Exploring protein networks with a semantic similarity measure across Gene Ontology

Zelmina Lubovac[1] Jonas Gamalielsson[1] Björn Olsson[1] Angelica Lindlöf[1]

[1] University of Skövde, School of Humanities and Informatics, P. O. Box 408, S-541 28 Skövde, Sweden

Abstract

Understanding the structure of protein interaction networks is useful as a first step towards revealing the underlying principles of the large-scale organisation of the cell. In this study, we analyse the yeast (*Saccharomyces cerevisiea*) protein-protein interaction network with a semantic similarity measure based on functional annotations from Gene Ontology. We use this measure to assess the functional relevance of modular formations in the interactome. Our results indicate the usefulness of this measure as a tool for exploring the functional similarity of protein interactions based on Gene Ontology.

Keywords: Modular organisation, protein networks, semantic similarity, gene ontology

1. Introduction

Molecular biology is facing the great challenge of explaining biological organization in the light of the potential existence of modules in various biological networks. A recent proposal that cellular function is carried out by modules [6] has fired a "modular era" of systems biology where the focus has been on studying modularity at different levels of cellular organization. A series of studies attempting to reveal the modules in cellular networks, ranging from metabolic [14] to protein networks [17][22] support the proposal that modularity is one of the design principles underlying biological organization.

However, looking for a unique set of modules might be an indefinable problem, since clearly separated networks in a cell do not exist [19]. Furthermore, even if we could devise a unique definition of modules and methods to identify them, the problem with noise in the data that is used to identify those modules still remains [12]. In the light of this problem, we agree with Poyatos and Hurst [12] that the focus of attention should be on methods for evaluation of potential modules rather than on module-defining methods.

In this work, we analyse the protein interaction network (PIN) from the Database of Interacting Proteins (DIP) with a semantic similarity measure based on the functional annotation in Gene Ontology [18]. Gene Ontology (GO) offers a vocabulary of molecular biology structured within two types of relationships between terms, namely the "is-a" and "part-of" relationships. The ontology is sub-divided into three aspects that describe a gene product: molecular function, biological process and cellular component.

Generally, the term 'module' refers to a cluster of physically or functionally connected molecules that work together to achieve a relatively distinct function [2]. Protein complexes are well-defined examples of modularity since they interact functionally and physically to form a robust unit, which, in turn, carries out some biological function [22]. In a network representation, modules appear as highly interconnected groups of nodes [2]. Nodes with a large number of links are called 'hubs', the presence of which is often related to modular structure in the network [2][8].

One of the characteristic features of a module is the functional homogeneity between the proteins that constitute the module. In this study, we use a semantic similarity measure [10] to quantify functional homogeneity of potential modules using functional annotation from GO. We apply a hub-based selection of modules, which are explored with a semantic similarity measure across the GO terms of their constituents. Semantic similarity is compared with the clustering coefficient, which measures a structural characteristic that might indicate the presence of modularity.

The semantic similarity measure of modules is based on the similarity between the central node in the module and its neighbours, and on the similarity between its connected neighbours. Those similarity values are averaged over the total number of connections in the module. However, the fact that some neighbours do not interact in the particular experiment that has been used to generate the PIN does not imply that they are not functionally related. To investigate whether we should also include the similarities between non-connected neighbours when we calculate the overall similarity score for a module, we compared the average semantic similarity between connected neighbours to the corresponding value for non-connected neighbours.

The methods presented in this study may be used not only for evaluating the function of modules

in protein interactomes, but also for suggesting potential roles of proteins with unknown functional classifications that are associated with modules.

