Bioinformation
www.bioinformation.net
open access
Hypothesis

Predicting highly-connected hubs in protein interaction networks by QSAR and biological data descriptors

Michael Hsing[1], Kendall Byler [2], Artem Cherkasov[2*]

[1]Bioinformatics Graduate Program, Faculty of Graduate Studies, University of British Columbia. 100-570 West 7th Avenue. Vancouver, BC, Canada. V5T 4S6; [2]Division of Infectious Diseases, Department of Medicine, Faculty of Medicine, University of British Columbia. D 452 HP, VGH. 2733 Heather Street. Vancouver, BC, Canada. V5Z 3J5; Artem Cherkasov - E-mail: artc@interchange.ubc.ca;
*Corresponding author

Received September 30, 2009; Accepted October 13, 2009; Published October 15, 2009

Abstract:
Hub proteins (those engaged in most physical interactions in a protein interaction network (PIN) have recently gained much research interest due to their essential role in mediating cellular processes and their potential therapeutic value. It is straightforward to identify hubs if the underlying PIN is experimentally determined; however, theoretical hub prediction remains a very challenging task, as physicochemical properties that differentiate hubs from less connected proteins remain mostly uncharacterized. To adequately distinguish hubs from non-hub proteins we have utilized over 1300 protein descriptors, some of which represent QSAR (quantitative structure-activity relationship) parameters, and some reflect sequence-derived characteristics of proteins including domain composition and functional annotations. Those protein descriptors, together with available protein interaction data have been processed by a machine learning method (boosting trees) and resulted in the development of hub classifiers that are capable of predicting highly interacting proteins for four model organisms: *Escherichia coli*, *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Homo sapiens*. More importantly, through the analyses of the most relevant protein descriptors, we are able to demonstrate that hub proteins not only share certain common physicochemical and structural characteristics that make them different from non-hub counterparts, but they also exhibit species-specific characteristics that should be taken into account when analyzing different PINs. The developed prediction models can be used for determining highly interacting proteins in the four studied species to assist future proteomics experiments and PIN analyses.

Availability: The source code and executable program of the hub classifier are available for download at: http://www.cnbi2.ca/hub-analysis/

Keywords: QSAR; biological data; descriptors; protein interactions; network

Background:
The accumulation of vast amount of protein sequences and interaction data has not only accelerated study of cellular processes, but also revealed the underlying complexity of protein interactions. Previous protein interaction network (PIN) studies have demonstrated a characteristic distribution of interacting proteins, where the majority of them have a low number of connections, while there are few proteins that are involved in the majority of network interactions **[1]**. Those highly-connected proteins are referred as 'hubs' - the name that emphasizes a central role of such proteins in cellular processes. Naturally, such hub proteins represent attractive study objects helping our understanding of cellular interactions and promising new and intriguing opportunities for therapeutics development. Identification of high PIN interactors is straightforward when there is a large amount of interaction data available for a given proteome. However, the task of finding hubs in species that lack protein interaction information appears to be difficult. Many computational methods have been proposed for predicting pairwise protein interactions **[2]**. Those predictors demonstrate different degrees of accuracy, but none of them have actually been focused on explicit identification of protein hubs. Thus, it remains mostly unknown which structure- and/or sequence-related features can distinguish hubs from other proteins.

In our previous work **[3]**, we have constructed a classifier capable of distinguishing hubs and non-hubs with reasonable accuracy by utilizing the Gene Ontology (GO) annotations for hub prediction. Although the developed hub classifier has demonstrated suitable performance (28% sensitivity and 90% specificity), it was limited to the extent of available GO data and relied on the expert knowledge on proteins. In continuation of that work we have also had some success in employing a QSAR (quantitative structure-activity relationship) approach that utilized 75 physical and chemical descriptors to predict hub proteins in Methicillin-Resistant *Staphylococcus aureus* MRSA252 proteome **[4]**. Although the study was focused on a small subset of MRSA252 proteins, it has demonstrated the possibility of relatively accurate hub characterization based on physicochemical properties. The aim of the current study is to further improve our understanding of hub proteins through the use of a comprehensive set of 1300 protein descriptors reflecting their physicochemical properties (quantified through QSAR descriptors), as well as their domain and fold composition, cellular function and sequence similarity. We focused our efforts on determining common and distinctive features of the protein hubs in four model organisms: *Escherichia coli*, *Saccharomyces cerevisiae*, *Drosophila melanogaster* and *Homo sapiens*. Furthermore we have demonstrated the feasibility of combining numerous protein descriptors to construct boosting trees-based hub classifiers.

