Dynamically organised modularity in protein interaction networks

Sumeet Agarwal[*]
Supervisors: Charlotte Deane[*], Nick Jones[*], Mason Porter[*]

October 2, 2008

Abstract

A key question in modern biology is how the apparent complexity of protein interaction networks relates to biological functionality. One way of understanding the set of proteins and their interactions (known as the interactome) is to look at them as a network of nodes connected by links. By studying the structure of this network, we may hope to learn something about the interactome's organisation. However, it is important to note that this interaction network is not static or uniform throughout the organism: at different times, and under different physiological conditions, it varies substantially, as expression levels of proteins vary. In this project our goal was to use gene expression data to try and infer which parts of the interactome would be 'active' at a given time and place, and thus to study the dynamics of its organisation. We attempt to look at the idea of static and dynamic modules (Komurov and White, 2007), based on a partitioning of the network into communities. Our results provide some, though not very strong, evidence for these. Another aspect we focus on is the function of highly connected nodes, or hubs, in the interactome. It has been proposed that hubs fall into two classes, 'date' and 'party', and that these play a key role in the modular organization of the yeast interactome (Han et al., 2004). This classification was made on the basis of the extent to which hubs are co-expressed with their interaction partners, but was then used to impute to them specific topological roles. We attempt to use purely topological measures to examine the extent to which these hubs really fall into the roles thus attributed. Based on a study of multiple yeast and human datasets, our results suggest that there is little evidence for a clear date/party distinction, but rather hubs in the protein interaction network seem to perform a variety of roles falling along a continuum. Finally, we look at how incomplete datasets and the experimental methods used to generate interaction data may influence what we observe.

1 Introduction

Advances in molecular biology in recent years have allowed us to acquire a vast store of information about the molecules which are the primary building blocks of life on earth: proteins. We now know much about their makeup, their structural forms, the levels at which they are expressed in various situations, and their bindings and interactions. However, there remains a major disconnect between this new knowledge and the traditional study of biology, where living organisms are analysed by breaking them down into organs and organ systems and studying their respective functions. The challenge which has been receiving much attention in the last few years is that of going from the biochemistry of tens of thousands of proteins to the physiology of a relatively small number of high-level functions and processes. A key step in making this connection is to understand how groups of proteins combine to carry out various tasks. Thus, there has been a lot of interest in the study of the interactome, i.e., the set of all physical protein-protein interactions. The interactome can tell us how proteins 'talk' to each other, and how coordination amongst them comes about. Given that even a relatively simple organism like baker's yeast (*Saccharomyces cerevisiae*) is thought to have nearly 18,000 protein-protein interactions (Yu et al., 2008), it is clear that a very complex system underlies the high-level biological functionality which we observe, and understanding how it comes about is a major challenge.

From a mathematical perspective, the interactome is a graph or network, where nodes represent proteins and unweighted, undirected links between them represent binary interactions. A study of the structure and organisation of this network is likely to provide insights which will aid in forming abstractions for higher-level understanding. In particular, we have tried to look at the extent to which protein interaction networks are modular and exhibit coherent community structure (New-

[*] sumeet.agarwal@dtc.ox.ac.uk, deane@stats.ox.ac.uk, Nick.Jones@physics.ox.ac.uk, porterm@maths.ox.ac.uk

man, 2006b). The mathematical concept of modularity quantifies the extent to which the number of links falling within groups exceed the number that would be expected in an equivalent random network (Newman and Girvan, 2004). Based on this quantity, we can attempt to partition a network into subnetworks such that 'modularity' is maximized: this is one of the standard techniques for community detection. The subgroups or communities thus formed have been found to be informative with regard to functionality in a number of social and biological networks (Girvan and Newman, 2002; Fortunato and Castellano, 2008). Here we run community detection algorithms on protein interaction networks and look at whether proteins with similar functions group together. This is assessed using annotations assigned to proteins with known functions, as per the Gene Ontology (GO) database (Ashburner et al., 2000).

Another important issue is the dynamic nature of the interactome. Protein interaction networks as constructed from data obtained via techniques like yeast two-hybrid screening do not capture the fact that the actual interactions occurring *in vivo* depend on the prevailing physiological conditions. For instance, the proteins that are being actively expressed vary from tissue to tissue in the body of an organism, and also change over time. Thus, the specific parts of the interactome which are active and the organisational form it takes is determined to an extent by where and when we are looking at it: this is what we refer to as 'dynamically organised modularity'. In order to incorporate such information, mRNA expression data from microarray experiments can be used to get measures of which protein pairs are co-expressed. Han et al. (2004) used such expression data to examine the extent to which hubs (defined by them as proteins with 5 or more interactions) in the yeast interactome are co-expressed with their interaction partners. Based on the averaged Pearson correlation coefficient (PCC) of expression over all partners, they found that hubs fall into two distinct classes: those with a low average PCC (called 'date' hubs) and those with a high average PCC ('party' hubs). They inferred that these two types of hubs play different roles in the modular organization of the network, with party hubs serving to coordinate a single function performed by a community of proteins all expressed at the same time, and date hubs serving as higher-level connectors between communities which perform varying functions and are active at different times. However, the validity of the date/party hub distinction has since been debated in a sequence of papers (Batada et al., 2006; Bertin et al., 2007; Batada et al., 2007), and there appears to be no consensus on the issue yet. The key points of debate have been whether the distribution of hubs is truly bimodal, rather than following a unimodal variation, and also whether the date/party distinction originally seen was an artefact of the dataset used rather than a general property of the interactome. Different statistical tests have seemingly suggested different answers.

Here we seek to take a different approach to the hub classification problem, by seeing if we can assign these different roles to hubs purely on the basis of network topology, rather than on the basis of expression data. The rationale behind this is that the roles, by definition, are essentially topological, and so should be detectable within the topology rather than having to be inferred from additional information. Once we have partitioned the network into a set of meaningful communities, it is possible to compute statistics for how connected each hub is both within its own community and to other communities. A method for using such statistics to assign roles to nodes in a metabolic network has been described by Guimerà and Amaral (2005), and we follow the same procedure for the hubs in our networks. We then compare how these roles match up with the date/party hypothesis.

