International Journal of Engineering, Mathematical and Physical Sciences
ISSN: 2517-9934
Vol:3, No:4, 2009

Connectivity Characteristic of Transcription Factor

T. Mahalakshmi, Aswathi B. L., Achuthsankar S. Nair

***Abstract*—**Transcription factors are a group of proteins that helps for interpreting the genetic information in DNA. Protein-protein interactions play a major role in the execution of key biological functions of a cell. These interactions are represented in the form of a graph with nodes and edges. Studies have showed that some nodes have high degree of connectivity and such nodes, known as hub nodes, are the inevitable parts of the network. In the present paper a method is proposed to identify hub transcription factor proteins using sequence information. On a complete data set of transcription factor proteins available from the APID database, the proposed method showed an accuracy of 77%, sensitivity of 79% and specificity of 76%.

***Keywords*—**Transcription Factor Proteins, Hub Proteins, Shannon Index, Transfer Free Energy to Surface (TFES).

I. Introduction

IT is well known that transcription factor (also known as sequence specific DNA binding factor) is a protein that binds to specific sequences of DNA and thereby initiates the transcription process and is vital to many important cellular processes. They are a group of proteins that read and interpret the genetic "blueprint" in the DNA. A defining feature of transcription factors is that they contain one or more DNA binding domains (DBD). The DBD attach to specific sequences of DNA adjacent to the genes that they regulate. They form one of the largest family of proteins.

One of the ways this family is classified is based on their regulatory functions – constitutively active transcription factor and conditionally active transcription factor. The constitutively active factor proteins are present in all cells at all times and the conditionally active transcription factor proteins require activation. Hence interaction of former with other proteins will be naturally much larger than that of the latter.

Various studies on protein-protein interaction networks (PPIN) and transcription-regulation interaction networks (TRIN) have revealed, in addition to many attributes, their significance in cellular process as well as understanding of proteins functions [1,2,3,4]. PPIN are represented in the form of graph with nodes representing proteins and edges representing interactions.

An example is given in Fig 1 [6], which shows a sub-network of Vaccinia virus proteins. This network contains 7 proteins which are denoted by circles (nodes) (see Fig 1). The edges or lines joining the proteins indicate the interactions between the proteins. A line joining the nodes G2 and UDG shows that they have interactions between them. The node G2 interacts with all the other 6 nodes. The node A49 interacts with only 5 of them. The number of interactions is generally termed as the degree of connectivity of the node in a network. Also evident from this is that interactions can be direct or indirect. For example G2 interacts with H5 which is a direct interaction. But A49 does not have a direct interaction with Viral DNA but has an indirect interaction through H5 or G2 or UDG or B1 or D5. Similar is the case between B1 and Viral DNA.

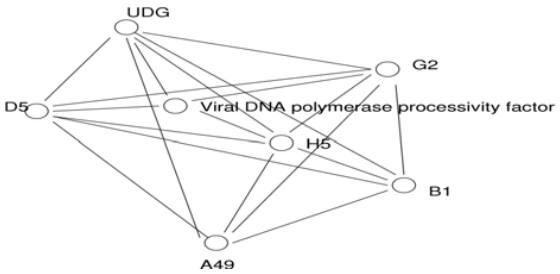

Fig. 1. A sample Network of PPIN [6]

These networks can be either considered in the form of a random networks or scale-free networks. In random networks the connections between the nodes are assumed to follow a Poisson distribution where as in the case of scale-free networks it follows a Power law distribution [4]. In a power law distribution there exits a few nodes known as hubs which have very high degree of connectivity. These highly connected nodes indicate, in the present context, that the interaction of some nodes with others is much larger.

Some researches have revealed that when PPIN is considered as scale-free networks, it follows a power law distribution [4,5]. That is, in PPIN some proteins are like hubs on a wheel with multiple spokes (interacting partners) attached. Even when a spoke is taken away the wheel will work. But if the hub is removed the wheel is useless [4]. In the same sense hub proteins are vital for the cellular functions of an organism. Hub proteins play an important role both

Dr. T. Mahalakshmi, Principal, Sree Narayana Institute of Technology, Vadakkevila, Kollam, Kerala, India, PIN – 691010 (phone: 91-9846315166; fax: 91-474-272-3156; e-mail: mlakshmi.t@gmail.com)
Aswathi B. L., Research Scholar, Centre for Bioinformatics, University of Kerala, Kariavattom, Thiruvananthapuram, Kerala, India, PIN – 695581, (e-mail: aswathi.bl@gmail.com)
Dr. Achuthsankar S. Nair, Director, Centre for Bioinformatics, University of Kerala, Kariavattom, Thiruvananthapuram, Kerala, India, PIN – 695581,
(e-mail: sankar.achuth@gmail.com)

evolutionary and physiologically and may constitute an important pool of attractive drug targets [3,4].

