Local modeling of global interactome networks

Denise Scholtens1,∗,†, Marc Vidal2 and Robert Gentleman3

1Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA, 2Center for Cancer Systems Biology and Department of Cancer Biology, Dana-Farber Cancer Institute, and Department of Genetics, Harvard Medical School, Boston, MA 02115, USA and 3Program in Computational Biology, Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

Received on January 31, 2005; revised on June 13, 2005; accepted on June 28, 2005
Advance Access publication July 5, 2005

ABSTRACT

Motivation: Systems biology requires accurate models of protein complexes, including physical interactions that assemble and regulate these molecular machines. Yeast two-hybrid (Y2H) and affinity–purification/mass-spectrometry (AP–MS) technologies measure different protein–protein relationships, and issues of completeness, sensitivity and specificity fuel debate over which is best for high-throughput ‘interactome’ data collection. Static graphs currently used to model Y2H and AP–MS data neglect dynamic and spatial aspects of macromolecular complexes and pleiotropic protein function.

Results: We apply the local modeling methodology proposed by Scholtens and Gentleman (2004) to two publicly available datasets and demonstrate its uses, interpretation and limitations. Specifically, we use this technology to address four major issues pertaining to protein–protein networks. (1) We motivate the need to move from static global interactome graphs to local protein complex models. (2) We formally show that accurate local interactome models require both Y2H and AP–MS data, even in idealized situations. (3) We briefly discuss experimental design issues and how bait selection affects interpretability of results. (4) We point to the implications of local modeling for systems biology including functional annotation, new complex prediction, pathway interactivity and coordination with gene-expression data.

Availability: The local modeling algorithm and all protein complex estimates reported here can be found in the R package apComplex, available at http://www.bioconductor.org

Contact: dscholtens@northwestern.edu

Supplementary information: Available at http://daisy.prevmed.northwestern.edu/~denise/pubs/LocalModeling

INTRODUCTION

Cellular systems depend on multiprotein complexes in which individual proteins assemble into functional modules (Alberts, 1998; Hartwell et al., 1999). Complexes may have stable or dynamic composition, and many share common proteins as members. Complex co-members may physically bind to each other, or instead directly interact with common bridge proteins. Large-scale experiments in Saccharomyces cerevisiae (Gavin et al., 2002; Ho et al., 2002; Ito et al., 2001; Uetz et al., 2000), Caenorhabditis elegans (Li et al., 2004), and Drosophila (Giot et al., 2003) continue to add to the vast compilation of data characterizing two different protein–protein relationships. Yeast two-hybrid (Y2H) detects binary physical protein–protein interactions, and affinity–purification/mass-spectrometry (AP–MS) identifies hit proteins that are co-members of complexes with bait proteins without regard to physical complex topology. Graph theoretic analyses have led to topological descriptions of the overall interactome (Jeong et al., 2001; Salwinski and Eisenberg, 2003), but current static models for these data neglect dynamic and spatial aspects of complex formation. Scholtens and Gentleman (2004) describe a statistical approach to local modeling of all complexes and their constituent proteins, thereby meeting one more requirement for a more complete functional characterization of the cell.

In this paper, we apply the methodology of Scholtens and Gentleman (2004) to estimate protein complex membership for two publicly available datasets and show how moving from static global interactome graphs to local models of protein–protein interactions provides a more realistic view of functional modules in the cell. We demonstrate the appropriate joint use of AP–MS and Y2H data, pointing out the need for increased data collection using both technologies. Recognizing the importance of integrated analyses of genomic and proteomic data, we examine the similarity of gene expression profiles for protein complex co-members. We also raise relevant AP–MS experimental design issues which await further research before they are resolved. The local modeling algorithm is summarized so that readers understand the nature of the results and the ensuing discussion.

To motivate the local modeling algorithm and formalize the conceptual differences between Y2H and AP–MS read-outs, we employ a hypothetical example of six proteins, P1, . . . , P6, composing two distinct complexes, C1 and C2. Suppose P4 is a part of complex C1 with P1, P2 and P6, and is also part of a different complex C2 with P1 and P3 with physical topologies shown in Figure 1a. In an idealized situation, the static graph obtained from 100% complete, sensitive and specific (100CSS) Y2H data would point to all binary physical interactions. However, the resulting model would fail to reflect the dynamic and/or spatial aspects of the system, i.e. P4’s presence in two distinct complexes (Fig. 1b, upper panel). Current ‘spoke’ and ‘matrix’ models (Bader and Hogue, 2003) of equally idealized 100CSS AP–MS data would generate networks of complex co-memberships. In the ‘spoke’ model, edges connect the bait protein in a purification to all corresponding hit proteins (Fig. 1b,
Fig. 1. Hypothetical protein complex scheme. (a) True physical topologies of two distinct complexes of which P4 is a member. (b) Static graphs of 100CSS Y2H and AP–MS data for the two complexes using the spoke model for the AP–MS data. (c) Local modeling of the AP–MS data in which maximal complete subgraphs differentiate between the two complexes, and Y2H data is used to ascertain physical interactions within the complexes.