Given the expression data for a set of proteins, we can compute a 'correlation matrix' where each entry is the PCC of expression between the corresponding pair of proteins. This can then be regarded as the adjacency matrix of another protein network, in which nodes are joined by weighted links, with weights proportional to coexpression. We attempt to analyse this network in a similar way to the interaction network, by running community detection algorithms to detect any 'coexpression communities'. We also examine combinations of the interaction and correlation matrices to see if the expression data can aid in finding better modules. Our results suggest that this is not the case, and the most functionally meaningful communities seem to come from looking at the interaction network alone. However, the correlation network does provide some interesting information, in that there seems to be some evidence for a distinction between 'static' and 'dynamic' communities (Komurov and White, 2007), defined by the variance in expression of the constituent proteins over a range of conditions.

The rest of this report is organised as follows: Section 2 discusses the concept of community detection and the methods we used for this. Section 3 introduces the protein interaction data we have used and gives the results of running a community detection algorithm on different datasets. Section 4 introduces the use of protein expression data, ways of combining this with interaction data and the idea of static and dynamic communities. Section 5 explains the method of node role assignment based on network topology and examines how our results match up to a date/party hub categorisation based on partner coexpression. Section 6 presents a comparison of some extant interactome datasets based on our analysis and discusses issues with data gathering techniques. Finally, Section 7 summarizes our observations and states possible directions for future study.

2 Modularity and Communities in Networks

2.1 Defining Modularity

A network consists of individual components (nodes) amongst which there exist interactions or connections of some sort (links). Many real-world networks, such as social, information and biochemical networks, are found to divide naturally into close-knit subnetworks, which are called communities or modules. The study of algorithms for detecting communities in networks has received a lot of attention in recent years (Fortunato and Castellano, 2008).

We have a fairly clear intuitive idea of what communities should be like: groups of nodes with many links within them and only sparse connections between groups. In order to devise algorithms to detect these groups automatically, we require a mathematical formalisation of this notion. One example of a metric which has been used for this purpose is called 'modularity', defined by Newman and Girvan (2004). Supposing an unweighted network with n nodes and m links is divided into N communities, denoted $C_1, C_2, ..., C_N$. Let k_i denote the degree (number of links) of node i, and let $A_{n\times n}$ be the adjacency matrix, such that $A(i,j)$ is 1 if nodes i and j have a link between them, and 0 otherwise. Then modularity Q is given by (Newman, 2006b):

$$Q = \frac{1}{2m}\sum_{k=1}^{N}\sum_{i,j\in C_k}(A_{ij} - \frac{k_i k_j}{2m}) \qquad (1)$$

Note that $k_i k_j/2m$ is the expected number of links between nodes i and j in a network with the same degree distribution where links are placed at random. The modularity metric is thus essentially capturing how many more links there are within the specified communities than one would expect to see by chance in a network with no modular structure. However, this is under the assumption of a particular null model, where we are explicitly preserving the degree distribution in the random setting; and it is possible to assume other null models, as we will do later in Section 4.1.

2.2 Community Detection

Detecting communities in this framework is reduced to a modularity maximisation problem over the space of all possible network partitions. Since the size of this space is huge for even modestly-sized networks, finding an exact solution is in general computationally intractable (Brandes et al., 2007). However, there exist a wide range of optimisation methods that can be used to compute approximate solutions.

The approach employed by us makes use of a physical interpretation of this problem as finding the ground state of a Potts spin glass (Reichardt and Bornholdt, 2006). The nodes can be thought of as spins, with the links representing ferromagnetic interactions and lack of link corresponding to an antiferromagnetic interaction. Then, under a natural choice of parameters, finding the ground state is equal to finding the maximum modularity partitioning, with spin states corresponding to communities. Thus, we can recast it as an energy minimisation problem, and then apply an appropriate optimisation algorithm. Here we have primarily used a spectral bisection algorithm (Newman, 2006a) on the interaction data, whilst we used a mix of this and a greedy algorithm (Reid, 2008) for the runs on the correlation networks, choosing the better of the solutions found by the two in each case.

3 The Organisation of the Interactome

3.1 Protein Interaction Datasets

There are several experimental methods which have been used to gather protein interaction data. Amongst these are high throughput yeast two-hybrid (Y2H) screening (Uetz et al., 2000; Ito et al., 2001; Fromont-Racine et al., 1997; Fromont-Racine et al., 2000); affinity purification of tagged proteins followed by mass spectrometry (AP/MS) to identify associated proteins (Ho et al., 2002; Gavin et al., 2002); curation of individual protein complexes reported in the literature (Mewes et al., 2002); and *in silico* predictions based on multiple kinds of gene data (von Mering et al., 2002). None of these methods is believed to give more than a partial picture of the interactome; for instance, a recent aggregation of high-quality Y2H datasets for *S. cerevisiae*, the best studied organism, was estimated to represent only about 20% of the whole yeast binary protein interaction network (Yu et al., 2008). Choosing which datasets to use for building and analysing the network is itself a major issue, dis-

Dataset name	Species	Nodes		Links		Source
		Total	MCC	Total	MCC	
Filtered yeast interactome (FYI)	*S. cerevisiae*	1,379	778	2,493	1,798	Han et al. (2004)
Filtered high-confidence (FHC)	*S. cerevisiae*	2,559	2,233	5,991	5,750	Bertin et al. (2007)
Database of Interacting Proteins core (DIPc)	*S. cerevisiae*	2,808	2,587	6,212	6,094	http://dip.doe-mbi.ucla.edu/ (Oct. 2007 version)
Structural Interaction Network v.1 (SIN)	*S. cerevisiae*	864	205	1,241	335	Kim et al. (2006)
CCSB Human Interactome v.1 (CCSB-HI1)	*H. sapiens*	1,549	1,307	2,611	2,483	Rual et al. (2005)

Table 1: Different protein interaction datasets used in this project. MCC refers to the maximal connected component.

cussed in more detail in Section 6 below. For our analysis, we chose to work mostly with networks consisting of multiply-verified interactions, i.e., those for which evidence has been found from at least two distinct sources. These datasets have high specificity, in the sense that they are unlikely to contain many false positives, but at the same time have low sensitivity and may have lots of false negatives, i.e., missing interactions. The specific datasets we use are summarized in Table 1, and described in more detail in Appendix A.

3.2 Communities in the Interactome

We ran spectral bisection based modularity optimisation on the various interaction networks to get an idea of their community structures. Figure 1 shows the results for the maximal connected component of the filtered yeast interactome (FYI) dataset (Han et al., 2004), with nodes coloured according to community. It is apparent that the network does have substantial 'modular' structure, in the sense defined in Section 2.1, and that the algorithm has done a fairly good job of finding it.

