BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 7(4)

Identification of hub proteins from sequence

Aswathi Balakrishnan Latha[1], Achuthsankar Sukumaran Nair[2]*, Athmaja Sivasankaran[3], Pawan Kumar Dhar[4]

[1]Centre for Bioinformatics, University of Kerala, Trivandrum 695581, India; [2]Centre for Bioinformatics, University of Kerala, Trivandrum 695581, India; [3]Christ University, Bangalore; [4]Centre for Systems and Synthetic Biology, University of Kerala, Trivandrum 695581, India; Achuthsankar S Nair – Email: sankar.achuth@gmail.com; Phone: 471-2308759; *Corresponding author

Received September 14, 2011; Accepted September 28, 2011; Published October 14, 2011

Abstract:
Identification of hub proteins from sequence is a challenge in molecular biology. Therefore, it is of interest to predict protein hubs in networks. We describe the prediction of protein "hub" using physiochemical, thermodynamic and conformational properties of amino acid residues in sequence. We have used twenty sequence based features to identify hub behaviour. Linear discriminant analysis and normalised Bayesian approach were utilized for identifying hub proteins solely using these sequence features in *E. coli*/*H. sapiens* datasets with accuracies of 99.5/98.6, 87.8/89.6 and 90.1/92.6, respectively.

Background:
Proteins, the work horse molecules of the cellular machinery, are accountable for a broad range of cellular functions. Proteins mostly function through their interactions with other proteins. Such protein-protein interactions are responsible for mediating vast majority of protein chores in living cells. A group of proteins and their interactions form a protein-protein interaction network (PPIN). In a PPIN, a node denotes a protein and a connecting edge represents a protein-protein interaction. Total number of interactions of a protein is its connectivity. Most proteins interact with very few other proteins while a relatively small number of proteins have a very large number of interacting partners **[1]**. Proteins with large number of interactions are called hubs and they literally 'hold the protein interaction networks together' **[2]**. Hub proteins are known to have high density of binding sites **[3]**, which enable multiple bindings. Due to their interactions with large number of other proteins and thus possible involvements in multiple pathways, study of hub proteins assumes critical importance **[3]**.

Previous studies have shown the relationship between the degree of connectivity of proteins in PPIN and their cellular essentiality properties **[4]**. When a hub node is deleted, it is more likely to be lethal to the organism than the deletion of those nodes which are connected with less number of proteins in a protein-protein interaction network **[3]**. Hub characterization is highly crucial for understanding cellular functions as well as identifying novel drug targets. Functions of many proteins are unknown and hence the identification of the physical interactions of these proteins could give an indication of their functions. Several well-known and extensively studied proteins including p53, p27, p21 and many others which are implicated in diseases are hubs **[1]**. Knowledge of the pathways, topologies and dynamics, of hub proteins should provide useful information for predicting side effects in drug discovery **[5]**.

There have been attempts to identify hub proteins in protein-protein interaction networks using various data such as gene proximity **[6, 7]**, gene fusion events **[8, 9]**, gene co-expression data **[10, 11]**, phylogenetic profiling **[12]**, interacting protein domains **[13, 14]** and gene ontology **[4]**. Most of such computational predictions have been centered on the identification of pairwise protein-protein interactions with varying degrees of accuracies. The lack of availability of the above data for the entire protein interaction network of an organism is a limiting factor in applying the methods generally. For instance, the hub classifier proposed by Michel Hsing *et al.* using gene ontology terms gives 84.96% accuracy, 34.41%

sensitivity and 90.27% specificity **[4]**. However, in order to predict whether a target protein is hub or not, this method requires gene ontology annotation of the target protein. The authors observe that "the performance of hub classifier will primarily rely on the number of gene ontology annotations available for each species"**[4]**. They relate the low sensitivity, to the lack of gene ontology annotations for certain proteins in their training sets. Generality of existing methods which use structural information are also severely limited as PDB structures and functional classifications are not available for many of the proteins.

