BIOINFORMATICS ORIGINAL PAPER Vol. 23 no. 9 2007, pages 1124–1131
doi:10.1093/bioinformatics/btm064

System biology

Clustering by common friends finds locally significant proteins mediating modules

Bill Andreopoulos[1,2,*], Aijun An[2], Xiaogang Wang[3], Michalis Faloutsos[4] and Michael Schroeder[1]

[1]Biotechnological Centre, Technische Universität Dresden, Germany, [2]Department of Computer Science and Engineering, [3]Department of Mathematics and Statistics, York University, Toronto, ON M3J1P3, Canada and [4]Department of Computer Science, University of California, Riverside, California, USA

Received on December 8, 2006; revised on January 31, 2007; accepted on February 17, 2007

Advance Access publication February 21, 2007

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Much research has been dedicated to large-scale protein interaction networks including the analysis of scale-free topologies, network modules and the relation of domain–domain to protein–protein interaction networks. Identifying locally significant proteins that mediate the function of modules is still an open problem.

Method: We use a layered clustering algorithm for interaction networks, which groups proteins by the similarity of their direct neighborhoods. We identify locally significant proteins, called mediators, which link different clusters. We apply the algorithm to a yeast network.

Results: Clusters and mediators are organized in hierarchies, where clusters are mediated by and act as mediators for other clusters. We compare the clusters and mediators to known yeast complexes and find agreement with precision of 71% and recall of 61%. We analyzed the functions, processes and locations of mediators and clusters. We found that 55% of mediators to a cluster are enriched with a set of diverse processes and locations, often related to translocation of biomolecules. Additionally, 82% of clusters are enriched with one or more functions. The important role of mediators is further corroborated by a comparatively higher degree of conservation across genomes. We illustrate the above findings with an example of membrane protein translocation from the cytoplasm to the inner nuclear membrane.

Availability: All software is freely available under Supplementary information.

Contact: williama@biotec.tu-dresden.de

Supplementary information:
http://www.cse.yorku.ca/~billa/MODULARPIN/

1 INTRODUCTION

The surge of high-throughput experiments for finding pairs of interacting proteins in a cell has led to the emergence of large protein interaction network (PIN) datasets. A PIN may contain thousands of proteins and interactions. A cluster is a set of proteins with similar interaction partners (neighborhoods), where a recorded interaction may be physical or logical (occurring through an unknown intermediary). In a cellular process, mediator proteins perform functionality that is required before mediated proteins' functionality can be carried out. Mediators can often be considered as parent-child relationships; childrens' functionality is dependent on their parent mediators, while parent mediators can function even in the absence of their childrens' functionality (Gavin *et al.*, 2006; Hollunder *et al.*, 2005; Jensen *et al.*, 2006; Ravasz *et al.*, 2002). A protein may be a mediator for a set of proteins, while it may be mediated itself by other proteins, thus resulting in hierarchical mediation. In this article, we propose clustering a PIN on the basis of neighborhood similarity, as a method for finding the mediators and modules that represent hierarchies of dependencies in a PIN. We detect a mediator protein for a cluster, as a protein that has recorded interactions with all of the cluster's member proteins. Mediators are 'locally' significant, since they are not necessarily the most highly connected proteins that affect the structure of the complete PIN, and cannot be detected based on degree.

Figure 1 shows, as example, the translocation of membrane proteins from the cytoplasm to the nucleus via nuclear pores (NUPs). Karyopherins are mobile transport receptors, also known as importins/exportins, that bind and mediate the translocation of cargoes through the NUP complexes (Marelli *et al.*, 2001). Blobel *et al.* recently elucidated some mechanisms by which karyopherins mediate translocation of membrane proteins (King *et al.*, 2006). Karyopherins mediate the translocation of membrane proteins heh1 and heh2 from the cytoplasm to the inner nuclear membrane (INM) through physical interactions with nuculear localization signal (NLS)-bearing substrates. The process of translocating the heh1 and heh2 membrane proteins is similar for both heh1 and heh2 and is mediated by the same karyopherins. *Binding sites*, such as the NLS, define the karyopherin mediators' interaction partners (Deng *et al.*, 2002; Hollunder *et al.*, 2005; Kim *et al.*, 2002; Morrison *et al.*, 2006; Wuchty, 2006). In turn, the karyopherins' functionality is dependent on Ran-GDP and Ran-GTP for binding and unbinding to membrane protein cargo, pointing to hierarchical levels of mediation (King *et al.*, 2006). Clearly,

*To whom correspondence should be addressed.

© 2007 The Author(s)
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/2.0/uk/) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

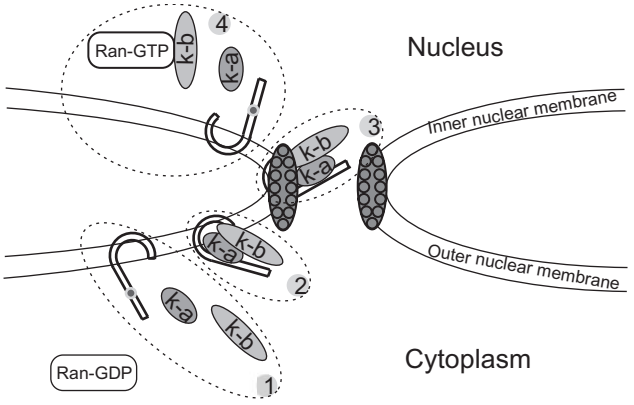

Fig. 1. The *U*-shapes represent membrane proteins, with the circles representing NLS-bearing substrates to which karyopherins bind. The translocation of membrane proteins from the cytoplasm to the inner nuclear membrane, via nuclear pores, has been shown to be mediated by karyopherin-*a* and karyopherin-*b*1 proteins. The karyopherins are in turn dependent on Ran-GDP and Ran-GTP as source of energy for binding and unbinding to membrane protein cargo. The process depends on nuclear pore organization and biogenesis (King *et al*., 2006).

the entire translocation process is dependent on the NUP organization and biogenesis. Given a yeast PIN dataset, we want to find the hierarchy of mediator proteins that are involved in membrane proteins' translocation.