1.1. Graph theoretic concepts

Here we describe some standard concepts that form the basis for the graph-theoretic definition of modules. A graph is generally defined as a set of points, called nodes or vertices, connected by edges. Let $G(V, E)$ represent a simple undirected graph where V is a nonempty set of nodes, and E is the set of edges connecting a subset of the nodes [16]. Two nodes u and v are adjacent if they are joined by an edge $e = \{u, v\}$ If node u is adjacent to node v, it is called a *neighbour* of v. Edge e is then called *incident* to the nodes u and v. The *degree* or connectivity of a node v in an undirected graph is the number of edges incident to v and is denoted by k_v. The *neighbourhood* $N(v)$ of a node v, consists of all neighbours of v, i.e. $N(v) = \{u \in V \mid \{u, v\} \in E\}$. By the *closed neighbourhood* $N[v]$ of node v, we mean $N(v) \cup \{v\}$. For the purpose of this work, we also define the set of edges connecting neighbours to a node v as $K(v)$ as $K(v) = \{\{a, b\} \mid \{a, v\} \in E \wedge \{b, v\} \in E \wedge \{a, b\} \in E\}$. The total number of edges that connect neighbours of node v to each other is K_v.

Recent studies focusing on complex networks have demonstrated that many real networks have a scale-free topology, which implies that the number of nodes with k links follows a power-law degree distribution, $P(k) \sim k^{-\gamma}$, where γ is degree exponent [1]. A scale-free topology of protein networks can coexist with a high clustering coefficient [20], and the presence of both these properties is a signature of potential modularity. The average clustering coefficient $\overline{C}$ for most real networks is considerably higher than for random networks of the same size [13][20]. Clustering coefficient for a node v is defined as [20]:

$$C_v = \frac{2K_v}{k_v(k_v - 1)} \qquad (1)$$

where K_v denotes the number of direct links between the k_v neighbours of node v. A high average clustering coefficient indicates the presence of modularity, i.e. the tendency of the network to form clusters.

2. Method

Information on protein interactions was downloaded from the Database of Interacting Proteins (DIP[1]) [21], which contains experimentally determined interactions between proteins in *Saccharomyces cerevisiae*. The complete set of 8063 protein-protein interactions (DIP-YEAST) that are described in DIP as of November 2001 was assessed in [3]. The majority of these interactions (≈ 6000) were identified with high-throughput yeast two-hybrid (Y2H) screens [7].

We analysed the subset of DIP-YEAST, denoted as CORE, which is the result of assessment with the Expression Profile Reliability Index (ERP Index) and the Paralogous Verification Method (PVM) (for further details, see [3]). After removing 195 self-interactions, the CORE subset contained 6 375 interactions. The total number of proteins is 2 231. To get an indication of the amount modularity in this network, we measured the average clustering coefficient $\overline{C}$ and compared it with the corresponding $\overline{C}$ for random networks. In previous work, it was found that random $\overline{C} \approx \overline{k}/N$, which here equals to 0.002. The measured $\overline{C}$ for the analysed yeast PIN is 0.34 which is significantly higher than the random $\overline{C}$. This indicates that the network possesses modular properties.

2.1. Module selection

We started by applying "hub-based" module selection, described in Box 1. A module includes a topological "hub", its neighbours, the connecting links between the hub and its neighbours, and the links between the neighbours. The term "hub" is loosely defined as a highly interconnected node [2]. In [5], hubs are defined as nodes with degree k greater than 5. For the purpose of this work, we initially set the threshold to $k > 2$. The terms "hub" and "central node" are interchangeable in this work.

procedure *Module_selection* (G: simple undirected graph)
- ▷ T_H: degree threshold for selection of hubs
- ▷ H : hub node vector
- ▷ $N[v]$: closed neighbourhood of node v
- ▷ $K(v)$: set of connecting edges between neighbours of node v
- ▷ M : module vector of node sets

$i := 0$
For each $v \in G$
  if ($k_v \geq T_H$) **then begin**
    $H_i := v$
    Identify $N[H_i]$
    Identify $K(H_i)$
    $M_i \leftarrow N[H_i] \cup K(H_i)$
    $i := i + 1$
  end

Box 1. Hub-based procedure for module selection.

2.2. Semantic similarity measure

We calculated semantic similarity with the information theoretic measure originally proposed in

[1] http://dip.doe-mbi.ucla.edu

[10] and later applied to Gene Ontology in [11]. In [11], this measure is correlated with sequence similarity, showing that a high similarity at the sequence level implies high similarity in GO annotation. We calculated similarity for modules as the average value of all hub-to-neighbour and neighbour-to-neighbour similarities. Similarity is calculated using the GO-terms assigned to the proteins.

