this data set 9 types of statistical analysis was applied to find the characteristics of functional groups of small proteins. The overall analysis technique is a two stage process. The first stage involves the conversion of chains, made up of characters, into a corresponding digital representation to obtain any hidden attribute as well as for the application of various statistical techniques. In the second stage from the digital representation of proteins in each functional group a corresponding characteristic is obtained.

A. PROPOSED METHODS

Method 1: This method uses Shannon index for finding attributes of chains belonging to various functional groups.

For each chain in the data set its Shannon index was obtained and the average of these values in each functional group was obtained, these values are depicted in column 2 of Table II. But this attribute did not reveal any obvious differences.

Table I. Frequency of Proteins in Each Function

Sl. No	Functions	Frequency of Chains	Chain length < 50	Balance chain	Group No.
1	BLOOD CLOTTING	13	3	10	I
2	CATALYSES	7			
3	CELL ADHESION	31	6	25	II
4	EXTRACELLULAR MODULE	2			
5	GROWTH FACTOR	33	0	33	III
6	HORMONE	7			
7	INHIBITOR	42	7	35	IV
8	MEMBRANE PROTEIN	21	2	19	V
9	RECEPTOR	27	1	26	VI
10	SIGNALING PROTEIN	25	1	24	VII
11	STRUCTURAL PROTEIN	2			
	TOTAL	210	20	172	7 groups
	Chains Considered for further analysis	210 – 18 = 192	192 – 20 = 172		

Method 2: In this method EIIP Electron Ion Interaction Potential) of amino acids was taken to convert the chain into a numerical series. The average of these EIIP values is considered as the characteristic value of that chain. So for all the 172 proteins the EIIP characteristic was obtained in this manner. The average of the mean of the members of the functional groups is considered as the characteristic of the group, the values of which are given in column 3 of Table II. The values range from 0.0442 to 0.0487, not much obvious difference was seen in this case also.

Method 3 : Similar to the previous method instead of EIIP values the hydrophobic values, H, are used to obtain the characteristic of each functional group, details of which are given in column 4 of Table II. In this a noticeable difference is seen in GROWTH FACTOR functional group when compared with others. All other groups also yielded good features for this attribute.

Column 1	Column 2	Column 3	Column 4	Column 5	Column 6

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 2, August 2013

Table II Statistical Analysis of Data

functional group	Method 1	Method 2	Method 3	Method 4	Method 5	Method 6	Method 7	Method 8	Method 9
blood clotting	2.8135	0.0487	0.6754	14.6245	0.9000	2.7609	-0.4808	13.75	0.0483
cell adhesion	2.8168	0.0480	-0.6020	18.1507	2.5000	2.8770	-0.1693	24.1927	0.0376
growth factor	2.8291	0.0459	-0.3834	20.3745	3.3000	2.7047	-1.1896	0.9427	0.0545
inhibitor	2.8167	0.0459	-0.5124	18.2547	3.5000	2.6540	-0.4062	18.4077	0.0470
membrane protein	2.7995	0.0442	-0.4930	19.0529	1.9000	2.8806	-0.1650	23.9862	0.0414
receptor	2.7942	0.0474	-0.5057	18.2187	2.6000	2.8196	-0.2583	19.1198	0.0452
signaling protein	2.8049	0.0469	-0.4292	19.7856	2.4000	2.8023	-0.3408	19.9728	0.0434
Distinguishable characteristic	NO	NO	YES	YES	YES	NO	YES	YES	NO

Method 4: In this method hydrophobic value with HP7 is used to extract the features of the functional groups. The characteristic values of each group are given in column 5 of Table II which reveals obvious difference in value of the GROWTH FACTOR functional group as in the method 3. Also for BLOOD CLOTTING functional group it is showing a value very much different from all others.

Method 5: In this method the Shannon index of each protein was obtained. For each members of the group the Fast Fourier Transform (FFT) was obtained. The mean of the histogram of the real part of FFT of the sequence was considered as the characteristic of each group. The characteristic values are given in column 6 of Table II which shows noticeable differences in the values. The largest value is for the INHIBITOR functional group and lowest is for BLOOD CLOTTING functional group.