In this example, a cluster of proteins may represent a module of functionality in the PIN, such as structural activity in the NUP complexes. For a cluster, there is usually at least one non-member mediator protein that plays a significant local role by interacting with all member proteins. In this example, the GTPase Ran, which coordinates the bidirectional transport of macromolecules across the nuclear envelope, would be expected to interact (physically or logically via intermediate proteins) with many of the proteins involved (King *et al*., 2006). A protein (e.g. GTPases) can mediate more than one cluster. A cluster (e.g. a functional module of structural activity) can be mediated by more than one mediator.

The main contribution of this article is to propose a novel methodology for clustering and analyzing PINs, which involves:

(a) We identify *clusters* of proteins with similar interaction partners, i.e. *neighborhoods*. Proteins in the same cluster have physical and/or logical interactions to an identical set of proteins. We show that clusters of proteins are enriched with homogeneous functional annotations and often comprise modules of functionality.

(b) We identify *mediator* proteins, where a mediator interacts with all of a cluster's protein members (according to the dataset and results). Mediators may have small degrees and may go unnoticed in other clustering approaches. We show that mediators of a cluster are enriched with diverse process and location annotations and often mediate a cluster's proteins' translocation in cellular processes and locations (Chen and Yuan, 2006; Chua *et al*., 2006; Espadaler *et al*., 2005; Okada *et al*., 2005; Pereira-Leal *et al*., 2004; Samanta and Liang, 2003; Spirin and Mirny, 2003). Mediators are more conserved than average, as shown by their orthologs across genome groups.

The rest of this article is organized as follows. Section 2 gives an overview of previous related work. Section 3 describes the PIN dataset, methods and the MULIC clustering algorithm. Section 4 discusses our experimental results. Section 5 concludes the article.

2 RELATED WORK

A PIN is an undirected graph, where the objects (nodes) represent proteins and the edges represent interactions. Successful PIN clustering applications are often based on graph theoretic techniques. Previous work has been successful in identifying clusters as tightly interacting groups of proteins. Predicting protein complexes may involve matching a cluster to a complex, such as the Restricted Neighborhood Search Clustering (RNSC) algorithm (Li *et al*., 2006). An early work on identifying protein complexes involved an application of the *k*-cores algorithm by Bader *et al*. (Bader and Hogue, 2003; Batagelj and Zavernik 2001). The *k*-core is computed by pruning all the nodes and their respective edges with degree (number of edges) less than k. That means that if a node u has degree m and it has n neighbors with degree less than k, then u's degree becomes $m-n$ and it will be also pruned if $k > m-n$. Consider a cluster of low-degree proteins $\{A, B, C\}$ that is a 2-core or 3-core, but not a 4-core, because A and B have three edges only; *k*-cores with $k=4$ cannot find this cluster. Our clustering, on the other hand, should still be able to find this cluster, if all proteins $\{A, B, C\}$ have edges to the same proteins.

Methods for predicting interactions and complexes in PINs often involve finding co-occurring domains in protein pairs believed to interact (Albrecht *et al*., 2005). Several articles have appeared on predicting protein–protein interactions based on their binding sites (Deng *et al*., 2002; Kim *et al*., 2002; Morrison *et al*., 2006; Sprinzak and Margalit, 2001; Wuchty, 2006). These methods generally associate a statistical score to the interaction probability between domains. These scores suggest which protein pairs are most likely to interact; then it is deduced that other protein pairs with these domains are likely to interact.

Similarly, Morrison *et al*. and Li *et al*. identify bipartite subgraphs in PINs, which arise from domain–domain interactions and motifs (Li *et al*., 2006; Morrison *et al*., 2006). Our approach is related: members of an *n*:*m* bipartite subgraph will result in n and m interacting members of two clusters. Our method generates 'approximate' bipartite graphs, revealing hierarchical cluster organization and modular mediation.

Ding *et al*. (2004) represent PINs based on an underlying bipartite graph model that allows generating the complex–complex association network. This representation allows viewing the PIN as consisting of protein complexes that share components.

Dunn *et al*. (2005) describe separating PINs into clusters of interconnected proteins, using Girvan and Newman's Edge-Betweenness algorithm (2002). Enriched gene ontology (GO) annotations are detected in clusters.

Yeh *et al*. (Segre *et al*., 2005; Yeh *et al*., 2006) present a clustering of drugs into functional classes on the basis of their interaction properties. Edges are colored to represent a drug pair's additive, synergistic or antagonistic effect. Then, a network of drugs is clustered into functional classes that interact with each other monochromatically.

Methods for finding functional modules in PINs often use node degrees to find dense areas (Chen and Yuan, 2006; Espadaler *et al*., 2005; Pereira-Leal *et al*., 2004; Spirin and Mirny, 2003). Some recently proposed methods predict functional modules based on how many common interaction partners two proteins share. Some of these methods have been shown to be effective in the presence of false positives (FPs) (Chua *et al*., 2006; Morrison *et al*., 2006; Okada *et al*., 2005; Samanta and Liang, 2003). We go a step beyond previous work, by examining the locally significant interaction partners (mediators) that mediate the functionality of a cluster's member proteins in processes and locations.

3 METHODS

3.1 Datasets

We used the yeast *Saccharomyces cerevisiae* PIN originating from Gavin *et al*. (2006) containing 93 881 interactions between 2551 proteins (all confidence levels included). We used the 'matrix' model for assigning binary interactions within purifications (Bader and Hogue, 2002).

We annotated yeast proteins with GO functional process and location annotations. The GO annotations were derived from the SGD yeast database (SGD Saccharomyces Genome Database).

