International Journal of Bioinformatics Research, ISSN: 0975–3087, Volume 1, Issue 2, 2009, pp-70-80

Prediction and disorderliness of hub proteins

Mahalakshmi T.[1]* and Achuthsankar S. Nair[2]
1* Sree Narayana Institute of Technology, University of Kerala, Kollam, Kerala, India– 691010
2 Centre for Bioinformatics, University of Kerala, Thiruvananthapuram, Kerala, India–695581

Abstract – Protein-Protein Interaction (PPI) networks are important as they provide clues about the functions of individual proteins as well as enable system level analyses of cellular processes. Predicting hub proteins, the highly connected proteins in PPI networks, is a challenging computational problem. This paper proposes a method for predicting hub proteins from sequence information with 76% accuracy, 84% sensitivity and 71% specificity. In this method, a biodiversity measure, Shannon Index, is used along with an amino acid attribute Transfer Free Energy to Surface (TFES) to distinguish hub proteins from non-hub proteins. Also an analysis of disorderliness in hub proteins revealed that some amino acids have higher composition in hub than in non-hub.
Keywords – Hub Proteins, Degree of Connectivity, Shannon Index, Transfer Free Energy to Surface (TFES), disorder proteins, globular proteins

Introduction

In Bioinformatics, one of the important data set is protein. They are the work horse of life whose importance can be understood from the following sentence:" Right time, Right place and Right quantity of protein production makes one healthy". Proteins are available in the form of character sequences these sequences contain many hidden attributes, revealing of which is one of the problems in Bioinformatics and is a major research area. Protein-Protein Interactions (PPI) are essential to most biological process and can aid significantly in identifying the function of new discovered proteins [1-3]. Both experimental methods and computational tools for identifying PPI pairs have given rise to huge amounts of data. Some such sources of data are BioGRID, DIP etc. Studies have revealed that abnormal interactions may have implications in a number of neurological syndromes [4], which again ratifies the importance of PPI.PPI are visualized in the form of a network (map) where each node represents a protein and each link represents an interaction between a pair of proteins (see Fig 1). One of the attributes that can be associated with a node of a network is its degree of connectivity (k), which gives the number of links connected with a node.

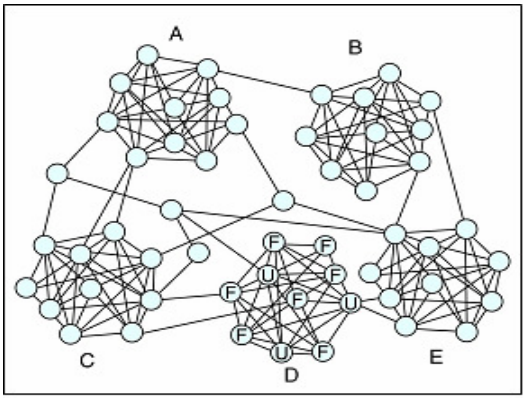

Fig. 1 A hypothetical protein-protein interaction network [5]

A notable feature of the PPI networks is their property of power-law and scale-freeness [6-8]. According to this model the probability of connecting to a new node in the network is proportional to the degree of connectivity of the existing node [6]. Thus a highly connected node has more chance of attracting a new node and thereby increasing its degree of connectivity. Such highly connected nodes are known as hubs and they are few in number in a power-law distribution. In other words, hubs are ubiquitous network elements with high connectivity [6]. Since PPI network follows a power-law distribution, as the studies have revealed, it follows that there are some proteins (nodes) with have high degree of connectivity than other proteins of the network. A protein (node) in a PPI network can be classified as a hub or as a non-hub based on the extent of connectivity [6]. Hub proteins are three times more essential than the non-hub proteins and they play an important role both evolutionary and physiologically [7, 9]. Hence they may constitute an important pool of attractive drug targets. They have a major place not only in information management within a network but also as regulatory molecules [6, 10]. Hence, the computational identification of hub proteins is one of the current topics in Bioinformatics. General techniques of PPI network maybe classified as experimental and computational. The experimental (large-scale proteomic experiment) techniques though they have vast coverage and sensitivity, do not give much information about the interacting residues. Computational analysis of PPI network is based on various attributes like gene proximity, gene fusion events, phylogenic profiling, identification of interacting protein domains and text mining techniques. Each of these approaches has its own strengths and weaknesses especially with regard to sensitivity and specificity. All the computational prediction of PPI techniques has focused on the identification of pair wise protein-protein interactions with varying degrees of accuracy. But none of them explicitly focuses on

