BIOINFORMATICS ORIGINAL PAPER

Vol. 24 no. 2 2008, pages 218–224
doi:10.1093/bioinformatics/btm565

Systems biology

Estimating node degree in bait-prey graphs

Denise Scholtens[1,*], Tony Chiang[2], Wolfgang Huber[2] and Robert Gentleman[3]

[1]Department of Preventive Medicine, Northwestern University Medical School, 680 N. Lake Shore Drive Suite 1102, Chicago, IL 60611-4402, USA, [2]EMBL Outstation – Hinxton, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK and [3]Fred Hutchinson Cancer Research Center, Computational Biology Group, 1100 Fairview Avenue North – M2-B876, P.O. Box 19024, Seattle, Washington 98109-1024, USA

Received on September 7, 2007; revised and accepted on November 7, 2007

Advance Access publication November 19, 2007

Associate Editor: Chris Stoeckert

ABSTRACT

Motivation: Proteins work together to drive biological processes in cellular machines. Summarizing global and local properties of the set of protein interactions, the *interactome*, is necessary for describing cellular systems. We consider a relatively simple per-protein feature of the interactome: the number of interaction partners for a protein, which in graph terminology is the *degree* of the protein.

Results: Using data subject to both stochastic and systematic sources of false positive and false negative observations, we develop an explicit probability model and resultant likelihood method to estimate node degree on portions of the interactome assayed by bait-prey technologies. This approach yields substantial improvement in degree estimation over the current practice that naïvely sums observed edges. Accurate modeling of observed data in relation to true but unknown parameters of interest gives a formal point of reference from which to draw conclusions about the system under study.

Availability: All analyses discussed in this text can be performed using the *ppiStats* and *ppiData* packages available through the Bioconductor project (http://www.bioconductor.org).

Contact: dscholtens@northwestern.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

Understanding the contribution of proteins to coordinated cellular systems requires knowledge of the various interactions they have with other proteins. Global and local statistics on the topology of *interactome* graphs aim to infer the nature and behavior of protein interactions, and can provide a basis for planning and interpretation of experiments. These measures capture simple but informative features of graphs, e.g. the number of interactors for each protein. There are caveats to the use of these measures, including the dynamic nature of the *in vivo* interactome as opposed to the static data yielded by currently available technologies. But these summary statistics capture characteristics of high-throughput observations in tractable form and accurate estimation is paramount to making correct conclusions about interactome behavior.

Affinity purification-mass spectrometry (AP-MS) technology provides data on complex co-membership interactions by using a series of bait proteins to detect the set of all prey proteins that share membership with the bait in at least one multi-protein complex. Yeast two-hybrid (Y2H) technology uses a fusion protein system to test for physical interactions between baits tagged with the DNA binding domain and prey with the activation domain. Hence AP-MS and Y2H technologies probe undirected, symmetric relationships in a directed manner from bait to prey. Inference on the features of AP-MS and Y2H graphs is complicated by this directedness, as well as sampling bias, incomplete coverage, and stochastic and systematic errors leading to both false positive (FP) and false negative (FN) observations. These issues have often been overlooked when analyzing protein interaction data. In this study, we apply a statistical modeling approach to ameliorate these difficulties for AP-MS, Y2H and other bait-prey technologies.

This report specifically demonstrates the use of statistical likelihood for estimating node degree to obtain a substantial improvement over the naïve approach in which degree is estimated by summing observed interactions. The implications are widespread since many other graph statistics, e.g. the clustering coefficient and node degree distribution, are functions of node degree and special biological interpretations are often assigned to nodes of particularly high degree. Furthermore, knowing the number of interactions for any protein is helpful for identifying its true interactors given a set of reported possibilities. Node degree estimation is one example of interactome-based statistical modeling. The paradigm we propose applies generally to other bait-prey graph statistics and is critical for accurately describing interactome behavior.

2 RESULTS

2.1 Multinomial model for node degree

In concept, the set of edges in a bait-prey graph can be divided into distinct sets of doubly, singly and untested edges. If all

*To whom correspondence should be addressed.