Methodology:
Figure 1 illustrates the entire process for the characterization and prediction of hub proteins.

Acquisition of protein-protein interaction data:
Experimental protein interaction data used in constructing hub classifiers were obtained from the IntAct database **[5]** for the following four species: *Escherichia coli* K 12, *Saccharomyces cerevisiae*, *Drosophila melanogaster*, and *Homo sapiens* (dated by Sep. 25th, 2007). Proteins were ranked based on their number of interactions within the same species, with the top 10% interactors considered as hubs (the same hub parameters in our previous studies **[3, 4]**).

Calculation of protein descriptors:
To fully characterize each protein in our dataset, over 1300 descriptors have been calculated and grouped into following five categories:

Gene Ontology (GO) annotations:
GO data **[6]** were obtained from the Uniprot Retrieval System **[7]** using unique UniProt protein accession numbers. To achieve a reasonable level in GO data hierarchy, we adapted a generic GO annotation level determined by the "GO slim" project. The resulting GO descriptor had a binary form, where '1' indicates that a protein has a certain GO annotation, while '0' indicates its absence. More details can be found in **[3]**.

Sequence conservation:
We compared the constituent proteins to 10 selected reference proteomes (2 Archaea, 4 Bacteria and 4 Eukaryota species) with protein sequences obtained from the UniProt database **[7]** and

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(4): 164-168 (2009)

RefSeq database **[8]**. The presence of similar sequences was determined based on the following BLAST criteria: e-value ≤ 10^5, similarity ≥ 50%, and alignment coverage ≥ 80%. Each of the 10 reference proteomes served as a binary protein descriptor, where '1' indicates that a given protein has a homolog in a reference proteome, while '0' indicates its absence.

Number of Pfam protein domains:
Each protein in the training set has been associated with Pfam domains **[9]** through the Uniprot Retrieval System **[7]** (dated by Oct. 1st, 2007). Descriptors were used to record the total number of Pfam domains for each protein. Among all identified Pfam domains a subset had the iPfam association representing interacting domains identified specifically from known protein complexes. The iPfam characteristics were obtained from the iPfam database **[10]** (version 20).

Protein threading scores and SCOP protein domain composition:
Descriptors in this category were derived in two steps. First, each protein sequence in the training set was compared against 8,539 PDB structural templates implemented by the THREADER 3.5 program **[11]** (threading template library obtained on Feb. 7th, 2008). The Z-score, which represents the fitness measure, was calculated for each pair of the query protein and each structural template. In the second step, each template in the library was linked to a specific SCOP domain classification number **[12]** (version 1.73). To achieve a manageable number of descriptors, the protein templates were grouped at their first (Class) and second level (Fold) levels of the SCOP classification. As a result, 8,539 PDB templates were classified into 1,105 folds resulting in 1,105 SCOP descriptors for this category.

Physicochemical properties:
A total of 75 sequence-based physical and chemical descriptors were calculated and used for training hub classifiers. Those descriptors quantify such protein properties as molecular weight, net charge, isoelectric point, hydrophobicity, surface area, solvent accessibilities, electronegativity, secondary structure composition, surface coils and flexibility, among others. The protein descriptors calculated by the QSAR approaches are described in more details in our previous work **[4]**.

Training and testing hub classifiers by boosting trees:
We have applied two rounds of training to select descriptors that were more capable of differentiating hubs and non-hubs. During both rounds each hub classifier was trained individually for each of the four species, and a four-fold cross-validation strategy was used (75% training and 25% testing). Each training and testing set had the constant hub to non-hub (1:9) ratio. The models were trained and tested by using the boosting trees method as implemented in STATISTICA version 8 **[13]**.

In the first round of training, all descriptors (~1300) were used, and the top 20% descriptors with the highest predictor importance values were selected for the second round of training. Four classifiers were built for each species (one for each of the four cross-validation samples) and compiled in the C++ language under Linux. A final consensus hub classifier was built for a given species (by a consensus voting method where a hub is predicted if it has two or more votes) and tested on the other three species, which were used as an external validation set.