We use the *Saccharomyces* Genome Database (SGD[2]) which contains GO annotations from all three sub-ontologies. In this work, the main focus is on the sub-ontology covering molecular function. The term-to-term relationships between nodes in the ontology denote *inheritance* or "is-a" relationships. Another important relationship between terms is the "part-of" relationship between part and whole, also called *aggregation*.

To calculate semantic similarity between gene products, the probability of each term assigned to the gene product is first derived. The probabilities reflect how many times the term or any of its descendants occurs in SGD. The procedure of calculating GO term probability, in agreement with [11], is described as follows. For each gene product in SGD, the probability is calculated by counting the number of times each term or its descendants occur in annotations divided by the total number of GO term annotations in SGD. The probability of each node increases as we move towards the root, which is defined as "molecular function" (GO:0003674) and has probability 1. Given these probabilities, there are several ways to calculate semantic similarity [9], [10], [15].

In this study, the semantic similarity between ontology terms is defined as [10]:

$$SS(t_1,t_2) = \frac{2\ln p_{ms}(t_1,t_2)}{\ln p(t_1) + \ln p(t_2)} \quad (2)$$

where $p(t_i)$ is the probability of term t_i and $p_{ms}(t_1,t_2)$ is the probability of the *minimum subsumer* for terms t_1 and t_2. The probability of the minimum subsumer for terms t_1 and t_2 is defined in [11] as the parent term with the lowest probability shared by those terms. As GO allows multiple parents for each term, two terms can share parents by multiple paths. Like [11], we used the average term-to-term similarity, since we are interested in similarity between proteins rather than ontology terms, and proteins can have several annotations.

A primary focus in this study is to investigate the functional relevance of the protein interaction sub-graphs, which motivates the use of the sub-ontology describing *molecular function*.

There are three types of semantic similarity that we calculate for each selected module. A module encloses a hub, all neighbours to the hub, the edges connecting the hub to its neighbours, as well as edges connecting pairs of neighbours to each other. Semantic similarity between a hub v and its neighbours is denoted by *SSH* and is defined as:

$$SSH_v = \frac{\sum_{\{v,u\}\in N[v]} SS_{vu}}{k_v} \quad (3)$$

where SS_{vu} is the semantic similarity between nodes v and u, $N[v]$ is the closed neighbourhood of hub v and k_v is the degree of v.

Semantic similarity between the neighbours of a hub v is denoted by *SSN* and defined as:

$$SSN_v = \frac{\sum_{\{l,j\}\in K(v)} SS_{l,j}}{K_v} \quad (4)$$

where $K(v)$ is the set of edges connecting the neighbours of v. The number of connecting edges is denoted by K_v.

Finally, we calculate semantic similarity for the whole module as an unweighted average similarity between all hub-to-neighbour and neighbour-to-neighbour similarities, i.e.:

$$SSM_v = \frac{\sum_{v,u\in N[v]} SS_{vu} + \sum_{\{l,j\}\in K(v)} SS_{l,j}}{k_v + K_v} \quad (5)$$

3. Results

3.1. The functional homogeneity of modules

As stated earlier, we aim to evaluate potential modules in terms of functional similarity. The underlying hypothesis is that a module is a functionally homogeneous unit. A starting point of the analysis is to quantify the functional homogeneity of modules based on GO-terms and compare it to the structural property of the modules measured with clustering coefficient. The purpose of the analysis is twofold: 1) inspecting the relationship between a knowledge-based (semantic similarity) and objective measure (clustering coefficient), and 2) verifying the potential of the semantic similarity measure for assessing the functional relevance of the modules.

To investigate the potential of the semantic similarity measure for modules (*SSM*), we compared this measure to the clustering coefficient, which reflects the degree of interconnectivity in the neighbourhood of each node (also called "cliquishness"). In the initial analysis, we correlated the clustering coefficient of the ten largest modules to their semantic similarity. The diagram in figure 2 shows significant correlation (correlation coefficient CC = 0.849) between these two properties, which

[2] http://genome-www.stanford.edu/Saccharomyces/

gets weaker with the decreasing connectivity. The high correlation coefficient between cliquishness and functional homogeneity indicates that the more connected the neighbours are in a particular module, the more similar are their annotations.

