Statistical Applications in Genetics and Molecular Biology

Volume 3, *Issue* 1 2004 *Article* 39

Making Sense of High-Throughput Protein-Protein Interaction Data

Denise Scholtens* Robert Gentleman†

*Northwestern University, dscholtens@northwestern.edu

†Fred Hutchinson Cancer Research Center, rgentlem@fhcrc.org

Making Sense of High-Throughput Protein-Protein Interaction Data*

Denise Scholtens and Robert Gentleman

Abstract

Accurate systems biology modeling requires a complete catalog of protein complexes and their constituent proteins. We discuss a graph-theoretic/statistical algorithm for local dynamic modeling of protein complexes using data from affinity purification-mass spectrometry experiments. The algorithm readily accommodates multicomplex membership by individual proteins and dynamic complex composition, two biological realities not accounted for in existing topological descriptions of the overall protein network. A likelihood-based objective function guides the protein complex modeling algorithm. With an accurate complex membership catalog in place, systems biology can proceed with greater precision.

KEYWORDS: protein-protein interactions, graph theory

---

*We thank M. Vidal for critical discussions. This work was done while both authors were with the Department of Biostatistics at Harvard University in Boston, MA. The authors were supported by grant P20 CA96470 (D.S.) and NIH grant #1R33 HG002708 (D.S., R.G.).

Introduction

Systems biology networks rely on functional macromolecules, often multiprotein complexes, to accomplish the work of the cell. A catalog of these complexes and their constituent proteins is a necessary, but recently unavailable, component of accurate systems biology modeling. To fully understand the interactive relationships of cellular modules, investigators need to account for both stable and dynamic complex composition, as well as multicomplex membership by individual proteins. Currently, large-scale topological descriptions of the overall protein network are available using protein-protein complex comemberships and binary physical interactions detected by affinity purification-mass spectrometry (AP-MS) and yeast two hybrid (Y2H) technologies, respectively (Jeong et al., 2001, Salwinski and Eisenberg, 2003). Scholtens et al. (2004) discuss additional computational methodology for ascertaining local dynamic models of precise protein complex membership from AP-MS data and outline the next steps required for integrating the different, but complementary, information offered by AP-MS and Y2H. In this paper, we develop in detail the likelihood-based objective function that drives the technology reported by Scholtens et al. (2004). Furthermore, we specifically note how biological particulars of AP-MS data fit into the graph theoretic and statistical paradigms that motivate the complex identification algorithm.

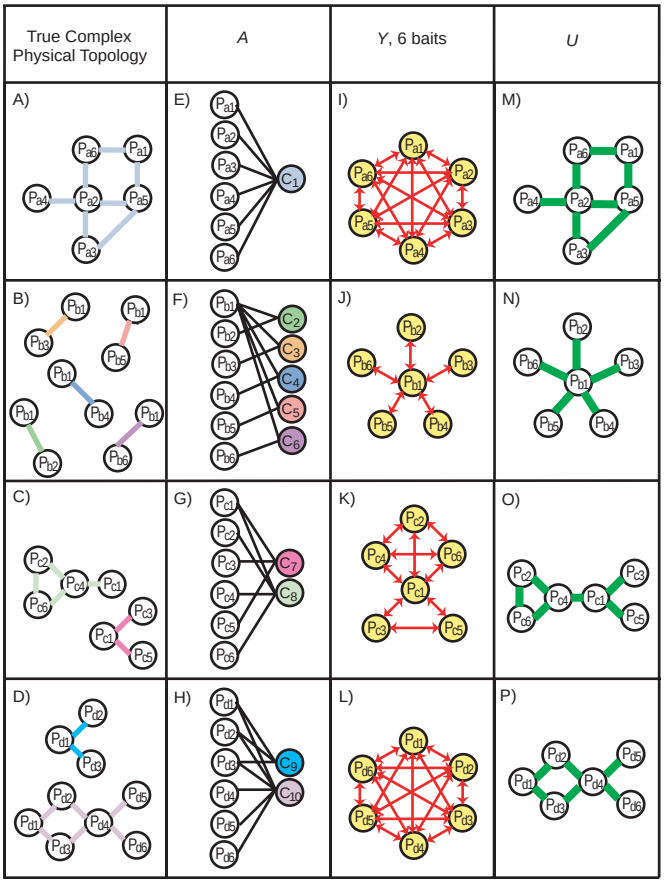
Graph Theoretic Paradigm

Graphs consisting of nodes and edges are particularly helpful for representing both the complex membership estimation problem and the available AP-MS and Y2H data. Figures 1A)-1D) contain four physical topology schemes for hypothetical protein complexes with nodes representing proteins and edges representing physical interactions. Separate complexes are denoted by different edge colors. We use these four schemes to demonstrate the important structural differences between the desired protein complex catalog and the different assays of protein-protein relationships.

Protein Complex Membership Graph, A

Our goal is to estimate A, a bipartite graph representing the desired protein complex catalog. One set of nodes in A, $V_J = \{P_1, ..., P_J\}$, represents J proteins and another set of nodes in A, $V_K = \{C_1, ..., C_K\}$, represents K protein complexes. An edge in A connecting P_j and C_k indicates membership

Figure 1: A, Y, and U for Four Hypothetical Protein Complex Schemes



of P_j in C_k for all $j = 1, ..., J$ and $k = 1, ..., K$. Since the number of complexes K is unknown, estimation of A implicitly requires estimation of K. Ideally, we would like to estimate A for all proteins and complexes in the cell, but we are restricted by the set of proteins involved in the AP-MS experiments and the set of complexes they are in. Figures 1E)-1H) depict A for the four complex schemes in Figures 1A)-1D), respectively.

The graph A accounts for two biological realities not accommodated in other automated algorithms for the analysis of AP-MS data (Jansen et al., 2003). First, it allows for multi-complex membership by individual proteins since a protein, a node in V_J, can be connected to multiple complexes, nodes in V_K. Second, dynamic complexes that allow different subunits at different times can be represented as separate nodes in V_K according to all possible compositions. These two allowances are of fundamental importance for procuring

DOI: 10.2202/1544-6115.1107

an accurate protein complex catalog, and are readily incorporated into the bipartite graph structure.

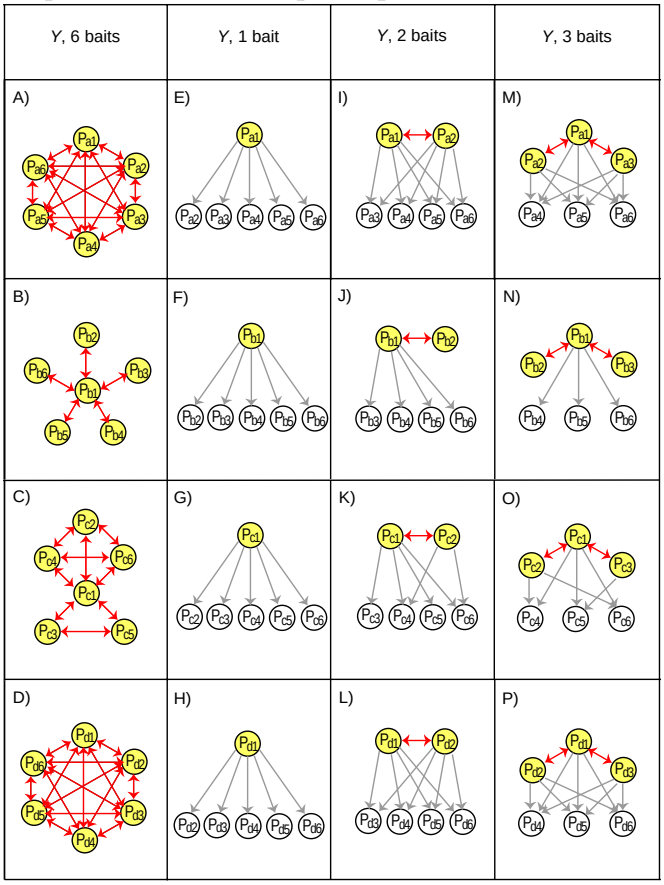
Protein-Protein Complex Comembership Graph, Y_{V_B}

AP-MS technology does not directly assay the edges in A. Instead, it assays the edges of Y_{V_B}, a graph of complex comembership, by finding all hits that are comembers in at least one complex with each bait. Y_{V_B} is related to A through a transformation we call $Y_{V_B} = (A \otimes A')_{V_B}$, where V_B is the subset of proteins in V_J that are used as baits in an AP-MS experiment, and $V_H = V_J \backslash V_B$ is the complementary subset of proteins in V_J that are found as hits but never used as baits. Y_{V_B} can be thought of as the graph of pairwise complex comemberships based on "ideal" observations from purifications with the set of baits V_B assuming perfectly sensitive and perfectly specific AP-MS technology. If $V_B = V_J$, then all proteins with at least one common complex membership are connected by reciprocated edges in Y_{V_B}, and proteins without common complex affiliations are unconnected. Sets of nodes in Y_{V_B} for which all pairwise (reciprocated) edges exist, i.e. *complete subgraphs*, correspond to sets of proteins that are all comembers in at least one complex. *Maximal complete subgraphs* in Y_{V_B} contain entire sets of proteins composing one complex. Figures 1I) - 1L) depict the corresponding Y_{V_B} graphs for Figures 1A)-1D) when all proteins are used as baits, presuming AP-MS technology with perfect sensitivity and specificity. While the mapping from A to Y_{V_B} is unique under the transformation $Y_{V_B} = (A \otimes A')_{V_B}$, the reverse mapping from Y_{V_B} to A is many-to-one. Maximal complete subgraphs resolve this lack of uniqueness, and so our complex membership estimation algorithm searches for collections of proteins in V_J which, given the observable data, resemble maximal complete subgraphs.