To overcome the limitations of availability of structural and ontology data which are slow in emergence, we have chosen to investigate whether hubness can be predicted from amino acid sequence information alone. Studying structural and functional phenomena from sequence information is not a new approach and has been widely used with the advent of bioinformatics approaches in genomics and proteomics. The author's research group itself has been applying this approach to various problems from gene finding **[15]** to protein subcellular localization **[16]** to protein allostery prediction **[17]**. The approach is of course a reductionist approach. Though the underlying hypothesis, that sequence information can predict structural and functional behaviors, is not yet completely accepted, the approach remains as a viable alternative to the data dependent methods at this point of time. We report here the extraction of twenty features based on amino acid sequence information which we have used in designing a hub prediction tool based on soft computing.

Methodology:

Dataset:

Two organisms, *E. coli* and *H. sapiens*, which are well annotated and have modest interaction information, were selected for this study. IntAct **[18]** database was used to extract the protein interaction data of the model organisms. These data were then curated to obtain the Uniprot IDs and corresponding connectivities of all the proteins. This non-redundant dataset included 10,578 *H. sapiens* PPIs and 2,047 *E. coli* PPIs. Using the uniprot IDs corresponding sequences of varying lengths were compiled from Uniprot **[19]**.

Determination of Connectivity Threshold for hubs:

The degree of connectivity of proteins in our PPI dataset ranged from 1 to 450. In order to classify a protein as hub, a connectivity threshold had to be determined. Review of literature revealed that, connectivity thresholds of hub proteins are species specific **[4]**. Nevertheless, there is no consensus on the exact connectivity threshold values for these proteins **[4]**. There are studies, which have taken the thresholds based on the accumulative protein interaction distribution plots **[4]**. Some other studies base it on fold change **[2]**. We have adopted the latter approach. The connectivity fold change was computed by taking the ratio of the connectivity value and average connectivity. In the case of *E. coli*, a node with fold change greater than or equal to 2 was considered as hub (cutoff, P-value < 0.03 using distribution of standard normalized fold change values in *E. coli*). In the case of *H. sapiens*, fold change greater than or equal to 10 (with P-value < 0.001) was the criterion applied for considering a protein as hub. Summary of protein interaction data of *E. coli* and *H. sapiens* used in this study is depicted in **Table 1 (see supplementary material)**.

The datasets of both organisms were divided into two sets, train and test, for both hub and non-hub proteins. Train datasets were used to develop a model to predict hubness of proteins and test datasets were used to evaluate the reliability of the model. To ensure stringent sieving of non-hubs, we considered only those proteins which have connectivity in a range between 1 and 5 for non-hub test and train sets in *H. sapiens* data. To minimize data variances, the datasets were equally divided into training and testing sets. **Table 2 (see supplementary material)** shows the train and test sets statistics.

Amino Acid Properties:

We examined a set of 28 diverse amino acid properties (physicochemical, thermodynamic and conformational) and these properties were extracted from most commonly used amino acid index databases, AAindex **[20]** and Protscale **[21]**. These properties are shown in **Table 3 (see supplementary material)**. They were normalized between 0 and 1 using the formulae, $P\ (i)_{norm} = (P\ (i) - P_{min})\ /\ (P_{max} - P_{min})$. Where, $P\ (i)$, $P\ (i)_{norm}$ are the original and normalized values of amino acid i for a particular property, and $Pmin$ and $Pmax$ are, respectively, the minimum and maximum values. For each protein, the average amino acid property was computed as the sum of amino acid indices for all residues divided by total number of residues for each property. For a short amino acid sequence “MAEKSLAMDG” having a length of 10 amino acids, we give the computed numerical values for chosen properties in **Table 3 (see supplementary material)**.