In order to assess how well these topological communities correspond to functional organisation, we used the Gene Ontology (GO) database. GO provides a controlled vocabulary for describing genes and gene products such as proteins, with a limited set of annotation terms, and actually consists of three separate ontologies, one each for biological process, cellular component and molecular function. We computed the p-value of the most enriched GO annotation term within each community, i.e., the term whose frequency within the community is highest relative to its background frequency in the entire network. For this we used the hypergeometric distribution, which corresponds to random sampling without replacement. The extent of enrichment can be gauged by a measure known as information content (IC), which is defined as $IC = -log_{10}(p-value)$ (Resnik, 1995). The results of calculating this measure for communities detected on two of the yeast interaction datasets are summarized in Table 2; a random partition of FYI into communities with the same size distribution as the actual ones is shown for comparison.

From these results, it is clear that there is on average very significant functional enrichment within the detected communities; in particular, it is far greater than could be expected by chance. It is also evident that the IC numbers vary widely over communities, and not all of them are equally enriched. There are some relatively vague communities (i.e., no single, specific GO term describes them very well), and others that show a very high functional coherence. In particular, more detailed inspection of the community composition revealed that proteins that are part of the large and small ribosomal subunit complexes had been almost perfectly grouped together, and several other communities comprised exclusively proteins that are known to be part of a given complex. Thus, the topology of the interaction network provides a great deal of information about the functional organisation of the proteome. However, it can only give us a static picture, and we know that the interactome is dynamic. Can we gain greater insight by using expression data to capture dynamic behaviour? The next section looks at this issue.

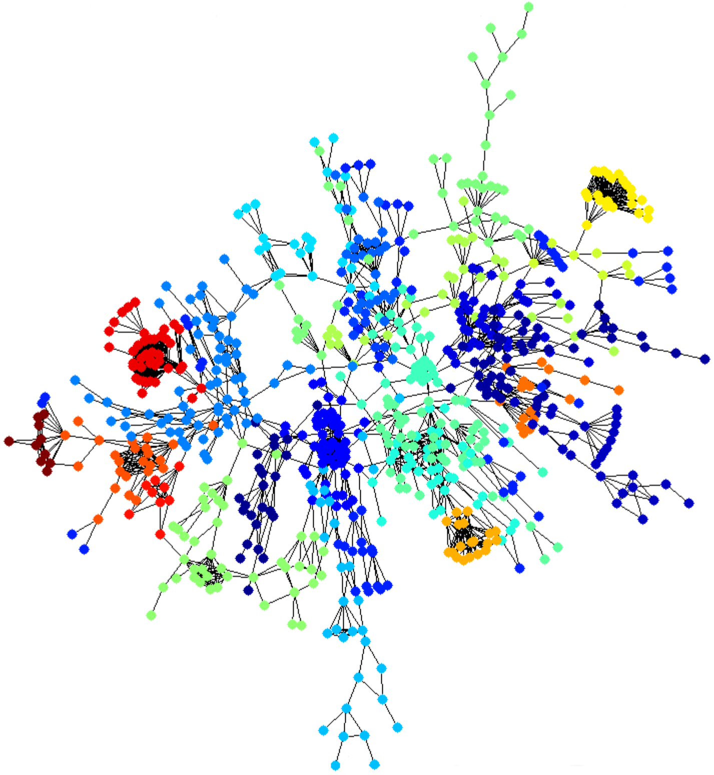


Figure 1: Community structure in the FYI network; the different colours correspond to different communities (25 in all). Visualisation generated using the Kamada-Kawai algorithm (Kamada and Kawai, 1989); code shared by Mason Porter.

Data set	**Commu-nities**	**Mol. Fn. IC**			**Cell. Comp. IC**			**Biol. Proc. IC**			**Best IC**		
		Min	Max	Avg	Min	Max	Avg	Min	Max	Avg	Min	Max	Avg
FYI	25	2.05	43.09	14.36	4.28	51.60	17.18	2.99	35.74	15.72	4.81	51.60	20.15
FYI	25 (rand.)	1.28	2.78	1.88	1.25	3.00	2.07	1.46	3.04	2.13	1.46	3.04	2.36
FHC	63	1.47	51.37	11.22	0.11	68.18	16.40	1.73	98.51	17.08	1.97	98.51	20.08

Table 2: Information Content of most enriched term for each of the 3 GO ontologies, and over all 3 ontologies ('Best IC'). Minimum, Maximum and Average are over all the communities detected in a given dataset; the random communities for FYI were generated with the same size distribution as the actual ones.

4 Interaction Dynamics

4.1 Analysing the correlation network

Different proteins are expressed differently at varying places and times, so the set of proteins and protein interactions in play varies across situations. Thus, it is important to think of the interactome as a dynamic network rather than a static one. However, most available protein interaction data does not indicate when and where those interactions actually occur *in vivo*. One way of attempting to infer this is by examining mRNA expression data, which has been gathered over a wide range of conditions via microarray experiments (Kemmeren et al., 2002). In particular, one can use such data to investigate which pairs of proteins are coexpressed, thus obtaining an extra type of information about protein association. Expression correlation was used as the basis for proposing a division of hubs into the date and party categories in the yeast interactome (Han et al., 2004).

Here, we investigated whether the expression data can provide information about community structure in any way, over and above what we see from the interaction network alone. We made use of two sources of expression data: for yeast, a set of microarray data measuring cell response to changing environmental conditions, with 174 data points for each gene (Gasch et al., 2000), and for humans, expression correlation coefficients for all pairs of proteins, computed over a range of experiments measuring expression levels, taken from COXPRESdb (Obayashi et al., 2008). For yeast too, we computed the PCC for each protein pair based on the expression dataset. Once we have a correlation value for each pair of proteins, we can model this data also as a network, albeit a weighted one. Since the interaction networks we were using had unweighted adjacency matrices with 0/1 entries, we chose to map the correlation values to this range as well by using the formula $A(i,j) = (1+PCC(i,j))/2$, the range of PCC values being between -1 and 1. Thus, we get an adjacency matrix for the 'correlation network', with a value of 1 representing perfect correlation in expression and 0 representing perfect anticorrelation. These matrices were calculated for the sets of proteins comprising the different datasets given in Table 1. The choice of our mapping is designed for the assumption that proteins with similar functionality are more likely to be coexpressed. Since it is possible that strongly anticorrelated proteins may also be linked, we also tried an alternative mapping of the form $A(i,j) = |PCC(i,j)|$, but the results obtained were of the same nature. Those presented here are for the former mapping.