The present paper proposes a method to predict hub transcription factor from the sequence information. A biodiversity measure Shannon-Weaver index (Shannon-Wiener index or Shannon index) [7,8] is used to map the protein sequences into a numerical value and this index is chosen as the hub characteristic for prediction.

Literature survey revealed the existence of a work in the similar area, known as hub classifier with high accuracy and specificity but low sensitivity [3]. It classifies a protein as hub or non-hub based on the Gene Ontology annotation [3]. The proposed method was applied on this data set obtained from [3] as well as on complete set of transcription factor proteins available from APID [9] database. On both these sets the proposed method has accuracy, sensitivity and specificity of around 80%. So the proposed method may be considered as a better attempt for predicting hub transcription factor proteins.

II. Data

For the proposed method the transcription factor proteins were obtained from the APID database that contains details about co-interacting proteins (proteins that have physical interaction). This database, in addition to other attributes, provides protein id's and the degree of connectivity for each protein. No sequence information is available in this database. All of the 776 transcription factor proteins were selected from this database, whose degree of connectivity ranged from 1 to 314. Out of 776 proteins sequence information about 774 proteins were obtained successfully from other databases, but for the remaining two no information was available.

Hub proteins are proteins that have high degree of connectivity in a network. Currently there is no consensus on exactly what is the degree of connectivity that a hub protein should have [3]. In this paper a protein is considered to be a hub if the degree of connectivity is at least four as per the convention followed in [5]. All of the 774 data were split into two sets depending on its connectivity as 422 hub and 352 non-hub.

A closer look at the two sets revealed that most of the non-hub transcription factor proteins belong to the class of conditionally active transcriptions factor proteins while most of the proteins in the hub set belong to the class of constitutively active proteins. Intuitively this has to be true since hub proteins are highly connected and the constitutively active proteins are always active in a cell.

The data obtained from literature [3] contained 2004 hub and 19104 non hub proteins. These data were taken from the organisms E.Coli, S.cerevisiae, D.melanogaster and H.sapiens. For this data the degree of connectivity that a protein should have to consider it as hub was taken as 20 for E.coli, 33 for S.cerevisiae, 16 for D.melanogaster and 13 for H.sapiens. The proposed method was experimented on this data set also.

III. Method

To find some hidden attributes of a protein sequence, which characterizes the protein as hub or non-hub, Shannon-Weaver index is used to map each protein into a numerical quantity. The calculation of this index depends on the frequency of each amino acid in the 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. The frequency of each character in 'abcdef' is one. Then Shannon-Weaver 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. For this sequence the frequency of each character is two. Then Shannon-Weaver 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 a_1, a_2, …. with frequency c_1, c_2, ….., with total length of S being n. Then Shannon-Weaver index corresponding to this sequence is given by

$((\log(n) * n) – (c_1 * \log(c_1) + c_2 * \log(c_2) + \dots))/ n.$

The value of this index will range from 0 to logarithm of the length of the sequence. From the formula it follows that Shannon–Weaver index depends on the sequence information and so can be considered as an attribute of the sequence.

For the 422 hub transcription factor proteins obtained from the APID database their Shannon-Weaver index value was found. The average of all these indices is considered as a characteristic of hub proteins. Similarly for all of 352 non-hub transcription factor proteins also their Shannon-Weaver index value is found and their average is considered as a characteristic of non-hub proteins.

The prediction of a target protein to be hub or non-hub depends on the nearness of the Shannon-Weaver index value of the target protein with characteristic value of the hub / non-hub proteins. Choosing the 422 hub proteins and 352 non hub proteins as the target proteins, this method for prediction yielded only 51% accuracy. So attempt was made to find some more characteristic of the sequence.