Dimensionality reduction of feature vector:

Feasibility of the chosen 28 amino acid features for the classification model was analyzed by designing a system of linear equations for both organisms. We derived a matrix of 28×28 feature vector elements by randomly choosing 28 amino acid sequences from hub dataset of both organisms. R_H is a matrix formed with each row representing the feature vector of a hub sequence in the train data. The entry r_{ij} represents the j^{th} feature vector of the i^{th} randomly selected sequence. Consider a coefficient matrix C_H to be determined such that, $R_H\ .C_H = I$, where, C_H has dimensions 28×28, and I is a unity vector with dimension 28×1. We computed C_H as $R_H^{-1}.\ I$. We then computed the average coefficient vector, $\overline{C_H}$, for hubs by taking the average of the modulus of each column of C_H. For dimensionality reduction, we dropped the least contributing features as dictated by the coefficients in $\overline{C_H}$. The dropped coefficients are conveniently numbered 21 to 28 in Table 3. Our classification model uses the reduced feature vectors of 20 elements, identified as mentioned above, since these are most contributing to classification vectors of hubness.

Hub classification model:

We developed a classification model based on linear discriminant analysis in combination with normal Bayesian approach **[22].** The twenty selected amino acid properties are the backbone of the classification model. Each protein sequence was encoded using all these features and we compiled a feature matrix for entire dataset of both organisms. We took 100 train data for both hub and non-hub and produced a matrix of

100×20 feature vectors, R_H and R_N, for hub and non-hub datasets respectively. Then the mean vectors $\overline{\mu_1}$ and $\overline{\mu_2}$ of each train dataset were calculated. The global mean, μ_G, was computed as the average of mean vectors $\overline{\mu_1}$ and $\overline{\mu_2}$. Then the mean corrected data for $\overline{R_H}$ and $\overline{R_N}$ were computed by subtracting μ_G from each row of R_H and R_N. The covariance matrix for each mean corrected data of hub dataset was calculated as, $\overline{C_H} = \overline{R_H}^T \times \overline{R_H}^T / N_H$. Similarly, for non-hub dataset, covariance matrix for each mean corrected data was calculated as, $\overline{C_N} = \overline{R_N}^T \times \overline{R_N}^T / N_N$. In the above equations, N_H and N_N are the total number of training data of hub and non-hub, which are 100 for each. Further, we generated the pooled covariance matrix, where element P_{ij}, is obtained as, $\overline{P_{ij}} = (N_H / (N_H + N_N) \times \overline{C_{Hij}}) + (N_N / (N_H + N_N) \times \overline{C_{Nij}})$

We applied the linear discriminant analysis formula for hubs as, $f_h = \mu_1 P^{-1} x^T - 1/2\ \mu_1 P^{-1} \mu_1^T + ln(P_i)$ where, μ_1 is the mean of hub train data, P is the pooled covariance , x^T is the transpose of the feature vector of the target data x and P_i is the prior probability which is imputed as 50% for both groups. Similarly for non-hub train data, the formula is, $f_n = \mu_2\ P^{-1}\ x^T - 1/2\ \mu_2\ C^{-1}\ \mu_2^T + ln(P_i)$ where, μ_2 is the mean of non-hub train data. In these formulae, the second terms, $\mu_1 P^{-1} \mu_1^T$ and $\mu_2 P^{-1} \mu_2^T$ are actually Mahalanobis distances, which is the distance to measure dissimilarity between several groups **[22]**. If $f_h > f_n$, the target data x will be assigned to hub category, otherwise to non-hub category.

Self-consistency test, jackknife test and independent data test were performed to evaluate the classification model. Same datasets are used for training and testing in self-consistency test. Hence the classification accuracy will be high. If the self-consistency of a method is good, it can be considered as a good classification method. In jackknife test, each protein in the training set is pulled out to make classification using the rest of the training set. Jackknife is considered as more objective and exhaustive than other tests. In independent test, different sets of training and testing datasets, which were randomly picked, were used for hub classification **[16]**. In independent test, we partitioned training and testing sets equally to minimize data discrepancies. We have used different measures to assess the accuracy of classifying hub and non-hub proteins. The formulae used are, Sensitivity = True Positive / (True Positive + False Negative); Specificity = True Negative/ (True Negative + False Positive); Accuracy = (True Positive +True Negative)/ (True Positive+ False Positive + True Negative + False Negative) where, True Positives are hubs identified as hubs, False Positives are non-hubs identified as hubs, True Negatives are non-hubs identified as non-hubs, and False Negatives are hubs identified as non-hubs, respectively.

