Recent Research in Science and Technology 2011, 3(2): 152-154
ISSN: 2076-5061
Available Online: http://recent-science.com/

Hub characteristics extraction of human proteins using tumor protein P53 – A case study

T. Mahalakshmi[1], M Sajeev[2], N. Nijil Raj[2] and N. Jiji[2]

[1]Principal, Sreenarayana Institute of Technology, Vadakkevila, Kollam 691010
[2]Department of Computer Science and Information Technology, Manonmaniam Sundaranar University, Tirunelveli-627012(TN)

Abstract

This paper addresses the characteristic extraction of hub protein based on Tumor Protein P53 whose properties are already established and known to have key functionalities. These characteristics can throw some light in the direction of hub classification in a cost effective manner. Current methods in this line use Gene Ontology database or sequence homology which are time consuming and complex. The proposed method uses a 420 element vector for the characteristic filtering of hub character from HPRD database and has shown some positive results.

Keywords: Hub protein, P53, PIN, Degree of Connectivity, HPRD

INTRODUCTION

Hub proteins are highly connected and active as the name suggested [1], [2]. These inevitable proteins are vital for the proper biological functioning of humans. The present work is an attempt to extract hub characteristics of Human Proteins using tumor protein P53 [24], [25], [26], [27] as the bench mark. The study checks whether the features available in the P53 Protein are responsible for the hub character of other human proteins. The work is a case study which explores the key characteristics of P53 which are responsible for its medicinal features. It is not mandatory that the characteristics responsible for the hubness in one protein should remain constant across other proteins, but the presence or absence of some key characteristics may play the big role for that.

Hub proteins are able to form a network of proteins [5]called PIN (Protein Interaction Network) due to their increased hub characteristics available in the surface. Figure-1 depicts a part of PIN derived from HPRD (Human Protein Reference Database) which gives clear perception that certain proteins such as BTRC, PAEP, UBE2EI carry more connections than other proteins called degree of connectivity.

PIN's are counted as either Random Networks or as Scale Free Networks where Scale Free Networks closely model most of the real world networks [3]. But still a lot of havoc exist about the scale free nature of human biological networks. The large and complex protein interactions direct most biological pathways and processes [3] and surprisingly most of these interactions are directed by hubs. This is the reason why they are considered as lethal proteins which are strategically located and if disturbed can lead to biological lethality [3]. Hence study of hub proteins is relevant to understanding the causes of diseases and provides efficient and cost effective solutions.

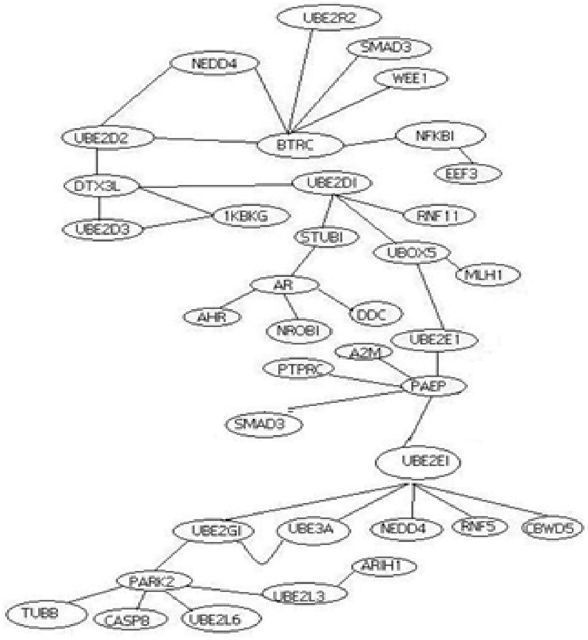

Fig 1. A sub-network of PIN using HPRD Database

Hub Proteins participate in significant number of protein interactions and play a vital role in the organization of cellular protein interaction networks [9, 10].This is the reason why Hub proteins are three times more essential than the non-hub proteins. So they could be of particular interest as drug targets [6].