Method 6: In this method which is very similar to previous method, instead of taking the histogram of the FFT of the sequence the average of the real part of the FFT of the sequence is considered and the characteristic values of each functional groups is given in column 7 of Table II. But obvious differences are less in this case.

Method 7: In this method the FFT of the hydrophobic values obtained in method 3 is used to distinguish between the attributes of the functional groups. The values obtained in this method are depicted in column 8 of Table II. These values are negative, the only attribute that gave negative value in the present data set. The values obtained are very distinct for each functional group and can be considered as a characteristic to distinguish between these groups. The smallest value is for the GROWTH FACTOR functional group.

Method 8: The hydrophobic value with HP7 obtained in method 4 is also used in this case. The average of the FFT of the values of each sequence in a functional group is considered as the attribute of that group. The values obtained are given in column 9 of Table II shows that the GROWTH FACTOR functional group seems to give a good characteristic of that group.

Method 9: In this last method the mean of the real part of FFT of EIIP values (obtained in method 2) is used to obtain the attribute of each functional group. Like in the case of method 3 not much obvious result is seen, the values of which are given in column 10 of Table II.

IV. DISCUSSIONS

The present paper depicts various statistical values of the sequences of small protein obtained from PDB. It is widely known that many of the revelations in the genomic data emerged from sequence information [13]-[18]. This influenced in using the approach presented in the paper.

Here IBM SPSS tool's hierarchical cluster analysis is used for data analysis. Hierarchical cluster analysis starts with each case as a separate cluster and then combines the clusters sequentially, reducing the numbers of clusters at each step until only one cluster is left.A hierarchical tree diagram called a dendrogram on SPSS is taken to show the most important result of cluster analysis. It lists all samples and indicates - at what level of similarity any two clusters were joined. The position of the line on the scale indicates the distance at which clusters were joined. The distance or similarity measure used here is squared Euclidean Distances and the hierarchical cluster analysis used the Ward's method of clustering algorithm.

The dendrograms obtained from cluster analysis for method 4 is given in fig I, as their result shows similarity to

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 2, August 2013

the results of proposed method 4. X-axis of the fig I represents the seven functional groups and the y-axis represents the similarity of two clusters that were merged. The seven functional groups are numbered from 1 to 7. In the explanation each functional group is followed by its number in brackets. The dendrogram in fig I show that initially there are 7 clusters. The clusters inhibitor (4), receptor (6), cell adhesion (2) and membrane protein (5), being combined at fusion value 1and forms a cluster4625. Similarly the clusters growth factor (3) and signaling protein (7) also combined at fusion value 1 and forms a cluster37. Clusters 4625 and 37 joined at fusion value 6 to form cluster462537 .The newly formed cluster combines with the cluster blood clotting (1) at fusion value 25. This shows that BLOOD CLOTTING functional group is showing characteristic value very much different from all others.

V. Conclusion

The present paper explores characteristics in the sequence of small proteins from the human organism in the PDB database. The data obtained consist of 136 proteins with 210 unique chains having a maximum length of 434 amino acids.

The first analysis on these 210 unique chains is about the frequency of absence of amino acids. It revealed the fact that more than half of the chains were made up of all the 20 amino acids. The second analysis is on the sequence length of the chains and it was found that out of the 210 chains, the chain lengths between 50 and 200 are found to be 171. The third analysis is to find the sequence not containing a particular amino acid and it revealed that 24% of them did not contain the amino acid W, 20% did not contain the amino acid M, 15% did not contain the amino acid H and almost all contained the amino acid A. The final analysis is to group the function domains of protein and was found that there were 70 function domains for the 210 chains and only few numbers of proteins under some domains. So the 70 function domains were manually grouped into 11 broader categories based on the synonymous names. This is one of the area were future modification is possible to obtain a more accurate mapping. From the 11 groups, 4 function domains were omitted as it contains only less than 10 proteins and the remaining7function domains with 172 chains were taken for further analysis.

