Creating in Silico Interactomes

October 3, 2006

1 Introduction

Understanding protein interactions is relevant to our understanding of complex biological processes and to deciphering the different roles played by genes and proteins. Since proteins seldom act alone, but generally form multi-protein complexes which carry out different molecular tasks. The presence of these complexes and their respective composition define an important level of organization within tissues or cells. In this paper we discuss the problem of estimating interactomes; in particular we concetrate on two different interactomes: one representing binary interaction data and the second representing protein complex composition. We propose treating these as estimation problems and apply a sound statistical approach. The estiamtes are based on experimental data as well as data from other sources, and we consider how and when to combine data from these different sources. There are several high through-put experiments that provide data on protein interactions and hence on the estimation problem we are interested in. These include the yeast two-hybrid system (Y2H) with data from Ito et al. [2001], Uetz et al. [2000], Giot et al. [2003] and high-throughput co-precipitation experiments such as [Gavin et al., 2002, Ho et al., 2002, Krogan et al., 2004].

While several authors have constructed, or estimated, interactomes by combining these, and other data sets, our approach is different (FIXME citations here). We create different interactomes, each relevant to a particular concept or type of interaction. The resultant interactomes can be combined by different users, often according to different criteria, to yield more comprehensive interactomes. We believe that there is value in keeping the different concepts separate. Binary physical interactions are quite distinct from multi-protein complex membership, since the relationships defined by the latter are not binary and not all members of a protein complex directly interact. Further, the data sources are quite different, and the experimental errors differ in type and magnitude according to the assays used. By keeping data types

separate we will be able to better assemble and model the observed data.

Data from high throughput experiments are expensive and time consuming to carry out. As a result relatively few statistical methods have been developed and those that have tend to be centered on a particular experiment and have not been broadly validated. These observations suggest that simulation models may be beneficially employed to identify the viability of experimental procedures, to aid in integrating data from disparate sources, and to provide a sound basis for designing experiments. In this paper we outline an approach in the creation of *in silico interactomes* that would be suitable for such simulation experiments as well as for a number of other analytic purposes.

We consider the problem of estimating *in silico* interactomes (ISI) for *Saccharomyces cerevisiae* and we describe five different roles where such a computational device can be used. First, the ISI provides a tool on which wet-lab experiments can be simulated; these *in silico* experiments can help to confirm the viability of the wet-lab experimental procedures by allowing investigators to alter experimental conditions (by the modification of both stochastic and systematic error rates) and then measure how the outputs change as the error rates do. Second, the ISI provides a tool from which multiple data sets can be generated, under different conditions, so that the statistical properties of different proposed estimation procedures can be directly compared. A third use for the ISI is to develop tools and strategies for small scale experiments that can probe the interactome at a very detailed level. For example, a reasonable strategy to investigate a set of protein complexes is to first select a small number of candidate genes and explore their interactions via co-precipitation experiments. Once the first experiment has been collected one would like to analyze the data and determine an optimal, or nearly optimal set of baits for future experiments. A fourth use of an ISI is to study the effect of perturbations in a network. Such perturbations could be caused by a mutation, gene deletion or treatment with a drug. The fifth use for such a tool is to explore and understand the effects of different sampling paradigms on the inferences that can drawn from different wet-lab experiments. We note that there are a number of recent papers [Barabási and Oltavi, 2004, de Silva and Stumpf, 2005, Stumpf et al., 2005] that raise many important issues in this regard.

An ISI is not without its limitations. It cannot establish new relationships between its members, but rather can only reflect the current state of biological knowledge. It can, however, be used in any of the five tasks listed above and the outputs of the computational experiments will often provide predictions of real interactions that can be tested. If these predictions are accurate there will be a resultant savings in cost as wet lab experiments can be directly tailored to the set of likely candidates. Further, one type of interaction can be used either to predict or to model other types of interactions and to provide information about the likelihood of interactions; the verification of interactions, however, remains the domain of experimental biology.

Protein interactions are dynamic by nature so that different interactions occur at different times and under different conditions. For example, some protein complexes such as those needed to facilitate cell division are often transient in nature appearing when needed and dissolving when no longer required, whereas other complexes such as the ribosome (and its subunits) are essentially permanently constituted. We will not directly address this concern but note that the tools and methods we are proposing have natural extensions that would cover many of the important scenarios that this observation suggests.