Orthologous analyses of the annotated ORFs in the yeast genome were parsed out from the clusters of orthologous groups (COGs) of proteins (ftp://ftp.ncbi.nih.gov/pub/COG) (Tatusov *et al*., 2001; Von Mering *et al*., 2002).

3.2 Multiple layer incremental clustering

The MULIC algorithm clusters together proteins with 'similar' interaction partners. Similarity implies many interactions with the same proteins and few interactions with different proteins. The motivation is that proteins interacting with the same proteins are likely to have similar binding sites and to belong to a functional module, which may be involved in diverse cellular processes and locations (Chua *et al*., 2006; Espadaler *et al*., 2005; Okada *et al*., 2005; Pereira-Leal *et al*., 2004; Samanta and Liang, 2003; Spirin and Mirny, 2003). Similar proteins are joined according to a threshold, which starts from a stringent value and is relaxed gradually. We identify one or more mediators for a cluster, which are locally significant and interact with all cluster members (Chen and Yuan, 2006).

Let *nei*(p) denote the set of interaction partners of protein p, i.e. p's edges to other proteins. Let *nei*(C) denote the union of the interaction partners of all of cluster C's protein members. The similarity of protein p to cluster C is the size of the intersection (overlap) between *nei*(p) and *nei*(C), divided by p's degree:

$$\text{similarity}(p, C) = |nei(p) \cap nei(C)|/|nei(p)|$$

With this relative similarity metric, the significance of a protein's similarity is relative to its degree. An overlap σ of interaction partners is more significant if the protein has small degree (few interaction partners) than if it is a 'hub' with many interaction partners. With this relative similarity, small-degree unclustered proteins are likely to be clustered first. The proteins clustered first have a relative similarity that is significant and influence the clusters most.

Input: a set S of proteins and their edges;
Output: a set $\mathcal{C}$ of clusters;
Variables: L : the sorted list of proteins in S;
p : a protein to be clustered;
p_{done} : boolean for each protein p;
C, C' : clusters (sets) of proteins;
ϕ : difference between p and C;
$\delta\phi$: increment for ϕ;
$increment$: boolean to increment ϕ;
ϕ_{max} : maximally allowed difference between p and C;
Initialisation:
1. $\phi = 1$; $\delta\phi = 1$; $\phi_{max} = |S|$;
2. $L = S$; Sort proteins in L from low to high degree;
3. For all p in L do: p_{done} = false;
4. $C = \{p\}$, where p is the first protein in L; remove p from L;
5. $\mathcal{C} = \{C\}$;
Main loop:
6. While there is a p in L with $p_{done} = false$ and $\phi \leq \phi_{max}$ do:
a. $increment = true$;
b. For each protein p in L with $p_{done} = false$ do:
$i.$ $p_{done} = true$;
$ii.$ Find $C \in \mathcal{C}$ such that $\max_{C' \in \mathcal{C}} similarity(p, C')$;
$iii.$ If $difference(p, C) \leq \phi$ then:
add p to $C \in \mathcal{C}$; $increment = false$;
Otherwise add a new cluster $C' = \{p\}$ to $\mathcal{C}$;
c. For each $C = \{p\}$ do: remove C from $\mathcal{C}$; $p_{done} = false$;
d. If $increment$ then $\phi = \phi + \delta\phi$.

Fig. 2. The MULIC clustering algorithm.

Figure 2 shows the MULIC clustering algorithm, as used in our application. The algorithm reads all proteins into set S. The proteins are clustered in order from low to high degree, for the reasons described above. The first protein is inserted into a new cluster. Each iteration considers all unclustered proteins. A protein p is either inserted into the cluster C to which it has the highest similarity or into a new cluster, depending on p's and C's difference, which is the number of p's different interaction partners.

Variable ϕ represents the maximum difference allowed between p and C, which is the number of p's different interaction partners:

$$\text{difference}(p, C) = |nei(p) - nei(C)| \leq \phi$$

If p's most similar cluster C is within range ϕ then p is inserted into C, else, p is inserted into a new cluster on its own. The variable ϕ is initialized to 1 and is incremented by $\delta\phi$ whenever no protein can be placed into an existing cluster with the current ϕ value. As clusters grow, a greater ϕ value becomes more suitable because the difference of proteins to existing clusters increases. A cluster consists of one or more 'layers' formed gradually, by relaxing the criterion ϕ for inserting proteins in clusters. The cluster layer in which a protein is inserted, which depends on ϕ, represents how high the protein's difference was to the cluster when the protein was inserted into the cluster. MULIC starts by inserting as many proteins as possible in top layers—such as layer 1—and then moves to bottom layers, creating them as ϕ increases. Bottom layers, such as 1000, correspond to high ϕ values.

Clusters of size 1 should not persist through the clustering. Only clusters of size greater than 1 should persist (since a module or complex consists of more than 1 protein). At the end of each iteration through all unclustered proteins, clusters of size 1 are removed and their proteins will be re-clustered at the next iteration.

One of the advantages of MULIC is that clusters are layered and objects clustered at top layers have more similar neighborhoods. Previous algorithms often do not consider the layered structure of protein complexes, creating instead a flat clustering.

Moreover, the focus of these algorithms is often on finding the most densely connected or largest hubs of a PIN, while MULIC focuses on finding proteins with similar interaction sets. Our clustering approach's goal is to analyze local topological properties in contrast to macro properties such as skewed degree distributions, which most of the previous analysis focuses on. We want to identify structure by way of similarity, by clustering proteins that interact with the same proteins. This is in contrast to previous studies that have focused in clustering according to (a) protein degree, (b) place in the PIN hierarchy and (c) local density, e.g. clustering coefficient (Bader and Hogue, 2003; Girvan and Newman, 2002; Yang *et al.*, 2006). Given the clusters, we find mediator proteins with locally significant roles, which interact with all of a cluster's members. Such microscopic analysis could not have been possible by just examining the degree of a protein or the clustering coefficient, since a small-degree protein may be locally significant for a PIN neighborhood.