|In this paper nine statistical methods have been proposed to characterize the functions of small proteins. Among them methods 3, 4, 5, 7 and 8 gives noticeable differences in the attribute values. As a support for the statistical analysis, the SPSS hierarchical cluster analyses were applied on the same data set. This again revealed the fact that a high characteristically different functional values were occurring in methods 4, 5, 7 and 8. This can be viewed from the representation, dendogram. Further research has to be done in this area to find a method to predict the functions of small proteins. The methods that showed noticeable differences were obtained from the two types of hydrophobic values of the amino acids in each sequence. This is not a surprise since many of the hidden attributes of genomic data have been revealed using the hydrophobic values [19].

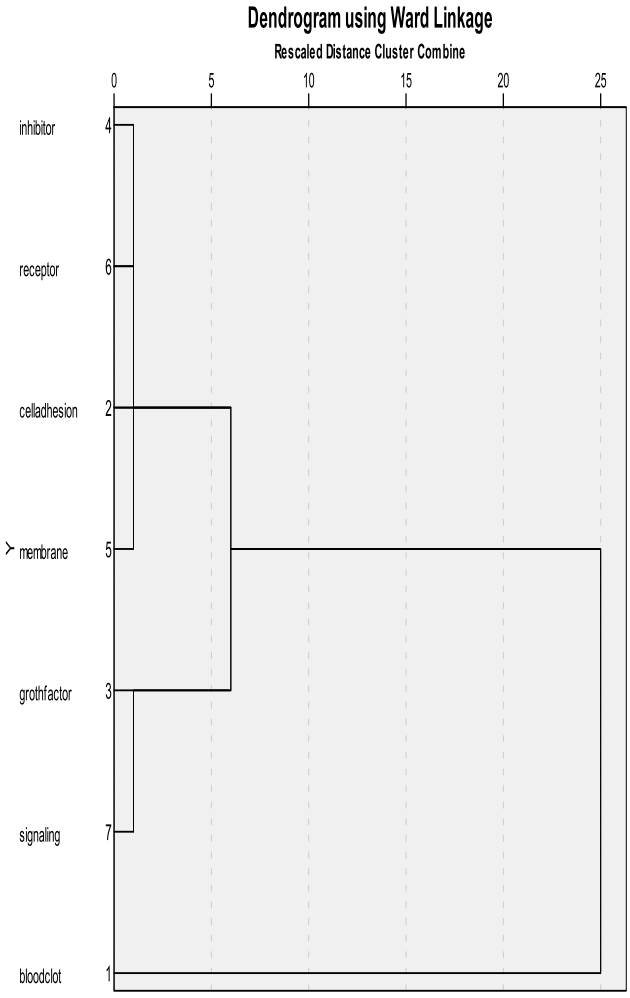

Fig I Dendrogram obtained for method 4.

REFERENCES

[1] Fanette Fontaine, Ryan T Fuchs and Gisela Storz, "Membrane Localization of Small Proteins in Escherichia Coli", The Journal of Biological Chemistry (JBC), September 16, 2011, 286, 32464-32474.

[2] Xiohnan Yang et al. "Discovery and annotation of small proteins using genomics, proteomics and computational approaches", Genome Research, March 2, 2011, doi: 10.1101/gr.109289.110.

[3] Errett C Hobbs, Fanette Fontaine, Xuefeng Yin and Gisela Storz, " An expanding universe of small proteins", Current Opinion in Microbiology, Science Direct, 2011, 14, 167-173.

[4] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C., "SCOP: a structural classification of proteins database for the investigation of sequences and structures", J. Mol. Biol. 2006, 359, 536-540.

[5] Sara Cheek, S. Sri Krishna and Nick V. Grishin, " Structural Classification of Small, Disulfide-rich Protein domains", J. Mol. Biol. 1995, 247, 215-237.