In actual AP-MS experiments, $V_J \neq V_B$, and the distinction between bait proteins and hit-only proteins is essential to knowing which complex comemberships are looked for, and which are not. The graphs in Figure 2 depict Y_{V_B} for the complexes in Figures 1A) - 1D) using different sets of bait proteins. Complex comemberships for pairs of bait proteins are tested twice, once during each respective purification, and are represented by red reciprocated edges in Y_{V_B}. For bait-hit-only pairs, complex comembership is only tested once since the hit-only protein is never used as a bait. In Y_{V_B}, gray unreciprocated edges from baits to non-baits represent existing singly tested complex comemberships. For pairs of hit-only proteins, complex comembership is never tested, and in Y_{V_B}, no edges connect hit-only proteins. The incomplete view of all possible pairwise complex comemberships prevents a direct mapping from

maximal subgraphs in Y_{V_B} to A. When only one protein is used as a bait, as in Figures 2E) - 2H), all four hypothetical complex schemes result in identical Y_{V_B} graphs, despite their very different true organization. Additional purifications help resolve complex comembership to an extent, but the topology among the remaining hit-only proteins is unknown. We have no basis on which to decide if the hit-only proteins form one complex with their common baits, separate complexes, or something in between since we cannot locate the maximal complete subgraphs in the partially tested Y_{V_B}.

Figure 2: Complex Comembership Graphs Y for Different Sets of Baits



To deal with the problem of indiscernable maximal complete subgraphs due to untested complex comemberships, we define a special type of maximal complete subgraph for AP-MS bait-hit data: a *BH-complete subgraph* is a collection of n_b nodes in V_B and n_h nodes in V_H for which all $n_b(n_b-1)$ reciprocated bait-bait edges exist, and all $n_b(n_h)$ unreciprocated bait-hit-only edges

DOI: 10.2202/1544-6115.1107

exist. A *maximal BH-complete subgraph* is a BH-complete subgraph that is not contained in any other BH-complete subgraph. We account for the precise set of observed data, including bait/hit protein status and tested/untested complex comemberships, by searching for maximal BH-complete subgraphs in Y_{V_B} and reporting these as complex estimates. Effectively, when using the data available from a partially observed Y_{V_B}, we assume that two hit-only proteins that are complex comembers with a common bait are also complex comembers with each other. This assumption should motivate further AP-MS experiments to test the actual complex comembership among pairs of hit-only proteins reported to be in the same complex by our algorithm. In general, maximal BH-complete subgraphs with a low proportion of bait proteins out of the total number of proteins represented may in fact compose more than one complex.

Observed AP-MS Graphs, Z_{V_B}

Using the set of bait proteins V_B, AP-MS technology assays the true complex comembership edges in Y_{V_B} , and we record these observations in the graph Z_{V_B}. For a purification using bait protein $P_b \in V_B$, directed edges are drawn from the node P_b to the nodes for all observed hit proteins. In practice, AP-MS technology is neither perfectly sensitive nor specific, resulting in both FN and FP assignment of edges. Since the doubly tested bait-bait edges are not necessarily consistently detected in both purifications, unreciprocated edges in Z_{V_B} may connect elements in V_B; in Y_{V_B}, unreciprocated edges between nodes in V_B are impossible. See Supporting Information for further discussion of FN and FP observations.

Protein-Protein Physical Interaction Graphs, U

Y2H technology does not measure the same complex comembership relationships as AP-MS technology. Instead, it detects direct binary physical interactions between pairs of proteins. Figures 1M)-1P) depict U for data resulting from hypothetical perfectly sensitive and specific Y2H experiments for the complexes in Figures 1A)-1D), respectively, with nodes representing proteins and edges representing physical interactions. In U, a missing edge between two proteins only implies lack of physical interaction, not lack of complex comembership. The topology required of complex comembers in U is that some edge path, possibly through other proteins, connects them. This necessary criteria is not sufficient for complex identification, however, since sets of proteins connected by edge paths in U do not always form complexes. In contrast,

for Y_{V_B}, sets of complex comembers in A form BH-complete subgraphs, and excepting nonidentifiable subcomplexes, maximal BH-complete subgraphs in Y_{V_B} translate back to A.

There is also a substantial difference in the interpretation of the terms FN and FP for AP-MS and Y2H technology, further underscoring the need to carefully distinguish between the two data types. If two proteins are complex comembers but do not directly interact, a FP Y2H observation of such a pair would be a true positive (TP) AP-MS observation. Similarly, a FN AP-MS observation of this same pair would be a true negative (TN) Y2H observation. While a TP Y2H observation is indicative of both physical interaction and complex comembership, a TN Y2H observation is uninformative regarding complex membership. Due to its direct relationship with A, AP-MS data is better suited for complex membership estimation. Once complex membership is known, an important next step would be to use complementary Y2H data to elucidate the actual physical connectivity among proteins.

Relating A, Y, and Z

In summary, the bipartite graph A can be transformed to $Y_{V_B} = (A \otimes A')_{V_B}$, the true complex comembership graph that is assayed by AP-MS technology. The observed portion of Y_{V_B} based on a set of bait proteins V_B, is then recorded in a graph Z_{V_B}, which is subject to both FN and FP assignment of edges. In short, the transition from true complex composition to the observed AP-MS data can be represented as $A \rightarrow Y_{V_B} \rightarrow Z_{V_B}$. Our complex membership estimation algorithm begins with the observed data Z_{V_B} and works toward an estimated $\hat{A}$, or $Z_{V_B} \rightarrow \hat{A}$, by searching for structures in Z_{V_B} that closely resemble maximal BH-complete subgraphs. Our algorithm incorporates user-specified sensitivity and specificity parameters that allow a group of proteins to be identified as a complex, even though there may be some missing edges from the maximal BH-complete subgraph. External evidence, such as cellular component data from the Gene Ontology (GO) Consortium, can also be included to lend credence to the existence of an edge, even though it is not observed in the data.

Matrix Representation

The three graphs A, Y, and Z can be represented as matrices. The bipartite graph A can be represented as an affiliation matrix, a data structure and concept frequently used in social networks analysis (Wasserman and Faust, 1999). The affiliation matrix A has J rows corresponding to the set V_J of J proteins and K columns corresponding to the set V_K of K complexes. An

DOI: 10.2202/1544-6115.1107

entry of 1 in the j^{th} row and k^{th} column of A, or $A[j,k]$, indicates membership of P_j in C_k ($j = 1, ..., J$, $k = 1, ..., K$), and an entry of 0 indicates lack of membership. For ease of notation, we assume that the first $N = |V_B| \leq J$ rows of A correspond to the N bait proteins.

Y is related to A through the product of A with its transpose, $Y = A \otimes A'$, under the Boolean algebra defined by $0 \times 0 = 0 \times 1 = 1 \times 0 = 0 + 0 = 0$ and $1 \times 1 = 0 + 1 = 1 + 0 = 1 + 1 = 1$. The symmetric matrix Y contains the set of true complex comemberships assayed by the AP-MS technology. For all $i, j \in \{1, ..., J\}$, if $A[i, k] = A[j, k] = 1$ for at least one $k \in \{1, ..., K\}$, then $Y[i, j] = 1$. If we let the J rows of Y represent baits and the J columns of Y represent hits, then $Y[i, j] = 1$ if bait protein i "ideally" finds hit protein j in its purification, and 0 otherwise. For these "ideal" observations, $Y[i, j] = Y[j, i]$. Actual AP-MS experiments only assay the first N rows of Y according to the set V_B of baits, or $Y[1{:}N,] = Y_{V_B}$. Y can be divided into three sections, corresponding to doubly tested, singly tested, and untested complex comemberships. 1) $Y[1{:}N, 1{:}N] = Y_{V_B}[, 1{:}N]$ records symmetric bait-bait complex comemberships that are tested twice. 2) $Y[1{:}N, (N{+}1){:}J] = Y_{V_B}[, (N{+}1){:}J]$ records bait-hit-only complex comemberships that are tested once. 3) $Y[(N{+}1){:}J, 1{:}J]$ represents possible AP-MS purifications using the set of proteins V_H that were never performed.

The matrix Z_{V_B} is an $N \times J$ matrix which represents the observed version of $Y_{Z_B} = Y[1{:}N,]$ using the actual data gathered in an AP-MS experiment. For $i = 1, ..., N$ and $j = 1, ..., J$ ($i \neq j$), $Z_{V_B}[i, j] = 1$ if bait protein i finds hit protein j, and 0 otherwise. We assign $Z_{V_B}[i, i] = 1$ for $i = 1...N$. Even though $Y_{V_B}[, 1{:}N]$ is necessarily symmetric, $Z_{V_B}[, 1{:}N]$ may be asymmetric due to the possibility of FN and FP observations.

For the discussion of the statistical model, we will refer to the matrix representations of A, Y_{V_B}, and Z_{V_B}, but the corresponding graph structures could also be used. For ease of notation, the V_B subscripts in Y_{V_B} and Z_{V_B} will be dropped.