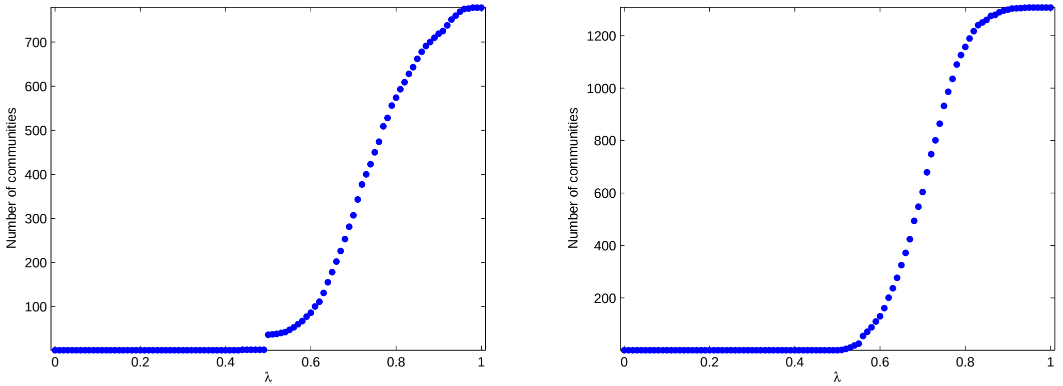
The idea of modularity has been generalised to weighted networks by counting sums of link weights rather than links, both within and between communities (Newman, 2006a). We ran the same community detection procedure on the correlation networks to see what communities we could detect in these. However, for these weighted networks, rather than preserving degree distribution as we did for the unweighted binary interaction networks, we followed Reid (2008) in choosing a uniformly connected null model, such that the quantity being maximised becomes:

$$Q = \frac{1}{2W} \sum_{k=1}^{N} \sum_{i,j \in C_k} (A_{ij} - \lambda), \tag{2}$$

where W is the sum of the weights of all links in the network, and λ is a constant between 0 and 1, meaning that our null model now is that all pairs of nodes are connected by a link of weight λ. By varying λ, we can change the resolution at which we are looking for community structure: the lower it is, the greater the tendency for nodes to clump together. In particular, $\lambda = 0$ corresponds to just one community, whereas $\lambda = 1$ will lead to each node being in its own community.

The results of running the algorithm for a range of λ values for two different datasets are depicted in Figure 2. We show the number of communities detected at each λ setting. As expected, the number increases with λ, but the notable thing is that there is apparently no long plateau in which the number of communities obtained is persistent. In fact what happens is that there is one giant community to start with, and nodes separate from it gradually, with a few nodes 'peeling' off for every increment in λ. Thus, this network by itself doesn't seem to tell us much about the proteome's functional organisation.

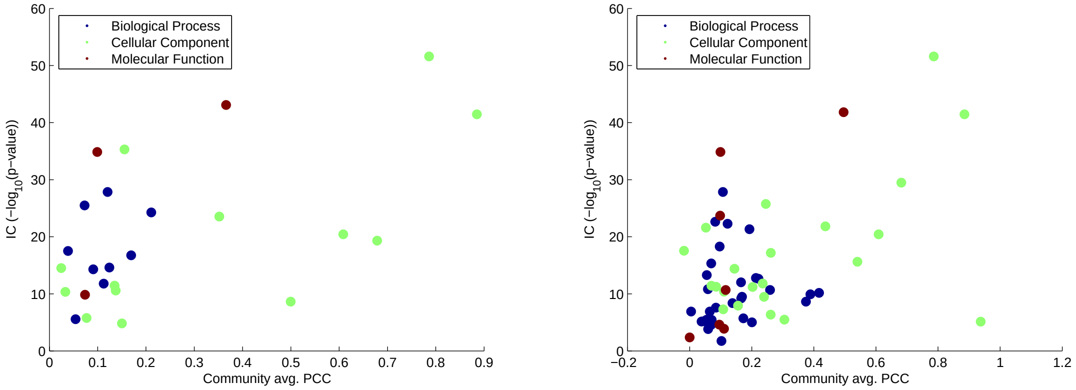
We then combined the correlation data with the interaction network. A natural way of doing so seemed to be to take a termwise product of the adjacency matrices of the two networks in order to get a new network. That is, we retain only those links in the weighted correlation network that correspond to interacting proteins, and remove all other links (i.e., set their weights to 0). When the community detection algorithm was run on this hybrid network, some community structure was seen to emerge. In order to assess its functional relevance, we computed information content (IC) based on GO term enrichment, as described in Section 3.2. The results for the hybrid network on the FYI dataset are depicted in Figure 3, along with the enrichment values for the original interaction network which were summarized in Table 2. Note that IC of the most informative GO term across all three



(a) FYI (778 nodes)

(b) CCSB-HI1 (1,307 nodes)

Figure 2: Number of communities detected at different settings of the resolution parameter, for the expression correlation networks of proteins in yeast and human datasets.



(a) FYI interaction network (25 communities, mean IC = 20.15)

(b) FYI hybrid network (59 communities, mean IC = 13.83)

Figure 3: Information content of most enriched GO term for each community versus community average PCC, computed over all pairs of nodes. Colours depict which ontology the most enriched term belongs to.

ontologies has been plotted against average community PCC, which is calculated by averaging expression PCC over all possible protein pairs in the given community. Colours depict which ontology the most informative term came from for each community.

From these results, it is apparent that the communities found from the interaction data alone are on average substantially better, in terms of functional coherence, than those detected on the hybrid network. Some of the most enriched communities are also those that have the highest coexpression amongst the constituents, and not surprisingly these are picked up well in both networks. However, many communities with large enrichment but low coexpression also seem to exist in the interaction network, and in the hybrid they appear to fragment into a larger number of less coherent subgroups. In terms of partitioning the proteome into functional modules, it seems that using interaction data alone works best. Another interesting observation from Figure 3 is that the communities with high coexpression tend to be best described by a term from the Cellular Compartment ontology, whereas those with lower coexpression are often most enriched in the Biological Process or Molecular Function ontologies. This suggests that perhaps different kinds of communities have varying expression dynamics, though this trend was much less apparent in the larger filtered high-confidence (FHC) (Bertin et al., 2007) dataset when the same analysis was repeated for it.