While the estimated interactome is itself of some great interest, we note that it is more important to describe the procedure by which an ISI is produced. There are many reasons why the estimate is of less importance, among them the fact that the methodology can be applied to many organisms and tissue types to yield specialized ISI estimates. We also note that the data used in the construction is constantly being updated and improved, hence users will want to regularly update any estimate as new data sources become available, or old ones are updated. Hence, to be viable there must be a mechanism for updating and refining the ISI. We also note that there are many different and disparate data sources and the system should be sufficiently flexible to allow users to select data sources appropriate for their studies as well as the ability to

incorporate additional data from new and potentially different experimental techniques. Finally we note that investigators will often want to modify the ISI, either with locally produced, though not yet public, data, or to satisfy personal beliefs. The paradigm we describe here allows investigators to create their own personalized estimates, to share interactomes and to revise and update interactomes.

We provide the complete set of tools used to create and manipulate the interactome in the ScISI package available through the Bioconductor project (`www.bioconductor.org`). Others are free to explore, extend, critique, and improve these methods.

2 Preliminaries

In our discussion we do not distinguish between the term gene and the term protein, since the available resolution is not fine enough and equating these concepts leads to a more comprehensible dialog. The methodology is sufficiently general to be able to address the more complicated situation of different protein/polypeptide variants should they arise.

2.1 Biological Interactomes

In the broadest sense, an interactome is a set of genes, or their related proteins, within a particular organism or cell type. The genes, or proteins have a variety relationships with each other and each of these different relationships can be modeled using some form of graph. We represent genes as nodes in a graph and then use edges to represent the different types of relationships that exist. Some relationships are binary, for example, protein p_1 is known to directly physically interact with some other protein, p_2. In other cases, the relations are not binary but rather are one to many or many to many. For example, all members of a complex are related to each other, but they need not directly interact and the relationship between the members of a multi-protein complex is not binary. A reasonable model for protein complex data is as either a hypergraph Berge [1973] or an equivalent bipartite graph. Proteins are grouped into sets according to whether they are in a complex and these sets constitute the hyperedges of the hypergraph. There is no need for the hyperedges to be disjoint and proteins can be in many or few hyperedges, depending on whether the protein is in many or few complexes. While some represent multi-protein complex data in the form of a binary graph we do not since such a representation imposes a strict loss of information; we prefer the hypergraph since it conserves all important information.

These two biological interactomes are related, but they are also distinct, and the use of the data in modeling and other downstream uses requires that some caution be used to not confuse the concepts. It is certainly the case that for our working definition of a protein complex (to be defined) the members of that complex will have some binary interactions with other members. But not all members have direct physical interactions with other members. These two examples highlight the subtle differences between two distinct interactomes and underscores the necessity to keeping distinct interactomes separate. We will refer to the interactomes modeling binary interactions as I_b, where the b subscript reinforces the notion that binary relationships are being modeled and the protein complex interactome as I_c where the c subscript reinforces the notion that protein complexes are being modeled.