Statistical Paradigm

The graph theoretic paradigm clarifies the nature of AP-MS and Y2H data and their relationship to protein complex membership recorded in A. The existence of FP and FN observations makes perfect detection of maximal BH-compelte subgraphs in Y_{V_B} impossible, thus motivating a statistical approach to estimating A from Z_{V_B}. Proposed estimates of A are evaluated according to the two-component objective function

$$P(Z|A,\mu,\alpha) = L(Z|Y = A \otimes A', \mu, \alpha) \times C(Z|A,\mu,\alpha), \quad (1)$$

where the first component, $L(Z|Y = A \otimes A', \mu, \alpha)$, is the likelihood of Z, given $Y = A \otimes A'$, and the second component, $C(Z|A,\mu,\alpha)$, is a probabilistic measure of the degree to which the proposed complexes in A each reflect their assumed underlying maximal BH-complete graph structure. Both $L(Z|Y = A \otimes A', \mu, \alpha)$ and $C(Z|A,\mu,\alpha)$ depend on μ and α, user-specified parameters set to reflect the believed specificity and sensitivity of the AP-MS technology. An extension for incorporating external similarity data is discussed in Supporting Information.

$L(Z|Y = A \otimes A', \mu, \alpha)$ (2), is the likelihood for independent Bernoulli observations of the existence of an edge under a logistic regression model. $L(Z|Y = A \otimes A', \mu, \alpha)$ (2) depends on A through $Y[1{:}N,] = (A \otimes A')[1{:}N,]$, and in our setting, the values of Y_{ij} are to be estimated. Specifically, define

$$L(Z|Y = A \otimes A', \mu, \alpha) = \prod_{i=1}^{N} \prod_{j=1, j\neq i}^{N} p_{ij}^{Z_{ij}} (1-p_{ij})^{1-Z_{ij}} \times \prod_{l=1}^{N} \prod_{m=(N+1)}^{(N+M)} p_{lm}^{Z_{lm}} (1-p_{lm})^{(1-Z_{lm})}, \quad (2)$$

where $\log\left(p_{ij}/(1-p_{ij})\right) = \mu + \alpha Y_{ij}$. If $Y_{ij} = 1$, then p_{ij} is the probability of observing an edge between proteins i and j, given that they are comembers in at least one complex, i.e. the sensitivity. If $Y_{ij} = 0$, then p_{ij} is the probability of observing an edge between proteins i and j, even though they are not actually in a complex together, i.e. the FP probability. We assume that any errors made in the observation of edges are independent of each other.

A range of values can be specified for μ and α, but we make two assumptions in the specifications. First, that $Pr(Z_{ij} = 0|\mu, \alpha, Y_{ij} = 0) > .5$, and $Pr(Z_{ij} = 1|\mu, \alpha, Y_{ij} = 1) > .5$, corresponding to sensitivity and specificity values each greater than .5. Second, that $Pr(Z_{ij} = 0|\mu, \alpha, Y_{ij} = 1) > Pr(Z_{ij} = 1|\mu, \alpha, Y_{ij} = 0)$, that is, the FN probability is higher than the FP probability. The purification step in AP-MS technology makes this assumption reasonable. In practice, we find that very small values of $Pr(Z_{ij} = 1|\mu, \alpha, Y_{ij} = 0)$ and moderate values of $Pr(Z_{ij} = 1|\mu, \alpha, Y_{ij} = 1)$ seem to yield the most biologically meaningful results.

Maximization of $L(Z|Y = A \otimes A', \mu, \alpha)$ to estimate the values of Y_{ij} is straightforward. The contribution of singly tested edges to the likelihood is maximized if $Y_{ij} = Z_{ij}$. For doubly tested edges, the contribution of these data to the likelihood is maximized if $(Y_{ij}, Y_{ji}) = \max(Z_{ij}, Z_{ji})$. Our algorithm begins by assigning the maximum likelihood estimates of Y_{ij} for $i = 1, ..., I$ and

DOI: 10.2202/1544-6115.1107

$j = 1, ..., (N + M)$, resulting in $\hat{Y}_{init}$. The initial estimate of A, $\hat{A}_{init}$, then consists of $\hat{K}_{init}$ columns recording protein membership in the $\hat{K}_{init}$ maximal BH-complete subgraphs that exist in the graph determined by $\hat{Y}_{init}$. The maximal BH-complete subgraphs represent a set of complex proposals that, based on the likelihood alone, appear to compose protein complexes. We present an algorithm for maximal BH-complete subgraph detection in Supporting Information; adaptations of other maximal subgraph algorithms are also a possibility (Bron and Kerbosch, 1973).

The initial estimate of complex membership based strictly on the likelihood does not allow missing edges from the BH-complete subgraphs between baits and hit-only proteins, and if only one bait–hit-only edge from an existing complex is a FN, the complex will be estimated as two complexes in $\hat{A}_{init}$. Since the thousands of individual edges in Y are tested at most twice, it is plausible that the maximum likelihood estimate $\hat{Y}_{init}$ may not accurately reflect reality; $C(Z|A, \mu, \alpha)$ offers a second criteria to further refine $\hat{A}$ and thus improve upon the estimate based solely on $L(Z|Y = A \otimes A', \mu, \alpha)$. The second part of $P(Z|A, \mu, \alpha)$ (1), namely $C(Z|A, \mu, \alpha)$, is designed to allow combinations of the complex estimates in $\hat{A}$ that increase $C(Z|A, \mu, \alpha)$ in favor of small decreases in $L(Z|Y = A \otimes A', \mu, \alpha)$.

Suppose the k^{th} column of $\hat{A}$ contains a proposed complex, c_k, consisting of a set b_k of n_k bait proteins and a set h_k of m_k hit-only proteins. Since the edges in our graph represent complex comembership, the true BH-complete subgraph for a set of proteins forming one complex should contain $n_k \times (n_k - 1)$ reciprocated edges connecting all pairs of baits and $n_k \times m_k$ unreciprocated edges connecting all baits to all hit-only proteins. Since the AP-MS technology is not perfectly sensitive, only a number, x_k, of the total number $t_k = n_k \times (n_k + m_k - 1)$ of edges might be observed, with $t_k - x_k$ unobserved. The consistency of x_k with the believed sensitivity of the AP-MS technology can be measured for each complex according to

$$\Gamma(c_k) = \left(\begin{array}{c} t_k \\ x_k \end{array} \right) \prod_{g \in b_k} \prod_{h \in b_k \cup h_k, h \neq g} \frac{e^{z_{gh}(\mu+\alpha)}}{1 + e^{\mu+\alpha}} \tag{3}$$

for $k = 1, ..., K$. $\Gamma(c_k)$ is the binomial probability for the number of observed edges in the proposed BH-complete subgraph for complex c_k, given the sensitivity of the AP-MS technology.

The allocation pattern of the missing edges is as important as the number of missing edges in a subgraph for a proposed complex. The randomness of the allocation of the missing edges that are observed in the data, denoted by $\Phi(c_k)$, can be measured by the cumulative probability of observing a particular

missing edge pattern for the edges in complex c_k or something more extreme using the hypergeometric distribution (e.g., a two-sided p-value from Fisher's exact test). We measure the randomness of the missing edge pattern based on indegree rather than outdegree. When an AP-MS tag is applied to a protein, in some cases, it may change the conformation of that protein, rendering it incapable of binding to some or all of its usual complex partners. If a bait protein does not find several complex partners, the measure of $\Phi(c_k)$ according to indegree does not penalize for systematic missing edges. If, however, a protein is found by few baits in the complex, then the protein seems not a member of the complex.

$C(Z|A,\mu,\alpha)$ is taken to be the product of $\Gamma(c_k)$ and $\Phi(c_k)$ for all K proposed complexes: $C(Z|A,\mu,\alpha) = \prod_{k=1}^{K} \Phi(c_k)\Gamma(c_k)$. Since $0 < \Phi(c_k) <= 1$ and $0 < \Gamma(c_k) < 1$, $C(Z|A,\mu,\alpha)$ tends to increase for a smaller number of high quality complexes.

The complex estimation procedure iteratively updates $\hat{A}$, beginning with the initial maximal subgraph estimate for the complexes $\hat{A}_{init}$ by proposing pairwise combinations of the complexes recorded in the columns of $\hat{A}$. If the set of proteins associated with two complexes in $\hat{A}$ increase $P(Z|A,\mu,\alpha)$ (1) when treated as one complex, then the combination is accepted. The acceptance criterion amounts to testing whether $\log P_{k*} - \log P_{k1,k2}$, the difference in the log of $P(Z|A,\mu,\alpha)$ when c_{k1} and c_{k2} are combined into c_{k*}, is greater than zero. Algebraic details are provided in Supplementary Information. Accepting refinements to $\hat{A}$ that increase $P(Z|A,\mu,\alpha)$ yields complexes that are both reflective of approximate maximal BH-complete subgraph structure for protein complexes and consistent with the observed AP-MS data. A tuning parameter could be used to further refine the contributions of $L(Z|Y = A \otimes A', \mu, \alpha)$ and $C(Z|A,\mu,\alpha)$ in the algorithm.

Complex Membership Estimation Algorithm

The entire complex membership estimation algorithm proceeds as follows and is executable using the R (Ihaka and Gentleman, 1996) package `apComplex` available at http://www.bioconductor.org. In practice, there may be a slight difference in the final complex estimates depending on the order in which proposals are made for complex combination. We recommend first ordering the initial complex estimates according to the number of bait proteins, and proposing combinations of these complexes with other complexes. The complex estimates with the highest number of baits are based on the largest amount of observed data, and are therefore potentially more reliable starting points for

DOI: 10.2202/1544-6115.1107

the combination procedure than the complex estimates with fewer baits.

1. Maximize $L(Z|Y = A \otimes A', \alpha, \mu)$ (2) by setting $\hat{Y}_{init_{ij}} = Z_{ij}$ for all singly tested edges and $(\hat{Y}_{init_{ij}}, \hat{Y}_{init_{ji}}) = \max(Z_{ij}, Z_{ji})$ for all doubly tested edges.