In addition, we inspected the relationship between all three types of semantic similarity (*SSM*, *SSH* and *SSN*) and clustering coefficient *C* across all modules. The degree threshold for module selection is set to 3, resulting in 948 modules. Further analysis is done by averaging semantic similarity over all modules whose clustering coefficient exceeds the average value ($\overline{C}$ = 0.34) and comparing them to the corresponding similarities for modules with *C* below the average.

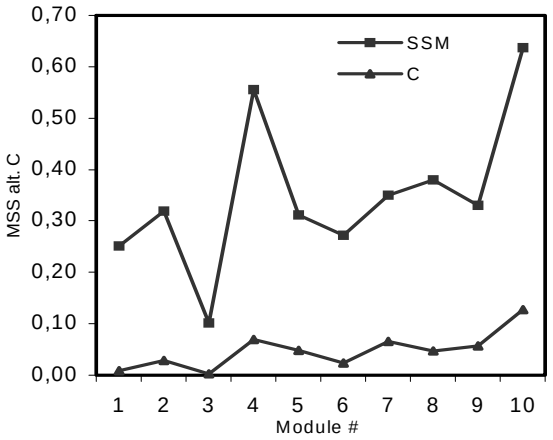

Fig. 2: Clustering coefficient *C* of the ten largest modules and their average semantic similarity (*SSM*). Correlation coefficient *CC* = 0.85.

The result can be viewed in figure 3. The average value for each of the three types of semantic similarity for the modules with *C* above the average is higher than the corresponding value for other modules. This also confirm that the more "cliquish" a module is, the more functionally related are its constituents.

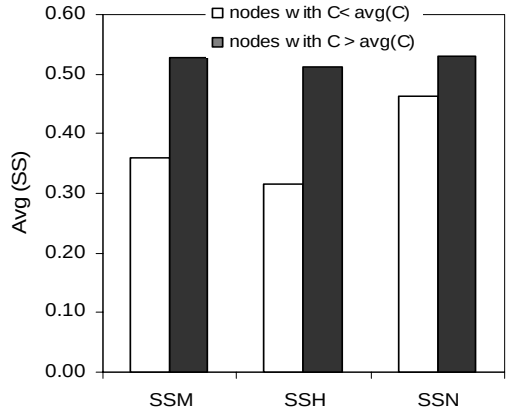

Fig. 3: Relationship between average clustering coefficient, avg (*C*), and average semantic similarities for corresponding modules. Avg (*C*) for nodes below average value is 0.19, whereas the corresponding value for nodes above average is 0.66. The average values of all three types of semantic similarity (*SSM*, *SSH* and *SSN*) for the modules with *C* above average (black bars) are higher than the corresponding similarities for the rest of the modules (white bars).

3.2. Validating semantic similarity

In this section, we aim to investigate the specificity of a semantic similarity measure as a tool for exploring and predicting the functional relevance of protein interactions. For each module, we compared the semantic similarity between the central node and its neighbours which are connected to each other (*SSN*, see equation 4) with the neighbours which lack such connections. Confirming that the *SSN* for the connected neighbours is higher than for non-connected neighbours is a first step in evaluation of the predictive power of the semantic similarity measure. We started by inspecting the difference in semantic similarities (*SSN*) for 1557 modules since we excluded the modules with degree $k <= 2$. There are 451 nodes that have no connected neighbours, which therefore have 0 or negative difference. Figure 4 shows how the minimum, maximum and average difference between *SSN* (connected) and *SSN* (non-connected) varies with the connectivity *k*. Nodes having no connecting neighbours were excluded.