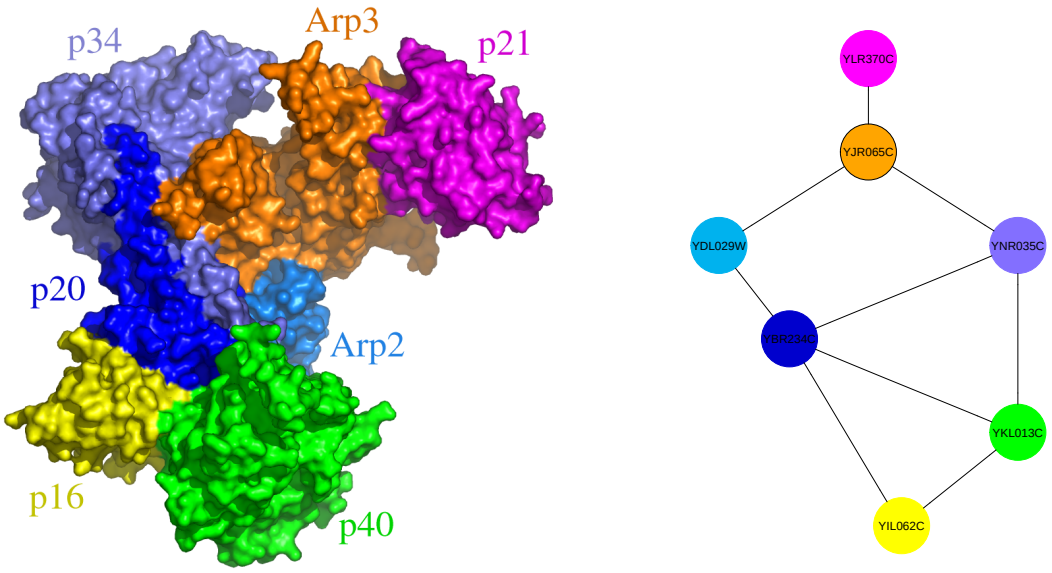
(a) Protein Binary Interaction Network

Figure 1: A protein binary interaction network as a graph and protein complex membership network rendered as a bipartite graph FIXME - Cannot use this as it is probably copyrighted, but wanted to put here as example...need to render one ourselves

There are of course many limitations to such an approach, however, there will be ample opportunity to refine and revise the models proposed here as our knowledge of the underlying biology improves and as the available computer technology improves. Among the limitations is the observation that some interactions will only take place under a specific set of conditions, and that other interactions, sometimes involving the same proteins, will occur under other conditions. Thus, the model representing these data will need to become richer. The models we propose can be extended in fairly straightforward ways to encompass these details, but the available data is currently to sparse to warrant such an approach.

2.2 Graphs and HyperGraphs

We require a small amount of graph theory to allow for succinct discussion of the relevant concepts. A graph G, consists of a pair of sets; the vertices, or nodes, V, and the edges, or relations, E. We write $G = (V, E)$. Graphs can be used to represent binary relationships and hence are an appropriate model and data structure for I_b. They cannot represent the more complicated situation of modeling protein complex co-membership. In this case we require the use of hypergraphs Berge [1973]. A hypergraph, H, consists of a set of nodes, V, and a set of hyperedges. A hyperedge is any subset of V. Thus, for modeling protein interaction data the vertex set is the set of proteins in the organism or tissue under study. To represent binary relationships we will use a graph, where the edges indicate the presence of a binary interaction and for complex co-membership data we will use a hypergraph, where the hyperedges are the identities of the proteins that constitute the complex. There is a one-to-one relationship between hypergraphs and bipartite graphs and we will use these terms inter-changeably in the report. Indeed it will be the case that bipartite graphs are much easier to render pictorially, and so we will defer to these graphs rather than the hypergraphs when needed.



(a) (b)

Figure 2: Some scientists might find the physical rendering of ARP 2/3 to be of more use while others might prefer a graphical representation. We note that the issue of identifiability presents itself as creator of the atomic rendering has chosen one version of the common names while the creator of the graph has chosen the systematic gene names. The coloring scheme provided allows for one to one matching of proteins from one rendering to the other.

There are a variety of different representations for both graphs and hypergraphs. We will mainly make use of the incidence matrix representation for H and the adjacency matrix representation of G. For the protein complex hypergraph the incidence matrix representation, M_H, is a $\{0,1\}$-matrix where proteins index the rows and protein complexes index the columns, with

$$(M_H)_{ij} = \begin{cases} 1 & \text{if protein } i \text{ is a member of complex } C_j, \\ 0 & \text{otherwise.} \end{cases} \tag{1}$$

We remark that this representation can be extended to include the multiplicities of protein membership by simply letting the entries $(M_H)_{ij} \in \mathbb{N}$. The non-negative integer would imply the multiplicity by which a protein appears in a given protein complex.

When studying binary interactions the representation will be of a graph, and then the adjacency matrix is typically square (or upper triangular) with the proteins indexing both the rows and the columns,

$$(M_B)_{ij} = \begin{cases} 1 & \text{if protein } i \text{ directly interacts with protein } j, \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

Gene Common Names	Corresponding Gene Systematic Names
YRF1	YER190W, YDR545W, YGR296W, YLR466W, YLR467W, YNL339C, YPL283C
YPK1	YKL126W, YJL093C, YNL307C
POT1	YIL160C, YKL198C
ZRG11	YIL046W, YOR030W
RTS1	YOR014W
Gene Systematic Names	Corresponding Gene Common Names
YPL084W	NPI3, ASI6, VPS31, LPF2
YIL148W	UB11, UBI1, CEP52A
YNL041C	SEC37, COD2
YGR028W	YTA4
YHR133C	NA
YPL170W	NA

Table 1: The correspondence between common names and systematic names in *S. cerevisiae*. Notice that the correspondence can be many to one in both instances and can also be undefined.