2. Find an initial estimate for A, $\hat{A}_{init}$, consisting of all maximal BH-complete subgraphs in $\hat{Y}_{init}$.

3. Order the columns of $\hat{A}_{init}$ according to the number of baits.

4. Set $k = 1$ and $\hat{K}_{init}$ = number of columns of $\hat{A}_{init}$.

5. Set $\hat{A} = \hat{A}_{init}$ and $\hat{K} = \hat{K}_{init}$.

6. For $c_k = \hat{A}[, k]$, find the set A_k of columns of $\hat{A}$, excluding c_k, that share at least one common entry of "1". Calculate $\log P_{k*} - \log P_{k1,k2}$ for c_k paired with all elements in A_k.

7. If at least one value of $\log P_{k*} - \log P_{k1,k2}$ for c_k and the elements of A_k is greater than 0, replace c_k with the union of c_k and $c_{A_{k,max}}$, the element of A_k with which c_k has the largest value of $\log P_{k*} - \log P_{k1,k2}$. Remove $c_{A_{k,max}}$ from $\hat{A}$, as well as any complexes that are now subsets of c_k. Set $\hat{K}$ = the number of columns of $\hat{A}$.

8. If none of the values of $\log P_{k*} - \log P_{k1,k2}$ for c_k and the elements of A_k are greater than 0, set $k = k + 1$ and return to step 5.

9. Repeat until $k = K$.

Our algorithm results in three types of complex predictions: 1) multi-bait-multi-edge (MBME) complexes containing more than one bait and more than one edge in the subgraph; 2) single-bait-multi-hit (SBMH) complexes containing one bait and a collection of hits; and 3) unreciprocated bait-bait (UnRBB) complexes containing two proteins, both used as baits, connected by one unreciprocated edge. The quality of the complex predictions from our algorithm depends on the number of baits in each complex since this relates directly to the amount of observed data for each complex. MBME complexes contain data from multiple purifications, allowing a more detailed estimation of the complex comembership structure among the hits. SBMH complexes may have more structure than reported since the edges among the hits are untested. The edges in UnRBB complexes may result from FP observations

since they are tested twice, observed once, and are not contextually confirmed by other edges. Investigators are encouraged to examine the MBME complexes as the most reliable outputs, and use the SBMH and UnRBB complexes to develop future experiments.

Simulation Study

To assess the accuracy of our complex identification algorithm, we performed a simulation study with $N = 200$ bait proteins, $M = 400$ hit-only proteins, and $K = 55$ complexes. To determine the size of the first 50 complexes, we generated 50 random variables from a Poisson distribution with $\lambda = 10$ resulting in complexes with a range of 2 through 17 proteins. For each complex, we then randomly selected a set of proteins equal to the complex size for membership in the complex. Since in reality, there is likely more overlap between protein complexes than expected by random chance, we hand-generated the last 5 complexes. For the 51^{st} complex, we selected an existing complex of size 10 and changed two of the hit proteins to other hits. For the 52^{nd} complex, we selected a different existing complex of size 10 and changed two of the bait proteins to other baits. The 53^{rd} complex was formed by taking an existing complex with 5 members, and switching one bait and one hit to a different bait and hit. Complex 54 is a subset of 3 bait proteins from a large 14-protein complex along with an extra hit. Complex 55 is a 6-protein complex consisting of 3 proteins from each of two different 7-protein complexes.

After the complexes were formed, we calculated $Y = A \otimes A'$, and then applied an error model to generate FN and FP observations. We replaced the set of TP edges with random Bernoulli observations with P(1)=.80. We replaced the set of TN edges with random Bernoulli observations with P(1)=.0035. The FP probability was chosen so that the number of observed edges in the data set was approximately 50% FP observations and 50% TP observations (von Mering et al., 2002). The FP probability used for a single edge is very small since the number of TNs far exceeds the number of TPs.

In an AP-MS experiment, some bait proteins may have a higher tendency to report FP interactions, that is, the bait might be "sticky". To simulate this, we random selected 10 bait proteins, and made their FP probability equal to .035, ten times the FP probability used for other proteins. We may also have proteins that, due to conformational changes by the AP-MS tag, are rendered incapable of binding to complex comembers when used as baits. To simulate this, we randomly selected 10 baits, and eliminated all observed edges from these baits to their respective hits.

DOI: 10.2202/1544-6115.1107

After simulation of the observed data, we applied our complex estimation algorithm. We used a measure of complex similarity, ω, to assess the accuracy of our predicted complexes with the true complexes recorded in A. For a true complex, say A_C, and a predicted complex, B_C, define $\omega(A_C, B_C) = \min(i/a, i/b)$, where i is the number of proteins in the intersection of A_C and B_C, a is the number of proteins in A_C, and b is the number of proteins in B_C. This measure finds the proportion of common proteins from the point of view of both complexes, and takes the minimum as a conservative measure of the overlap.

We applied the complex estimation procedure with values of .60,.70, .80 and .90 for the sensitivity. The results of the simulation for these four values are reported in Tables 1 and 2. Table 1 records the number of complex estimates with $\omega > .70$ for all true complexes in A, organized by complex size and the number of baits. For larger complexes, a sensitivity value of .90 prevented the combination of columns of $\hat{A}$ since there were too many missing edges to be consistent with such a high sensitivity. For example, four complex estimates had values of $\omega > .70$ for the true complex with five bait proteins and 17 total proteins. As the sensitivity parameter decreased, more missing edges were acceptable in the model, and these four disjoint complex estimates were combined into two complexes for sensitivity values of .80 and .70, and finally one complex for a sensitivity of .60. In other cases when the sensitivity was set to .90, such as the protein complex with eleven proteins and seven baits, the complex estimates were forced to remain so disjointed that values of $\omega > .70$ were not achieved by any of the complex estimates. Once the sensitivity values were decreased, enough proteins were included in the estimate to be uniquely associated with that complex.

For larger complexes, more permissive sensitivity values were effective at uniquely mapping estimates to true complexes, but for smaller complexes, too many proteins were often permitted in the estimate, thus overestimating complex composition. For the complex with seven proteins and three baits, the sensitivity value of .60 allowed several additional proteins to enter the complex, and the complex was too big to be uniquely identified with the a true complex. This same complex had several proposals as unique estimates for the other sensitivity parameters.

We were unable to uniquely identify any of the five complexes with only one bait for two main reasons. First, we specified a .20 probability of not observing an edge, given it did truly exist. Several of these bait-hit edges were not observed in the data, and since there was only one attempt at finding them in the experiment, this prevented the possibility of predicting complex comembership for these proteins. Second, the bait proteins involved in these five

Table 1: Simulation Values for $\omega = \min(\frac{i}{a}, \frac{i}{b})$

$\frac{\exp\{\mu+\alpha\}}{1+\exp\{\mu+\alpha\}} = .60$

Complex	Number of Bait Proteins							
Size	1	2	3	4	5	6	7	8
2	0	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-
4	-	-	0	-	-	-	-	-
5	-	1,0,1	-	-	-	-	-	-
6	0	-	1,0	1	-	-	-	-
7	-	1	0	1,1	-	-	-	-
8	0,0	1	1,0,1	-	-	-	-	-
9	-	1	1,1	-	-	-	-	-
10	0	0,1	1,0	-	1	1,1	-	-
11	-	0,0	1	-	-	1	1	-
12	-	-	1	1,1	-	1	-	-
13	-	-	1	-	1	1	1	-
14	-	-	0	-	1	1	-	1
15	-	-	-	-	-	-	1,1	1
16	-	-	-	-	-	-	-	-
17	-	-	-	-	1	-	1	-

$\frac{\exp\mu+\alpha}{1+\exp\mu+\alpha} = .70$

Complex	Number of Bait Proteins							
Size	1	2	3	4	5	6	7	8
2	0	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-
4	-	-	0	-	-	-	-	-
5	-	1,1,0	-	-	-	-	-	-
6	0	-	1,0	1	-	-	-	-
7	-	1	3	1,1	-	-	-	-
8	0,0	1	1,1,2	-	-	-	-	-
9	-	1	1,1	-	-	-	-	-
10	0	0,0	1,0	-	1	1,1	-	-
11	-	1,1	3	-	-	2	2	-
12	-	-	1	1,1	-	1	-	-
13	-	-	2	-	1	1	1	-
14	-	-	2	-	2	1	-	1
15	-	-	-	-	-	-	1,1	1
16	-	-	-	-	-	-	-	-
17	-	-	-	-	2	-	2	-

$\frac{\exp\{\mu+\alpha\}}{1+\exp\{\mu+\alpha\}} = .80$

Complex	Number of Bait Proteins							
Size	1	2	3	4	5	6	7	8
2	0	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-
4	-	-	0	-	-	-	-	-
5	-	1,1,0	-	-	-	-	-	-
6	0	-	1,1	1	-	-	-	-
7	-	1	3	2,1	-	-	-	-
8	0,0	2	2,1,2	-	-	-	-	-
9	-	1	1,1	-	-	-	-	-
10	0	0,0	1,0	-	1	1,1	-	-
11	-	2,1	3	-	-	3	2	-
12	-	-	1	2,0	-	2	-	-
13	-	-	2	-	1	2	2	-
14	-	-	3	-	2	2	-	2
15	-	-	-	-	-	-	1,1	2
16	-	-	-	-	-	-	-	-
17	-	-	-	-	2	-	2	-

$\frac{\exp\{\mu+\alpha\}}{1+\exp\{\mu+\alpha\}} = .90$

Complex	Number of Bait Proteins							
Size	1	2	3	4	5	6	7	8
2	0	-	-	-	-	-	-	-
3	-	-	-	-	-	-	-	-
4	-	-	0	-	-	-	-	-
5	-	1,1,0	-	-	-	-	-	-
6	0	-	1,1	1	-	-	-	-
7	-	1	4	2,1	-	-	-	-
8	0,0	2	2,1,2	-	-	-	-	-
9	-	1	2,2	-	-	-	-	-
10	0	0,0	0,0	-	1	1,1	-	-
11	-	2,1	4	-	-	0	0	-
12	-	-	2	0,0	-	1	-	-
13	-	-	3	-	0	2	1	-
14	-	-	4	-	2	3	-	4
15	-	-	-	-	-	-	2,3	2
16	-	-	-	-	-	-	-	-
17	-	-	-	-	4	-	3	-

complexes were also involved in other complexes. The SBMH complexes remaining at the end of the complex estimation algorithm included components of the actual SBMH complexes as subsets. The lack of other purification information for these complexes prevented their unique identification. In general, increasing complex size, and an increased number of bait proteins, facilitated more unique mappings from an estimated complex to a true complex.