3.3 Evaluation of results

Our results' evaluation involves assessing the effectiveness of our methodology for identifying modules of functionality as well as the proteins mediating their translocation and involvement in processes.

3.3.1 Overlap of our clusters and mediators with the complexes by Gavin et al. (2006) We compared our results to the modularity results of Gavin *et al.* (2006). They partitioned proteins in yeast PIN complexes into two types: *core* components that are present in most isoforms and attachments present in only some of them. Among the attachments, sometimes two or more proteins were always together and present in multiple complexes, which they call *modules*. We evaluated the overlap of our mediators with their cores and our clusters with their modules. To evaluate this correlation, we computed the precision and recall to find how accurately our mediators match their cores and our clusters match their modules. Precision measures the extent to which each cluster matches its most similar complex. Recall measures the extent to which each cluster's objects are spread out over many complexes.

3.3.2 Annotation enrichment of clusters and mediators We evaluate the enrichment of a cluster's proteins with functional annotations using *P*-values (King *et al.*, 2004). We compare this to the enrichment of a cluster's mediators with process and location annotations. The *P*-value for a cluster of size C containing $k \leq C$ proteins with annotation X is:

$$P = 1 - \sum_{i=0}^{k-1} \frac{\binom{C}{i}\binom{G-C}{n-i}}{\binom{G}{n}}.$$

This is the likelihood that the cluster would have k or more proteins annotated with X, if the cluster's contents were drawn randomly from the set of known proteins. The size of the set of known proteins is G and contains $n \leq G$ proteins with annotation X. This *P*-value metric is also used to evaluate the enrichment of a cluster's mediators with annotation X; in this case, C is the number of mediators of the cluster, $k \leq C$ of which have annotation X.

Enrichment of a cluster or a cluster's mediators with X means a *P*-value $< 1\%$ for annotation X.

4 RESULTS AND DISCUSSION

4.1 Overlap with complexes

The MULIC algorithm produced 274 clusters on the PIN by Gavin *et al.* (2006). The cluster sizes ranged from 2 to 50 proteins. We identified 985 mediator proteins. Figure 3 shows that mediators are widely dispersed across all degree ranges. The number of proteins that have degree 1 in our dataset is 171 out of 2551 proteins in total (6.7%). Even if some of these are FPs, such a small number of FPs will not greatly affect our results. We evaluated if a cluster and its mediators overlap with modules and cores, respectively, in the yeast complexes by Gavin *et al.* (2006). We found a 71% precision and 61% recall. Clearly, there is a high correlation between our mediators clusters and the core modules. Figure 4 shows such a typical example of overlap between module cluster and core mediators.

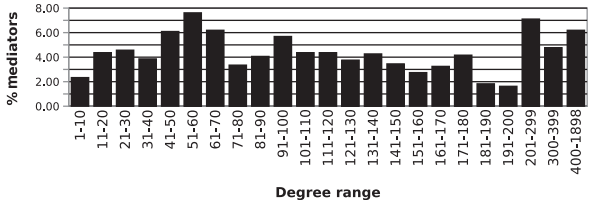

Fig. 3. Number of mediators of various degrees (Gavin *et al.*, 2006; Von Mering *et al.*, 2002).

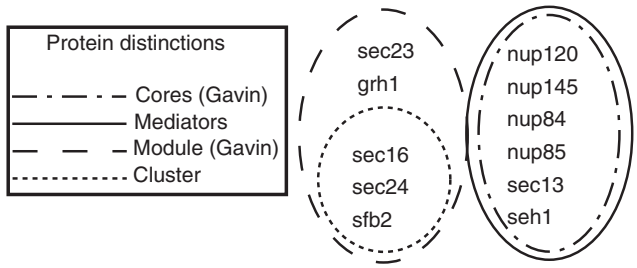

Fig. 4. Correlation of core module in (Gavin *et al.*, 2006) and mediator cluster.

We tried various $\delta\phi$ values and selected $\delta\phi = 1$. To select the best $\delta\phi$ value, we clustered the PIN originating from Von Mering *et al.* (2002) and matched the MULIC clusters to known yeast complexes from the MIPS database (Mewes *et al.*, 2002). The matching criteria, described in (King *et al.*, 2004), are roughly that a cluster should either have ~60% of its proteins overlapping with a complex of a similar size, or ~90% of its proteins overlapping with a complex of a larger size. We generated 85 clusters of minimum size 4, of which 46 matched a known MIPS complex. In comparison, King *et al.*, who used the same yeast dataset and matching criteria for this experiment, generated 28 clusters of minimum size 4, of which 23 matched MIPS complexes (King *et al.*, 2004). Bader and Hogue generated 209 predicted yeast complexes, of which 54 match the MIPS database in at least 20% of their proteins (Bader and Hogue, 2003).

We repeatedly changed the order of proteins within a degree class of the Gavin dataset, for degree classes 1–31. Then, we re-clustered the dataset and evaluated the precision and recall to the initial clustering. The clusterings were always identical.

We randomly removed $x\%$ of edges in the Gavin dataset. A protein was removed if all of its edges were removed.

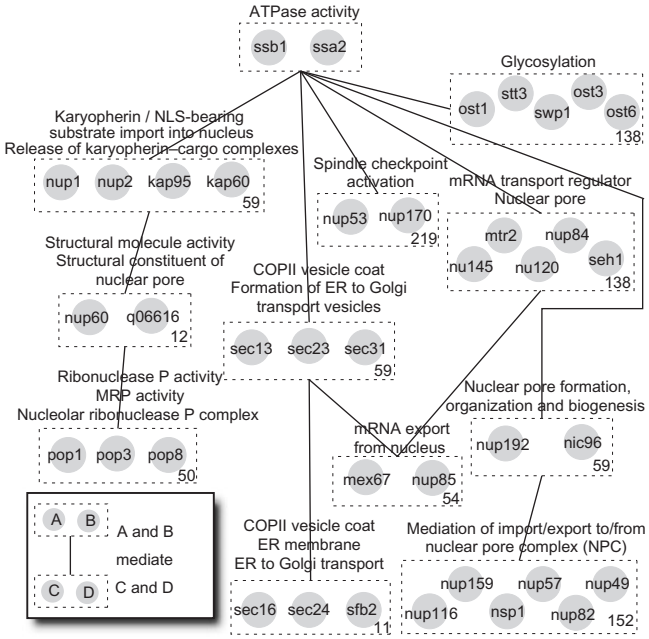

Fig. 5. Circles are proteins and dotted rectangles are clusters. The cluster numbers in Supplementary information are indicated at the lower right.