Background

P53 protein is known to have some key functionalities in the direction of cancer prevention. The protein acts as lead player in the P53 pathway, which leads to the prevention of cancer by controlling the growth of cells. This is the reason behind the selection of P53 as the benchmark for the study. P53 (also known as tumor protein

*Corresponding Author
Email: m.lakshmi@gmail.com

53 or protein 53), is a tumor suppressor protein that in humans is encoded by the gene TP53[23].

Human P53 is 353 amino acids long and has seven domains. Residues from 100-300 contribute to DNA Binding Domain (DBD) which contains one zinc atom and several arginine amino acids. Arginine is highly hydrophobic in nature and contributes heavily to a protein's hub characteristics [24]. This DBD information is made use in the proposed method for the extraction of the major characteristics.

One of the methods for the prediction of interaction between proteins is using the amino acid sequence information alone [13], [14], [15]. In all these computational prediction techniques significant priority is given to the identification of pair wise protein-protein interactions with varying degrees of accuracy [6], [16].

High end computation is required to unearth the features of PIN using gene proximity analysis, gene fusion events, phylogenic profiling [6] etc. These tests have both strengths and weeknesses in terms of computational complexity.

For exploring a protein for its hub characteristics its Gene Ontology annotation is required. Michael Hsing *et al.* [6] have stated that the performance of the algorithm is directly proportional to the availability of Gene Ontology annotations and this is found to be the reason for the low sensitivity score.

In the light of these methods, it appears that the characteristics under study are either not economical or adequate for perfect hub prediction. The best way around is to extract feasible characteristics from the amino acid sequence itself and is overlooked that the same can be made use to categorize the protein according to their medicinal effects.

Data Set

Human Protein Reference Database (HPRD) [17] explores maximum number of human proteins. The database provided the data in the form of binary interactions which reduced the data preparation overheads.

This Database contained 27080 human proteins. Among them 9630 have interactions with others and that information is presented in the form of binary interactions in the data base. This information was used to derive a connectivity metric known as degree of connectivity of the protein which ranged from 0 to 267. The below given table shows the information derived out of this data.

Table I. Degree of Connectivity Vs Protein Frequency in HPRD

Degree of Connectivity (k)	Number of proteins
0	17450
1	2237
2	1424
3	1009
4	759
5	618
6	468
7	422
8	287
>> 8	2406
Total	27080

It is evident from the table that the value of k is inversely proportional to the number of proteins and using this information the measure of central tendency, mean is calculated and found as 8.0557. According to this statistical view it was found that 2406 proteins satisfied the condition k > 8 and considered as positive data which account to 25% of the total interacting sequences and the rest of 7224 proteins opposed that condition and considered as the negative data.

Proposed Method

The proposed method is a two stage procedure. In the first stage a vector containing some numeric attributes are derived from the sequence information. The second stage performs a similarity check through the whole sequence and the result is saved for final analysis.

Initially the amino acid count and dimmer count for all 20 amino acids comprising of 420 attributes are generated out of the DBD [24] of P53 protein of size 201 amino acids. This information is stored in a vector. Then divide the target sequence into subsequence of length 201 and obtain the attributes for this subsequence. Given below are two sample vectors generated from P53 and a target protein subsequence. Here A, C … Y are the 20 amino acids and AC, AD… ZZ are the dimmers for these 20 amino acids.

```
                A   C D E F …. Y   AC AD AE ……….YY
            i=0    1    2 3 4 …. 20 21 22 23 ………. 420
P53 - P [12 11 8 9 14       32 36 43 23                      33]
Seq1- S [21 7 18 9 12       30 39 47 29                       40]
```

Here if | P[i] – S[i] | <= 25% of P[i] then S[i] is a true estimation of attribute P[i] for Seq1. This is a kind of fuzzification for a near optimal solution since a perfect similarity is below expectation.

Repeating this process for all 420 attributes, the Seq1 is considered as a hub protein if at least 315 numbers of attributes are a true estimation of Seq1. If any of the subsequence of a protein is found to be a hub by satisfying the above condition then the algorithm proceeds to the next protein.

The complete set of both positive and negative data in terms of hubness is tested and significant results obtained.

RESULTS AND DISCUSSIONS

One of the recent methods developed to predict protein interactions [22] is purely based on the sequence and this throws light on the fact that there is some hidden information in the protein sequence at the amino acid level itself. Hence chances are high that predicting hub using sequence information based on proven hub proteins.