The largest amount of unique mappings to true complex estimates occurred for the sensitivity value of .70. The improved estimates for a lower sensitiv-

DOI: 10.2202/1544-6115.1107

Table 2: Summary Statistics for Simulation Complexes

$P(0\|1)$	total # of estimated complexes	# identified with true complex	# UnRBB complexes	# SBMH complexes
.60	281	35	116	86
.70	333	53	117	136
.80	386	65	117	155
.90	463	71	117	166

ity value than the actual true positive probability reflects the modal binomial distribution. The true positive probability of .80 generally removes approximately 20% of the edges from the complete graph, but many of the graphs lose more than this amount and are still consistent with the .80 sensitivity. The use of a lower sensitivity rate than actually expected for the AP-MS technology in the complex estimation algorithm will accomodate such complexes.

For the first of the hand-crafted complexes, we were able to uniquely identify this 10-protein-6 bait complex, as well as its original counterpart using all four sensitivity parameters. For a large complex with a high proportion of baits, the algorithm is fairly robust to the parameter specification. The second hand-crafted complex was identical to another 10-protein complex with 3 baits, only with 2 different hits. Due to the nature of the bait-hit data, we would not expect to be able to uniquely identify this complex. In fact, the connectivity for this 52^{nd} complex happened to be quite low, and we were unable to identify a complex that closely resembled this complex using any of the sensitivity specifications. The 53^{rd} complex consisted of the same proteins as a different 5 protein complex, but with one bait and one hit switched. The original complex was successfully identified for all parameter specifications, but due to poor connectivity, the 53^{rd} complex only had a successfully overlapping estimate for a sensitivity value of .60. The 54^{th} complex consisted of a subset of 3 baits from a larger complex, with one extra hit. This complex was reported as a subset of these three baits with all other hits from the original complex, as well as the extra hit. If this extra hit had in fact been a bait protein, we likely would have been able to uniquely identify this complex. The 55^{th} complex was a combination of 3 baits and 4 hits from two other complexes, and was uniquely identified for sensitivity values of .80 and .90. For the more permissive parameter specifications, these proteins were combined with several of the extra hits from the original complexes. The re-

sults for these five hand-designed complexes demonstrate the dependence of the complex estimates upon the bait/hit status of the protein, the sensitivity value specification, and the amount of overlap with other complexes. Since one sensitivity value does not necessarily work better uniformly over all complex sizes, it may be advisable to perform the estimation procedure with a few different values, and then examine the results for a sense of the connectivity of the proteins in the predicted complexes.

Table 2 reports the total number of estimated complexes, the number of complexes identified with one of the true complexes in A, the number of UnRBB complexes, and the number of SBMH complexes for all sensitivity values. A predicted complex was determined to be associated with a true complex if $\omega > .7$. This criteria requires an overlap of at least 70% of the proteins from the perspective of both the predicted complex and the true complex. The complexes that are uniquely identified with a true complex tend to be complexes with more than one bait, with increasing accuracy as the total complex size increases. The large number of UnRBB complexes are primarily due to the generous portion of FP observations entered into the simulation, and the SBMH complexes in general contain complexes that lack adequate edge data for unique identification. The sensitivity value of .70 most successfully mapped 53 complex estimates to the 55 true complexes.

Results using Publicly Available AP-MS Data

Both Gavin et al. (2002) and Ho et al. (2002) published data sets resulting from high throughput AP-MS experiments. The two groups have somewhat different lab procedures that affect the number of observed bait-hit pairs. Gavin et al. use a tandem-affinity purification (TAP) process that consists of two purification steps and then identify the hits using peptide mass fingerprinting by matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) MS. In a method they term high-throughput mass spectromic protein complex identification (HMS-PCI), Ho, et al. overexpress the bait proteins, perform a one-step purification procedure, and then identify the hits using ultrasensitive liquid chromatography (LC)-tandem MS. A comparison of the two techniques is made in von Mering et al. (2002) and Bader and Hogue (2002).

The TAP data consist of 589 'raw' purifications, available in Supplementary Table 1 of Gavin et al. at http://www.nature.com. Excluding homodimers, there are 455 bait proteins and 909 hit-only proteins. Gavin et al. group the purifications into 232 annotated "yTAP" complexes, available in Supplementary Table 3 of Gavin et al. at http://www.nature.com and at

DOI: 10.2202/1544-6115.1107

http://yeast.cellzome.com. The exact method used to perform this grouping is not described, but it is said to be made on the basis of substantial overlaps between the purifications (p. 143 in Gavin et al.).

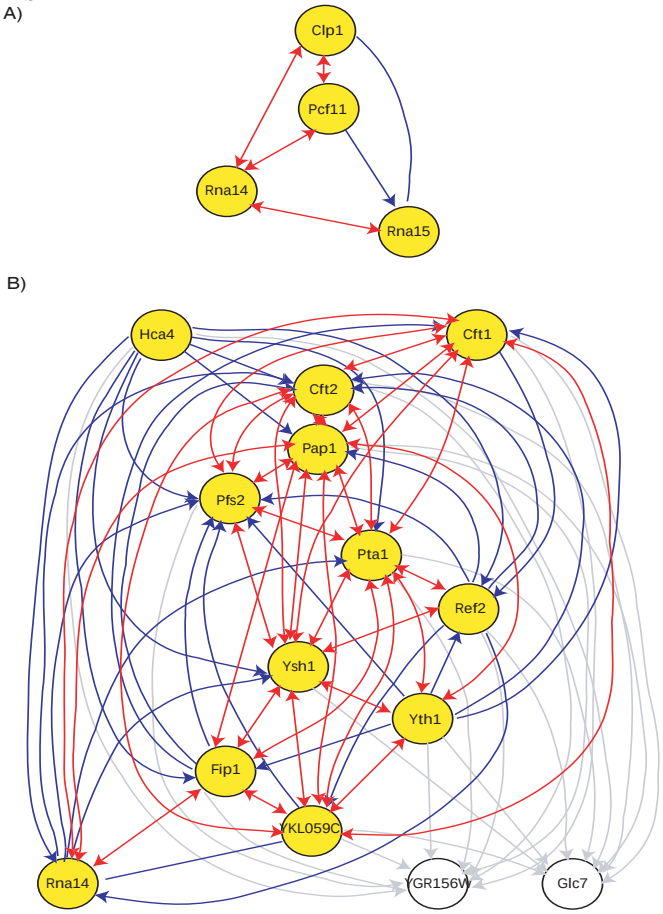
For our analysis of the TAP data, we specified a sensitivity of 0.75 and specificity of .995, and used a GO-based similarity measure in an extended logistic regression model with $\beta = -0.2$ (see Supporting Information). Our complex identification algorithm predicted a total of 708 complexes, including 123 UnRBB pairs, 325 SBMH complexes, and 260 MBME complexes. We compared several of the 260 MBME complexes with the yTAP complexes and found that our estimates more closely reflected well-characterized protein complexes. Complexes such as Arp2/3, Orc, PP2A, and RNA polymerases I, II, and III demonstrate the accuracy of our results. Readers are encouraged to investigate all 708 complexes at http://www.bioconductor.org/Docs/Papers/2003/apComplex. As a specific example, here we discuss the results for the messenger RNA cleavage and polyadenylation factors as they compare to the yTAP predictions.

Messenger RNA cleavage and polyadenylation requires the cooperativity of cleavage factor I (CFI), polydenylation factor I (PFI), and Pap1 (poly(A) polymerase). Current belief is that CFI is composed of two subunits, CFIA (Rna14, Rna15, Pf11, Clp1) and CFIB (Hrp1). PFI is believed to consist of CFII, a complex of four units (Cft1, Cft2, Ysh1, and Pta1) and the additional proteins Pfs1, Pfs2, Fip1, Yth1, Mpe1, and Pti1 (Gross and Moore, 2001, Skaar and Greenleaf, 2002, Zhao et al., 1997, Vo et al., 2001, Russnak et al., 1995). Our algorithm distinguishes between the two distinct CFI and PFI complexes, as demonstrated in Figure 3. We estimate that Rna14 is part of the PFI complex; perhaps it serves in a communicative role between PFI and CFI. Gavin, et al. report CFI and PFI as part of the same complex in yTAP-C162 (see Supporting Information).