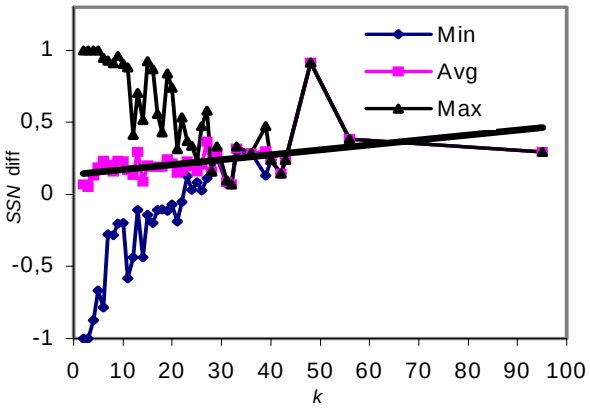

Fig. 4: Relationship between the *SSN* difference (the semantic similarity for connected neighbours vs. semantic similarity for non-connected neighbours) and the connectivity.

The difference increases with degree and increasing number of connections between neighbours to the central node. The negative difference is concentrated around the area of $k <= 20$) and mostly positive for nodes with many connected neighbours. We analysed further what this result depends on. Since we have outliers in each degree interval (negative differences), we averaged the values of *SSN* for different degrees. The averaged *SSN* difference shows clearly that connected neighbours are more

semantically similar than non-connected neighbours for the majority of modules. The difference increases with increasing degree, which depends on both increasing average *SSN* values for the connected neighbours and decreasing average *SSN* values for non-connected neighbours. This result is shown in figure 5.

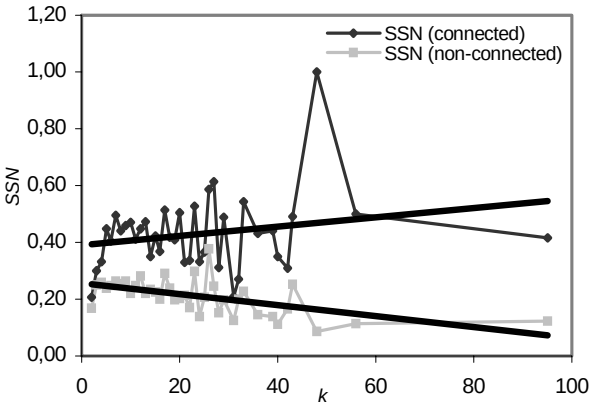

Fig. 5: *SSN* values for connected versus non-connected neighbours averaged for all degrees *k*.

Since the dataset we used here, even though it has been verified with two different methods, probably contains some false interactions, it is not yet possible to say whether we can use the semantic similarity measure to completely distinguish between protein interactions and non-interactions. However, we can state that it is better to calculate *SSN* based on connected proteins, since including absent (but potential) connections would lower the semantic similarity.

4. Discussion

The purpose of this paper is to systematically evaluate the functional homogeneity of sub-graphs in the yeast protein network by applying semantic similarity measures based on GO terms. For that purpose, we use the molecular function aspect and annotations of all kinds of evidence. We compared the semantic similarity of modules to node neighbourhood 'cliquishness' measured with clustering coefficient. The comparison of those two potential measures of modularity shows that modules with a clustering coefficient above average value tend to be more functionally related.

We also investigated the usability of semantic similarity by studying the differences in *SSN* between connected and non-connected neighbours in modules. The averaged *SSN* difference shows clearly that connected neighbours are more semantically similar than non-connected neighbours for the majority of nodes. This motivates exclusion of non-connected proteins in the calculation of average semantic similarity for potential modules.

This approach can be used to suggest potential functions of unknown gene products as well as confirm the functional relevance of derived modular formations. Since currently available PINs contain many false positives, we can use this method to improve the accuracy of the interactions.

Future work will explore the evaluation of PINs with semantic similarity focused on other aspects of GO, such as the process and cellular location sub-ontologies. We also aim to apply this method to regulatory networks with the purpose of exploring the modularity at different levels, e.g. regulatory networks versus protein networks.

Semantic similarity might be useful in different contexts when the biological plausibility of the interactions needs to be assessed. An example of such an application is the modular decomposition tool developed in [4], which is based on a graph theoretic definition of modules. By integrating the strength of the interactions in terms of functional similarity, we can achieve a more biologically plausible representation of modular decomposition.