lower panel), and in the ‘matrix’ model, additional edges connect all pairs of hits. In the hypothetical example in Figure 1b, the matrix model would place edges between all six proteins since they would all be found in the purification using P4 as the bait. Even with 100CSS AP–MS data, the matrix model generates false positive (FP) complex co-memberships, e.g. P1 and P3 would be connected by an edge even though they are not complex co-members. For this reason, we work with the spoke model which, in the 100CSS setting, accurately records the complex co-memberships assayed by AP–MS. Despite its correct representation of observed AP–MS data, the static spoke model does not recognize that a protein might be a part of different complexes in vivo, either in the same cell or in different cells of a population.

Even for 100CSS data, both Y2H and spoke or matrix modeling of AP–MS would fail to generate the necessary information to extract a correct functional wiring diagram. This is a limitation for systems biology research since valid assessment of both the functional activity and interactivity of protein complexes depends on their accurate description as physically distinct modules. Local modeling addresses this limitation and points to the next steps for ultimately understanding the operation and regulation of cellular systems.

SYSTEM AND METHODS

Modeling idealized data

Local modeling exploits the direct relationship between ‘maximal complete subgraphs’ in spoke AP–MS graphs and ‘complex membership’. In a graph of nodes and edges, complete subgraphs are sets of nodes for which all pairwise edges are present (e.g. P1, P2 and P3), and maximal complete subgraphs...
are complete subgraphs not contained in any other complete subgraph (e.g. $P_1$, $P_2$, $P_4$ and $P_6$; Fig. 1c, lower panel). Although the commonly used static spoke AP-MS representation only reflects complex co-membership, maximal complete subgraphs in the spoke network capture complex membership, represented in a bipartite graph (Fig. 1c, lower panel). In this example, maximal complete subgraphs lead to strict assignment of complex membership with represented in a bipartite graph (Fig. 1c, lower panel). In this example, maximal complete subgraphs correspond to the correct local model. In this case, maximal BH-complete subgraphs correspond to the correct local model. Further purifications could result in a more accurate complex membership estimate.

In the hypothetical AP-MS spoke graph in Figure 2a, $P_1$, $P_2$ and $P_3$ are baits and $P_4$, $P_5$ and $P_6$ are hit-only proteins. Noting the two types of proteins helps distinguish between three types of edges in the AP-MS spoke graph: (1) tested and present, (2) tested and absent and (3) untested. All possible directed edges originating at baits $P_1$, $P_2$ and $P_3$ are tested. If complex co-membership is detected the edges are present, e.g. the edges between $P_1$ and $P_2$ and from $P_1$ to $P_4$, and if co-membership is not detected the edges are absent, e.g. the absent edges between $P_1$ and $P_2$ and from $P_1$ to $P_6$. On the contrary, all possible edges among hit-only proteins are untested. The edges between $P_4$ and $P_5$ and $P_6$ are missing, not because they are known not to exist, but because complex co-membership for these proteins is never directly assayed. It would be inappropriate to assume the existence of any edges originating at hit-only proteins since they are never tested. Local modeling acknowledges these missing data and only models the observed data.

A mapping from maximal complete subgraphs in the spoke models to the bipartite graph is not reasonable for incomplete AP-MS data owing to these missing data and only models the observed data. A mapping from maximal complete subgraphs in the spoke models to the bipartite graph is not reasonable for incomplete AP-MS data owing to these missing data and only models the observed data. A mapping from maximal complete subgraphs in the spoke models to the bipartite graph is not reasonable for incomplete AP-MS data owing to these missing data and only models the observed data.

Modeling actual interactome data

Actual AP-MS experiments are typically not genome-wide and therefore, not 100% complete, i.e. not all complex co-memberships are tested. At the start of AP-MS experiments a collection of proteins is prespecified to be baits, one for each purification. Each bait finds a set of hit proteins that are co-members with itself in at least one complex. The hits may be baits for other purifications, or they may be hit-only proteins that are detected as hits but never used as baits; therefore, all proteins in AP-MS datasets can be designated as either baits or hit-only proteins. The resultant set of co-memberships is not unbiased and is limited to those that were directly tested (i.e. determined by the bait proteins). A subtle but important point: under 100CSS conditions a bait will not find proteins as hits that are not complex co-members with itself. These non-co-memberships are meaningful features in AP-MS data which help distinguish between distinct complexes.