The HMS-PCI data set, available at http://www.mdsp.com/yeast, contains 493 bait proteins, and 1085 hit-only proteins. We applied our complex estimation algorithm to these data with the same GO-based similarity measure, β value of -.2, and sensitivity of .75, but specified a slightly higher false positive probability of .01 compared to the .005 probability used for the TAP analysis. The HMS-PCI lab process only includes one round of mild purification, and Bader and Hogue (2002) suggest that LC-tandem MS, while very sensitive, does have the propensity to identify low level background proteins that are not specifically associated with other proteins in the purification.

Our analysis resulted in a total of 1008 complexes including 329 UnRBB complexes and 437 SBMH complexes. The remaining 242 MBME complexes were in some cases similar to those found in the TAP data. For example, the

Figure 3: Our Complex Estimates for Cleavage Factor I and Polyadenylation Factor I Proteins



COPI coatomer complex is known to consist of seven proteins: Cop1, Ret2, Ret3, Sec21, Sec26, Sec27, and Sec28 (Duden et al., 1998). Using the TAP data, we identify a complex containing all seven, along with an extra protein Mrpl10. Using the HMS-PCI data, we identify a seven-subunit complex containing all COPI coatomer proteins except Ret3 and an additional protein Prb1 (see Supporting Information). Ret3 was not among the set of proteins reported with the HMS-PCI data. The identification of this protein complex using both data sets confirms the reproducibility of results for overlapping baits in different experiments. The observations of Mrpl10 and Prb1 interacting with this complex may be spurious, or they may suggest avenues for further investigation.

DOI: 10.2202/1544-6115.1107

The set of proteins included in the HMS-PCI data set was quite different from that of the TAP data set with only 81 common bait proteins, thus enabling the identification of complexes that were otherwise unobserved in the TAP data. One example is the three-subunit Rad50-Mre11-Xrs2 complex involved in several biological processes including homologous recombination and DNA damage signaling (Trujillo et al., 2003) (see Supporting Information). We identify this complex using the HMS-PCI data, but we do not identify the complex using the TAP data since none of the three were used as baits in the TAP experiment.

A large scale comparison of our TAP complexes and Gavin, et al.'s yTAP groupings with a list of 267 curated protein complexes available at MIPS (ftp://ftpmips.gsf.de/yeast/catalogues/ complexes/complex_130603) shows that we accurately identify approximately twice the number of previously characterized complexes with very high compositional accuracy compared to Gavin, et al. The details of this comparison are available in Scholtens et al. (2004). High correspondence with previously characterized protein complexes confirms the general accuracy of the complex membership catalog estimated by our algorithm.

Conclusion

Local dynamic models of protein complex membership offer a characterization of functional modules that compose cellular systems at a finer level of detail than overall network topology descriptions. We apply graph theoretic and statistical principles to accommodate biological particulars of the dynamic modeling problem, as well as the available Y2H and AP-MS data, resulting in local models that closely reflect well-documented complexes. The accuracy with which we can estimate previously characterized protein complexes suggests the utility of our results for designing new biological investigations. We can predict new complexes, suggest complex involvement on behalf of uncharacterized proteins, and investigate communication between cellular systems by ascertaining complex comembership for proteins known to be associated with different pathways. Joint analyses of protein complexes with other data, such as Y2H physical interactions, binding domains, and gene expression profiles are a promising next step for investigation into the physical mechanics and transcriptional control of the modular systems responsible for cellular activity. With a catalog of functional macromolecules in place, systems biology modeling can proceed with greater precision.

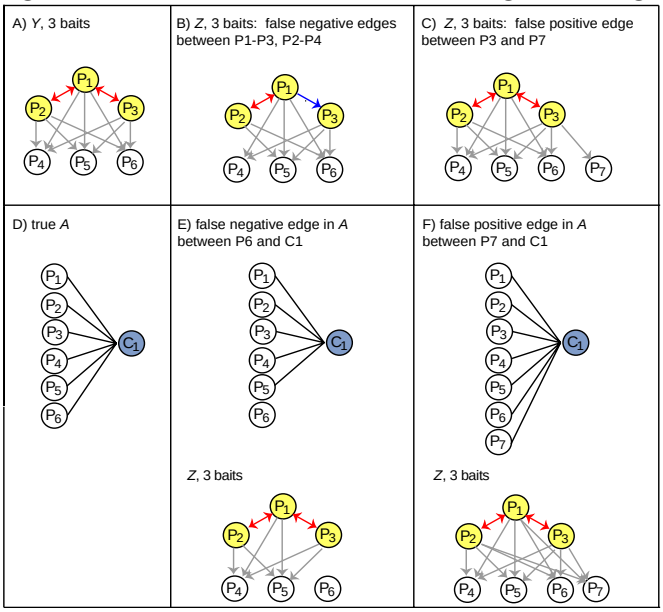
Supporting Information

Scholtens and Gentleman, Making Sense of High-Throughput Protein-Protein Interaction Data

Figure 4: False Positive and False Negative Edges

We note here that the notions of false negative (FN) and false positive (FP) apply to observations of the edges in Y_{V_B}, not the edges in A. If Y_{V_B} is as depicted in Figure 4A), then we might actually observe FN edges like those missing in Z_{V_B} between $P_2 - P_4$ and $P_1 - P_3$ in Figure 4B). We might also observe FP edges, such as the edge between $P_3 - P_7$ in Figure 4C). If a FN edge was assumed to be at the level of A, as between P_6 and C_1 in Figure 4E), then Z_{V_B} would have three missing edges between $P_1 - P_6$, $P_2 - P_6$, and $P_3 - P_6$. If a FP was assumed to occur in A, as between P_7 and C_1 in Figure 4F), then three FP edges would be observed between $P_1 - P_7$, $P_2 - P_7$, and $P_3 - P_7$. Since AP-MS technology directly assays the graph $Y_{V_B} = (A \otimes A')_{V_B}$, not A, we account for random errors in the edges of Z_{V_B} as they reflect Y_{V_B}.

Figure 4: False Positive and False Negative Edges



DOI: 10.2202/1544-6115.1107

Complex Combination Acceptance Criteria

The complex estimation procedure iteratively updates $\hat{A}$, beginning with the initial maximal subgraph estimate for the complexes $\hat{A}_{init}$ by proposing pairwise combinations of the complexes recorded in the columns of $\hat{A}$. If the set of proteins associated with two complexes in $\hat{A}$ increase $P(Z|A,\mu,\alpha)$ when treated as one complex, where

$$P(Z|A,\mu,\alpha) = L(Z|Y = A \otimes A', \mu, \alpha) \times C(Z|A,\mu,\alpha), \tag{4}$$

then the combination is accepted. The acceptance criteria is quite simple to test. Suppose complexes c_{k1} and c_{k2} with subgraphs $G_{c_{k1}}$ and $G_{c_{k2}}$, respectively, are treated as one complex c_{k*} with subgraph $G_{c_{k*}}$. Let $S_{new} = \{(g,h) : e_{gh} \in G_{c_{k*}} \backslash (G_{c_{k1}} \cup G_{c_{k2}})\}$ where e_{gh} is the edge connecting bait protein g to hit protein h; that is, S_{new} is the set of all edges that are part of $G_{c_{k*}}$ that were not originally part of either $G_{c_{k1}}$ or $G_{c_{k2}}$. Then the difference in the log of $P(Z|A,\mu,\alpha)$ when c_{k1} and c_{k2} are combined into c_{k*}, P_{k*}, versus uncombined, $P_{k1,k2}$ is

$$\begin{aligned}
\log P_{k*} - \log P_{k1,k2} &= \log\Phi(c_{k*}) - \log\Phi(c_{k1}) - \log\Phi(c_{k2}) \\
&\quad + \log\Gamma(_{k*}) - \log\Gamma(c_{k1}) - \log\Gamma(c_{k2}) \\
&\quad + \sum_{S_{new}} [\alpha z_{gh} - \log(1+e^{\mu+\alpha}) + \log(1+e^{\mu})] \\
&= \log\Phi(c_{k*}) - \log\Phi(c_{k1}) - \log\Phi(c_{k2}) \\
&\quad \log\begin{pmatrix} t_{k*} \\ x_{k*} \end{pmatrix} - \log\begin{pmatrix} t_{k1} \\ x_{k1} \end{pmatrix} - \log\begin{pmatrix} t_{k2} \\ x_{k2} \end{pmatrix} \\
&\quad + \sum_{i \in b_{k*}} \sum_{j \in b_{k*} \cup h_{k*}, j \neq i} [z_{ij}(\mu+\alpha) - \log(1+e^{\mu+\alpha})] \\
&\quad - \sum_{q \in b_{k1}} \sum_{r \in b_{k1} \cup h_{k1}, r \neq q} [z_{qr}(\mu+\alpha) - \log(1+e^{\mu+\alpha})] \\
&\quad - \sum_{u \in b_{k2}} \sum_{v \in b_{k2} \cup h_{k2}, v \neq u} [z_{uv}(\mu+\alpha) - \log(1+e^{\mu+\alpha})] \\
&\quad + \sum_{S_{new}} [\alpha z_{gh} - \log(1+e^{\mu+\alpha}) + \log(1+e^{\mu})]. \qquad (5)
\end{aligned}$$

In our algorithm, if $\log P_{k*} - \log P_{k1,k2}$ (5) is larger than zero, then the combination of c_{k1} and c_{k2} into c_{k*} is accepted.