After re-clustering the dataset with missing edges, we evaluated the precision/recall to the initial clustering. The average precision/recall over 100 trials with $x = 10$ was 0.72 and 0.65. The average precision/recall over 100 trials with $x = 20$ was 0.6 and 0.51. This shows that the results are relatively reproducible even in the case of 20% missing edges. Nonetheless, the limitations of this method are obvious in light of the low connectivity and sparseness of PIN datasets.

4.2 Hierarchical modules of mediators

4.2.1 Protein translocation to nucleus Figure 5 gives a detailed example from our results. We observed a hierarchy of mediators and clusters. A cluster of proteins is often mediated by a 'parent' cluster and in turn mediates a 'child' cluster. After mapping the proteins to their annotations, we retrieve a hierarchy of functions/processes; often clusters are enriched with annotations, and parent clusters represent processes that are required for their child clusters. In other words, proteins collaborate to form modules of functionality, which in turn mediate subordinate modules; parent modules mediate the cellular involvement of child modules (Chen and Yuan, 2006; Samanta and Liang, 2003).

The most obvious processes that are described by the results involve translocation of biomolecules and transmembrane proteins between the cytoplasm and the INM, through NUPs. Translocation of membrane proteins has long been an enigma but it has recently been shown to require energy and nuclear pore complexes (NPCs) together with karyopherins are involved. This translocation is believed to be karyopherin-mediated through NLS binding sites (King *et al.*, 2006). The mediator hierarchy in Figure 5 shows proteins' reliance on receptor-mediated transport through NPCs, as described by Blobel *et al.* (2006). Several mediator dependencies are shown as involved in protein translocation between the cytoplasm and nucleus via NPCs. Next, we discuss this mediator hierarchy, and how the hierarchy's annotations are in accordance with Sacchanomyces Genome Datebase (SGD) and other literature (King *et al.*, 2006; SGD Saccharomyces Genome Database). The top-level mediators include some highly connected ATPases in the yeast PIN that are involved in releasing energy that drives chemical reactions. The proteins at the lowest levels have more granular functions.

The mediator hierarchy involving clusters 59, 12, 50 in Figure 5 shows the karyopherins' involvement in translocation. Kap95 is karyopherin-β1, which forms a dimeric complex with kap60 (karyopherin-α), and they mediate nuclear import of cargo proteins via NLS signals (King *et al.*, 2006). Nup1 and nup2 are described by SGD as involved in nucleocytoplasmic transport, release of karyopherin–cargo complexes after transport across NPC, and as binding to either the nucleoplasmic or cytoplasmic faces of the NPC depending on Ran-GTP levels; this is in accordance with the results described by Blobel *et al.* (King *et al.*, 2006; SGD Saccharomyces Genome Database). An INM protein in the cytoplasm is bound to two karyopherin proteins that act as mediators; one of these proteins recognizes the INM protein and the other probably mediates the interactions with the NPC to drive translocation of the INM protein to the nuclear side (King *et al.*, 2006). Nup60 and q06616 are NPC subunits, which anchor nup1 and nup2 to the NPC (SGD Saccharomyces Genome Database). These mediator proteins mediate the export from the nucleus of nucleus-encoded ribonuclease P complexes (pop1, pop3, pop8). The nucleus-encoded ribonuclease is transported to mitochondria for mitochondrial RNA processing, a well-known process (Chang and Clayton, 1987).

The mediator hierarchy involving clusters 59, 152 shows NPC formation, organization and biogenesis as a prerequisite for import/export to/from the nucleus. SGD describes the cluster 152 as a subcomplex of the NPC consisting of nsp1-nup57-nup49-nic96. Nup57 is an essential subunit of this NPC, functioning as the organizing center. On the other hand, nup159 is a subunit of the NPC that is found exclusively on the cytoplasmic side, forms a subcomplex with nup82 and nsp1, and is required for mRNA export. Nsp1 is an essential component of the NPC, mediating nuclear import and export. Nup82 also interacts with nup116 and is required for proper localization of nup116 in the NPC. Nup116 is a subunit of the NPC that interacts with karyopherin kap95 (King *et al.*, 2006). However, since kap95 did not interact with nup116 in the dataset we clustered, kap95 cannot mediate nup116 in our results.

The mediator hierarchy involving clusters 59, 138, 11, 54 shows the formation of endoplasmic ruticulum (ER) to Golgi transport vesicles as a prerequisite for coat protein complex II (COPII) vesicle coat and ER to Golgi transport. The sec13 protein mediates two distinct clusters involved in known processes: (a) mex67, nup85: NUP components involved in mRNA export from nucleus, and subunits of the Nup84 subcomplex, respectively. (b) sec24, sfb2, sec16: involved in ER to Golgi transport. SGD describes sec13 as important for

the formation of ER to Golgi transport vesicles, and as a component of both the Nup84 subcomplex (nup84, nup85, nup120, nup145, sec13, seh1) and of the COPII complex (sar1, sec13, sec16, sec23, sec24, sec31(web1), sfb2, sfb3). Sec13 was probably separated from nup84 complex because, as described above, sec13 is important for both the Nup84 subcomplex and the COPII complex. SGD describes mtr2 as an mRNA transport regulator, and mex67 is a NUP component that is involved in nuclear mRNA export (SGD Saccharomyces Genome Database). Mex67 and mtr2 form a nuclear mRNA export complex, which binds to RNA.