2.3 Protein complexes: Definition, Representation and Characteristics

We define a protein complex to be two or more proteins that combine to form a single, connected functional unit for the purposes of carrying out some biological objective. There is no assumption that all members of a complex physically interact with one another, but there must be sufficient physical binary interactions so that all proteins form a connected multi-protein structure. In essence, we define a multi-protein complex as a collection of proteins, and in some cases polypeptides, that are connected via binary interactions.

2.3.1 Protein (Complex) Identifiability

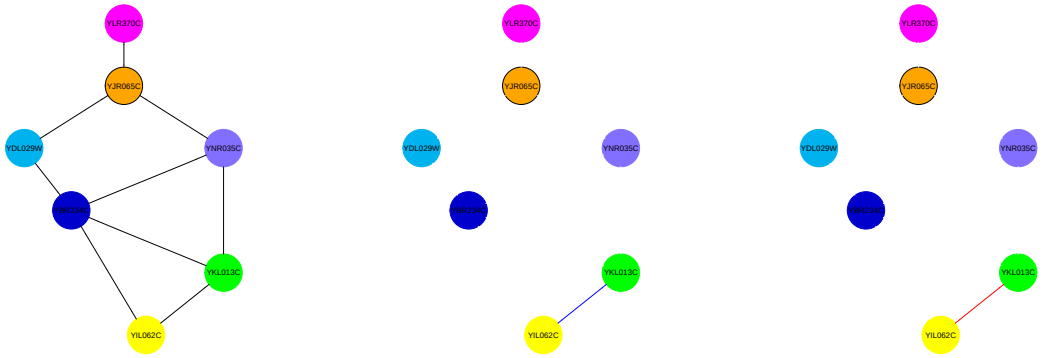
For a protein complex to be uniquely identifiable, it is necessary for each of the constituent protein members to be uniquely identifiable. Unfortunately, this is not always possible. Genes, and their corresponding proteins, are often referenced using database or investigator specific names and mappings from one nomenclature to another are not always one to one. We have used the approach of mapping all identifiers to the set of yeast systematic names and all references will be made using systematic names. When the mappings were not unique we selected the first name in the list. Table 2.3.1 gives presents some examples of the problems that arise when mapping between systematic and common names.

3 Methodology

The creation of an interactome is based on the integration of a number of different data sources. These data sources need to be identified, the likely sources of variability and the nature and extent of incompleteness needs to be assessed. Once these issues have been addressed we can then proceed to develop methods for combining the data into a single overall estimate. Many methods for combining data have an intrinsic assumption that the error rates are similar for

	MIPS	GO	Gavin et al	Ho et al	Krogan et al	IntAct	Miller et a
Data Type	CM	CM	CM	CM	CM	PPI	PPI
Methods	Various	Various	TAP	HMS-PCI	TAP	Y2H	Y2H (Mer
Update	Bi-annually	Bi-annually	None (Static)	None (Static)	None (Static)	Monthly	None (Sta

Table 2: This table details the data sources from which we have derived both the complex membership (CM) data as well as the protein-protein interaction (PPI) data. The rows (in descending order) detail the gene association type, the method of detection of the association type, and the time-table of updates.



(a) (b) (c)

Figure 3: These figures show the difference between the true atomic structure determined by crystallography and by Y2H experimentation as well as crystallography and computationally inferred interactions. The missing edges from the Y2H graph are indeterminate since missing edges can either be due to the relationship not being tested or due to false negative observations.

the experiments being combined. While the use of graph models makes some of the steps easier, their use does not preclude the application of sound statistical practice.

We identify three distinct sources of data that can be used to construct the protein complex interactome and then two other sources of data that can be used to create the binary interactome. We emphasize the construction of the protein complex interactome as it is quite different from other published protein interaction data bases, while the binary interaction estimate is more similar to other constructs and tends to differ from them in important, but less dramatic ways.