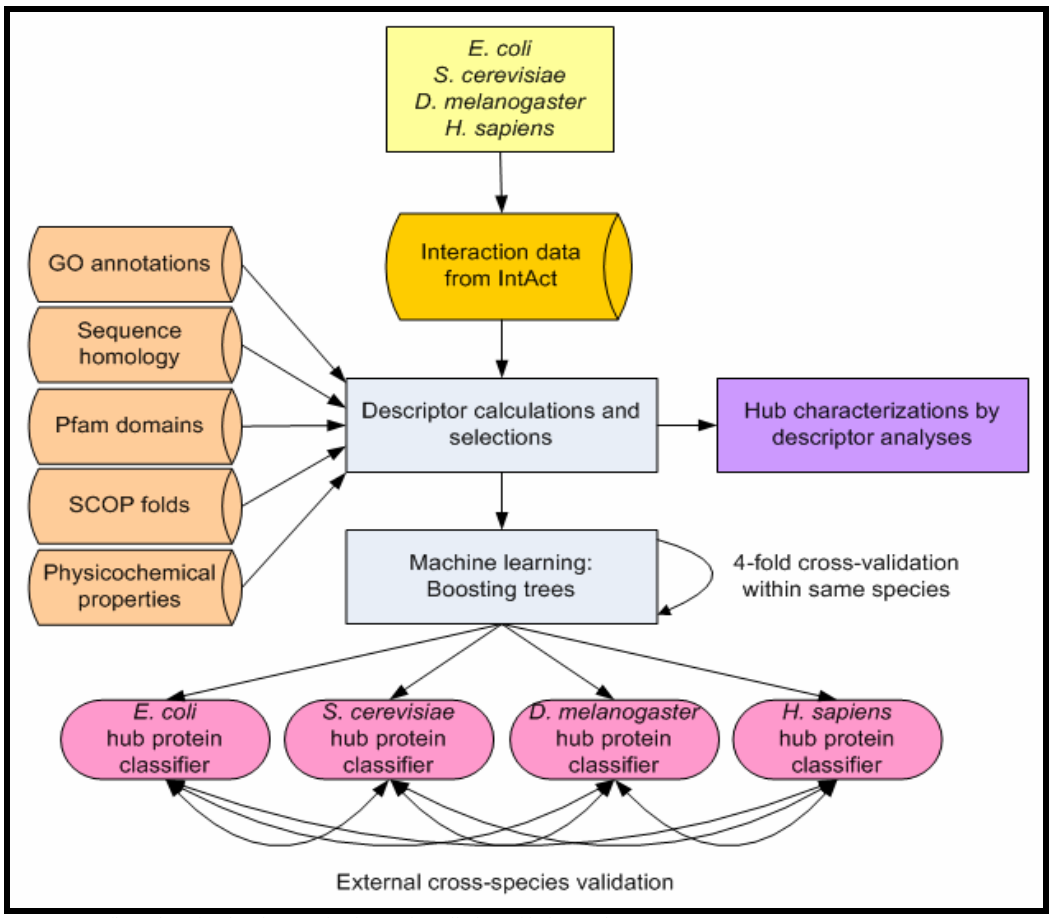

Figure 1: A flow chart for the characterization and prediction of hub proteins.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(4): 164-168 (2009)

Discussion:

Prediction performance of the hub prediction classifiers

Table 1 (see supplementary material) show the classification statistics for the hub classifiers trained and tested on *E.coli*, *S. cerevisiae*, *D. melanogaster*, and *H. sapiens* proteomes respectively. When the classifiers were tested on the same species using the four-fold cross-validation procedure, they all have shown improved prediction performance in comparison to our previous GO-based classifier. For instance, the *E. coli* hub classifier has achieved 51.40% sensitivity and 32.59% PPV and the *S. cerevisiae* hub classifier has reached 62.99% sensitivity and 33.37% PPV on the testing sets; these values are significantly higher than 28.10% sensitivity and 22.00% PPV that we have reported in of our previous work **[3]**.

However, when a classifier trained for one species is applied to another one, its prediction accuracy decreases. On one hand, these results suggest that hub proteins within the same species have enough commonality so the species-specific hub predictors demonstrated reasonable accuracy. On the other hand, the diversity among hubs from different species appears to be much greater; therefore, it is not always feasible to apply the hub classifier trained from one species to another.

Comparing hub prediction with a traditional approach based on sequence conservation

We have investigated a traditional hub prediction approach based on the idea that if a hub protein from one species is conserved in the second species, then it is likely that the conserved protein is also a hub. By examining the degree of sequence conservation in the studied data sets, we found conserved proteins do not necessarily behave like hubs in the other species. Our analysis has shown that the percentage of conserved hubs (homologous proteins that are hubs in both species) is relatively low (1.05% ~ 10.00%). Thus, such a hub prediction strategy solely based on sequence conservation has produced a very low sensitivity (0.48% ~ 11.59%). We suspect that there are two contributing factors: 1) lack of complete protein interaction data and 2) natural differences in PIN of the species that might have occurred as the result of evolutionary network rewiring.