4.2.2 Mitochondrial replication Table 1 shows, as another example, the enriched annotations in the mediators-cluster pair 106 in our results. These results are in accordance with knowledge that mitochondrial replication is controlled by chromosomes in the nucleus. Proteins in cluster 106 comprise the 'nuclear origin of replication recognition complex', which is transported to the mitochondria via mediators; there the complex guides mitochondrial replication. As shown, the mediators of cluster 106 are enriched with annotations on 'mitochondrial transport' of energy-related biomolecules, such as ATP activity through the electron transport chain. Proteins in cluster 106 are additionally enriched with annotations on DNA replication and its initiation. The proteins of the mediators-cluster pair 106 have complementary roles in the nucleus and mitochondrial processes that are involved in mitochondrial replication.

4.2.3 Discussion: Hierarchical modules of mediators Analyzing a PIN dataset on the basis of common friends can be more effective than degree or density-based analysis for some purposes. The hierarchy of mediators and clusters that is shown in Figure 5 can be explained as a result of an underlying PIN organization that approximates a hierarchy of bipartite graphs. Figure 6 shows such a PIN organization. If any two proteins are connected in the PIN through a path of physical interactions, then physical interactions are conceptualized as occurring along two dimensions: x-dimension interactions connect proteins within the same module, while y-dimension interactions connect proteins between modules at different levels of the hierarchy. A direction worth pursuing as future work is to examine if such an organization is inherent to PINs, or if it also exists in other scale-free networks. If this organization is inherent to PINs, then it could help to find the FP and false negative (FN) edges that result from the matrix model of interpreting experimental results (Bader and Hogue, 2002).

Our results for the Gavin *et al*. dataset indicate that the root of the hierarchy includes the ATPases, GTPases and heat-shock proteins (HSPs). The rationale is that when considering indirect mediations of clusters, such that cluster A mediates B which in turn mediates C, we notice that these hubs mediate most of the other clusters directly or indirectly. Thus, the hierarchy of mediation can be considered as starting at the hubs.

If domain–domain interactions (DDIs) produce the PIN structure of Figure 6, then modules are produced by DDIs along dimension x, while DDIs along dimension y connect modules into a hierarchical PIN structure. As an example of a cluster involving domains, the following yeast proteins were placed in the same cluster, even though there was no apparent interaction between them in the experimental results: ADH1, EIF3B, SPAPB1E7.07. According to the SCOPPI database, six known DDIs occurred between these proteins within the cluster (see Supplementary information). These DDIs involved domains with the following related functional annotations: oxidoreductase activity, transferase activity, magnesium ion binding, purine nucleotide biosynthesis. An oxidoreductase is an enzyme that catalyzes the transfer of electrons from one molecule to another, and may be involved in magnesium ion binding.

Table 1. Enriched annotations in cluster 106 and its mediators

	P-value
Enriched annotations in cluster 106 and its mediators	
RNA ligase (ATP) activity	0.0001
Iron ion transporter activity	0.0006
Mitochondrial iron ion transport	0.0006
Mitochondrial transport	0.0006
Carrier activity	0.0006
Additional enriched annotations in cluster 106	
Pre-replicative complex formation and maintenance	1E−8
DNA replication initiation	2E−7
Nuclear origin of replication recognition complex	1.6E−11
pre-replicative complex	2.6E−8
Chromatin silencing at silent mating-type cassette	5.6E−8
Homologous chromosome segregation	0.0008
DNA replication origin binding	4.7E−10
Mannose-1-phosphate guanylyltransferase activity	0.0008

4.3 Annotation enrichment of clusters and mediators

We computed the *P*-values of the GO annotations in each of our 274 clusters. We also computed the *P*-values of the GO annotations in the set of mediators of each cluster. The *P*-values point out the enriched annotations in a cluster or a cluster's mediators, considering the annotations' distributions over all proteins (see Section 3.3.2).

Fifty five percent of the mediators are enriched with a set of diverse process and location annotations, often related to transport of biomolecules across cellular locations.

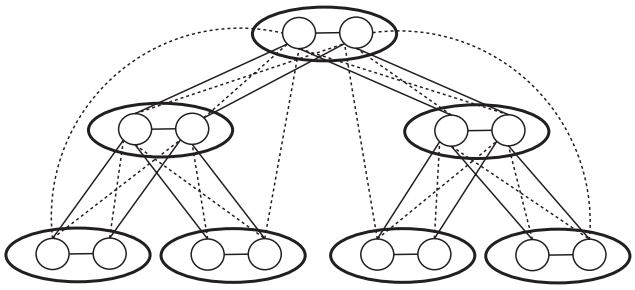

Fig. 6. A hypothetical PIN that approximates a hierarchy of bipartite graphs. Proteins (circles) are grouped (ovals) based on common interaction partners. The matrix model of interpreting experimental mass spectrometry results is known to produce FP edges (broken lines).

After replacing the mediators' granular annotations with their ancestors at GO depth 7, we found that all clusters' mediators are enriched with more than one generic process/location annotations. Complementarily, 82% of clusters are enriched with one or more functional annotations with low *P*-values, sometimes near 0.0. These clusters are enriched with functional annotations in addition to being enriched with similar process/location annotations as their mediators.

As an example, cluster 202's mediators are enriched with both processes 'endocytosis' and 'exocytosis', while cluster 202 is additionally enriched with functions 'mRNA binding' and 'RNA polymerase II transcription factor activity'. Endocytosis is often coupled with exocytosis (Batteya *et al*., 1999). Molecular mechanisms of exocytosis and endocytosis have extensively been investigated, but the coupling mechanism between these two events is unknown. Several clusters' mediators are also enriched with 'snRNA/tRNA/mRNA/rRNA export from nucleus', 'ribosomal/snRNP protein import into nucleus', 'nuclear pore organization and biogenesis' (see Supplementary information).