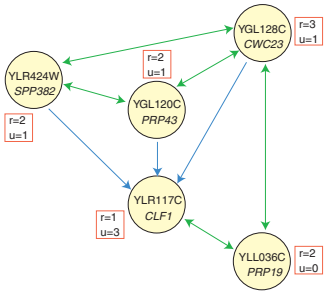

Fig. 1. A subgraph of VBP nodes from the Gavin *et al.*'s (2002) AP-MS data. Green reciprocated edges are tested twice and observed twice. Blue unreciprocated edges are tested twice and observed once. All edges not shown are tested twice and not observed twice.

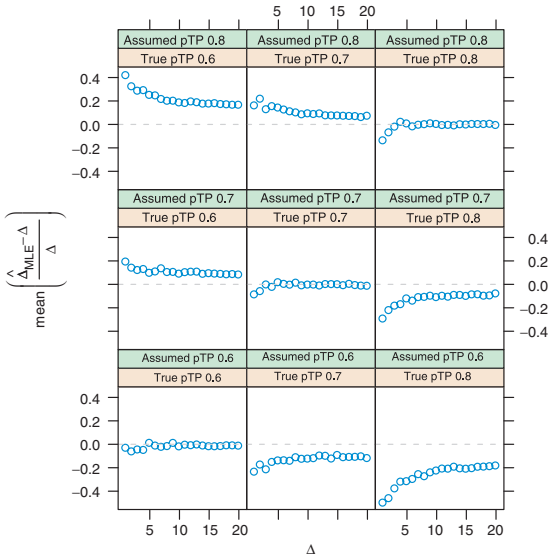

Fig. 2. Mean relative bias for $\hat{\Delta}_{MLE}$ for 500 observations from graphs with 1000 nodes calculated as the mean difference between $\hat{\Delta}_{MLE}$ and Δ, divided by Δ. True p_{TP} values are those used to generate observations from the simulated graphs and the assumed p_{TP} values are those used in estimation of degree. Incorrect, assumed p_{TP} values are used to study the accuracy of $\hat{\Delta}_{MLE}$ under p_{TP} misspecification. In all simulations here, $p_{FP} = 0.001$ and is correctly specified for estimation purposes. Each panel represents a single simulation. Assumed p_{TP} is constant within rows, and true p_{TP} is constant down columns.

experiments for each bait work properly and all proteins in the cell are available for detection as prey, then all edges between pairs of baits are tested twice, all edges extending from baits to non-bait proteins are tested once, and all edges between pairs of non-baits are untested. Under realistic experimental conditions, some proteins fail as baits and some proteins are not detectable as prey, and so the distinctions between these collections of edges blur. Chiang *et al.* (2007) discuss a viable-bait-prey (VBP) subgraph of a full set of bait-prey experimental data induced by the subset of bait proteins that detect at least one prey and are detected as prey by at least one other bait. By focusing on the set of proteins with direct evidence of viability as both bait and prey and by eliminating proteins prone to systematic bias (see Section 3.2), the VBP graph only includes edges for which two bait-prey assays running in opposite directions between a pair of proteins can reasonably be viewed as replicate observations on the same underlying true edge. Different experimental conditions, and other factors, dictate that VBP nodes may differ even for experiments using the same original bait set (see Section 2.4).

Despite replicate testing of all edges in the VBP graph, the observations in each direction are not necessarily consistent due to measurement error. For each VBP node there is an observed number of reciprocated edges, r, an observed number of unreciprocated edges, u (Fig. 1), and a true but unknown degree Δ. The joint probability of observing specific values of r and u for any given Δ can be written as a function of Δ, the TP probability (p_{TP}) and the FP probability (p_{FP}) using Multinomial models for both TP and FP observations; full model development is reported in Section 3.3. After adjusting the probability statement for only observing non-zero counts of interactions, statistical maximum-likelihood estimation (MLE) can be used to arrive at a degree estimator, $\hat{\Delta}_{MLE}$, that accounts for restrictions on observed data in the VBP graph as well as p_{TP} and p_{FP}. Current practice is to estimate degree naïve to p_{TP}, p_{FP} and subtleties in data collection, specifically $\hat{\Delta}_{naïve} = r + u$.