predicting hub proteins. The proposed method gives attention to prediction of hub proteins, the proteins with very high connectivity, from the amino acid sequences. The classical point of view on protein function claims that the functionality of a protein requires the presence of a well-defined three-dimensional structure [11]. But experimental evidence have pointed out that there is a large number of proteins that do not require a stable structure even under physiological conditions in order to fulfill their biological role [11-15]. Such proteins are known as intrinsically disordered proteins (IDP) or intrinsically unstructured proteins (IUP). It has been recently suggested that IDP play an important role in PPI [12, 14, 16]. Literature review has revealed that intrinsic structural disorder is a distinctive and common characteristic of eukaryotic hub proteins, and that disorder may serve as a determinant of protein interactivity [16]. Recent evidence points to the preponderance of structural disorderliness in hub proteins compared with non-hub proteins [6, 16, 17]. In this paper a method is described to predict hub proteins from its sequence information. The proposed method when applied on a set of proteins of rat was found to have accuracy, sensitivity and specificity of around 76%. One more work done on this paper is an analysis of disorderliness in hub proteins. Literature review has revealed that for a disordered protein its amino acid (AA) composition is different from that of the normal proteins. This analysis revealed that amino acids C and P have lower composition in hub than in non-hub.

Data

The data used for this method is chosen from APID [17] database that contains details about co-interacting proteins (proteins that have physical interaction). APID (Agile Protein Interaction Data Analyzer) is an interactive tool that allows exploration and analysis of main currently known information about PPI.. It provides an open access frame where all known experimentally validated protein-protein interactions from various data bases like BIND, BioGRID, DIP, HPRD, IntAct etc are unified in a unique web application that allows an agile exploration of the interactome network and includes certain calculated parameters that weight the reliability of a given interaction like degree of connectivity, cluster coefficient etc [17]. In this paper the attribute degree of connectivity (k) that provides number of interactions of each proteins is used for obtianing characterisitics of hubness of proteins. The dat set was obianed by searching for protiens of ‘rat’ from this data base. Such a search was made since it is well known that generally for all clinical experiments rats are used initially, the success of which leads the researchers to do experiment on humans. The search result yeilded 4760 proteins ID’s from various organisms like Human, Yeast, Rat, Ecoli etc. These ID’s were downloaded and the correponding protein sequences were obatined from PDB (Protein Data Bank). All of the 4760 proteins, with connectivity ranging from 1 to 315 were selected from this database. Of the 4760 data it was possible to get sequence of 4753 only. Currently there is no consensus on exactly how many interactions a hub protein should have [9]. In this paper a protein is considered to be a hub if the number of interactions is at least four as per the convention followed in [8]. All of the 4753 data were split into two sets depending on its connectivity as 1630 hub and 3123 non-hub.The attributes of the data used in the proposed method are given in Table I. The first row of the table gives information about number of proteins in the test data set of hub, train data set of hub, test data set of non-hub, train data set of non-hub and total number of proteins in the data set. The second row specifies the number of interactions in each of the four data sets and the total number of interactions. The third, fourth and fifth row gives the information about the minimum, maximum and average degree of connectivity (Min k, Max k, and Avg k) of proteins in the four data sets - test data set of hub, train data set of hub, test data set of non-hub, train data set of non-hub. From the last row, average connectivity, it can be seen that train and test set have similar attributes in terms of connectivity. Another attempt that is made in this paper is the analysis of disorderliness in these proteins. Literature review revealed that disordered proteins have a specific amino acid composition that does not allow the formulation of a stable well-defined structure [19, 20]. Composition of amino acid of globular proteins which have a stable well-defined structure is given in [11] (see table II) and is used for analysis of disorder in the data set.

Method

It is widely known that many of the revelations in the genomic data emerged from sequence information [21-26]. The method proposed in this paper, to predict hub proteins, takes as input the sequence information of proteins. In this method a biodiversity measure Shannon index (Shannon-Wiener index) [27] is used to reveal the characteristics of hubness of a protein. Shannon index is considered as an important tool in information theory and statistical Physics and is well known as ‘sequence information extractor’ [28, 29]. This index gives a measure of relative diversity among organisms present in different ecosystems. It may be viewed as a measure of average degree of “uncertainty” in predicting to what extent an element may belong to a given set. The Shannon index technique has been adopted here to map each protein sequence to a

numerical value which is used as an index measure of that sequence. A few examples are given below to illustrate how the index is obtained from a sequence of characters.