In the hypothetical AP-MS spoke graph in Figure 2a, $P_1$, $P_2$ and $P_3$ are baits and $P_4$, $P_5$ and $P_6$ are hit-only proteins. Noting the two types of proteins helps distinguish between three types of edges in the AP-MS spoke graph: (1) tested and present, (2) tested and absent and (3) untested. All possible directed edges originating at baits $P_1$, $P_2$ and $P_3$ are tested. If complex co-membership is detected the edges are present, e.g. the edges between $P_1$ and $P_2$ and from $P_1$ to $P_4$, and if co-membership is not detected the edges are absent, e.g. the absent edges between $P_1$ and $P_2$ and from $P_1$ to $P_6$. On the contrary, all possible edges among hit-only proteins are untested. The edges between $P_4$ and $P_5$ and $P_6$ are missing, not because they are known not to exist, but because complex co-membership for these proteins is never directly assayed. It would be inappropriate to assume the existence of any edges originating at hit-only proteins since they are never tested. Local modeling acknowledges these missing data and only models the observed data. A mapping from maximal complete subgraphs in the spoke models to the bipartite graph is not reasonable for incomplete AP-MS data owing to these missing data and only models the observed data. A mapping from maximal complete subgraphs in the spoke models to the bipartite graph is not reasonable for incomplete AP-MS data owing to these missing data and only models the observed data. A mapping from maximal complete subgraphs in the spoke models to the bipartite graph is not reasonable for incomplete AP-MS data owing to these missing data and only models the observed data.
P1 and P2 to P5. The potential edge connecting hit-only proteins P1 and P2 is untested, and as the definition states, untested edges are permitted in BH-complete subgraphs. Similarly, in the P1, P3, P4 subgraph, the tested edges from P1 to P3 and P1 to P4 are present, and the potential edge connecting P3 and P4 is untested. All other subgraphs in Figure 2a are not maximal BH-complete, either because they are not maximal or because they are not BH-complete. For example, the subgraph P2, P4, P5 is not maximal; it is BH-complete since all tested edges are present, but it is contained in the subgraph P1, P2, P3, P4, P5 which is also BH-complete. As another example, the subgraph P3, P5, P6, P8 is not BH-complete because the tested edge from P3 to P6 is absent. Similarly, the subgraph P1, P2, P3, P5, P8 are not BH-complete since all tested edges originating at bait proteins are not all present.

In Figure 2a, maximal BH-complete subgraphs in the AP–MS spoke model correspond to a bipartite graph reflecting exact complex membership. Importantly, different baits may result in different estimates. For example, if P2, P3 and P4 were used as baits, our algorithm would detect the two correct complex estimates, as well as a third irrelevant complex (Fig. 2b). Without further AP–MS experiments, it is impossible to determine actual complex membership among the hit-only proteins. It would be difficult to compare two experiments that use different bait proteins as the set of tested co-memberships could be quite different. This raises interesting questions of experimental design and suggests that it is essential for the investigator to critically evaluate the set of tested co-memberships so that the resultant complex estimates pertain to the cellular networks of interest and are as complete as possible.

In addition to being incomplete, AP–MS data are neither 100% sensitive nor 100% specific, resulting in false negative (FN) observations (a complex not present in the AP–MS technology), and false positive (FP) observations (a complex present in the AP–MS technology). To accommodate this, local modeling first locates maximal BH-complete subgraphs in the stalk-modeled data and then locates maximal BH-complete subgraphs in the estimated AP–MS graph after likelihood maximization.

**ALGORITHM**

The local modeling algorithm of Scholtens and Gentleman (2004) assumes that the errors in the (imperfect) observation of these edges are independent, and uses a logistic regression model for the probability

\[
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \mu + \alpha Y_{ij},
\]

where \( Y_{ij} = 1 \) if the edge truly exists and 0 otherwise, \( e^{\mu + \alpha Y_{ij}} / (1 + e^{\mu + \alpha Y_{ij}}) \) is the sensitivity of the AP–MS technology and \( 1 / (1 + e^{\mu}) \) is the specificity. The algorithm first considers the likelihood \( L \), the product of \( e^{(\mu + \alpha Y_{ij})Z_{ij}} / (1 + e^{\mu + \alpha Y_{ij}}) \) for all tested edges, where \( Z_{ij} = 1 \) if protein \( j \) is found as a hit using protein \( i \) as a bait and 0 otherwise. The algorithm begins by maximizing \( L \) to estimate each \( Y_{ij} \) and then locates maximal BH-complete subgraphs in the estimated AP–MS graph and records these as the initial complex estimates.