2.2 Estimating degree when p_{TP} and p_{FP} are known

2.2.1 Local performance under misspecification of p_{TP} Calculation of $\hat{\Delta}_{MLE}$ depends on values of p_{TP} and p_{FP} and since true values of p_{TP} and p_{FP} are not generally known for any particular technology, they must also be estimated. Given the potential for misspecification of p_{TP} and p_{FP}, we studied the accuracy of $\hat{\Delta}_{MLE}$ under deviations from the truth for these parameters. Since p_{TP} applies to a small number of true edges relative to the total number possible, its effects are most evident at the per-node level.

For nodes with Δ ranging from 1 to 20, 500 observations were generated from graphs with 1000 nodes, doubly tested edges, $p_{TP} = 0.6$, 0.7 and 0.8, and $p_{FP} = 0.001$. This p_{FP} parameter results in a mean of 2 FP observations per node with FP observations occurring in either direction. To simulate the VBP paradigm, observations of zero incident edges were excluded from analysis. Given the simulated observations, $\hat{\Delta}_{MLE}$ was estimated under the true p_{TP} and p_{FP} parameters as well as incorrect assumed values for p_{TP}. Figure 2 demonstrates the mean relative error for $\hat{\Delta}_{MLE}$ using both correct and misspecified values of p_{TP}. For these simulation parameters, relative bias in the estimator corresponds roughly with the amount of under- or overestimation of p_{TP}; when p_{TP} is under- or overestimated by 0.10 (0.20), degree is estimated on average within ten (twenty) percent of the true degree. Greater relative bias is apparent for nodes of lower degree.

2.2.2 Global performance under correct specification of p_{TP} *and* p_{FP} *for varying graph topologies* A series of Erdös–Renyi (ER) random graphs containing 1000 nodes and 2000 edges were examined to explore global performance of $\hat{\Delta}_{MLE}$ versus $\hat{\Delta}_{naïve}$. For 100 of these ER graphs, edges were doubly tested and observed with $p_{TP} = 0.6$ and 0.7 and $p_{FP} = 0.0008$ and 0.001. Table 1 records mean, minimum and maximum

Table 1. Mean(minimum,maximum) values of RMSE for $\hat{\Delta}_{naïve}$ and $\hat{\Delta}_{MLE}$ for 100 ER graphs with 1000 nodes and 2000 edges

		$p_{TP}=0.60$	$p_{TP}=0.70$
	p_{FP}		
$\hat{\Delta}_{naïve}$	0.0008	1.77 (1.61, 1.91)	1.86 (1.66, 2.00)
	0.001	2.11 (1.91, 2.30)	2.24 (2.04, 2.42)
$\hat{\Delta}_{MLE}$	0.0008	1.63 (1.50, 1.75)	1.44 (1.34, 1.52)
	0.001	1.76 (1.62, 1.88)	1.48 (1.37, 1.58)

observations of the square root of the mean squared errors (RMSEs) for the naïve and MLE estimates on each of the 100 generated graphs. Interestingly, as p_{TP} increases from 0.6 to 0.7 (i.e. the technology is more sensitive), RMSE for $\hat{\Delta}_{naïve}$ also increases for equal values of p_{FP}. The naïve approach to degree estimation simply adds FPs to TPs, hence an increase in sensitivity can lead to overestimation of degree depending on the number of FPs per node. On the other hand, as p_{TP} increases, RMSE for $\hat{\Delta}_{MLE}$ decreases, indicating improved estimation of degree for more sensitive technology as would be expected. Also of interest in Table 1 is the notable increase in RMSE for $\hat{\Delta}_{naïve}$ as p_{FP} increases from 0.0008 to 0.001 for equal values of p_{TP}. In contrast, $\hat{\Delta}_{MLE}$ accounts for these FP observations and shows only a modest increase in RMSE for the larger value of p_{FP}.