Consider a sequence 'abcdef' of length 6 made up of different characters. Then Shannon index for this sequence is: (6*log 6- (1*log 1+1*log1+1*log1+ 1*log1+1* log1+1 *log1)) / 6. Consider a sequence 'aabbcc' of length 6 made up of 3 different characters. Then Shannon index for this sequence is (6*log 6-(2*log 2+2*log2+2*log2)) / 6. To generalize, assume that a sequence S is made up of alphabets a1, a2, …. with frequency c1, c2, ….. and the total length of S is n. Then Shannon index corresponding to this sequence is given by:

$$\text{Shannon index of } S = ((\log(n) * n) - (c1 * \log(c1) + c2 * \log(c2) + \dots))/ (c1 + c2 + \dots\dots)$$

Obviously, the value of this index measure depends on the frequency of each amino acid in a protein sequence. It will range from 0 (the worst case when the sequence is made up of only one alphabet) to logarithm of the length of the sequence (the best case when the sequence is made up of all different characters). From the formula it follows that Shannon index depends on the sequence information and so it can be considered as an attribute of the sequence. In the proposed method this index measure is used to predict if that protein can belong to a particular set or not. The proposed method is a two stage process. In the first stage the characteristics of hub proteins are found from a train set of hub proteins. Similarly in the case of non-hub proteins train set also the characteristics were obtained. For this purpose the hub and non-hub data each where split into two sets – train and test sets. Shannon index value was obtained for each of the protein sequence in the training sets of data. The average of the respective training set of hub and non-hub is taken as their characteristic value which is used for hub prediction. A target protein, which is to be classified as hub or non-hub, is an element of the test set. For target protein also the Shannon index value was computed. The distance between index value of target protein and characteristic values of the training set was chosen as the basis for the target protein to be a hub or non-hub protein. If target protein is nearer to that of the hub characteristic value then target protein was considered as a hub other wise as a non-hub. When this characteristic alone was used for prediction, the accuracy obtained was 43% only. Hence it was necessary to find some additional characteristics for prediction. For this purpose, a few of the attributes of amino acids in the AAindex database [30] were selected and computational experiments were conducted. Among them the attribute Transfer Free Energy to Surface (TFES) was found to be able to increase the percentage of prediction. A classification of TFES using k-means clustering tool provided in C-REx [31] is given in Table III which was used in the proposed method. The training sequences were subjected to this classification and then, on the new sequence so obtained Shannon index was calculated. The average of this measure of all hub and non-hub training set gave rise to another characteristic of hub proteins. When TFES was used for hub prediction the accuracy was found to be 52% only. But when a combination of Shannon's Index and TFES was used, it was found to be a better predictor and the accuracy obtained was 76%. The combination condition used for a target protein to be a hub protein was to check the nearness of the target protein with the two characteristic values obtained as described above. All data from the test set were subjected to this process and the number of correctly predicted hub and non-hub proteins is provided in Table IV. The accuracy, sensitivity and specificity of the proposed method on the data set is given in table V. The proposed method was applied on the whole data set obtained on searching for 'rat' in the database APID. The sensitivity obtained was 84% where as specificity was 71%. The low specificity may be due to the failure of the training data set to give an exact characterization. The accuracy of the proposed method is 76%. The proposed method was also applied on the data set of hub classifier, obtained from literature survey [9], which is a large set of data consisting of 21108 sequences from eukaryotes and prokaryotes. It was found to yield a sensitivity of 74%, specificity of 81% and an accuracy of 80%. The proposed method's sensitivity is far greater than what is stated in the literature which is only 34.41% where as the accuracy and sensitivity are less in the proposed method than what is stated in the literature which are 84.96% and 90.27% respectively. It follows that the proposed method can be considered as a better procedure for prediction of hub proteins.Table V also gives the sensitivity, specificity and accuracy obtained when experiment was conducted using the two different procedures individually and not as a combination, on both the set of data, data from APID and that from literature. It is very much evident from Table V that one characteristic alone is not sufficient for predicting the hub proteins where as a combination of both characteristics yields a higher percentage of prediction. From the results it can be seen that hub classification using proposed method is more suitable for predicting hub proteins. A plot of the data set when subjected to the proposed method is given in Figure 1. There are four quadrants in each figure which correspond to that of non-hub training, hub training, non--hub test and hub test data. The x-axis indicates each protein in a set and y axis indicates the numerical value of a