A motif based string search could achieve this much accuracy only by considering the 420 element vector alone compared with GO method discussed in section 3 which use more than 1200 parameters. So the results have shown that considerable reduction in computational time is achieved through this method.

CONCLUSION

Protein interactions are ever-present in nature and essential for cellular functions in organisms. It is already proved that all the physical protein-protein interactions for a given cell or organism is complex bio-molecular network mapped as Protein interaction network. Predicting the hub proteins of this network is a promising computational problem.

Many methods have been suggested for the hub characterization of proteins. But the challenge always lied in the computational complexity of the method. In the proposed method the results have favored in this direction. But still a lot of confusion exists about a minority of unidentified proteins in the positive data set. Works needed to be done to achieve maximum accuracy. This can be done by increasing the vector size by adding more attributes. Experiments are targeted in this direction. The method can also be tested with other databases of different organisms for consistency of the results

REFERENCES

[1] Chad Haynes, 2006. “Intrinsic Disorder is a Common Feature of Hub Proteins from Four Eukayotic Interactomes”, PLOS Computational Biology, Vol 2, Issue 8.

[2] Rual JF, 2005. “Towards a proteome scale-map of the human protein-protein interaction network”, Nature, 437, 1173-1178.

[3] http://en.wikipedia.org/wiki/P53 dated 18/9/2011.

[4] http://www.ncbi.nlm.nih.gov/ dated 18/9/2011.

[5] Alexei Vazquez, Elisabeth E Bond, Arnold J Levine, Gareth L Bond, 2008. “The genetics of the p53 pathway, apoptosis and cancer therapy”, Nature Reviews Drug Discovery Volume: 7, Issue: 12, Publisher: Nature Publishing Group PP: 979-987.

[6] Carol Prives1, Peter A. Hall, 1999. “The p53 pathway”, The Journal of Pathology, Special Issue: Molecular and Cellular Themes in Cancer Research, 187(1):112–126.

[7] Nizar N. Bataba, Laurence D. Hurst and Mike Tyers, 2006.“Evolutionary and Physiological Importance of Hub Proteins”, PLOS Computational Biology, 2, 7, 748:756

[8] Sriganesh Srihari et al, 2008. “Detecting Hubs and Quasi Cliques in Scale-free Networks”, IEEE,

[9] Barabasi, A. L. and Oltvai Z N, 2004. “Network biology: understanding the cell's functional organization”, Nat Rev Genet 5(2):101-113.

[10] Albert R., 2005. “Scale-free networks in cell biology”, J Cell Sci., 118 (21): 4947-4957.

[11] Michael Hsing, Kendall Grant Byler and Artem Cherkasov, 2008. “The use of Gene Ontology terms for predicting highly-connected 'hub' nodes in protein-protein interaction networks”, BMC Systems Biology, 2:80

[12] Rual JF, 2005. “Towards a proteome scale-map of the human protein-protein interaction network”, Nature, 437: 1173-1178.

[13] Matlashewski G, Lamb P, Pim D, Peacock J, Crawford L. and Benchimol, S., 1984. "Isolation and characterization of a human p53 cDNA clone: expression of the human p53 gene". EMBO *J.* 3 (13): 3257–62.

[14] Hamid Shateri Najafabadi, and Reza Salavati, 2008. “Sequence-based prediction of protein-protein interactions by means of codon usage”, Genome Biology, 9:87.

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

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

[17] Qi, Y, Bar-Joseph, Z. and Klein-Seetharaman, J., 2006. “Evaluation of different biological data and computational classification methods for use in protein interaction prediction”, Proteins, 63(3):490-500.

[18] Kim, PM, Lu, LJ, Xia, Y, Gerstein, MB., 2006. Relating three-dimensional structures to protein networks provides evolutionary insights”, Science 314, 2006 1938–1941.

[19] http://www.hprd.org/ Release 9 dated May 24, 2010

[20] Raghuraj Rao., 2009.“Amino-acid residue association models for large scale protein-protein interaction prediction”, In Silico Biology, 9: 0015.