Discussion:

Our results show that meaningful amino acid features can produce signature features for differentiating hubs from non-hubs. For different performance tests including self-consistency, jackknife and independent data tests, our hub classifier gave comparable accuracies and the results are shown in **Table 4 (see supplementary material)**. Part (a) of the table reports the results on the test data summarized in **Table 2 (see supplementary material)**. Part (b) of the table reports the results on a broader interaction data from APID **[23]**, which consists of 12,053 sequences for *H. sapiens* and 2,997 sequences for *E. coli*. For the original test data which is depicted in **Table 2 (see supplementary material)**, the best accuracies, sensitivities and specificities in different performance tests are close to 90%. For the broader APID dataset the best results in different performance tests are close to 87%, which is yet impressive. Since the two different datasets gave prediction results on a par, we anticipate that the tool would be useful to provide strong hypothesis on hubness of proteins. The beta version of our tool is available online at http://hubcentre.in/.

The biological significance of the selected amino acid properties in this work for hubness identification are yet to be explained, even though there are some results which match with our extracted feature list **[24]**. It would be interesting to investigate the significance of these properties in the formation of PPINs. Recent works have indicated basic flaw in the concept of high connectivity hubs in PPINs **[25]**. Chung-Jung Tsai suggests that hubs could be multi conformation proteins and each conformation is to be considered as a separate molecule **[25]**. Lack of comprehensive structure data prevents us from testing this hypothesis at this point of time.

Acknowledgement:

Funding from Department of Information Technology, Govt. of India (DIT/R&D/B10/15(23)2008, dated 07/09/2010), is acknowledged. Athmaja S. acknowledges the internship support from State Inter-University Centre for Bioinformatics, University of Kerala.

References:

[1] Patil A *et al. Int J Mol Sci.* 2010 **11**: 1930 [PMID: 20480050]
[2] Tun K *et al. Syst Synth Biol.* 2009 **2**: 75 [PMID: 19399641]
[3] He X & Zhang J. *PLoS Genet.* 2006 **2**: e88 [PMID: 16751849]
[4] Hsing M *et al. BMC Syst Biol.* 2008 **2**: 80 [PMID: 18796161]
[5] Keskin O *et al. Chem Rev.* 2008 **108**: 1225 [PMID: 18355092]
[6] Dandekar T *et al. Trends Biochem Sci.* 1998 **23**: 324 [PMID: 9787636]
[7] Overbeek R *et al. Proc Natl Acad Sci U S A.* 1999 **96**: 2896 [PMID: 10077608]
[8] Marcotte EM *et al. Science* 1999 **285**: 751 [PMID: 10427000]
[9] Enright AJ *et al. Nature* 1999 **402**: 86 [PMID: 10573422]
[10] Ge H *et al. Nat Genet.* 2001 **29**: 482 [PMID: 11694880]
[11] Pellegrini M *et al. Proc Natl Acad Sci U S A.* 1999 **96**: 4285 [PMID: 10200254]
[12] Matthews LR *et al. Genome Res.* 2001 **11**: 2120 [PMID: 11731503]
[13] Reiss DJ & Schwikowski B. *Bioinformatics* 2004 **20**: i274 [PMID: 15262809]
[14] Qi Y *et al. Proteins* 2006 **63**: 490 [PMID: 16450363]
[15] Achuthsankar SN & Sreenadhan SP. *Journal of the Computer Society of India* 2006 **60**: 66
[16] Cherian BS & Nair AS. *Biochem Biophys Res Commun.* 2009 **391**: 1670 [PMID: 20036215]
[17] Saritha Namboodiri *et al. Syst Synth Biol.* 2011 **271**: 280
[18] http://www.ebi.ac.uk/intact/main.xhtml
[19] http://www.uniprot.org
[20] http://www.genome.jp/aaindex
[21] http://expasy.org/tools/protscale.html