protein. The two colors –red and blue- indicate the values obtained under Shannon index and TFES procedures mentioned above. The lines show the average value under each procedure. The minimum and the average numerical value of each set are also given in the figure. It can be seen from Figure 2 that non-hub average of train and test data are 0.9592 & 0.9334 and that of hub is 0.9629 & 0.9595. These values also throw light on the similarity of train and test data sets chosen.

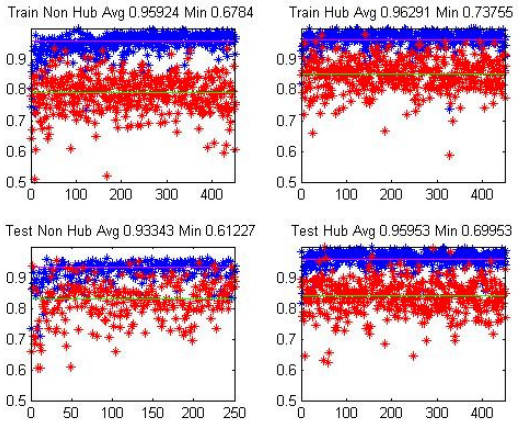

Fig 2. Plotting of 4 sets of Data from APID

Based on literature evidence that hub proteins are more disordered than nonhub, an analysis was conducted on the amino acid (AA) composition of data set. For analysis purpose the AA) compositions of the twenty amino acids were found in each of the protein of the data set. These values where compared with AA composition of globular proteins in three different ways as follows.

Comparison 1:

AA composition of hub / non hub < AA composition of globular proteins

For each sequence in hub it is checked whether the AA composition of each amino acid is less than that of the globular proteins. The same process is repeated with that of non-hub sequences. Then the percentage of sequences which satisfies the required condition is evaluated. These results are given in table VI. The first column of the table lists the 20 amino acids. The second column lists the number of non-hub proteins sequences (out of 3123) for which the AA composition is less than that of globular proteins. For example the percentage of amino acid 'a' in globular protein is 7.67 % as seen from table II. Among the 3123 non-hub proteins 1908 of them have the percentage of amino acid 'a' less that 7.67%. The third column gives this information in the form of percentage, 1908/3123 = 60.99 %. The fourth and fifth column gives the same information in the case of hub proteins. The total hub proteins are 1630. The last column is the difference between the percentage values of non-hub and hub satisfying the required condition. It can be seen from the table that for non-hub proteins the percentage of sequences satisfying the given condition is highest for the amino acid 'Y' and lowest for 'V'. The same is the case of hub proteins is 'W' and 'V' respectively.

Considering the last column diff with values >= 2 or < = -2 it is seen that

List of AA that is higher in hub when compared with Non hub – d, e, i, n, q, s

List of AA that is lower in hub when compared with Non hub – c, f, p, t, w

Comparison 2:

| AA composition of hub / non hub - AA composition of globular proteins | < 0.5

In this case for each sequence in hub it is checked whether the absolute difference of AA composition of each amino acid and that of the globular proteins is less than 0.5. The same process is repeated with that of non-hub sequences. Then the percentage of sequences that satisfies the required conditions is evaluated. These results are given in table VII.

The columns of this table are similar to that of the previous table. It can be seen from the table that for both non-hub and hub proteins the percentage of sequences satisfying the given condition is highest for the amino acid 'H' and lowest for 'V'.