To provide some measure of concreteness to our discussion we will consider the Arp 2/3 complex. Figure **??** provides two different views of the Arp 2/3 complex. In panel b) we see a graph that represents the known topology of protein interactions in Arp 2/3. Not all proteins interact with each other and in fact the graph is quite sparse. Of the 21 possible edges there are 8, and the minimum number of edges that would connect all proteins is 6.

3.1 Estimation of the Protein Complex Interactome

The explicit details of the procedures and processes are provided in the supplementary data. In brief, we assembled data from available public databases (GOA and MIPS) that report information on known, or predicted protein complexes and assembled that with protein complexes predicted from experimental data [Ho et al., 2002, Gavin et al., 2002, Krogan et al., 2004] using the methodology of Scholtens and Gentleman [2004], Scholtens et al. [2005]. In order to make the data as comparable as possible we chose to process all experimental data uniformly, and attempted to exclude all predicted complexes from GOA and MIPS that were based on the high-throughput experimental co-precipitation experiments.

	MIPS	GO	Gavin	Ho	Krogan
Distinct Complexes	163	225	260	242	82
Expressed Genes	2479	1273	1363	397	200

Table 3: This table details the number of complexes derived from each data repository or high-through-put experiment as well as the number of expressed genes participating in some none trivial protein complex.

3.1.1 Merging of the Data-Sets

After the protein complex membership data-sets have been obtained and transformed into an incidence matrix representation, the next step is the construction of an aggregate protein complex interactome by combining the separate incidence matrices into one single non-redundant bipartite graph incidence matrix. The process of merging the smaller incidence matrices requires that we compare each predicted protein complex with all other protein complexes. Before proceeding, we introduce a small lemma that simplifies the merging process.

Lemma 1. *Let $C_1, \ldots, C_n$ be non-trivial finite sets which need not be disjoint. Comparing all the sets $(C_1, \ldots, C_n)$ and removing all sets C_{l_m} properly contained in any C_i $\forall i \in [1, n]$ is an associative process, so it can be reduced to pairwise comparisons in any order.*

Proof. The proof of the lemma is a clear application to the fact that set inclusion is a transitive property. For instance, suppose we have three sets $\{C_i, C_j, C_k\}$. We may assume that $C_j \subsetneqq C_i$ for otherwise, we do not delete C_j. If we first compare C_j with C_i, we would immediately delete C_j when we see the proper inclusion. Now we can also suppose that $C_k \subsetneqq C_j$. Then we must have $C_k \subsetneqq C_i$, and so it still be removed. We have shown comparisons to be independent of ordering and therefore, is an associative process. □

With the an application to the lemma, we can merge the graph incidence matrices iteratively by combining them pairwise. As an example, we detail how the protein complexes of GO, I_g, and of MIPS, I_m, are merged. We define the merging of two interactomes as not only combining the protein complexes from I_g to I_m, but also removing protein complexes which are common to both interactomes. We note that we can also remove a protein complex of I_g which is sub-complex to a complex of I_m (or vice versa), and the process remains independent of ordering. Once the matrices I_g and I_m are merged into I_{gm}, say, we can repeat the process

described above by merging I_{gm} with another incidence matrix. This process is repeated until all estimates are merged. We will refer to the resulting aggregate interactome as I_c.

3.2 Binary Interaction Interactome

To estimate a binary interaction interactome, I_b, we made use of available data, largely based on the Y2H [Fields and Sternglanz, 1994]. We considered two distinct sources of data, one being experimental data and the other being binary interaction predictions made by **?**. We used two sources for experimental data: the `IntAct` repository and the experimental data obtained from (cite Miller et al). We choose to use the IntAct database as the primary source for Y2H data since it is quite complete, well documented and provides the data in a form that is relatively easy to process. The traditional Y2H methods fail to detect interactions between membrane bound proteins; (cite Miller et al) developed a variant of the Y2H methodology to overcomes these difficulties.

A fundamental distinction needs to be made regarding the lack of edges within I_b. An edge is missing between two proteins p_1 and p_2 in I_b if either p_1 never interacts with p_2, a true negative interaction, or p_1 has never been tested with respect to p_2, an untested interaction. Problems arise in interpreting the structure I_b because the bait population and the prey population are seldom fully documented. In this way, the distinction between true negative edges and untested edges becomes indeterminate.