Much debate has centered on whether graphs exhibit node degree distributions with heavier tails or higher variability than ER random graphs (Li *et al.*, 2006). To explore the performance of $\hat{\Delta}_{MLE}$ in this setting, we generated a series of graphs with 1000 nodes and 1000 edges according to the preferential attachment model of Barabási and Albert (1999) and observed edges according to $p_{TP}=0.7$ and $p_{FP}=0.001$. Log–log plots depicted in Supplementary Figure 1 demonstrate that the distribution of $\hat{\Delta}_{MLE}$ more closely resembles the true distribution than that of $\hat{\Delta}_{naïve}$. Goodness-of-fit of the naïve and MLE distribution estimates was also assessed using RMSE, this time comparing the true probability mass function with the estimated values at each degree. The mean ratio of RMSE for the $\hat{\Delta}_{MLE}$ and $\hat{\Delta}_{naïve}$ estimates on each graph was 0.716 (minimum = 0.598, maximum = 0.796), indicating a consistent reduction in RMSE between 20% and 40% when using the MLE approach. Large variability of degree is a general property of graph data (Li *et al.*, 2006), and these simulations suggest that the MLE approach improves coverage of the full range of true node degrees, particularly in the extremes of the distribution.

2.3 Estimating p_{TP} and p_{FP} using a gold standard

For real data analysis, estimation of Δ requires estimates of p_{TP} and p_{FP} since their true values are generally unknown. A variety of techniques have been discussed for estimation of p_{TP} and p_{FP} (Collins *et al.*, 2007; Deng *et al.*, 2003; D'haeseleer and Church, 2004; Hart *et al.*, 2006) but these do not directly account for bait-prey viability in an experiment. The method we develop here for AP-MS data first aligns protein complex viability in a gold standard set with the data under study. The observed interactions are then compared to the viable complexes in the gold standard to estimate values of p_{TP}. Results are reported here for five AP-MS data sets on *Saccharomyces cerevisiae*: Gavin02 (Gavin *et al.*, 2002), Ho02 (Ho *et al.*, 2002), Krogan04 (Krogan *et al.*, 2004), Gavin06 (Gavin *et al.*, 2006) and Krogan06 (Krogan *et al.*, 2006).

Table 2. Number of VBP nodes, estimated values of p_{TP} and p_{FP}, and mean number of FPs per node for the five AP-MS data sets

	Number of VBP nodes	p_{TP}	p_{FP}	Mean number of FPs per node
Gavin02	268	0.63 (0.57,0.70)	1.0E−3 (9.0E−06,1.8E−3)	0.54 (4.8E−3,0.98)
Ho02	226	0.67 (0.54,0.79)	3.6E−3 (2.7E−3,4.3E−3)	1.6 (1.2,1.9)
Krogan04	149	0.84 (0.74,0.94)	3.7E−3 (1.3E−3,5.6E−3)	1.1 (0.4,1.7)
Gavin06	852	0.70 (0.66,0.74)	7.9E−4 (6.1E−4,9.6E−4)	1.4 (1.0,1.6)
Krogan06	1505	0.52 (0.46,0.59)	8.9E−4 (6.3E−4,1.1E−3)	2.7 (1.9,3.3)

Numbers in parentheses below the p_{TP} estimates are 95% confidence intervals using the variance estimate for $\hat{p}_{TP}$ discussed in Section 3.4. Numbers in parentheses below the p_{FP} estimates are corresponding method of moments estimates of p_{FP} for the estimated range of p_{TP} estimates. The expected number of FPs per node is roughly the product of p_{FP} and the number of VBP nodes multiplied by two to account for FPs occurring as either in- or out-edges.