Considering the last column diff with values >= 2 or < = -2 it is seen that

1. List of AA that is higher in hub when compared with Non hub - nil
2. List of AA that is lower in hub when compared with Non hub – c, m, p, r

Comparison 3:

| AA composition of hub / non hub - AA composition of globular proteins | < 1.0

In the third case for each sequence in hub it is checked whether the absolute difference of AA composition of each amino acid and that of the globular proteins is less than 1.0. The same process is repeated with that of non-hub sequences. Then the percentage of sequences that satisfies the required condition is evaluated. These results are given in table VIII. The columns of this table are similar to that of the previous table. It can be seen from the table that for both non-hub and hub proteins the percentage of sequences satisfying the given condition is highest for the amino acid 'H' and lowest for 'V'

Considering the last column diff with values >= 2 or < = -2 it is seen that

1. List of AA that is higher in hub when compared with Non hub - f, k, v, y

2. List of AA that is lower in hub when compared with Non hub – c, e, m, p, q, r, w

In summary the conclusions that can be drawn from the above three comparisons are given in Table IX.

Conclusions

For predicting whether a target is a hub protein or not in a PPI networks, a method is proposed which gives more than 76% accuracy for data set. The method was applied to a random set of data from eukaryotes and prokaryotes, obtained from literature survey, was also found to have almost similar accuracy. It was primarily Shannon index which was used in the proposed method to map a protein sequence to a numerical value. This index measure depends on the count of each amino acid in a protein sequence. To improve the accuracy of the method, another characteristic known as Transfer Free Energy was also used for classification. With a combination of these two characteristics, the proposed method could produce an accuracy of 76%, sensitivity of 87% and specificity of 73% with an exhaustive protein data of 'rat'. Even on a data set containing both eukaryotes and prokaryotes (data from literature [9]) the proposed method was able to predict with an accuracy of more than 74%. One of the findings in [9] was that different organisms have different hub connectivity threshold. Application of the proposed method on the various sets of data also indicates the same. Also the authors are of the opinion that if the exact hub connectivity threshold could be found out for each organism, the accuracy of the proposed method can be improved tremendously. Literature review failed to find any method belonging to this category. Only data set for prediction of Hub was obtained. That data set was used in the proposed method gave better sensitivity than that reported in the literature. Another work reported in on this paper is the analysis of AA composition in hub and non-hub proteins by comparing it with the AA composition of globular proteins. The analysis revealed that some amino acids have higher composition in hub than in non-hub. In particular it is found that the amino acids C and P have higher composition in non-hub than in hub. This result may also be incorporated to predict hub proteins in future works.

References

[1] Jingkai Yu, Russell L. Finley (2009) Bioinformatics, Vol. 25, pp 105-111.

[2] Ideker T, Sharan R (2008) Genome Res, 18, 644-652.

[3] Utez P, Finely R L (2005) FEBS Lett, 579, 1821-1827.

[4] Nazar Zaki, Sanja Lazarova, Wassim El-Haji, Piers Campbell (2009) BMC Bioinformatics, 10 :150, ISSN 1471-2105.

[5] http://genomebiology.com/content/figures/gb-2005-6-3-210-1-l.jpg

[6] K. Tun, R. Rao, L. Samavedham, H. Tanaka, and P. Dhar (2009) Systems and Synthetic Biology. http://dx.doi.org/10.1007/s11693-009-9024-9.

[7] Nizar N. Bataba, Laurence D. Hurst and Mike Tyers (2006) PLOS Computational Biology, 2, 7, 748:756.

[8] Diana Ekman, Sara Light, Asa K Bjorklund and Arne Eloffsson, (2006) Genome Biology, 7:R45.

[9] Michael Hsing, Kendall Grant Byler and Artem Cherkasov, (2008) BMC Systems Biology, 2:80.

[10] Rodriguez-Caso C, Medina MA, Solé RV, (2005) FEBS J 272:6423–6434.

[11] Mészáros B, Simon I, Dosztányi Z (2009) PLoS Comput Biol 5(5): e1000376.

[12] Wright PE, Dyson HJ, (1999) J Mol Biol 293: 321–331.

[13] Dyson HJ, Wright PE, (2005) Nat Rev Mol Cell Biol 6: 197–208.

[14] Dunker AK, Lawson JD, Brown CJ, Williams RM, Romero P (2001) J Mol Graph Model, 19: 26–59.

[15] Tompa P(2002), Biochem Sci 27: 527–533.

[16] Haynes C, Oldfield CJ, Ji F, Klitgord N, Cusick ME (2006) PLoS Comput Biol 2(8): e100.

[17] Singh GP, Ganapathi M, Dash D, (2007) Proteins 66:761–765.