Thus, the above results indicated that our prediction models based on the protein descriptors are more capable of determining hub proteins than the traditional sequence conservation approach.

Characteristics of hub and non-hub proteins

The utilization of the 1300 protein descriptors has enabled not only construction of hub classifiers but also allowed characterization of hub proteins in each of the four species: *E. coli*, *S. cerevisiae*, *D. melanogaster*, and *H. sapiens.*

Notably, different descriptor categories have shown different contributions to the final descriptors set. For example, only a small percentage of Gene Ontology descriptors (5~12%) and SCOP fold descriptors (17~18 %) have been selected, while Pfam descriptors (25~100%) and physicochemical parameters (60~75%) were among the most relevant ones. The sequence similarity descriptors also showed a high contribution rate. In the next sections, we report the similarities and differences between hubs and non-hubs among the four species by focusing on some of the more relevant protein descriptors.

Gene Ontology (GO) annotations

The advantage of using the GO terms in hub prediction is in availability of GO information for proteins in hundreds of species, and that they can provide useful insight into their cellular functions and locations. However, we have observed that the level of GO annotations could vary greatly among the species. For instance, among the four studied species, *S. cerevisiae* had the highest percentage of the proteins with GO annotations (87.8%), while only 48.2% of the proteins in *E. coli* could be related to a particular GO term.

Our results have indicated that certain GO terms tend to appear more frequently for hubs. For instance, the GO term, [protein binding], appears to be the most frequent one. Several others such as [cytoplasm], [protein complex], [nucleolus], [nucleobase, nucleoside, nucleotide and nucleic acid metabolic process], [nucleoplasm] and [response to endogenous stimulus] are also associated with highly interacting proteins.

Sequence similarity

In this category of the descriptors, one of the most notable observations is that hubs in general have higher sequence conservation compared to non-hubs. We have conducted two sample t-tests comparing hubs and non-hubs in terms of their occurrence of homologous proteins in the 10 reference proteomes, and the difference is significant within *E. coli*, *S. cerevisiae,* and *H. sapien*, except for *D. melanogaster*.

Number of Pfam protein domains

Previous studies have suggested that the presence of multiple interaction interfaces on a protein surface can be an indication of its hub role **[14]**. In this study, we have observed that hubs in *E. coli* and *S. cerevisiae* indeed contain more Pfam domains than non-hubs (p-values = 0). However, the difference in the number of Pfam domains between hubs and non-hubs in *D. melanogaster* and *H. sapiens* appeared to be not significant.

Protein threading scores and SCOP protein fold composition

One of the main advantages of SCOP classification as descriptors is in the general availability of these structural parameters for any protein sequence as they can be obtained easily from the threading program. It is notable that many SCOP descriptors have shown a good ability to distinguish hubs and non-hubs. However the majority of such SCOP descriptors did not follow a simple trend across all of the 4 species but instead have complex characteristics where hubs might be associated with a certain SCOP fold more frequently than non-hubs in some species but not in the others. Two notable examples are the 'PDZ domain-like' and 'beta-Grasp (ubiquitin-like)' folds. The hubs are associated with the two folds more often than the non-hubs in *E. coli* and *H. sapiens*, but not in *S. cerevisiae*, and *D. melanogaster.*

Physicochemical properties

The descriptors in this category have contributed greatly to the discriminative power of the hub classifiers due to the unbiased characterization of all the studied proteins in terms of their physical and chemical properties. One of the observed trends is that the average hydrophilicity of hubs is significantly higher than that of non-hubs in all four species. This can be explained perhaps by the fact that hubs tend to contain more polar residues on the surfaces, which facilitate protein-protein interactions.

Another protein descriptor, fraction of flexible coil residues, also demonstrated higher values for hubs in all four species. This observation agrees with previous studies **[15]** that suggested that structural disorder is a common feature of hub proteins. The hubs/non-hubs discriminative power of others physicochemical descriptors such as surface area and average surface polarizability is also high. However, the observations are more complex for these descriptors because their values are higher for hubs than non-hubs in some species, but lower in the others. It is also notable that the absolute values of average hydrophilicity of *E. coli* proteins are the smallest among the studied species. This in turn can be related to the fact that *E. coli* proteins have the smallest average surface area among the studied proteins. The above results showed that the physicochemical descriptors allowed us to derive meaningful observations, interpretable from the stand point of protein biology. In addition, these descriptors also demonstrated the complexity of hub characteristics among the different species.