[22] http://people.revoledu.com/kardi/tutorial/LDA/

[23] http://bioinfow.dep.usal.es/apid/index.htm

[24] Mahalekshmi T & Nair AS. *International Journal of Bioinformatics* 2009 **70**: 80

[25] Tsai CJ *et al*. *Trends Biochem Sci*. 2009 **34**: 594 [PMID: 19837592]

Edited by P. Kangueane

Citation: Latha *et al*. Bioinformation 7(4): 163-168 (2011)

Supplementary material:

Table 1: Protein-Protein Interaction dataset compiled from databases intact **[18]** and Uniprot **[19]**.

Item	*E.coli*	*H.sapiens*
Total Proteins	2,047	10,578
Total Interactions	15222	53120
Average Connectivity	7.982	9.534
Fold change cutoff to decide hubness	≥2	≥10
HubThreshold	16	53

Table 2: Train and Test datasets of *E. coli* and *H. sapiens* used in the present study.

Species	**Item**	**Train**		**Test**		**Total**
		Hub	**Non-hub**	**Hub**	**Non-hub**	
E.coli	Total Proteins	100	100	100	100	400
	Total Interactions	3280	197	3365	190	7032
	Min. Connectivity	16	1	16	1	----
	Max. Connectivity	139	15	132	15	----
H.sapiens	Total Proteins	100	100	100	100	400
	Total Interactions	6718	320	6957	300	14295
	Min. Connectivity	53	1	53	1	----
	Max. Connectivity	450	5	448	5	----

Table 3: Amino acid indices compiled from AA index [20] and Protscale [21] for feature vector formation in the current study. The extreme right column shows the computed values of the properties for the sample sequence "MAEKSLAMDG". (The row values are shown, they are used after normalization.)

Sl.No.	**Amino acid Properties**	**Computed numerical values for sample sequence**
1	Free energy of transfer to surface	0.45
2	Coil confirmation	0.93
3	Relative mutability	86.12
4	Hydrogen bond donors	0.51
5	Alpha helix index	1.23
6	Beta strand index	0.86
7	A.A Composition	2.25
8	Hydrophobicity index	-0.09
9	VanderWaals parameter	5.71
10	Refractivity	12.76
11	Molecular Weight	0.07
12	Electron_ion interaction poteintial	0.05
13	Reduced distance	0.98
14	Recognition factor	83.81
15	Bulkiness	13.07
16	Transmembrane Index	-0.49
17	Atomic weight	0.47
18	Flexibility	1.01
19	Polarity	8.73
20	Molecular weight	121.31
21	Alpha helix frequency	2.01
22	Charge transfer index	1.92
23	Beta strand frequency	1.45
24	Vander waals Volume	2.69
25	Transfer energy	4.99
26	isoelectric point	1.98
27	Absolute entropy	2.90
28	Residue Volume	6.98

Table 4: Summary of results of hubness predictor

(a) Test results on original test data of 100 amino acid sequences of *H. sapiens* and *E. coli* compiled from intact **[18]**.

Species	Performance parameters	Performance Tests		
		Self-consistency	Jackknife	independent
E.coli	Accuracy	99.5	87.8	90.1
	Sensitivity	100	88	90
	Specificity	99	89.2	90.2
H.sapiens	Accuracy	98.6	89.6	92.6
	Sensitivity	99	85.8	89.6
	Specificity	98.2	89.2	95.5

(b) Test results on a wider dataset of 12,053 sequences for *H. sapiens* and 2,997 sequences for *E. coli* compiled from APID **[23]**.

Species	Performance parameters	Performance Tests		
		Self-consistency	Jackknife	independent
E.coli	Accuracy	88	87.3	87.8
	Sensitivity	84.5	88.4	87.1
	Specificity	91	87.2	86.8
H.sapiens	Accuracy	87	88	90
	Sensitivity	86.6	88.8	89.4
	Specificity	87.3	86.2	90