[18] Prieto C. and De Las Rivas J. (2006) Nucl. Acids Res., 34: W298-W302.

[19] Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) J Mol Biol 347: 827–839.

[20] Dosztanyi Z, Csizmok V, Tompa P, Simon I (2005) Bioinformatics 21: 3433–3434.

[21] Hamid Shateri Najafabadi and Reza Salavati (2008) Genome Biology, 9:R87.

[22] Bock J R and Gough D A (2001) Bioinformatics (Oxford England), 17:455-460.

[23] Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y and Jiang H, (2007) Proceedings of the National Academy of Sciences of the USA 2007, 104: 4337-4341.

[24] Achuthsankar S. Nair and T. Mahalakshmi (2006) In Silico Biology, International Journal of Computational Molecular Biology, IOS Press, 6, pp 215-222.

[25] Achuthsankar S. Nair and Sivarama Pillai Sreenadhan (2006) Bioinformation 1(6), pp. 197-202.

[26] Achuthsankar. S. Nair and Sreenadhan.S (2006) Journal of the Computer Society of India, Vol. 36, No.1, pp 60-66.

[27] Keylock C.J., (2005) Earth and Biosphere Institute and School of Geography, UK. OIKOS, 109:1
[28] Yanan Yu, Mya Breitbart, Pat McNairnie and Forest Rohwer (2006) BMC Bioinformatics, 7:57
[29] Florent Angly, Beltran Rodriguez-Brito, David Bangor (2005) BMC Bioinformatics, 6:41
[30] www.genome.ad.jp/aaindex
[31] CREX http://cbi.keralauniversity.edu dated December 15th 2008.

Table I- Attributes of Data Set

ITEMS	HUB TEST	HUB TRAIN	NON HUB TEST	NON HUB TRAIN	TOTAL
No.of. Proteins	814	816	1558	1565	4753
No. of. Interactions	12407	12434	2340	2343	29524
Min. k	4	4	1	1	--
Max k	287	315	3	3	--
Avg k	15.24	15.23	1.5	1.5	--

Table II- AA composition in Globular proteins

Sl. No	AA	AA % in globular protein
1	A	7.67
2	C	2.43
3	D	4.92
4	E	5.43
5	F	3.19
6	G	8.46
7	H	2.00
8	I	6.35
9	K	6.37
10	L	8.22
11	M	1.84
12	N	4.69
13	P	4.89
14	Q	3.86
15	R	3.68
16	S	8.05
17	T	6.35
18	V	3.86
19	W	1.76
20	Y	3.86

Amino acid composition of the reference globular protein dataset comprised of all the amino acids in the longer chains of the ordered complexes dataset. AA denotes amino acid and F denotes the fraction of the respective amino acid expressed as a percentage. [doi:10.1371/journal.pcbi.1000376.t001] [11]

Table III- Amino Acid Classification

Group #	**1**	**2**	**3**	**4**	**5**
TFES	ADQ	RH	NG	CILKMFPSTWYV	E

Table IV- Data Test Sequences Prediction Result

	DATA SET	
	Hub	**Non Hub**
Number of Sequences	814	1558
Correctly Predicted	685	1102
Incorrectly Predicted	129	456

Table V- Result

Method	DATA SET	Sensitivity (%)	Specificity (%)	Accuracy (%)
Proposed Method	DATA from APID	84	71	76
Proposed Method	Data from hub classifier	74	81	80
Shannon Index	DATA from APID	60	35	43
Shannon Index	Data from hub classifier	41	54	62
TFES	DATA from APID	55	51	53
TFES	Data from hub classifier	61	52	52