If parameter specification is done according to the suggestions in Scholtens and Gentleman (2004) where sensitivity is assumed to be less than specificity, then \( L \) will be maximized when doubly tested but singly observed edges (i.e. unreciprocated edges connecting bait proteins) are estimated to exist. If so, the initial maximal BH-complete subgraphs may include doubly tested but unreciprocated edges, e.g. P3 and P5 in the graph of observed AP–MS data in Figure 3. All singly tested edges must still be observed; for example, P3, P4, P5, P6 is not maximal BH-complete in Figure 3 since the subgraph P1, P2, P3, P4, P5 contains all tested edges originating at bait protein P1 and is BH-complete. For example, the subgraph P1, P3, P4, P5, P6 is not BH-complete since all tested edges originating at bait protein P1 are absent. Similarly, the subgraph P1, P2, P3, P4, P5 and the entire graph P1, P2, P3, P4, P5, P6 are not BH-complete since all tested edges originating at bait proteins are not all present.

As another example, the subgraph P1, P3, P4, P5, P6 and the entire graph P1, P2, P3, P4, P5, P6 are not BH-complete since all tested edges originating at bait proteins are not all present.

In Figure 2a, maximal BH-complete subgraphs in the AP–MS spoke model correspond to a bipartite graph reflecting exact complex membership. Importantly, different baits may result in different estimates. For example, if P2, P3 and P4 were used as baits, our algorithm would detect the two correct complex estimates, as well as a third irrelevant complex (Fig. 2b). Without further AP–MS experiments, it is impossible to determine actual complex membership among the hit-only proteins. It would be difficult to compare two experiments that use different bait proteins as the set of tested co-memberships could be quite different. This raises interesting questions of experimental design and suggests that it is essential for the investigator to critically evaluate the set of tested co-memberships so that the resultant complex estimates pertain to the cellular networks of interest and are as complete as possible.

In addition to being incomplete, AP–MS data are neither 100% sensitive nor 100% specific, resulting in false negative (FN) observations (a complex not present in the AP–MS technology), and false positive (FP) observations (a complex present in the AP–MS technology). To accommodate this, local modeling first locates maximal BH-complete subgraphs in the stalk-modeled data and then locates maximal BH-complete subgraphs in the estimated AP–MS graph after likelihood maximization.

**ALGORITHM**

The local modeling algorithm of Scholtens and Gentleman (2004) assumes that the errors in the (imperfect) observation of these edges are independent, and uses a logistic regression model for the probability

\[
\log \left( \frac{p_{ij}}{1 - p_{ij}} \right) = \mu + \alpha Y_{ij},
\]

where \( Y_{ij} = 1 \) if the edge truly exists and 0 otherwise, \( e^{\mu + \alpha Y_{ij}} / (1 + e^{\mu + \alpha Y_{ij}}) \) is the sensitivity of the AP–MS technology and \( 1 / (1 + e^{\mu}) \) is the specificity. The algorithm first considers the likelihood \( L \), the product of \( e^{(\mu + \alpha Y_{ij})Z_{ij}} / (1 + e^{\mu + \alpha Y_{ij}}) \) for all tested edges, where \( Z_{ij} = 1 \) if protein \( j \) is found as a hit using protein \( i \) as a bait and 0 otherwise. The algorithm begins by maximizing \( L \) to estimate each \( Y_{ij} \) and then locates maximal BH-complete subgraphs in the estimated AP–MS graph and records these as the initial complex estimates.

If parameter specification is done according to the suggestions in Scholtens and Gentleman (2004) where sensitivity is assumed to be less than specificity, then \( L \) will be maximized when doubly tested but singly observed edges (i.e. unreciprocated edges connecting bait proteins) are estimated to exist. If so, the initial maximal BH-complete subgraphs may include doubly tested but unreciprocated edges, e.g. P3 and P5 in the graph of observed AP–MS data in Figure 3. All singly tested edges must still be observed; for example, P3, P4, P5, P6 is not maximal BH-complete in Figure 3 since the singly tested edges from P3 to P4, P3 to P5, and P3 to P7 are not present.

Although AP–MS experiments do result in high-dimensional data, there are, in fact, at most two observations of each edge when the purifications are performed only once. FN observations will break a true complex subgraph into multiple maximal BH-complete subgraphs. The local modeling algorithm incorporates a second criterion, called \( C \), to assess combinations of the maximal BH-complete subgraphs. For a proposed complex \( c_k \), the binomial probability for the number of observed edges \( x_c \) out of the number of tested edges \( n_c \) is

\[
\Gamma(c_k) = \binom{n_c}{x_c} \left( \frac{e^{\mu + \alpha Y_{ij}}}{1 + e^{\mu + \alpha Y_{ij}}} \right)^x_c \left( \frac{1}{1 + e^{\mu + \alpha Y_{ij}}} \right)^{n_c - x_c},
\]

is used to assess the consistency of the observed data in the subgraph for \( c_k \) with the sensitivity of the AP–MS technology. If the majority of missing edges in \( c_k \) are directed at only a few proteins, then these proteins are probably not a part of \( c_k \). The randomness of the allocation of the missing edges in the subgraph for \( c_k \) is quantified by letting \( \Phi(c_k) \) equal the two-sided \( P \)-value from Fisher’s exact test for the distribution of observed incoming edges for each protein.
C is then taken to be the product of $\Gamma(c_k) \times \Phi(c_k)$ for all proposed complexes $c_k$.