Search for another characteristic lead to the amino acid attribute Transfer Free Energy to Surface (TFES) [10]. Using the k-means clustering tool provided in CREX [11], 20 amino acids were classified into 5 groups – ADQ, RH, NG, E and CILKMFPSTWYV based on their TFES values. Each of the 422 proteins in the hub and 352 proteins in the non-hub set was subjected to this classification, which yielded a new sequence made up of five characters. For this new sequence also Shannon-Weaver Index value was calculated. The average of all elements in hub and non-hub sets gave rise to another characteristic each for hub and non-hub proteins. Only 52% accuracy was obtained on the transcription factor protein data set, when it is used as the target proteins, and the average value obtained from TFES attributes was used as the

prediction characteristic.

But a combination of the above two characteristic on the transcription factor data set was able to yield an accuracy of 80%, sensitivity 82% and specificity 77%. 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 from the above two methods.

From the above discussion it follows that for the proposed method there are two stages – obtaining the characteristic value of hub and non-hub proteins and prediction of whether a target protein is hub or non-hub. The sequence of steps involved in the proposed method for identifying the characteristic values of hub and non-hub are given below:

1. Obtain the Shannon-Weaver index value of all hub proteins in the given data set.
2. Obtain the average of the numerical measures obtained in step 1, say HM1.
3. Find the Shannon-Weaver index of non-hub proteins in the given data set.
4. Find the average of the numerical measure obtained in step 3 , say NM1.
5. Convert each of the hub proteins into a sequence, which is a combination of 5 characters only as per the classification - ADQ, RH, NG, E and CILKMFPSTWYV.
6. Obtain the Shannon-Weaver index of the hub proteins obtained in step 5.
7. Obtain the average of the numerical measured obtained in step 6, say HM2.
8. Repeat steps 5 and 6 for non-hub proteins.
9. Obtain the average of the numerical measures obtained in step 8, say NM2.

The sequence of steps involved in the second stage to predict whether a target protein is hub or non-hub are given below:

1. Obtain the Shannon-Weaver index value of the target protein say E1.
2. Obtain the Shannon-Weaver index value of the target protein on which the TFES classification has been applied say, E2.
3. The prediction of target protein is a hub or non-hub is based on the distance of target protein from the trained value and is given by the following condition:

if abs(E1-NM1) >= abs(E1-HM1)
and abs(E2-NM2) >= abs(E2-HM2)
then
target protein is a hub protein
else
target protein is a non hub protein

Fig. 2 portrays the plotting of transcription factor data set when subjected to the proposed method. There are two plots in this figure which correspond to that of non-hub and hub set. The blue colors indicate the values obtained under first and red color that of the second method mentioned above and the lines show the average value under each method. The x axis indicate the proteins and y axis indicate the Shannon-weaver Index value of each protein. In the figure for each set the minimum and average numerical value is also given. In the case of non-hub set the average is 0.95167 and that of hub is 0.95762. The respective minimum in each case is 0.56846 and 0.57087.

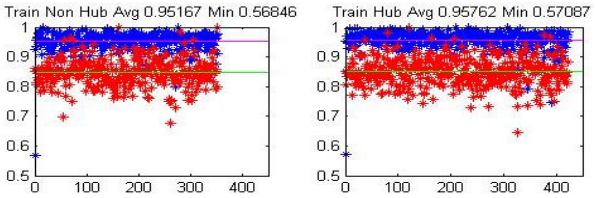

Fig. 2 Plotting of 2 sets of Data of Transcription Factor proteins

For confirming the result obtained from the proposed method, the transcription factor protein data set of hub and non-hub were spilt equally into two sets – train and test. There were 175 elements in the non-hub train set and 176 in non-hub test set. Similarly in the case of hub they were 211 and 212 elements in train and test respectively. The proposed method was applied to the hub and non-hub train sets to obtain the corresponding characteristic values. Each sequence in the test set was considered as the target protein. The condition specified in the proposed method was applied to the target protein to find whether it is hub or non-hub. In this case the accuracy, sensitivity and specificity obtained were 77%, 79% and 76% respectively.

Fig. 3 portrays the plotting of the proposed method on the new data set. There are four quadrants in this figure which corresponds to that of train and test of non-hub and hub set. In the case of non-hub train set average is 0.95732 and that of test is 0.9523. In the case of hub the respective values are 0.958 and 0.95875. These averages indicate that the train and test set selected are of the same nature. The red color, blue color, x axis and y axis information are similar to that given in Fig 2.

As mentioned earlier, DNA binding domains (DBD) are one of the distinctive features of transcription factor proteins. These domains are functionally similar and hence expected to have similar set of amino acids. The Shannon-Weaver index value depends on the frequency of amino acids in a sequence. Total of DBD are found to be higher in constitutively active proteins than in conditionally active proteins. It may be considered that DBD's are higher in hub proteins than in non-hub proteins. Again the classification of sequence using TFES forms clusters of amino acids based on their TFES values. So both these attributes are certain to throw light on the hub characteristic nature of transcription factor proteins.