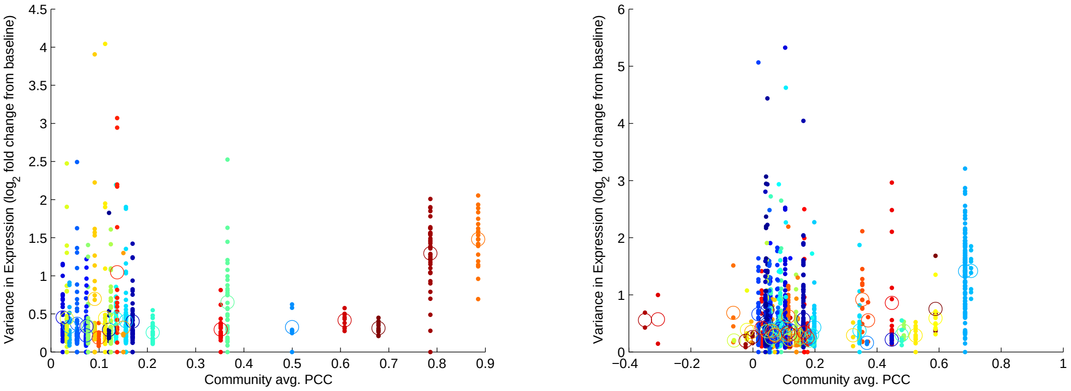
4.2 Static and Dynamic Communities

We see in Figure 3(a) that most of the structural communities have high functional coherence, but they show a wide range of average expression correlation values. The natural question that arises is how do proteins manage to interact and carry out a given function if they are not being expressed at the same time and place? One possible answer may lie in the idea of static and dynamic communities, proposed by Komurov and White (2007). They used the concept of 'expression variance' (EV), i.e., how much the expression level of a given protein changes with time and over different physiological conditions. They found that proteins tend to interact preferentially with other proteins with similar EV values, and based on this proposed that modules in the network could be broadly categorized as static or dynamic. In static modules, the proteins all have a low EV, so that their expression does not change much. In dynamic modules, all proteins show high variation in expression, and tend to be strongly co-expressed, implying that the entire community becomes 'active' only under certain conditions. Static modules have relatively low expression correlation, but they are 'on' all the time and thus still able to interact.

In order to examine if this categorisation can be applied to our communities, we used the microarray dataset for yeast to compute the variance of the expression vector for each protein (we could not do this for the human dataset, as we only had the correlation values for pairs of proteins, not the actual expression values). We then plotted these values against the communities ordered by average pairwise expression correlation. The results for two yeast datasets are shown in Figure 4. In each of the two cases, we see that a couple of communities with the highest average coexpression also have much higher expression variance on average, and seem to be 'dynamic modules'. However, there is little relationship between coexpression and expression variance for the remaining communities, and even those with low coexpression seemingly contain some 'dynamic' proteins. Of course, these may be due to mismatches between the communities detected here and the actual functional modules. On the whole, there is some evidence for a static/dynamic distinction, but it does not appear to provide a comprehensive answer. Also note that even for those communities with a seemingly low average expression PCC, the number is in most cases higher than the average PCC on the whole network. For FYI, the whole-network average is 0.0743, whilst for FHC it is 0.0492; in both cases the distribution is approximately normal. It is thus clear that there is a strong tendency for the structural communities from the interaction network to have higher expression correlation than would be expected at random, even though the value varies quite widely across communities.

5 Node Roles and the Date/Party Distinction

Interactome dynamics were first studied in the context of network hubs by Han et al. (2004). They used expression data to calculate the coexpression (PCC) of hubs (which they defined as nodes with degree 5 or more) with their interaction partners, and found that the average PCC for hubs follows a bimodal distribution in the yeast interaction network (FYI). Based on this, they proposed the existence of two kinds of hubs, *date* and *party*, and inferred certain roles for them in the network: party hubs interact with all their partners at once, and serve as coordinators of a particular task or module, whereas date hubs interact with their partners at different times and/or places, and function as higher-level organisers, linking together multiple modules. Thus, they used expression data to make inferences about the topological roles played by hub nodes in the interactome.



(a) FYI (778 nodes, 25 communities)

(b) FHC (2,233 nodes, 63 communities)

Figure 4: Expression variance of proteins by community. Each colour represents a community, and larger circles are community averages.

Given that we have found some functionally meaningful communities based on interaction data alone, and in fact that attempting to add in coexpression data seemingly makes the communities worse, it seemed prudent to examine whether something like the date/party distinction could also be observed based only on interaction data. Guimerà and Amaral (2005) have proposed a scheme for classifying nodes into roles in a modular network, according to their pattern of intra- and inter-module connections. This classification uses two statistics: within-module degree and participation coefficient. The within-module degree is normalized to a z-score; for the i^{th} node:

$$z_i = \frac{\kappa_i - \bar{\kappa}_{s_i}}{\sigma_{\kappa_{s_i}}}, \tag{3}$$

where κ_i is the number of links of node i to other nodes in the same module s_i, $\bar{\kappa_{s_i}}$ is the average of κ for all nodes in s_i and $\sigma_{\kappa_{s_i}}$ is the standard deviation of κ in s_i. The participation coefficient measures how a node's links are distributed amongst different modules. It is defined as:

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{\kappa_{is}}{k_i} \right)^2, \tag{4}$$

where N_M is the number of modules, κ_{is} is the number of links of node i to nodes in module s, and k_i is the total degree of node i. The participation coefficient approaches 1 if the links of node i are uniformly distributed amongst all modules, and is 0 if they are all within its own module.