After the initial estimate, pairwise unions of complex estimates are investigated. The union which leads to the highest increase in $P = L \times C$ is accepted. The pairwise union process is then repeated using the new complex estimates. The algorithm stops when no union proposals increase $P = L \times C$ (Fig. 3). Since $C$ is the product of probabilities ranging from 0 to 1 over all complex estimates, it tends to increase for a smaller number of complexes that reasonably reflect the underlying BH-complete subgraph structure. An increase in $C$ may lead to a decrease in $L$; the product $P = L \times C$ balances the contribution of each term.

If desired, externally derived similarity measures between proteins can be incorporated into an extended logistic regression model:

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

where $s_{ij}$ is a similarity measure between proteins $i$ and $j$, and $\beta$ is a user-specified parameter which controls the extent to which the similarity measure supports the existence of an edge. Complete specification of the model, algebraic details of the complex combination acceptance criteria, the initial maximal BH-complete subgraph finding algorithm and a simulation studying the robustness of the algorithm to parameter specification are found in Scholtens and Gentleman (2004). The algorithm can be run using the R (Ihaka and Gentleman, 1996) package apComplex, available at http://www.bioconductor.org.

Local modeling may result in three types of complex estimates: (1) multibait-multiedge (MBME) complexes containing multiple baits and multiple edges; (2) single bait–multihit (SBMH) complexes containing one bait and multiple hits and (3) unreciprocated bait–bait (UnRBB) complexes containing two proteins, both used as baits, connected by one unreciprocated edge (Fig. 3). MBME complexes contain data from multiple purifications, allowing a more detailed view of complex co-membership among the hits, and are, therefore, the most accurate local models afforded by the data. Any hits found by a common bait are reported in a SBMH complex if they do not appear together in a MBME complex. SBMH complexes may contain proteins from multiple complexes since edges between hit-only proteins are untested, but they are crucial outputs that prevent any loss of information after refined local modeling represented in the MBME complexes. The edges in UnRBB complexes may be FPs since they are tested twice, observed once and not contextually confirmed by other edges, or the unreciprocated edge may be an FN observation of an edge that does exist but was not observed experimentally. MBME complexes are believed to be the most reliable outputs and are the basis for our ensuing discussion of the implications of local modeling for systems biology. SBMH and UnRBB complexes, although less reliable, can be used to design future experiments. It is important to note that a large portion of protein complexes may be missed since the complex estimates are strictly limited to the cell types and experimental conditions used for the purifications.

**IMPLEMENTATION**

**Detecting previously characterized protein complexes in the global AP–MS network**

We applied our algorithm to 589 raw tandem affinity purifications (TAP) from Saccharomyces cerevisiae (Gavin et al., 2002). Excluding homodimers, there are 455 bait proteins and 909 hit-only proteins in these data. The raw purifications were manually organized and originally released as 232 annotated ‘yTAP’ complexes (Krause et al., 2003). Our algorithm predicted 708 complexes containing 260 MBME, 325 SBMH and 123 UnRBB complexes available for use and examination at http://www.bioconductor.org/Docs/Papers/2003/apComplex and in the apComplex package.

To reflect the previously estimated sensitivity of TAP, we used a sensitivity of 0.75. Previous studies suggest that 50% of reported protein–protein interactions may be FPs (von Mering et al., 2002). In the TAP data, 455(454 + 909) = 620 165 edges were tested and 3420 were observed. If 1720 are FPs, then an estimate of the specificity of TAP is $1 - 1720/(620 165 - 1720 \times 1.25) \approx 0.997$. For our analysis, we conservatively used a specificity parameter of 0.995. For analyses of new data, investigators are encouraged to perform repeated purifications using well-characterized proteins as baits to estimate the sensitivity of their AP–MS technology for detecting known co-memberships. Specificity can then be estimated using a procedure similar to ours. If parameter specification in this manner is not possible, investigators may want to start with the values used here for the TAP data and observe changes in the complex estimates for differing values of specificity and sensitivity. The robustness of our algorithm to parameter specification is discussed in Scholtens and Gentleman (2004).

yTAP complexes are unions of all hits for hand-picked sets of baits linked together by spoke edges, neglecting the fact that hits in an individual purification may belong to different complexes. Since our local modeling algorithm accounts for this biological reality, MBME complexes reflect better in vivo complexes. For example, all hits from six purifications using Arc15, Arc18, Arc35, Arc40, Arp2 and Arp3 as baits were grouped together in yTAPC153 (Fig. 4b).