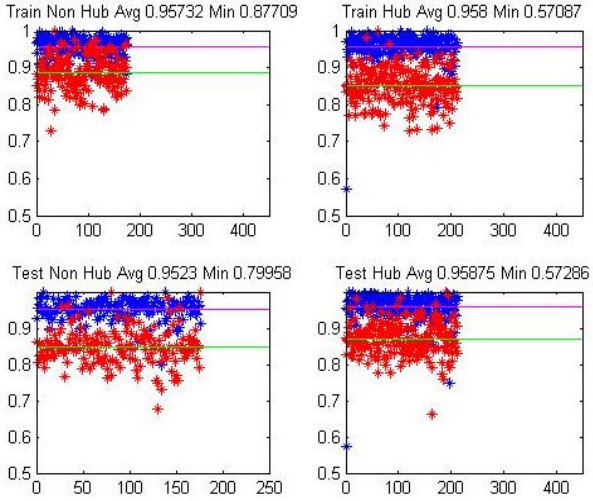

Fig. 3 Plotting of 4 sets of Data – Transcription Factor proteins

The application of the proposed method on the data set obtained from literature survey had accuracy of 80%, sensitivity 74% and specificity 81%. This new data set had 21108 proteins and was a mixture of all organisms with various hub connectivity threshold values for each organism. The results reported in [3] had accuracy 84.96%, sensitivity 34.41% and specificity 90.27%. In comparison, the proposed method shows a better result not only on the data set of transcription factor proteins but also on the set of proteins of all organisms. For the proposed method to predict a target protein, it is necessary to know its sequence information and the hub connectivity threshold of the organism.

The sensitivity, specificity and accuracy reported in this paper are obtained using the following formulae in which TP indicates True Positive, TN indicates True Negative, FP indicates False Positive and FN indicates False Negatives

Sensitivity TP / (TP + FN)
Specificity TN / (TN + FP)
Accuracy (TP + TN) /(TP + TN + FP + FN)

IV. Conclusion

Transcription factor proteins are very important family of proteins. Most of the hub transcription factor proteins belong to the class of constitutively active proteins. A method is proposed in this paper to predict hub transcription factor proteins from its sequence information. On the complete set of transcription factor proteins from APID database the proposed method has accuracy, sensitivity and specificity of around 80%. On a set of data obtained from the literature also the proposed method yielded a very similar result. So the proposed method can be considered as a stepping-stone for predicting hub transcription factor protein or possibly for all organisms. It is a very simple method which helps to characterize the nature of transcription from its sequence itself.

References

[1] Esti Yeger-Lotem, Shmuel Sttath, et.al, "Network motifs in integrated cellular networks of transcription – regulation and protein-protein interaction", PNAS, 2004, 101, 16, 5934:5939.
[2] Ideker T, Sharan R "Protein Networks in Disease", Genome Res, 2008, 18, 644-652.
[3] Michael Hsing, Kendall Grant Byler and Artem Cherkasov, "The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks", BMC Systems Biology, 2008, 2:80
[4] Nizar N. Bataba, Laurence D. Hurst and Mike Tyers, "Evolutionary and Physiological Importance of Hub Proteins", PLOS Computational Biology, 2006, 2, 7, 748:756
[5] Diana Ekman, Sara Light, Asa K Bjorklund and Arne Eloffsson, "What properties characterize the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae?", Genome Biology 2006, 7:R45.
[6] www.biomedcentral.com dated 21st March 2009
[7] Keylock C.J., "Simpson diversity and the Shannon/Wiener index as special cases of a generalized entropy", Earth and Biosphere Institute and School of Geography, UK. OIKOS, 2005, 109:1
[8] Yanan Yu, Mya Breitbart, Pat McNairnie and Forest Rohwer, "FastGroupII: A web-based bioinformatics platform for analyses of large 16S rDNA libraries", BMC Bioinformatics, 2006, 7:57
[9] Prieto C. and De Las Rivas J. , "APID: Agile Protein Interaction DataAnalyzer." Nucl. Acids Res. 2006, 34: W298-W302.
[10] www.genome.ad.jp/aaindex
[11] CREX http://cbi.keralauniversity.edu dated December 15th 2008