4.4 Cluster-mediator cross talk of processes–locations

Functions and processes are sometimes known to occur in specific cellular locations. We find the annotations that cross talk most frequently between mediators and clusters. The cross talk frequency is the number of mediator-cluster pairs in which the location annotation occurs in the mediator and the functional or process annotation in its mediated cluster. Table 2 shows some cross-talk frequencies between mediators' location annotations and clusters' functional or process annotations. Some examples of known associations that cross talk frequently include: mitochondria/fermentation, nucleus/DNA-dependent regulation of transcription, ER/glycosylation and vacuole/telomere maintenance.

4.5 Mediators and evolution

We mapped the mediator proteins in yeast to their orthologs across several groups of genomes, including *archaea*, *eukaryota*, *bacteria*, *Gram−positive* bacteria (excluding actinobacteria), Gram-positive *actinobacteria*, Gram-negative *proteobacteria* (excluding alpha and gamma), Gram-negative *alphaproteobacteria*, Gram-negative *gammaproteobacteria* and *chlamydiae bacteria*. Actinobacteria have been considered as possible ancestors for archaeans and eukaryotes.

We computed the percentage of yeast mediators that have orthologs in these groups of genomes, according to the COG orthologs database. Then, we compared this to the percentage of all yeast proteins (Gavin *et al*., 2006) that have orthologs in these groups of genomes. Figure 7 shows that for all groups of genomes, there is a higher percentage of orthologs in our set of yeast mediators than in the set of all yeast proteins. Especially with respect to ancient genome groups such as archaea and actinobacteria, the percentage of orthologs in our yeast mediators is higher than the corresponding percentage in the set of all yeast proteins. Mediator proteins, which are locally significant in the yeast PIN and are often of low degree, tend to be more evolutionarily conserved than any average protein, as suggested by their orthologs across genome groups. This may hint that mediators play a role in the evolution of functional modules (clusters).

Since mediators interact with all members of a cluster (according to the dataset), mediators may contribute to new protein additions through evolution, similar to the preferential attachment model of an evolving scale-free PIN [4]. Figure 8 illustrates a high-degree central hub and three mediator proteins; a new protein's addition to this hypothetical PIN is connected to the hub and one of the mediators. Over time, a few hub proteins will have many connections, while more locally significant mediators will have fewer connections. PIN evolution may involve a type of synergism effect, where one or more hubs cooperate with mediator proteins. Synergistic

Table 2. Frequent cross talks between mediators' location annotations and clusters' functional or process annotations

Cluster functional or process annotations	Cross talk
Mediator location: endoplasmic reticulum (ER)	
Membrane	6
Vesicle-mediated transport	6
Vesicle fusion	5
Mannosyl-oligosaccharide glucosidase activity	5
Mediator location: vacuole	
Golgi to vacuole transport	7
Protein targeting to vacuole	6
Telomere maintenance	6
Mediator location: mitochondria	
Fermentation	8
Mediator location: nucleus	
DNA-dependent regulation of transcription	18

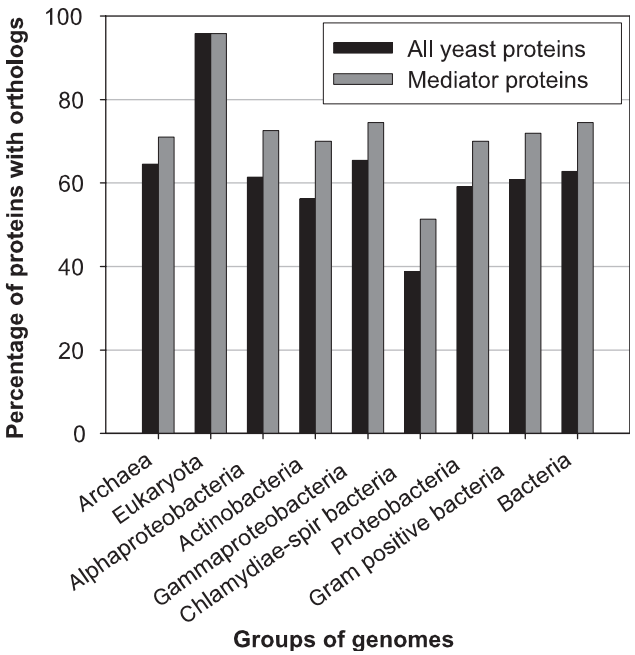

Fig. 7. For all groups of genomes, there is a higher percentage of proteins with orthologs in our set of yeast mediators than in all yeast proteins (Gavin *et al*., 2006).

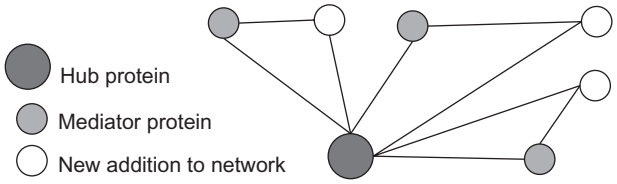

Fig. 8. A new protein connects to a mediator and a hub.

selection often means that a single new protein added to an evolving PIN is likely to be filtered out as irrelevant, while an addition of several interacting new proteins are more likely to be preserved. A PIN backbone traditionally involves hub proteins of high degree, while our set of mediators consists mostly of low-degree locally significant proteins.

5 CONCLUSIONS

Proteins with similar interaction partners often comprise a functional module, which is supported by the homogeneity of functional annotations within our clusters. Cluster involvement in processes is mediated by locally significant mediator proteins that may be of low degree. A cluster may be both a mediator and mediated itself by other cluster(s), resulting in a hierarchy of clusters and mediators. Our cluster-mediator yeast results match the modular complexes of Gavin *et al*. 2006. Mediators are more evolutionarily conserved than other proteins, as shown by orthologs across genomes.