Incorporating External Similarity Data

A similarity measure $0 \leq s_{ij} \leq 1$, based on data external to the AP-MS experiment for each pair of proteins i and j, can be included in the logistic regression model for p_{ij} with an amount of influence determined by a parameter β. Let $p_{ij} = Pr(Z_{ij} = 1|\mu, \alpha, \beta, Y_{ij}, S_{ij} = s_{ij})$ and

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \mu + \alpha Y_{ij} + \beta s_{ij}. \tag{6}$$

Specification of the β parameter is best done by examining the contribution of the bait-bait observations to the likelihood. If a bait-bait pair connected by one unreciprocated edge has a high similarity measure $s_{ij} = s_{ji}$, then the contribution of the $(0,1)$ or $(1,0)$ observations to $L(Z|Y = A \otimes A', \mu, \alpha, \beta, S)$ should approach that of the $(1,1)$ edges. When $Y_{ij} = Y_{ji} = 1$ the difference in the log likelihood contribution ℓ for $(1,1)$ edges and $(0,1)$ (or $(1,0)$) edges is

$$\ell_{(1,1)} - \ell_{(0,1)} = \mu + \alpha + \beta s_{ij}. \tag{7}$$

For $\ell_{(1,1)} - \ell_{(0,1)}$ (7) to decrease in s_{ij}, it must be that β is less than 0. The assumption that $P(Z_{ij} = 1|Y_{ij} = 1, S_{ij} = s_{ij}) > .5$, indicating that the AP-MS technology has sensitivity greater than .5 independent of the similarity measure, further requires that β is greater than $-(\mu + \alpha)$.

High similarity measures for $(0,0)$ bait-bait observations, or for 0 bait-hit observations, should lend credence to the existence of the edge, even though it is missing according to the AP-MS data. The difference in the log likelihood contribution for an unobserved edge when $\hat{Y}_{ij} = 1$ and $\hat{Y}_{ij} = 0$ is

$$\ell_{\hat{Y}_{ij}=1} - \ell_{\hat{Y}_{ij}=0} = -\log(1 + \exp\{\mu + \alpha + \beta s_{ij}\}) + \log(1 + \exp\{\mu + \beta s_{ij}\}). \tag{8}$$

For each unobserved edge, $\ell_{\hat{Y}_{ij}=1} - \ell_{\hat{Y}_{ij}=0} < 0$, but for increasing values of s_{ij}, $\ell_{\hat{Y}_{ij}=1} - \ell_{\hat{Y}_{ij}=0}$ increases. When combining two complexes with fairly high connectivity and/or high similarity measures, the negative contribution of a few unobserved edges with high similarity scores may well be overruled by the strength of the evidence of the other edges in the subgraph.

The value for p_{ij} that depends on s_{ij} can also be incorporated into the $\Gamma(c_k)$ values in $C(Z|A, \mu, \alpha)$. Since the value of p_{ij} decreases with increasing s_{ij} values, a lower proportion of edges in the graph must then be observed for the complex to be consistent with the maximal BH-complete graph structure.

The simple decision criteria for combining two complexes can be expanded to include the similarity measure as follows:

DOI: 10.2202/1544-6115.1107

$$
\begin{aligned}
\log P_{k*} - \log P_{k1,k2} &= \log \Phi(c_{k*}) - \log \Phi(c_{k1}) - \log \Phi(c_{k2}) \\
&\quad + \log \Gamma(_{k*}) - \log \Gamma(c_{k1}) - \log \Gamma(c_{k2}) \\
&\quad + \sum_{S_{new}} [\alpha z_{gh} - \log(1 + e^{\mu+\alpha+\beta s_{gh}}) + \log(1 + e^{\mu+\beta s_{gh}})] \\
&= \log \Phi(c_{k*}) - \log \Phi(c_{k1}) - \log \Phi(c_{k2}) \\
&\quad \log \begin{pmatrix} t_{k*} \\ x_{k*} \end{pmatrix} - \log \begin{pmatrix} t_{k1} \\ x_{k1} \end{pmatrix} - \log \begin{pmatrix} t_{k2} \\ x_{k2} \end{pmatrix} \\
&\quad + \sum_{i \in b_{k*}} \sum_{j \in b_{k*} \cup h_{k*}, j \neq i} [z_{ij}(\mu + \alpha + \beta s_{ij}) - \log(1 + e^{\mu+\alpha+\beta s_{ij}})] \\
&\quad - \sum_{q \in b_{k1}} \sum_{r \in b_{k1} \cup h_{k1}, r \neq q} [z_{qr}(\mu + \alpha + \beta s_{qr}) - \log(1 + e^{\mu+\alpha+\beta s_{qr}})] \\
&\quad - \sum_{u \in b_{k2}} \sum_{v \in b_{k2} \cup h_{k2}, v \neq u} [z_{uv}(\mu + \alpha + \beta s_{uv}) - \log(1 + e^{\mu+\alpha+\beta s_{uv}})] \\
&\quad + \sum_{S_{new}} [\alpha z_{gh} - \log(1 + e^{\mu+\alpha+\beta s_{gh}}) + \log(1 + e^{\mu+\beta s_{gh}})]. \qquad (9)
\end{aligned}
$$

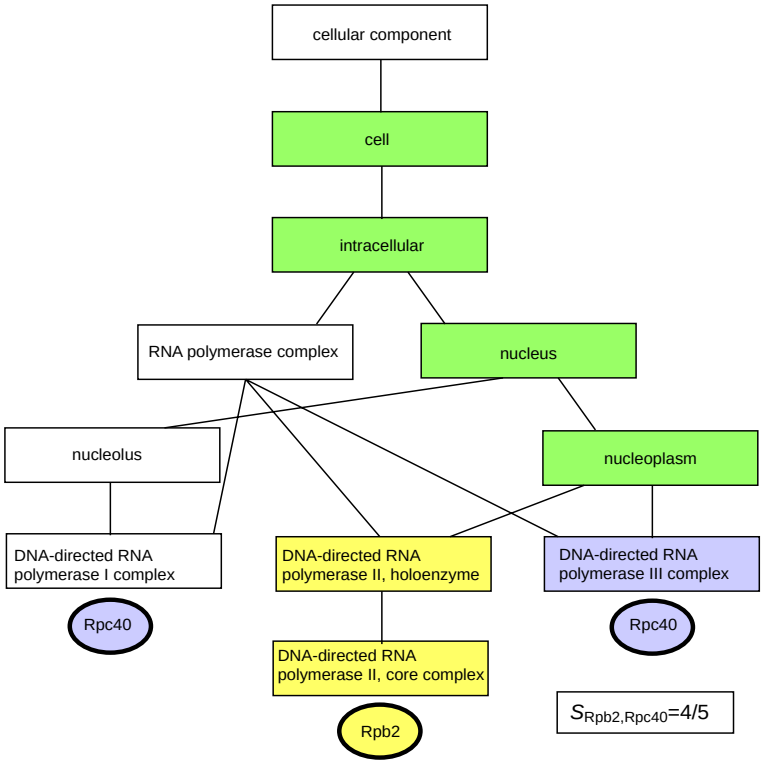
If $\log P_{k*} - \log P_{k1,k2}$ (9) is greater than zero, then c_{k1} and c_{k2} are combined to form c_{k*}.

Figure 5: Example of Similarity Measure using GO Cellular Component Annotation

In our analysis of publicly available data, we used a similarity measure for the proteins based on cellular component data available from GO. GO annotation is based on a rooted acyclic directed graph in which nodes represent annotation terms, and directed edges connect one annotation term to a more specific, but related, annotation term. The GO graph allows for several paths from the root to the annotation node for a protein. Furthermore, proteins may be annotated at different nodes, also resulting in multiple paths through the graph for a particular protein. For two proteins i and j, we calculated all of the paths from the root of the GO graph to the node(s) at which the proteins were annotated. We then calculated $s_{ij} = \max(g_{max}/p_i, g_{max}/p_j)$ where p_i is the number of nodes on the path for protein i and p_j is the number of nodes on the path for protein j that maximally overlap at g_{max} nodes. We used this measure for analyses of the publicly available AP-MS data with $\beta = -.2$. Our GO-based similarity measure for proteins Rbp2 and Rpc40 is demonstrated

in Figure 5. Rpb2 is annotated at the GO tree node "DNA-directed RNA polymerase II, core complex" and Rpc40 is annotated at both "DNA-directed RNA polymerase I complex" and "DNA-directed RNA polymerase III complex". The similarity measure for these two proteins is 4/5 using s_{ij} defined as above.

Figure 5: Example of Similarity Measure using GO Cellular Component Annotation



DOI: 10.2202/1544-6115.1107

Maximal BH-Complete Subgraph Identification Algorithm

Efficient maximal subgraph finding for undirected graphs has long been a topic of research among statisticians and computer scientists. The methods that exist for undirected graphs assume (necessarily) that the adjacency matrix corresponding to the graph is both square and symmetric. The available adjacency matrix for AP-MS data is the rectangular $N(N+M)$ matrix Z, the observed form of $Y[1:N,]$. The algorithm we apply for maximal subgraph identification begins after initial estimates of (Y_{ij}, Y_{ji}) for the unreciprocated doubly tested edges have been made, and $Y_{init}[1:N, 1:N]$ is symmetric. Since the connectivity between proteins that were only found as hits and never used as baits is unknown, we effectively assume that all hits connected to a common bait are also connected to each other. This permits hits to be identified as part of the same maximal subgraph, and eventually part of the same protein complex, as long as their connectivity to the baits resembles that of the other hits.

Given input of $Y = Y_{init}[1:N,]$ the algorithm proceeds as described in Supporting Information. The resultant maximal BH-complete subgraphs are represented as an affiliation matrix called M with $N+M$ rows and K_{MG} columns where K_{MG} is the number of maximal BH-complete subgraphs.