Table VI- AA composition less than in Globular proteins

AA	No. of sequences less than globular protein in Nonhub	% of no. of sequences less than globular protein in Nonhub	No. of sequences less than globular protein in Hub	% No. of sequences less than globular protein in Hub	Difference in % (Col.5– Col.3)
(1)	(2)	(3)	(4)	(5)	(6)
A	1908	60.9974	1011	61.9865	0.9891
C	2459	78.6125	1337	81.9742	3.3617
D	1248	39.8977	569	34.8866	-5.0111
E	776	24.8082	325	19.9264	-4.8818
F	1097	35.0703	638	39.1171	4.0468
G	2488	79.5396	1325	81.2385	1.6989
H	1353	43.2545	702	43.0411	-0.2134
I	2491	79.6355	1264	77.4985	-2.137
K	1728	55.243	875	53.6481	-1.5949
L	1071	34.2391	532	32.618	-1.6211
M	868	27.7494	421	25.8124	-1.937
N	2225	71.1317	1061	65.0521	-6.0796
P	1583	50.6074	867	53.1576	2.5502
Q	1536	49.1049	758	46.4746	-2.6303
R	463	14.8018	259	15.8798	1.078
S	1795	57.3849	903	55.3648	-2.0201
T	2536	81.0742	1383	84.7946	3.7204
V	203	6.4898	132	8.0932	1.6034
W	2530	80.8824	1391	85.2851	4.4027
Y	2553	81.6176	1337	81.9742	0.3566

Table VII- AA composition in test sequence – AA composition in Globular proteins is less than 0.5

AA	|globular - test protein | < = .5	|globular - test protein | < = .5	|globular - test protein | < = .5	|globular - test protein | < = .5	Difference in % (Col.3–Col.5)
	No. of sequences of Nonhub	%of no. of sequences of Nonhub	No. of sequences of Hub	% of no. of sequences of Hub	
(1)	(2)	(3)	(4)	(5)	(6)
A	485	15.5051	266	16.309	-0.8039
C	740	23.6573	317	19.4359	4.2214
D	826	26.4066	404	24.7701	1.6365
E	501	16.0166	229	14.0405	1.9761
F	914	29.2199	499	30.5947	-1.3748
G	348	11.1253	167	10.2391	0.8862
H	1191	38.0754	618	37.8909	0.1845
I	468	14.9616	274	16.7995	-1.8379
K	507	16.2084	280	17.1674	-0.959
L	461	14.7379	237	14.531	0.2069
M	1137	36.3491	549	33.6603	2.6888
N	684	21.867	383	23.4825	-1.6155
P	665	21.2596	301	18.4549	2.8047
Q	813	25.991	417	25.5671	0.4239
R	450	14.3862	191	11.7106	2.6756
S	449	14.3542	240	14.7149	-0.3607
T	546	17.4552	264	16.1864	1.2688
V	243	7.7685	140	8.5837	-0.8152
W	783	25.032	388	23.7891	1.2429
Y	599	19.1496	326	19.9877	-0.8381

Table VIII-. AA composition in test sequence – AA composition in Globular proteins is less than 1.0

AA	|globular - test protein | < = 1.0	|globular - test protein | < = 1.0	|globular - test protein | < = 1.0	|globular - test protein | < = 1.0	Difference in % (Col.3–Col.5)
	No. of sequences of Nonhub	%of no. of sequences of Nonhub	No. of sequences of Hub	% of no. of sequences of Hub	
(1)	(2)	(3)	(4)	(5)	(6)
A	944	30.179	491	30.1042	0.0748
C	1479	47.2826	672	41.2017	6.0809
D	1494	47.7621	749	45.9227	1.8394
E	987	31.5537	482	29.5524	2.0013
F	1628	52.046	904	55.4261	-3.3801
G	699	22.3465	348	21.3366	1.0099
H	2090	66.8159	1114	68.3017	-1.4858
I	951	30.4028	491	30.1042	0.2986
K	1031	32.9604	585	35.8676	-2.9072
L	941	30.0831	495	30.3495	-0.2664
M	2007	64.1624	996	61.0668	3.0956
N	1372	43.8619	727	44.5739	-0.712
P	1297	41.4642	584	35.8063	5.6579
Q	1515	48.4335	754	46.2293	2.2042
R	893	28.5486	430	26.3642	2.1844
S	918	29.3478	472	28.9393	0.4085
T	1090	34.8465	551	33.783	1.0635
V	502	16.0486	304	18.6389	-2.5903
W	1816	58.0563	895	54.8743	3.182
Y	1194	38.1714	680	41.6922	-3.5208

Table IX- Summary of the result of three comparisons

	AA that is higher in Hub when compared with Nonhub	AA that is lower in Hub when compared with Nonhub
No. of Sequences Less than globular	D, E, I, N, Q, S	C, F, P, T, W
|globular - test protein | < = .5	---	C, M, P, R
|globular - test protein | < = 1.0	F, K, V, Y	C, E, M, P, Q, R, W
Common AA	----	C, P,