In order to make the data more comparable we restricted our analysis to the XXX data sets reported in IntAct that appeared to be genome wide (i.e. at least 80% of the genome was tested as prey). We set a lower bound for the number of baits used at the threshold of twenty proteins and note that due to the manner in which the data are reported we will not know about baits for which no interactions were detected. In this case, armed with knowledge of the bait proteins we can correctly interpret missing edges between known baits and prey as tested but not observed and those between prey only proteins as not tested.

We further used a number of different statistical tests, explicit details are given in the supplementary details section, to assess the observed error rates in the different experiments. If experiments have vastly different error rates combining them has the potential to dilute the good data more than we improve it.

4 The estimates

We have now created two fundamentally different *in silico* interactomes, I_c and I_b by using the resources obtained from the R package ScISI. From these two estimates, we can generate a list of general summary statistics that offer a more qualitative view of these interactomes.

4.1 Summary Statistics for the Protein Complex Interactome

Our estimate of I_c has 1702 unique genes participating within 908 non-trivial protein complex. We do note that this estimate will continue to expand as more data are included in it.

Not limited to the number of expressed genes and protein complexes, we can investigate both the distribution of the size of the protein complexes as well as the distribution of the number of complexes to which each protein belongs. Figure 4 shows gives us these distributions.

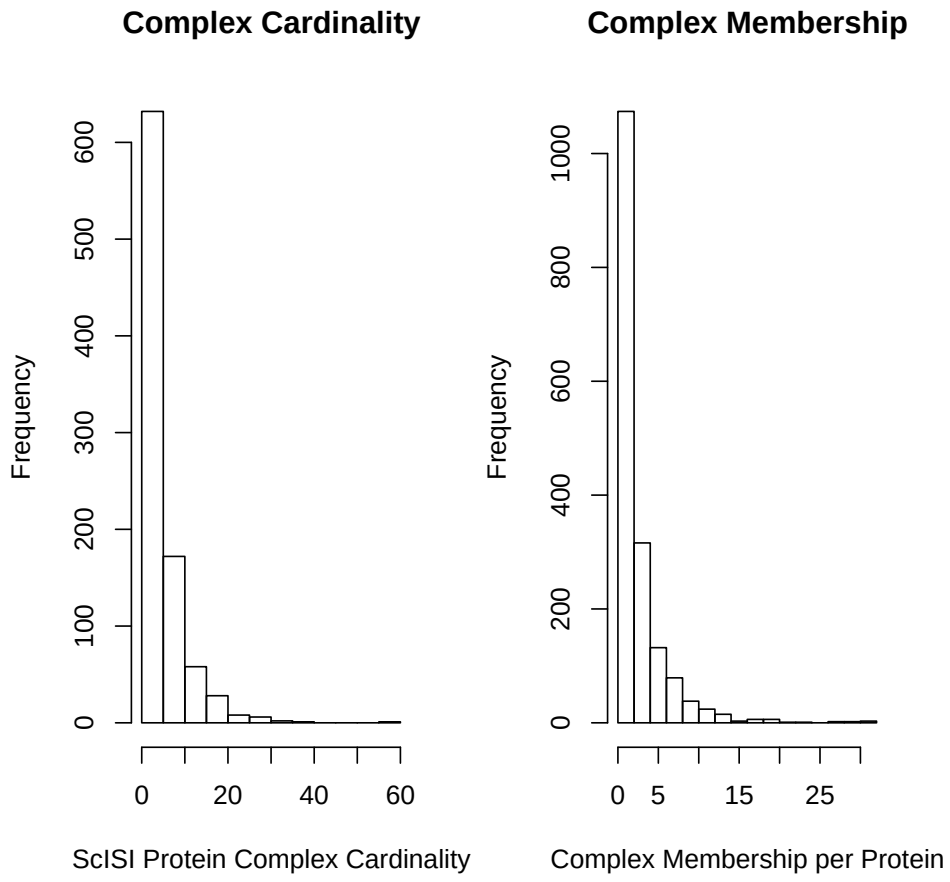


Figure 4: Distribution of Complex Cardinality and Distribution of Complex Membership per Protein

We can see from each histogram that the vast majority of proteins participate in relatively few protein complexes and also that most of predicted protein complexes contain relatively few proteins.