Algorithm

Bait-Hit Maximal Subgraphs(Y)
Argument: Y the $N \times (N+M)$ bait-hit adjacency matrix with $Y[1{:}N, 1{:}N]$ symmetric;
Function: *REDUCEMAT* removes columns of a matrix for which all elements are less than all elements of another column in the matrix;

```
1   begin
2     M := Y[1,]
3     for each i ∈ {2, ..., N} do
4       Set g = repeat(0, N + M)
5       g[i : (N + M)] = Y[i, i : (N + M)]
6       G = ∅
7       V = {columns of M : ∀ v ∈ V, v[i] = 1 and v[j] > g[j] for some j ∈ {i + 1, ..., N + M}
8       if (V = ∅) then continue
9       else do
10        for each v ∈ V do
11          M = M\v
```

$v_1 = v$
$v_1[i] = 0$
$G = G \cup v_1$
$v_2 = v$
$v_2[(i+1):(N+M)] = \min(v[(i+1):(N+M)], g[(i+1):(N+M)])$
$G = G \cup v_2$
$G = G \cup g$
$\mathcal{M} = \mathcal{M} \cup G$
$\mathcal{M} = REDUCEMAT(\mathcal{M})$
return $\mathcal{M}$

DOI: 10.2202/1544-6115.1107

Figure 6: yTAP-C162, Gavin, et al.(2002)

Below is a graph representation of the cleavage factor I (CFI) and polyadenylation factor I (PFI) complexes as reported by Gavin, et al. While these two complexes are functionally dependent, they are physically separate in the cell. We were able to distinguish between the two complexes using our algorithm and the TAP data.

Figure 6: yTAP-C162, Gavin et al. (2002)

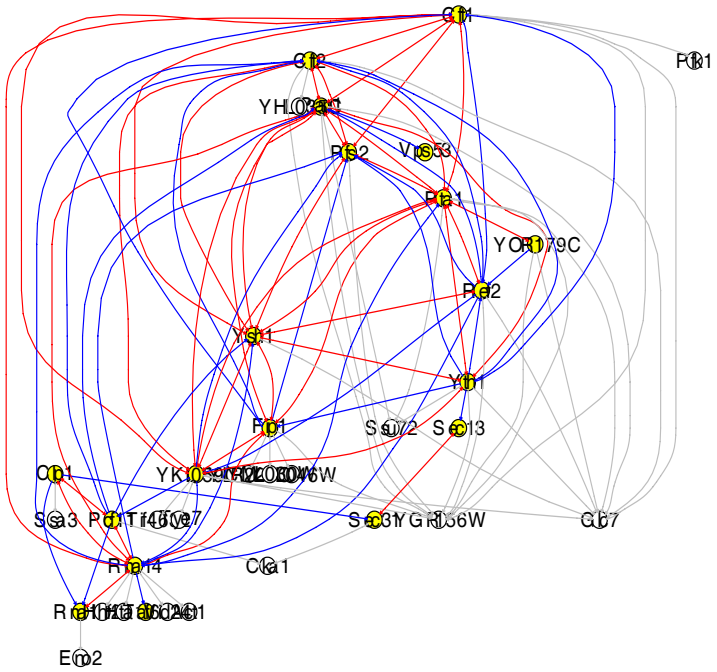
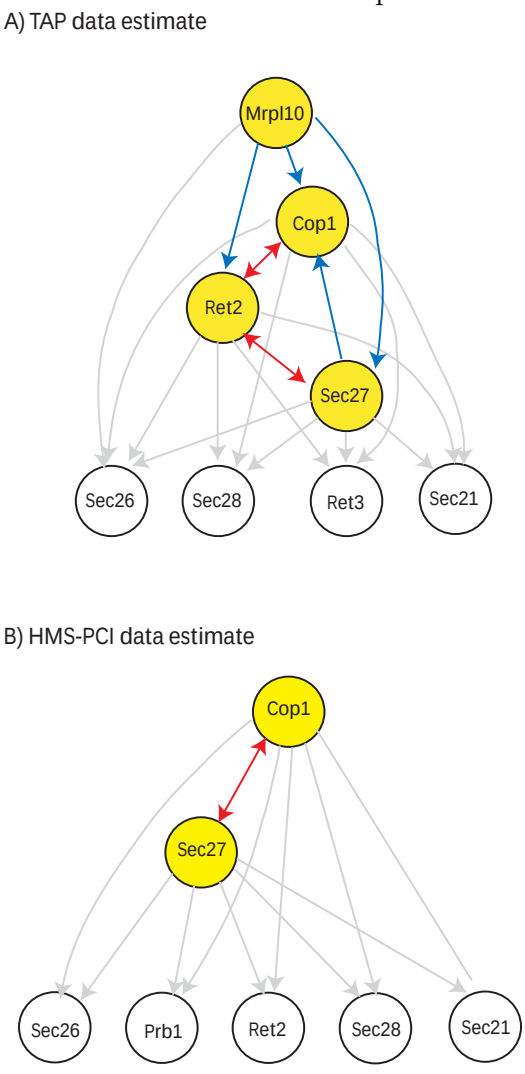


Figure 7: COPI Coatomer Complex Estimates

Below is a graph representation of the COPI coatomer complex estimates using both the TAP data and the HMS-PCI data. Seven proteins are known to compose the COPI coatomer complex: Cop1, Ret2, Ret3, Sec21, Sec26, Sec27, and Sec28. Ret3 was not reported in the HMS-PCI data set. The interactions of Mrpl10 and Prb1 with this complex may be spurious observations, or they may suggest areas of further research.

Figure 7: COPI coatomer Complex Estimates

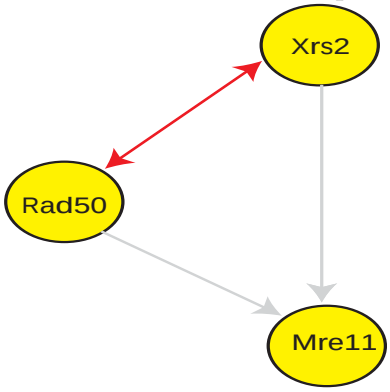


DOI: 10.2202/1544-6115.1107

Figure 8: Rad50-Mre11-Xrs2 Complex Estimate

Below is a graph representation of the Rad50-Mre11-Xrs2 complex, detected using the HMS-PCI data but not the TAP data. These three proteins were not used as baits in the TAP data.

Figure 8: Rad50-Mre11-Xrs2 Complex Estimate



References

Bader, G.D. and Hogue, C.W.V. Analyzing yeast protein-protein interaction data obtained from different sources. *Nature Biotechnol.*, 20:991–997, 2002.

Bron, C. and Kerbosch, J. Finding all cliques of an undirected graph [H]. *Communications of the ACM*, 16(9):575–577, 1973.

Duden, R., Kajikawa, L., Wuestehube, L., and Schekman, R. epsilon-COP is a structural component of coatomer that functions to stabilize alpha-COP. *EMBO J.*, 17:985–995, 1998.

Gavin, A.-C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., and others, . Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415:141–147, 2002.

Gross, S. and Moore, C. Five subunits are required for reconstitution of the cleavage and polyadenylation activities of *Saccharomyces cerevisiae* cleavage factor I. *Proc. Nat. Acad. Sci. U.S.A.*, 98(11):6080–6085, 2001.

Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., and others, . Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature*, 415:180–183, 2002.

Ihaka, R. and Gentleman, R. R: A language for data analysis and graphics. *J. Comput. Graph. Stat.*, 5:299–314, 1996.

Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., and Gerstein, M. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302:449–453, 2003.

Jeong, H., Mason, S.P., Barabási, A.-L., and Oltvai, Z.N. Lethality and centrality in protein networks. *Nature*, 411:41–42, 2001.

Russnak, R., Nehrke, K.W., and Platt, T. REF2 encodes an RNA-binding protein directly involved in yeast mRNA 3'-end formation. *Mol. Cell. Biol.*, 15(3):1689–1697, 1995.

Salwinski, L. and Eisenberg, D. Computational methods of analysis of protein-protein interactions. *Curr. Opin. Struct. Biol.*, 13:377–382, 2003.

DOI: 10.2202/1544-6115.1107

Scholtens, D., Vidal, M., and Gentleman, R. Local modeling of global interactome networks. *Submitted*, 2004.

Skaar, D.A. and Greenleaf, A.L. The RNA polymerase II CTD kinase CTDK-I affects pre-mRNA 3' cleavage/polyadenylation through the processing component Pti1p. *Mol. Cell*, 10:1429–1439, 2002.

Trujillo, K.M., Roh, D.H., Chen, L., Van Komen, S., Tomkinson, A., and Sung, P. Yeast Xrs2 binds DNA and helps target Rad50 and Mre11 to DNA ends. *J. Biol. Chem.*, 278(49):48957–48964, 2003.

Vo, L.T.A., Minet, M., Schmitter, J.-M., Lacroute, F., and Wyers, F. Mpe1, a zinc knuckle protein, is an essential component of yeast cleavage and polyadenylation factor required for the cleavage and polyadenylation of mRNA. *Mol. Cell. Biol.*, 21(24):8346–8356, 2001.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., and Bork, P. Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417:399–403, 2002.

Wasserman, S. and Faust, K. *Social Network Analysis*. Cambridge University Press, New York, 1999.

Zhao, J., Kessler, M.M., and Moore, C.L. Cleavage factor II of *Saccharomyces cerevisiae* contains homologues to subunits of the mammalian cleavage/polyadenylation specificity factor and exhibits sequence-specific, ATP-dependent interaction with precursor RNA. *J. Biochem.*, 272(16):10831–10838, 1997.