We model the seven known subunits of the Arp2/3 complex (Winter et al., 1997) and assign the five loosely-connected proteins to other complexes with their corresponding baits (Fig. 4a and c).

Another striking example involves the RNA polymerase complexes (PolI, PolII and PolIII). We detect three MBME complexes corresponding to the distinct Pol complexes known to exist in vivo (Archambault and Friesen, 1993). Importantly, their overlapping members (e.g. Rpo40 is a known member of both Pol II and PolIII) are correctly modeled. In contrast, the yTAP analysis collapses Pol II and PolIII together in yTAPC154 and includes several extraneous proteins in the estimate of Pol II in yTAPC145 (Fig. 5). A correct description of cellular systems demands recognition of these functionally dependent but physically distinct RNA polymerase complexes.

In another example, the heterotrimeric composition of protein phosphatase 2A (PP2A) was reported for the PP2A proteins Tpd3, Pph21, Pph22, Cdc55 and Rts1 (Jiang and Broach, 1999). All four, along with cell-cycle regulators Zds1 and Zds2, are combined into the 46-subunit yTAPC151 (Fig. 6b). Our local model distinguishes between the four known PP2A trimers and furthermore, demonstrates the exclusive association of Zds1 and Zds2 with the Cdc55/Pph23/Tpd3 trimer (Fig. 6b–g). Such precise local modeling is not presented in the manually compiled yTAP complexes nor is it immediately evident in the static spoke representation of the AP–MS data for these proteins (Fig. 6a). Local modeling recovers crucial information regarding the dynamic composition of PP2A as well as interactivity of Zds1 and Zds2 with a specific PP2A trimer.

For a large-scale assessment of the detection of known protein complexes in publicly available datasets, we compared our models
Local modeling of global interactome networks

Fig. 4. (a) Local model for Arp2/3 complex. (b) yTAPC153. Arp2/3 proteins with five extra proteins collapsed into one complex. (c) Bipartite graph representing local complex membership for all proteins reported in yTAPC153.

to 267 hand-curated complexes available at MIPS (ftp://ftp.mips.gsf.de/yeast/catalogues/complexes/complex.130603) using a similarity measure $\Omega$. For a true complex, $T_C$, and a predicted complex, $P_C$, define $\Omega(T_C, P_C) = \min(i/a, i/b)$, where $i$ is the number of proteins in the intersection of $T_C$ and $P_C$, and $a$ and $b$ are the number of proteins in $T_C$ and $P_C$, respectively. This measure finds the proportion of overlapping proteins from the perspective of both complexes and takes the minimum as a conservative similarity measure. We similarly compared the yTAP complexes to the MIPS complexes. We used the proteins common between the MIPS and TAP data, leaving 129 MIPS multiprotein complexes for comparison. Using $\Omega > 0.70$ to guarantee close correspondence between complex estimates, we mapped 80 of our complexes to 62 MIPS complexes (Table 1). MIPS complexes with multiple mappings from our complexes contained a high percentage of common core elements. For example, six of our complex estimates mapped to the casein kinase II complex containing Cka1, Cka2, Ckb1 and Ckb2 reported by MIPS. Our TAP analysis showed these four proteins interacting with other proteins in distinct complexes, and the similarity criterion $\Omega > 0.70$ mapped all six to their common core. Only 38 yTAP complexes mapped to 36 MIPS complexes.

The large-scale comparison and the specific examples just described confirm that our local models better resemble well-characterized complexes. The local modeling algorithm provides an automated and repeatable analysis of raw AP–MS data that accommodates multicomplex membership by individual proteins and membership in different complexes for hits in a single purification. Accounting for these biological realities is not to be underestimated. In our TAP analysis, 341 of the 669 proteins composing all MBME estimates are members of two or more complexes and 296 of the 455 purifications detect hits that are members of different MBME complexes. For large scale systems investigations, local modeling is readily executable and describes functional modules at a finer level of detail than is otherwise available.

Combining AP–MS and Y2H data
Another large-scale assessment of our complexes and the yTAP estimates involved publicly available Y2H physical interactions, specifically the Ito core (Ito et al., 2001) and Uetz data (Uetz et al., 2000). For a complex to exist, all subunits must physically bind to at least one subunit. We calculated the ‘overlap proportion’ of proteins in our complexes that are reported in the combined Ito/Uetz data
Fig. 5. (a–c) Local models for RNA polymerases (PolI, PolII and PolIII, respectively). Overlapping members are correctly assigned, e.g. Rpc40 is a member of both PolI and PolIII. (d) yTAPC154, PolI and PolIII are collapsed into a single complex. (e) yTAPC145, PolII with several additional loosely connected proteins collapsed into a single complex.