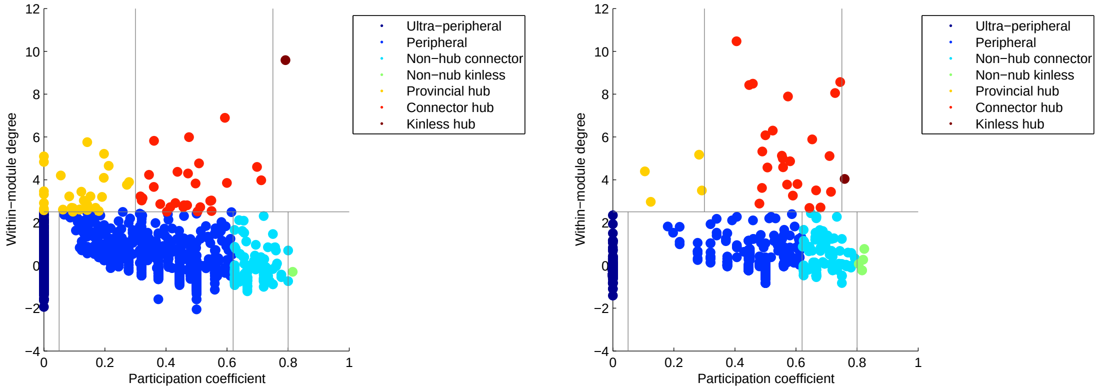
When we plot all nodes in a modular network in a two-dimensional space with their coordinates determined by the two measures above, we can divide the space into regions that correspond to node roles. The boundaries between regions are of course arbitrary, but we have used the same cut-offs given by Guimerà and Amaral (2005). They first make a distinction between 'module hubs' and 'non-hubs', defining the former as those nodes with $z \geq 2.5$. Note that the term 'hub' as used by them refers only to high within-module degree, and even their 'non-hubs' may have high overall degree. These two categories are further partitioned on the basis of participation coefficient P as follows:

- Non-hubs: ultra-peripheral nodes ($P \leq 0.05$ - virtually all links within own module), peripheral nodes ($0.05 < P \leq 0.62$ - most links within own module), non-hub connector nodes ($0.62 < P \leq 0.80$ - links to many other modules) and non-hub kinless nodes ($P > 0.80$ - links homogeneously distributed amongst all modules).

- Module hubs: provincial hubs ($P \leq 0.30$ - vast majority of links within own module), connector hubs ($0.30 < P \leq 0.75$ - many links to most other modules) and kinless hubs ($P > 0.75$ - links distributed amongst all modules).

Thus the space is divided into 7 role boxes. These are depicted in Figure 5, which shows the node roles for yeast and human datasets, based on communities detected as per the method of Section 2.

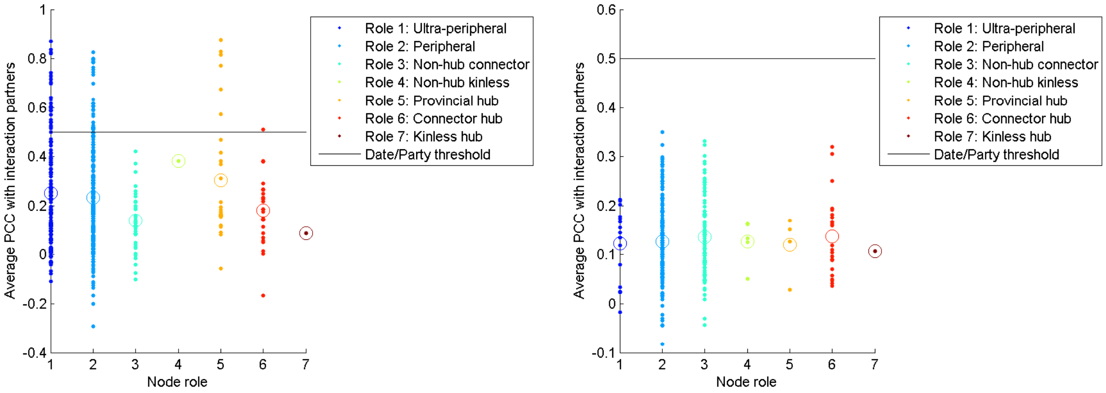
It would appear that some of the universal roles found by this method are similar to the roles ascribed to date/party hubs. For instance, party hubs should be 'provincial hubs', which have many links within their



(a) FHC (2,233 nodes, 63 communities)

(b) CCSB-HI1 (1,307 nodes, 38 communities)

Figure 5: Node role assignments for yeast and human interaction datasets.



(a) FHC (553 hubs with min. degree = 7)

(b) CCSB-HI1 (326 hubs with min. degree = 4)

Figure 6: Node role versus average expression correlation with partners, for hubs in yeast and human networks. Larger circles are averages over all nodes in a given role. Note that ‘hub’ as used in the role names refers only to within-module hubs, but all of the nodes shown here are hubs in the sense of being nodes with high degree. The minimum degree to qualify as a hub was determined so that approximately the top 20% most connected nodes in each case are hubs. The date/party PCC threshold of 0.5 was fixed for yeast by Bertin et al. (2007).

module but few or none outside. Date hubs could be 'non-hub connectors' or 'connector hubs', both of which have links to several different modules; they could also fall into the 'kinless' roles, though very few nodes are actually found in these categories. We sought to examine the relationship between the date/party classification and the universal role classification. Figure 6 plots the hubs (defined, as per Bertin et al. (2007), as the top 20% most connected nodes) in two interaction networks according to node role and average expression PCC with interaction partners. The horizontal lines correspond to an average PCC of 0.5, which was the threshold used to distinguish date and party hubs in the yeast interactome (Bertin et al., 2007).

One immediate observation from these results is that the PCC threshold clearly does not carry over to the human data; in fact all the hubs in the latter have average PCC well below 0.5. Also, in the human network there is little difference in the PCC distribution across roles, suggesting that at least for this dataset no meaningful date/party categorisation can be made. This may be because the human dataset likely represents only a small fraction of the actual interactome, and is derived from only one technique (Y2H) and therefore not multiply-verified like the yeast datasets; see further discussion of data issues in Section 6.

For yeast, we see that hubs below the threshold line (date hubs) include virtually all of those falling into the 'connector' roles, but also many of the 'provincial hubs'. On the other hand, those above the line (party hubs) include mainly the provincial hub and peripheral categories. Whilst there is a difference in role distributions above and below the threshold, it is not very clear-cut, and in particular the date hubs include nodes in all 7 roles. From these figures it would appear that even for yeast, the distribution of hubs is not bimodal (the original statistical analysis has already been disputed by Batada et al. (2006, 2007)), and the topological properties attributed to date and party hubs by Han et al. (2004) do not seem to correspond very well with their actual roles as estimated here, which are more diverse.

6 Data Incompleteness and Experimental Bias

Multiple methods have been employed to gather the data studied here, some of which were mentioned in Section 3.1. In a recent paper, Yu et al. (2008) examined the properties of interaction networks derived from different sources, and suggested that experimental bias may play a key role in determining which properties are observed in any given dataset. In particular, their findings seem to indicate that Y2H tends to detect key interactions between protein complexes and Y2H datasets contain a high proportion of date hubs, whereas AP/MS largely detects interactions within complexes and hubs in AP/MS-derived networks are predominantly party hubs.