Conclusion:

The use of a comprehensive set of sequence- and structure-derived descriptors has enabled in-depth characterization of hub proteins. On one hand, hubs in *E. coli*, *S. cerevisiae*, *D. melanogaster*, and *H.*

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(4): 164-168 (2009)

sapiens generally have demonstrated higher sequence conservation, tend to be more hydrophilic and possess higher fraction of flexible coil residues. On another hand, hubs have exhibited several species-specific characteristics including certain associated GO terms, occurrence of Pfam domains, more frequent SCOP folds, and surface area.

Thus, the descriptor analyses have indicated (somewhat surprisingly) that highly connected proteins are much more different across the species than previously anticipated. Thus, it has become clear that hub proteins cannot be characterized easily by only a few descriptors, but the use of a complex set of protein descriptions is required. Nonetheless, by integrating both bioinformatics and physiochemical protein descriptors, the species-specific hub classifiers have still shown improved prediction accuracy over their GO term-based and sequence similarity-based predecessors. We anticipate that the developed hub classifiers can be used for determining new hubs in the four studied species to assist future proteomics experiments and PIN analyses.

Acknowledgement:
MH was supported by the Michael Smith Foundation for Health Research (MSFHR) and the Natural Sciences and Engineering Research Council (NSERC). MH, KB and AC were funded by Genome Canada and Genome BC through the PRoteomics for Emerging PAthogen REsponse (PREPARE) project.

Reference:

[1] AL Barabasi *et al.*, *Nat Rev Genet*. **5:** 101 (2004) [PMID:14735121]
[2] Y Qi *et al.*, *Proteins*. **63:** 490 (2006) [PMID:16450363]
[3] M Hsing *et al.*, *BMC Syst Biol*. **2:** 80 (2008) [PMID:18796161]
[4] K Byler *et al.*, *QSAR & Combinatorial Science* **28:** 509 (2009)
[5] H Hermjakob *et al.*, *Nucleic Acids Res*. **32:** D452 (2004) [PMID:14681455]
[6] E Camon *et al.*, *Nucleic Acids Res*. **32:** D262 (2004) [PMID:14681408]
[7] UniProt Consortium, *Nucleic Acids Res*. (2009) **37:** D169 [PMID: 18836194]
[8] KD Pruitt *et al.*, *Nucleic Acids Res*. **35:** D61 (2007) [PMID:17130148]
[9] RD Finn *et al.*, *Nucleic Acids Res*. **36:** D281 (2008) [PMID:18039703]
[10] RD Finn *et al.*, *Bioinformatics* **21:** 410 (2005) [PMID:15353450]
[11] http://bioinf.cs.ucl.ac.uk/threader/
[12] AG Murzin *et al.*, *J Mol Biol*. **247:** 536 (1995) [PMID:7723011]
[13] http://www.statsoft.com/
[14] PM Kim *et al., Science* **314:** 1938 (2006) [PMID:17185604]
[15] C Haynes *et al.*, *PLoS Comput Biol*. **2:** e100 (2006) [PMID:16884331]

Edited by P. Kangueane
Citation: Hsing *et al.*, Bioinformation 4(4): 164-168 (2009)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(4): 164-168 (2009)

Supplementary material:

Table 1: Prediction performance of hub classifiers based on the 4-fold cross validation testing.

	sensitivity	specificity	accuracy	PPV	NPV
E. coli	51.40%	88.19%	84.51%	32.59%	94.23%
S. cerevisiae	62.99%	86.16%	83.86%	33.37%	95.49%
D. melanogaster	41.24%	83.86%	80.00%	20.28%	93.48%
H. sapiens	26.61%	88.78%	82.93%	19.76%	92.10%

Sensitivity = TP / (TP + FN), Specificity = TN / (TN + FP), Accuracy = (TP + TN) / (TP + TN + FP + FN), PPV (Positive Predictive Value) = TP / (TP + FP), NPV (Negative Predictive Value) = TN / (TN + FN) where TP = number of true positive, FP = number of false positive, TN = number of true negative, and FN = number of false negative.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 4(4): 164-168 (2009)