We can also inspect each pair of protein complexes from each of the data repositories. In Table 4, we have the number of protein complexes which are common between each pair of databases. For instance, the number of common protein complexes between the GO and MIPS repositories is 38, and this is a relatively small overlap between protein complexes which have been verified. It comes as no surprise that the overlap of putative protein complexes estimated from the high through-put data via apComplex with each other as well as with GO and MIPS is particularly small. While this fact does not repudiate these putative protein complexes, it does signal the need for further validation.

	MIPS	GO	Gavin	Ho	Krogan
MIPS	4	38	1	2	0
GO	38	0	5	2	1
Gavin	1	5	0	0	1
Ho	2	2	0	0	0
Krogan	0	1	1	0	0

Table 4: Number of repetitive protein complexes.

	MIPS	GO	Gavin	Ho	Krogan
MIPS	39	52	21	0	2
GO	34	14	19	1	7
Gavin	10	26	0	5	1
Ho	12	14	9	0	1
Krogan	11	8	14	4	0

Table 5: Number of protein sub-complexes: the rows indicates the source of the sub-complex while the column indicates the source of the complex containing the sub-complex. For example, 51 MIPS complexes are sub-complexes of GO complexes.

Table 5 accounts for the number of protein complexes from one database which are proper sub-complexes of protein complexes in any database. These cross-references show one limitation discussed by [Scholtens and Gentleman, 2004]; the fact that apComplex is highly sensative the false negative observations in the AP-MS technology. From Table 5 we can observe on anomalie in these data. The Gavin and Krogan estimates are roughly symmetric with respect to GO and MIPS, but Ho is not. Many Ho estimates are sub-complexes of GO and MIPS complexes, but almost no GO and MIPS complexes are sub-complexes of Ho estimates.

4.2 Summary Statistics for the Protein Binary Interactome

Much like the analysis for I_c, we begin the dissection of I_b by investigating the number of baits sampled. From the restrictions described above, we made use of only 7 of the experimental

data-sets, and from these, a total of 1922 unique proteins were tagged with binding domains and became functional bait proteins. While the prey population is genome wide, only 3852 unique proteins interacted with some particular bait protein(s) and were observed as prey.

4.3 PPI Predictions

Liu et al. [2005] report 20088 pairwise predictions between proteins in *S. cerevisiae*, with a presumed false positive rate of $3E-4$ and a presumed false negative rate of 0.85. The predictions are accompanied by a probability and one can select only those predicted interactions according to this probability. We have chosen to use only those interactions where the probability is at least 0.5. This greatly reduces the number of predicted interactions and also greatly reduces the number of proteins that were observed.

When using these data, only proteins that are predicted to be involved in at least one interaction are *observable*. This changes as the probability cut-off changes and can be interpreted as the sensitivity of the analysis. This phenomenon is no different than that observed with different experimental procedures, not all interactions can be detected with all procedures and one must make some efforts to determine which interactions can be detected.

Once the observable proteins have been determined they can be use to compute the different summary statistics discussed in Section **??**.

For each protein complex, C, we divide its constituent proteins into two sets, those that are observable, denoted U and those that are not, $\overline{U} = C$
U. If $P \in U$, then we expect there to be an edge between P and some other member of C, but we will only observe this edge if it is to another member of U.

4.4 Putting it together

Are we going to put the predictions and the binary data together to get I_b? It seems like we should...

5 Validation

Perhaps the only truly contentious aspect of our analysis is the use of the methodology of Scholtens and Gentleman [2004], Scholtens et al. [2005] to provide a majority of the predicted complexes in I_c. And we now turn our attention to a validation of that approach. Our basic premise is that the complex is a collection of physically connected proteins. While, in some cases this may not be strictly true and there could be two or more subcomponents that do not physically interact such cases will have little impact on the validation methods we employ and hence are, to a large extent, covered by our proposals. To validate our estimates we will compare them directly to the curated complexes of MIPS and GO. If the distribution of different test statistics looks the same for our data as for MIPS and GO estimated complexes then this observation provides some validation of our approach and suggests that for the criteria tested there is little difference between our estimates and the more heavily curated protein complex estimates from MIPS and GO.