ACKNOWLEDGEMENTS

The authors are grateful for the comments of Christof Winter, the financial support of the Natural Sciences and Engineering Research Council (NSERC), the Ontario Graduate Scholarship Program (OGS), the Canada Foundation of Innovation (CFI), the Ontario Innovation Trust (OIT) and the EU projects REWERSE and Sealife. Funding to pay the Open Access charges was provided by EU projects Sealife and REWERSE.

Conflict of Interest: none declared.

REFERENCES

Albrecht,M. *et al*. (2005) Decomposing protein networks into domain-domain interactions. *Bioinformatics*, **21**, 220–221.

Bader,G. and Hogue,C. (2002) Analyzing yeast protein-protein interaction data obtained from different sources. *Nat. Biotechnol.*, **20**, 991–997.

Bader,G. and Hogue,C. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**.

Barabasi,A.L. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.*, **5**, 101–113.

Batagelj,V. and Zavernik,M. (2001) *Cores Decomposition of Networks. Recent Trends in Graph Theory, Algebraic Combinatorics, and Graph Algorithms*, Slovenia.

Batteya,N.H. *et al*. (1999) Exocytosis and Endocytosis. *Plant Cell*, **11**, 643–660.

Chang,D. and Clayton,D. (1987) A mammalian mitochondrial RNA processing activity contains nucleus-encoded RNA. *Science*, **235**, 1178–1184.

Chen,J. and Yuan,B. (2006) Detecting functional modules in the yeast protein-protein interaction network. *Bioinformatics*, in press.

Chua,H.N. *et al*. (2006) Exploiting indirect neighbours and topological weight to predict protein function from protein-protein interactions. *Bioinformatics*, **22**, 1623–1630.

Deng,M. *et al*. (2002) Inferring domain-domain interactions from protein-protein interactions. *Genome Res.*, **12**, 1540–1548.

Ding,C. *et al*. (2004) Multi-protein complex data clustering for detecting protein interactions and functional organizations. In *Interface 2004: Computational Biology and Bioinformatics*, Baltimore, MD, USA.

Dunn,R. *et al*. (2005) The use of edge-betweenness clustering to investigate biological function in protein interaction networks. *BMC Bioinformatics*, **6**, 39.

Espadaler,J. *et al*. (2005) Detecting remotely related proteins by their interactions and sequence similarity. *PNAS*, **102**, 7151–7156.

Gavin,A.C. *et al*. (2006) Proteome survey reveals modularity of the yeast cell machinery. *Nature*, **440**, 30.

Girvan,M. and Newman,M. (2002) Community structure in social and biological networks. *PNAS*, **99**, 7821–7826.

Hollunder,J. *et al*. (2005) Identification and characterization of protein subcomplexes in yeast. *Proteomics*, **5**, 2082–2089.

Jensen,L. *et al*. (2006) Co-evolution of transcriptional and posttranslational cell cycle regulation. *Nature* , in press.

Kim,W. *et al*. (2002) Large scale statistical prediction of protein-protein interaction by potentially interacting domain pair. *Genome Inform.*, **13**, 42–50.

King,M.C. *et al*. (2006) Karyopherin-mediated import of integral inner nuclear membrane proteins. *Nature*, **442**, 1003–1007.

King,A.D. *et al*. (2004) Protein complex prediction via cost-based clustering. *Bioinformatics*, **20**, 340–348.

Li,H. *et al*. (2006) Discovering motif pairs at interaction sites from protein sequences on a proteome-wide scale. *Bioinformatics*, **22**, 989–996.

Marelli,M. *et al*. (2001) The dynamics of karyopherin-mediated nuclear transport. *Biochem. Cell Biol.*, **79**, 603–612.

Mewes,H.W. *et al*. (2002) Mips: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

Morrison,J. *et al*. (2006) A lock-and-key model for protein-protein interactions. *Bioinformatics*, **22**, 2012–2019.

Okada,K. *et al*. (2005) Accurate extraction of functional associations between proteins based on common interaction partners and common domains. *Bioinformatics*, **21**, 2043–2048.

Pereira-Leal,J.B. *et al*. (2004) Detection of functional modules from protein interaction networks. *Proteins*, **54**, 49–57.

Ravasz,E. *et al*. (2002) Hierarchical organization of modularity in metabolic networks. *Science*, **297**, 1551–1555.

Samanta,M.P. and Liang S. (2003) Predicting protein functions from redundancies in large-scale protein interaction networks. *PNAS*, **100**, 12579–12583.

Segre,D. *et al*. (2005) Modular epistasis in yeast metabolism. *Nat. genet.*, **37**, 77–83.

SGD Saccharomyces Genome Database: http://www.yeastgenome.org/.

Sharan,R. *et al*. (2005) Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J. Comput. Biol.*, **12**, 835–846.

Spirin,V. and Mirny,L.A. (2003) Protein complexes and functional modules in molecular networks. *PNAS*, **100**, 12123–12128.

Sprinzak,E. and Margalit,H. (2001) Correlated sequence-signatures as markers of protein-protein interaction. *J. Mol. Biol.*, **311**, 681–692.

Tatusov,R.L. *et al*. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.*, **29**, 22–28.

The Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **1**, 258–261.

Von Mering,C. *et al*. (2002) Comparative assessment of large-scale datasets of protein-protein interactions. *Nature*, **417**, 399–403.

Wuchty,S. (2006) Topology and weights in a protein domain interaction network. *BMC Genomics*, in press.

Yang,Q. *et al*. (2006) Evolution versus ''Intelligent Design'': *Comparing the Topology of Protein-Protein Interaction Networks to the Internet*. CSB2006, Stanford, CA, USA, pp. 299–310.

Yeh,P. *et al*. (2006) Functional classification of drugs by properties of their pairwise interactions. *Nat. Genet.*, **38**, 489–494.