to physically interact with at least one other complex member. For Arp2/3, the combined Ito/Uetz data reported one physical interaction between Arc15 and Arc19, resulting in an overlap proportion of $\frac{2}{7} = 0.286$ for our complex estimate. (Note that this single physical interaction suggests that the Ito/Uetz data are not comprehensive since one physical interaction is insufficient to compose seven Arp2/3 proteins into one complex.) For the 12-subunit yTAPC153 (Fig. 4b), the proportion is $\frac{2}{12} = 0.167$. Excluding complexes with no representation in the Ito/Uetz data, for our MBME complexes, the mean overlap proportion was 0.412 (25th and 75th percentiles 0.25 and 0.50, respectively). For the yTAP complexes, the mean overlap proportion was 0.318 (25th and 75th percentiles 0.14 and 0.39, respectively).
Local modeling of global interactome networks

Fig. 6. (a) Static spoke model for PP2A proteins, Zds1 and Zds2. (b–f) Local models for four PP2A heterotrimers. Zds1 and Zds2 associate exclusively with the Cdc55/Pph22/Tpd3 form of PP2A. (g) Bipartite graph for complex membership of all five PP2A proteins, as well as Zds1 and Zds2. (h) yTAPC151, all five PP2A proteins and Zds1 and Zds2 are collapsed into a single 46-subunit complex.

respectively. Even though the publicly available Y2H data are not 100% accurate, our complex estimates are more structurally substantiated than the yTAP estimates.

If available Y2H data were more comprehensive, local modeling could be carried out to further completion. For example, we identify Clp1, Pcf11, Rna14 and Rna15 as an MBME complex; these four proteins are known to compose cleavage factor IA (CFIA). The Ito/Uetz data report physical interactions between Pcf11 and Clp1, Rna14 and Rna15 (Fig. 7). The identification of these four proteins as composing a distinct complex, followed by structural prediction using the Ito/Uetz Y2H physical interactions matches the behavior and mechanical assembly reported by Gross and Moore (2001). Unfortunately, sparse Y2H data prevent local modeling to its fullest extent for many complexes, e.g. the aforementioned lack of physical interactions reported for Arp2/3. Continued and improved Y2H data collection will mitigate the effects of apparent FN and
short, local modeling provides a much needed platform for identify-
both individual proteins and potentially to a new protein complex. In
complex suggests many avenues for further research, pertaining to
umented as complex co-members (Supplementary Figure 1b). This
and the remaining three, Bud20, Sda1 and YLR106C are not doc-
Kre32, YNL182C and YHR197W have incomplete GO annotation,
 generation.
in which they work at a minimum allows informed hypothesis
incorrect to confer the specific functionality of Pph9 and Spt5 on

CFIA using Ito/Uetz Y2H data.

Fig. 7. (a) Local model for CFIA using TAP data. (b) Physical structure of

previously reported FP rates (von Mering et al., 2002) on local

Provisional functional annotation prediction

Many MBME predictions are not documented in the literature, and
our analyses suggest provisional functional annotation for many
uncharacterized proteins. We predict a four-subunit complex con-
taining Ppb9, SpT5, YNL201C and YBL046W (Supplementary
Figure 1a). YNL201C and YBL046W have incomplete Gene Onto-
logy (GO) annotation (Dwight et al., 2002). Although it may be
incorrect to confer the specific functionality of Ppb9 and SpT5 on
YNL201C or YBL046W, an understanding of a protein complex
in which they work at a minimum allows informed hypothesis
generation.

Similarly, for a seven-subunit complex prediction, YCR072C,
Kre32, YNL182C and YHR197W have incomplete GO annotation,
and the remaining three, Bud20, Sda1 and YLR106C are not doc-
dumented as complex co-members (Supplementary Figure 1b). This
complex suggests many avenues for further research, pertaining to
both individual proteins and potentially to a new protein complex.
In short, local modeling provides a much needed platform for identif-
ing specific follow-up experiments regarding functional annotation
after global network data collection.

Pathway interactivity

Communication between cellular systems may also be investigated
through local models of proteins associated with different pathways.
An example involves the five translation initiation factor 3 (eIF3) core
complex proteins: Prt1, Rpg1, Nip1, Tif34 and Tif35. We predict the
co-activity of eIF3 proteins with Tif5 (Phan et al., 1998) and Sua7
(Supplementary Figure 2a). Prt1, Rpg1 and Nip1 are also known to
form a stable subcomplex of eIF3. We predict that these three pro-
teins interact separately with Cbk2, Hat1 and Cdc95 (Supplementary
Figure 2b–d). These observations may be used to confirm the sugges-
tion that Prt1, Rpg1 and Nip1 perform several different translation
initiation functions (Phan et al., 2001).