If all physical interactions between proteins were known, then the problem would be somewhat straightforward to address. One would simply take a predicted complex and determine

whether the constituent elements interact and deem the complex *viable* if it could be configured as a single connected component. However, very few interactions are known, and it is this incompleteness that requires us to make inference about whether the prediction is likely to be valid. Unfortunately the topology of the complexes (the set of actual true interactions) is unknown and hence cannot itself be used to assess the observed data. We do not know how many edges (physical interactions) there should be, and whether the complex follows a spoke topology, a matrix topology or any of the many other possibilities.

We also note that one of the alternatives that some have used, which is to compare their predictions to the published, and to some extent verified complexes described in GO and MIPS is not available to us. We used those sources in construction of our estimate and hence cannot validate against them. Instead we make use of two other sources of data. One being the set of observed yeast two-hybrid (Y2H) experiments available from IntAct and the other being a set of predicted protein-protein interactions from Liu et al. [2005].

Before describing the data we briefly discuss the statistical issues involved. In most cases only a subset of the genes were assayed, or considered and it will be important to restrict the computations to those genes that were assayed. A working hypothesis is that every protein complex is a single connected component, so that when the complex is functional it is a single unit. When considered in terms of binary interactions this requires that there be at least $k-1$ edges for a complex containing k proteins, but not any set of $k-1$ edges will suffice. There can be no more than $k \cdot (k-1)$ edges in total. It does seem that most complexes tend to have fewer, rather than more edges.

5.1 Y2H Validation

We will make use of I_b to provide us with an estimate of the binary interactome. These data give us some sense of the known binary interactions but the logic underlying their use requires some care in application. APMS data indicates the consituent members of the complex but gives no information about the true topology of the binary interactions.

Since Y2H data are somewhat notorious for high false positive and false negative rates we have chosen to restrict the data to more informative subset. Since the Y2H reactions occur in the nucleus

Hence we require only that for a given complex, any member used as a bait pro unknown we cannot test whether the observed Y2H

only be obtained for complex estimates that contain one or more Y2H bait proteins. And since the actually physical interactions are not known, the topology of the Y2H graph is also not known.

Any bait protein in a Y2H experiment should find one or more complex co-members provided those co-members were available as prey in the Y2H experiment. This proviso indicates one of the true weaknesses of Y2H data as they are often reported since only information about the prey that were detected and not the prey that were tested is reported. Thus, the absence of an edge is not as informative as it could be. We do not know if it was tested and not found or not tested, and these are very different things.

6 Discussion

What did we learn,

References

A-L Barabási and Z. N. Oltavi. Network biology: Understanding the cell's functional organization. *Nature Reviews: Genetics*, 5:101–113, 2004.

C. Berge. *Graphs and Hypergraphs*. North-Holland, Amsterdam, 1973.

E. de Silva and M. P. H. Stumpf. Complex networks and simple models in biology. *J. R. Soc. Interface*, 2:419–430, 2005.

S. Fields and R. Sternglanz. The two hybrid system: an assay for protein-protein interactions. *Trends in Genetics*, 10:145–150, 1994.

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

L. Giot et al. A protein interaction map of *Drosophila melanogaster*. *Science*, 302:1727–1736, 2003.

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

T. Ito, T. Chiba, R. Ozawa, et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. of the U.S.A.*, 98:4569–4574, 2001.

N. J. Krogan et al. High-definition macromolecular composition of yeast rna-processing complexes. *Molecular Cell*, 13:225–239, 2004.

Y. Liu, N. Liu, and H. Zhao. Inferring protein-protein interactions through high-throughput interaction data from diverse organisms. *Bioinformatics*, 21:3279–3285, 2005.

D. Scholtens and R. Gentleman. Making sense of high-throughput protein-protein interaction data. *Statistical Applications in Genetics and Molecular Biology*, 3:Article 39, 2004.

D. Scholtens, M. Vidal, and R. Gentleman. Local dynamic modeling of global interactome networks. *Bioinformatics*, 21:3548–3557, 2005.

M. P. H. Stumpf, C. Wiuf, and R. M. May. Subnets of scale-free networks are not scale-free: Sampling properties of networks. *Proc. Natl. Acad. Sci. of the U.S.A.*, 102:4221–4224, 2005.

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