5. References

[1] A-L. Barabási, and R. Albert, "Emergence of Scaling in Random Networks", *Science*, pp. 509-512, 1999.

[2] A-L. Barabási, and Oltvai, Z.N., "Network biology: understanding the cell's functional organization", *Genetics*, pp. 101-113, 2004.

[3] C.M. Deane, L. Salwinski, I. Xenarios, and D. Eisenberg, "Protein interactions: two methods for assessment of the reliability of high throughput observations", *Molecular and Cellular Proteomics*, pp. 349–356, 2002.

[4] J. Gagneur, R. Krause, T. Bouwmeester, and G. Casari, "Modular decomposition of protein-protein interaction networks", *Genome Biology,* 2004.

[5] J-D.J. Han, N. Bertin, T. Hao, D.S. Goldberg, G.F. Berriz, L.V. Zhang, D. Dupuy, A.J.M. Walhout, M.E. Cusick, F.P. Roth, and M. Vidal, "Evidence for dynamically organized modularity in the yeast protein-protein interaction network", *Nature*, pp. 88-93, 2004.

[6] L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, "From molecular to modular cell biology" *Nature*, pp. 47-52, 1999.

[7] T. Ito, T. Chiba, R. Ozawa, M., Yoshida, M. Hattori, and Y. Sakaki, "A comprehensive two-hybrid analysis to explore the yeast protein interactome", *Proceedings of the National Academy of Sciences of the USA*, pp. 4569-4574, 2001.

[8] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A-L. Barabási, "The large-scale organization of metabolic networks", *Nature*, pp. 651-654, 2000.

[9] J.J. Jiang, and D.W. Conrath, "Semantic similarity based on corpus statistics and lexical taxonomy", *Proceedings of International Conference on Research in Computational Linguistics*, pp. 19-33, 1998.
[10] D. Lin, "An information-theoretic definition of similarity", *Proceedings of the 15th international conference on machine learning*, pp. 296-304, 1998.
[11] P.W. Lord, R.D. Stevens, A. Brass, C.A. Goble, "Investigating semantic similarity measures across the gene ontology: the relationship between sequence and annotation", *Bioinformatics*, pp. 1275-1283, 2003.
[12] J.F. Poyatos, and L.D. Hurst, "How biologically relevant are interaction-based modules in protein networks?", *Genome Biologi*, R93, 2004.
[13] E. Ravasz,.and A-L. Barabási, "Hierarchical organization in complex networks", *Physical review*, pp. 1-7, 2003.
[14] E. Ravasz,. A.L. Somera, D.A. Mongru, Z.N. Oltvai, and A-L. Barabási, "Hierarchical Organization of Modularity in Metabolic Networks", *Science*, pp. 1551-1555, 2002.
[15] P. Resnik, "Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language", *Journal of Artificial Intelligence Research*, pp. 95-130, 1999.
[16] K.H. Rosen, *Discrete mathematics and its applications*, McGraw-Hill, 1995.
[17] V. Spirin, and L.A. Mirny, "Protein complexes and functional modules in molecular networks", *Proceedings of the National Academy of Sciences of the USA*, pp. 12123-12128, 2003.
[18] The Gene Ontology Consortium, "Creating the gene ontology resource: design and implementation", *Genome Res*., pp. 1425-1433, 2001.
[19] S. Tornow, and H.W. Mewes, "Functional modules by relating protein interaction networks and gene expression", *Nucleic Acids Research*, pp. 6283-6289, 2003.
[20] D.G. Watts, and S.H. Strogatz, "Collective dynamics of ´small worlds´networks", *Nature*, pp. 440-442, 1998.
[21] I. Xenarios, D.W. Rice, L. Salwinski, M.K. Baron, E.M. Marcotte, and D. Eisenberg, "DIP: The Database of Interacting Proteins", *Nucleic Acids Research*, pp. 289-91, 2000.
[22] S-H. Yook, Z.N. Oltvai, and A-L. Barabási, "Functional and topological characterization of protein interaction networks", *Proteomics*, pp. 928-942, 2004.