Bait selection

We analyzed another AP–MS dataset, known as high-throughput
mass spectrometry–protein complex identification (HMS–PCI) data
(Ho et al., 2002), containing 493 baits and 1085 hit-only proteins.
Only 81 of the 493 HMS–PCI baits also found hits in the TAP exp-
eriments, and a comparison of the estimated complexes from each
dataset demonstrates the great influence of the set of baits on the
interpretable data. As suggested previously, when designing AP–MS
experiments, investigators must take care to select baits that probe the
cellular networks of interest as completely as possible. We applied
our algorithm to the HMS–PCI data with a sensitivity of 0.75, but a
higher FP probability of 0.01. The HMS–PCI lab process includes
one round of mild purification, and it has been suggested that their MS
method may identify background proteins (Bader and Hogue, 2002).
We doubled the FP probability to conservatively accommodate the
potential increase in FP observations for these data.

Our HMS–PCI analysis resulted in 1008 complex predictions,
including 329 UnRBBs and 437 SBMHs. Several of the 242 MBMEs
were similar to those found in the TAP data, despite the different baits
in the two experiments. For example, our HMS–PCI estimate of the
Arp2/3 complex contained six of the seven known elements (Supple-
mentary Figure 3). In the HMS–PCI data, only two Arp2/3 proteins
were used as baits, rather than six as in the TAP data. Arc15, the
element we excluded, was detected in the Arp2 purification along
with 21 other hits, but was missed in the Arc40 purification. There
were insufficient data to differentiate Arc15 from the 21 other Arp2
hits for inclusion in the Arp2/3 estimate.

The different HMS–PCI baits did enable complex identifica-
tions that were unobserved in the TAP data. One example is an
eight-subunit prediction, six of which (Fyv10, Gd7, Gd8, Rmd5,
Vid28 and Vid30) carry out catabolic degradation of fructose-1,6-
biphosphatase (FBPase) in Saccharomyces (Regelmann et al.,
2003) (Supplementary Figure 4). The relationships of the other two pro-
teins, Cct2 and YBL049W to the FBPase proteins may represent
co-involvement in other complexes.

Local modeling and the cell cycle

Joint analyses of AP–MS and other high-throughput data will provide
a solid platform for new systems biology investigations (Vidal, 2001).
For example, yeast cells synchronized according to cell-cycle peri-
odicity have been used to study genome-wide mRNA transcript
fluctuation. We investigated the concordance of our complex pre-
dictions with genes exhibiting similar cell-cycle-controlled expres-
sion profiles, using data in which 3000 genes were grouped into
30 k-means clusters (Tavazoie et al., 1999). For the 578 genes in com-
mon between the cell-cycle and TAP data, there were 578/577/2 =
166,753 pairs of genes, 6619 of which shared membership in one of
the 30 clusters. For each of our MBMEs and the yTAP complexes, we
assessed the statistical significance of the total number of intraclust-
er pairs using the hypergeometric distribution. We used the overlapping
proteins and any remaining complexes with more than two subunits,
resulting in 101 MBME and 107 yTAP complexes for evaluation.
We found that 42/101 = 0.42 and 31/101 = 0.31 of our MBMEs had $P$-values < -0.05 and 0.01, respectively. For the yTAP data, we found 32/107 = 0.30 and 21/107 = 0.20 had $P$-values < -0.05 and 0.01, respectively. Our MBMEs contain a higher proportion of proteins with similar gene-expression profiles than the yTAP complexes, and, therefore, better reflect previous reports of co-regulated interacting proteins (Ge et al., 2001; Grigoriev, 2001; Kemmeren et al., 2002). Three proteins Smc1, Smc3 and Mcd1 that participate in mitotic chromosome assembly during cell division (Hirano, 1999) were locally modeled in one MBME and reported to share cell-cycle cluster membership (Supplementary Figure 5). Several other predicted complexes demonstrated similarly significant numbers of intralcluster pairs. Local modeling done under specific experimental conditions could clarify the coordination of gene expression and protein complex formation for systems operation. Important areas of further research also include the study of complexes with groups of proteins involved in different expression clusters and a comparison of the expression profiles for proteins found in many complexes with those involved in only one.

DISCUSSION

A comprehensive catalog of multiprotein complex membership is a foundational element of systems biology. The local modeling algorithm proposed by Scholtens and Gentleman (2004) accommodates biological realities, such as multicomplex membership by individual proteins, and exploits subtle differences between AP–MS and Y2H data types. In this paper, we use local modeling to move from global static interactome graphs to a rigorous characterization of protein macromolecules at a finer level of detail than previously available using AP–MS data. Y2H-detected physical interactions complete the local interactome model by providing a view of the mechanics of assembly into functional units. Local modeling provides a platform for explicit hypothesis development regarding functional annotation and pathway interactivity. Joint analyses of protein complexes with other data, such as gene-expression profiles, promise salient insight into the systems of modular networks responsible for cellular activity.

Conflict of Interest: none declared.

REFERENCES