With AP/MS, there is also the issue of converting protein complex data into interaction data. Tandem-affinity purification (TAP) involves using a 'bait' protein to 'capture' other proteins which bind to it to form complexes. Once these complexes have been obtained and the proteins in them identified via mass spectrometry, there are two ways of assigning protein-protein interactions, known as the spoke and matrix models (Hakes et al., 2007). The spoke model only counts interactions between the bait and each of the proteins captured by it, whereas the matrix model counts all possible pairwise interactions in the complex. The actual topology of the complex may well be different from both of these representations. So in terms of detecting binary interactions, Y2H is likely to be more accurate, and due to the washing steps involved in AP/MS, Y2H is also better at detecting transient interactions. On the other hand, AP/MS may be more reliable at finding permanent associations, and two-hybrid approaches also do not seem to be particularly suited for characterizing protein complexes, giving rise to the view that complex formation is more than the sum of binary interactions (Gavin et al., 2002). Thus, the two major techniques are in a sense orthogonal and cover different subspaces of the interactome, and the differences between datasets from these sources perhaps correspond mostly to false negatives rather than false positives.

In order to look at how well the properties such as communities and node roles computed here are preserved across different datasets, we compared results on four different yeast interaction datasets: FYI, FHC, DIPc and SIN (see Table 1). For each one, we examined only the largest connected component, and in pairwise comparisons, counted the number of nodes and links in common. For the overlapping portions, we computed the extent of overlap in node roles and community structure. For the latter we used a measure known as Jaccard distance. If a node is part of set A of nodes in one network and set B in the other, then Jaccard distance $J = 1-(A \cap B)/(A \cup B)$. A distance of 0 corresponds to identical communities, whereas for very different ones J approaches 1. By averaging J over all nodes, we can get an estimate of the similarity of two partitions on the same set of nodes. Table 3 presents the results of our binary comparisons of the yeast datasets.

The comparisons show that there are large variations amongst the different networks reported in the litera-

Datasets compared **(no. of nodes)**	**Common** **nodes**[1]	**Links in** **overlap**[2]	**Communities**[3] **Jaccard distance**	**Role**[3] **overlap**[4]
FYI (778) vs. FHC (2233)	687	FYI-1416; FHC-1968; Both-1154	0.76	317 (46%)
FYI (778) vs. DIPc (2587)	616	FYI-1167; DIPc-1567; Both-885	0.77	244 (40%)
FHC (2233) vs. DIPc (2587)	1373	FHC-3350; DIPc-3243; Both-2139	0.84	696 (51%)
FYI (778) vs. SIN (205)	115	FYI-95; SIN-178; Both-63	0.66	72 (63%)

Table 3: Comparisons of analysis results on different yeast interaction datasets.

[1] Proteins occurring in both networks for which expression data was also available in the Gasch et al. (2000) dataset.
[2] Links amongst the common nodes as counted in the previous column, individually in either network and common to both networks.
[3] Communities and node roles computed over entire maximal connected component in each dataset.
[4] The number of nodes with the same role classification in both networks.

ture. FYI, FHC and DIPc are all supposed to be 'high-quality' datasets, yet there are many interactions they do not share. SIN is a smaller set of structurally verified interactions, but only about a third of its links occur in FYI (63/178). These differences lead to very different community structure as well: Jaccard distance for each pairwise comparison amongst the 3 bigger networks is around 0.8, so on average the intersection of communities for the same node covers only about a fifth of their union. Since node roles are computed based on modular structure, it is not surprising that the role overlap too is not very high. Thus we cannot really make any general inferences from our results: they are largely dependent on the dataset we are looking at, which in each case represents only part of the overall picture of the interactome. However, if we are able to replicate some of our observations on a big dataset that is a union of all of these smaller ones, or independently on different sets obtained from individual sources like Y2H and AP/MS, then our confidence in them will increase; so this is something we intend to try soon.

7 Conclusions

In this project, we have tried to explore ways of combining protein interaction and expression data to analyse interactome dynamics, in particular the issues of modular organisation and the roles of hubs in the network. We showed that partitioning the interaction network based on maximising modularity leads to largely functionally coherent communities. The expression data can be used to compute pairwise expression correlations, and this can also be seen as a weighted network; however, this network does not seem to have any persistent community structure, and combining it with the interaction network also appears unhelpful. However, the communities from the interaction network do tend to show higher than average coexpression, and there is also some indication that certain communities may be dynamic, in that their components show high variance in expression levels, whereas others are static. It would be interesting to examine the expression variance over multiple expression datasets to see if we can get more evidence for a static/dynamic distinction. Also, the communities used by us are obtained using a particular algorithm and a particular null model, but there exist many others, and in order to make our results more robust it is important to see how sensitive they are to the choice of algorithm and resolution parameters.

Previous work looking at interactome dynamics had largely focused on hubs in the network. Here we used the community structure found from the interaction network to study the properties of hub nodes. Our results show that hubs are found across the entire spectrum of structural roles, and there is little to suggest a definitive date/party classification. Coexpression of a hub with its partners is not necessarily a strong predictor of its role, and coexpression properties also appear to be quite different for the yeast and human datasets we examined. Our feeling is that a date/party dichotomy is an oversimplification; in fact there appear to be many kinds of hubs with a variety of properties, and any categories we put them into will probably be largely arbitrary.

A key issue with existing interaction networks is that they are very incomplete, and we have compared some available yeast datasets and shown that they differ widely. Protein interaction data is gathered using several experimental techniques, and these appear to preferentially pick up different kinds of interactions. The datasets we used here all consisted of interactions taken from multiple sources, so it is not possible to isolate the effect of this factor on our observations. In order to do so, as a next step we would like to repeat our analyses for networks consisting of data from only one experimental source. It is also important to examine data from a number of species to come to general conclusions, although at present there is not a great deal of it available for organisms other than yeast. As the quantity, quality and diversity of protein interaction and expression datasets increases, we should be able to enhance our understanding of the organisational principles of the interactome.

Acknowledgments

Apart from my supervisors, I would like to thank Patrick Kemmeren for providing me with yeast expression datasets; Pao-Yang Chen and Waqar Ali for pointing me to relevant online data repositories; Stephen Reid for sharing community detection code; and Anna Lewis for several very useful discussions, as well as providing GO annotation data.