[6] Flory, P. J.," Theory of elastic mechanisms in fibrous proteins", J. Am. Chem. Soc., 1956, 78, 5222–5235.

ISSN: 2277-3754
ISO 9001:2008 Certified
International Journal of Engineering and Innovative Technology (IJEIT)
Volume 3, Issue 2, August 2013

[7] Thornton, J. M, "Disulphide bridges in globular proteins". J. Mol. Biol., 1981, 151, 261–287.

[8] Orengo C.A., Flores T.P., Taylor W.R. and Thornton J.M., "Identification and classification of protein fold families", *Protein Engineering Design and Selection*, 1993, 6, 485-500.

[9] Holm, L., & Sander, C., "The FSSP database of structurally aligned protein fold families", Nucl. Acids Res., 1994, 22, 3600-3609.

[10] Tim J. P. Hubbard, Bart Ailey1, Steven E. Brenner3, Alexey G. Murzin1 and Cyrus Chothia2, "SCOP: a Structural Classification of Proteins database," Nucl. Acids Res., 1998, 27, 254-256.

[11] Macro Punta, Yanay Ofran, " The rough guide to in silico function prediction or How to use sequence and structure information to predict protein function", PLoS Computational Biology, 2008, 4, 10, e1000160.

[12] http://www.rcsb.org dated 17th January 2012.

[13] Hamid Shateri Najafabadi and Reza Salavati, "Sequence-based prediction of protein-protein interactions by means of codon usage", Genome Biology, 2008, 9:R87.

[14] Bock J R and Gough D A, " Predicting protein-protein interaction from primary structure", Bioinformatics (Oxford England), 2001, 17:455-460.

[15] Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y and Jiang H, "Predicting protein-protein interactions based only on sequence information", Proceedings of the National Academy of Sciences of the USA 2007, 104: 4337-4341.

[16] Achuthsankar S. Nair and T. Mahalakshmi " Are categorical periodograms and indicator sequences of genomes spectrally equivalent?, In Silico Biology, International Journal of Computational Molecular Biology , IOS Press, 6(2006), pp 215-222.

[17] Achuthsankar S. Nair and Sivarama Pillai Sreenadhan, A coding measure scheme employing electron-ion interaction pseudopotential (EIIP), Bioinformation 1(6), 2006, pp. 197-202.

[18] Achuthsankar. S. Nair and Sreenadhan.S, An improved digital filtering technique using nucleotide frequency indicators for locating exons. Journal of the Computer Society of India, Vol. 36, No.1 , 2006, pp 60-66.

[19] T. Mahalakshmi , M Sajeev, N. Nijil Raj and N. Jiji, "Hub characteristics extraction of human proteins using tumor protein P53 – A case study", Recent Research in Science and Technology 2011, 3(2): 152-154, ISSN: 2076-5061, Available Online: http://recent-science.com/.

AUTHOR'S PROFILE

Jiji N is working as Associate Professor in Department of Department of Computer Science and Engg. , Younus College of Engg. And Technology, Kollam, Kerala, Iindia. She is pursuing her PhD in Manonmaniam Sundaranar University, Tirunelveli, Tamilnadu. Her fields of interest are Data Mining, Computational Biology, Computer Networks, Artificial Neural Networks, etc. She received her Master Degree in Computer and Information Technology from the Department of CITE, Manonmaniam Sundaranar University, Tirunelveli. She is a member of ISTE and IEEE. She has published 3 technical papers in national and international conferences and journals.

Dr T. Mahalakshmi is working as a Principal of Sree Narayana Institute of Technology managed by Sree Narayana Educational Society, Kollam, Kerala, India. After taking Master Degree in Computer Science from a US University in 1985, she worked for two decades in the IT arena along with Mr. R. P. Lalaji, the computer guru of Kerala. Later in 2007 her academic pursuits made her to take a Doctorate in Computer Science in Kerala University under the guidance of Dr. Achuthsankar S Nair. She has authored Computer Science text books and presented 29 papers in National and international fora.