A Protein Interaction datasets

The interaction datasets used by us are summarized in Table 1. Below are more details about how they were compiled:

- Filtered Yeast Interactome (FYI): Compiled by Han et al. (2004). Was created by intersecting data generated by different methods, including Y2H, AP/MS, literature curation, *in silico* predictions and the MIPS (http://mips.gsf.de/) physical interactions list. Contains 2,493 interactions observed by at least two different methods.

- Filtered High-Confidence (FHC): Was generated by Bertin et al. (2007) by filtering a dataset called high-confidence (HC) compiled by Batada et al. (2006). The filtration was done by applying criteria similar to those used for FYI, to obtain 5,996 independently verified interactions. HC consisted of 9,258 interactions taken from published literature-curated and high-throughput datasets, which were also supposed to be multi-validated. However, Han et al. (2004) claimed that many interactions in HC had in fact been derived from a single experiment reported in multiple publications, and thus removed such instances from it to generate FHC.

- Database of Interacting Proteins core (DIPc): Obtained from the DIP website (http://dip.doe-mbi.ucla.edu/). DIP is a large database of protein interactions compiled from a number of sources. The 'core' subset consists only of the most reliable interactions, as judged manually by expert curators and also automatically using computational approaches.

- Structural Interaction Network version 1 (SIN): Published by Kim et al. (2006). The interaction network was compiled via a consensus from various sources, and low-confidence interactions were filtered out by statistical analysis. The remaining interactions were mapped to Pfam (http://pfam.sanger.ac.uk/) domains and thereby onto known structures of protein interactions. Only those interactions in which both partners or their homologs could be found in a 3-D structure of a protein complex were retained in SIN. The interactions were then annotated structurally and classified as 'simultaneously possible' or 'mutually exclusive', depending on the protein interfaces used. We used the complete set of SIN interactions for our analysis.

- Center for Cancer Systems Biology Human Interactome version 1 (CCSB-HI1): This was obtained by Rual et al. (2005) by means of a high-throughput yeast two-hybrid system, which was used to test pairwise interactions among the products of about 8,100 human open reading frames. Nearly 2,800 interactions were detected, and the dataset was found to have a verification rate of 78% in an independent co-affinity purification assay.

References

M. Ashburner et al. Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nature Genet.*, 25:25–29, 2000.

Nizar N. Batada et al. Stratus not altocumulus: A new view of the yeast protein interaction network. *PLoS Biology*, 4(10):1720–1731, 2006.

Nizar N. Batada et al. Still stratus not altocumulus: Further evidence against the date/party hub distinction. *PLoS Biology*, 5(6):1205–1210, 2007.

Nicolas Bertin et al. Confirmation of organized modularity in the yeast interactome. *PLoS Biology*, 5(6):1202–1205, 2007.

Ulrik Brandes et al. On finding graph clusterings with maximum modularity. In *Proc. 33rd Intl. Workshop Graph-Theoretic Concepts in Computer Science (WG'07)*, volume 4769 of *LNCS*, pages 121–132. Springer-Verlag, 2007.

Santo Fortunato and Claudio Castellano. *Encyclopedia of Complexity and System Science*, chapter Community structure in graphs. Springer, 2008.

Micheline Fromont-Racine, Jean-Christophe Rain, and Pierre Legrain. Toward a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nature Genet.*, 16(3):277–282, 1997.

Micheline Fromont-Racine et al. Genome-wide protein interaction screens reveal functional networks involving sm-like proteins. *Yeast*, 17(2):95–110, 2000.

Audrey P. Gasch et al. Genomic expression programs in the response of yeast cells to environmental changes. *Molecular Biology of the Cell*, 11:4241–4257, 2000.

Anne-Claude Gavin et al. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415: 141–147, 2002.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl Acad. Sci. USA*, 99(12):7821–7826, 2002.

Roger Guimerà and Lúis A. Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433:895–900, 2005.

Luke Hakes, David L. Robertson, Stephen G. Oliver, and Simon C. Lovell. Protein interactions from complexes: A structural perspective. *Comp. Funct. Genomics*, 2007:49356, 2007.

Jing-Dong J. Han et al. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430:88–93, 2004.

Yuen Ho et al. Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868): 180–183, 2002.

Takashi Ito et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl Acad. Sci. USA*, 98(8):4569–4574, 2001.

Tomihisa Kamada and Satoru Kawai. An algorithm for drawing general undirected graphs. *Information Processing Letters*, 31:7–15, 1989.

Patrick Kemmeren et al. Protein interaction verification and functional annotation by integrated analysis of genome-scale data. *Molecular Cell*, 9:1133–1143, 2002.

Philip M. Kim, Long J. Lu, Yu Xia, and Mark B. Gerstein. Relating three-dimensional structures to protein networks provides evolutionary insights. *Science*, 314:1938–1941, 2006.

Kakajan Komurov and Michael White. Revealing static and dynamic modular architecture of the eukaryotic protein interaction network. *Mol. Sys. Bio.*, 3:110, 2007.

H. W. Mewes et al. MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, 30(1):31–34, 2002.

M. E. J. Newman. Finding community structure in networks using the eigenvectors of matrices. *Phys. Rev. E*, 74:036104, 2006a.

M. E. J. Newman. Modularity and community structure in networks. *Proc. Natl Acad. Sci. USA*, 103(23):8577–8582, 2006b.

M. E. J. Newman and M. Girvan. Finding and evaluating community structure in networks. *Phys. Rev. E*, 69:026113, 2004.

Takeshi Obayashi, Shinpei Hayashi, Masayuki Shibaoka, Motoshi Saeki, Hiroyuki Ohta, and Kengo Kinoshita. COXPRESdb: a database of coexpressed gene networks in mammals. *Nucleic Acids Res.*, 36:77–82, 2008.

Jörg Reichardt and Stefan Bornholdt. Statistical mechanics of community detection. *Phys. Rev. E*, 74:016110, 2006.

Stephen Reid. Legislatures as spin glasses. Master's thesis, Oxford University, 2008.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. 14th Int'l Joint Conf. Artificial Intelligence*, pages 448–453, 1995.

Jean-François Rual et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–1178, 2005.

Peter Uetz et al. A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. *Nature*, 403:623–627, 2000.

Christian von Mering et al. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.

Haiyuan Yu et al. High-quality binary protein interaction map of the yeast interactome network. *Science*, Express: 10.1126/science.1158684, 2008.