Our source of candidate gold standard complex co-memberships is a collection of 335 protein complexes culled from the Munich Information Center for Protein Sequences (MIPS) (Mewes *et al.*, 2004) and Gene Ontology (GO) (The Gene Ontology Consortium, 2000), specifically excluding complex estimates based on the high-throughput AP-MS data sets under investigation in this text (see Section 3.4). All pairs of proteins jointly annotated in one of these 335 complexes could, in principle, be detected as interactors by AP-MS technology, as long as all members of the complex are viable proteins in the experiment under consideration. For each AP-MS experiment, we determine the subset of the 335 candidates whose constituent members are all reported in the data set as viable prey and, when applicable, viable baits. Given the resultant number of true complex co-memberships and the observed data, a slightly modified version of the probability statement in Equation (1) can be used to estimate p_{TP} and its variance for each data set (see Section 3.4). Specific estimates of p_{TP} for each experiment are reported in Table 2. Supplementary Figure 2 demonstrates the trend in estimated p_{TP} as the gold standard set of complexes centers on the set of viable proteins in each experiment.

The lack of a robust set of high confidence protein complex '*non*-comemberships' prohibits estimation of p_{FP} in the same manner as p_{TP}; however, with an estimate of p_{TP} in place, a corresponding value for p_{FP} can be calculated using a method of moments approach (see Section 3.5). Table 2 records the

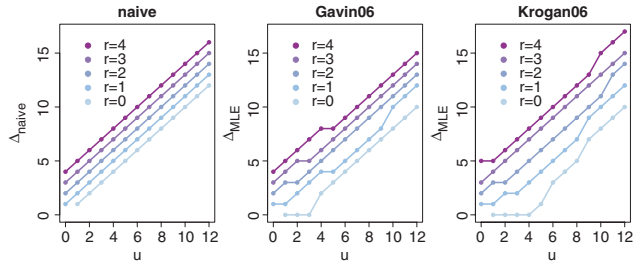

Fig. 3. Estimated degree versus u for increasing values of r. The left-most panel plots $\hat{\Delta}_{\text{naïve}}$ versus u. The middle and right panels plot $\hat{\Delta}_{\text{MLE}}$ for the same values of r and u using the p_{TP} and p_{FP} parameters for the Gavin06 and Krogan06 data sets, respectively.

method of moments p_{FP} estimates for the five AP-MS data sets. The effects of the specific estimates on node degree are discussed in Section 2.4.

2.4 AP-MS data results

We estimated node degrees via the statistical likelihood for the five *S.cerevisiae* AP-MS data sets. As stated in Section 2.1, the Multinomial model assumes that baits prone to various types of systematic bias have been excluded from analysis and only stochastic errors, globally applicable to all VBP proteins, remain. We diagnosed systematically biased VBP proteins by examining the distribution of unreciprocated in- and out-edges as in Chiang *et al.* (2007), eliminating those with severe imbalance from further analysis. Supplementary Table 1 records the number of baits and prey originally reported for each data set as well as the number of VBP nodes both pre- and post-filtering for systematic bias. MLE estimates reported in Sections 2.4.1 and 2.4.2 were computed using the p_{TP} and p_{FP} estimates in Table 2.

2.4.1 Local analysis Figure 3 plots estimated degree versus u for increasing values of r. The left panel reveals the exact linearity of $\hat{\Delta}_{\text{naïve}}$ as r and u increase. The middle and right panels relevant to the Gavin06 and Krogan06 data, respectively, demonstrate the additional dependence of $\hat{\Delta}_{\text{MLE}}$ on p_{TP} and p_{FP}. In practical terms, this means that the same numbers of observed reciprocated and unreciprocated interactions in different data sets can map to different values of $\hat{\Delta}_{\text{MLE}}$. The naïve approach equates observations across all experiments without regard to divergent error probabilities.

Figure 4 illustrates the differences in local node degree estimates for the MLE and naïve techniques. Actual numeric estimates are reported in Supplementary Table 3. In these figures, colors map to negative or positive values of $\hat{\Delta}_{\text{MLE}} - \hat{\Delta}_{\text{naïve}}$. These figures illustrate important points regarding between-experiment variability in node degree estimation. First, the pattern with which estimated degree is higher or lower than the naïve sum varies dramatically from experiment to experiment. While the Gavin02 data set had a low mean estimate of 0.541 FPs per node, it also only had moderate sensitivity. Estimates of Gavin02 degree remain largely consistent with the naïve approach in the lower range, and then consistently increase. On the other hand, the Krogan04 data had high sensitivity as well as a mean of 1.1 FPs per node,

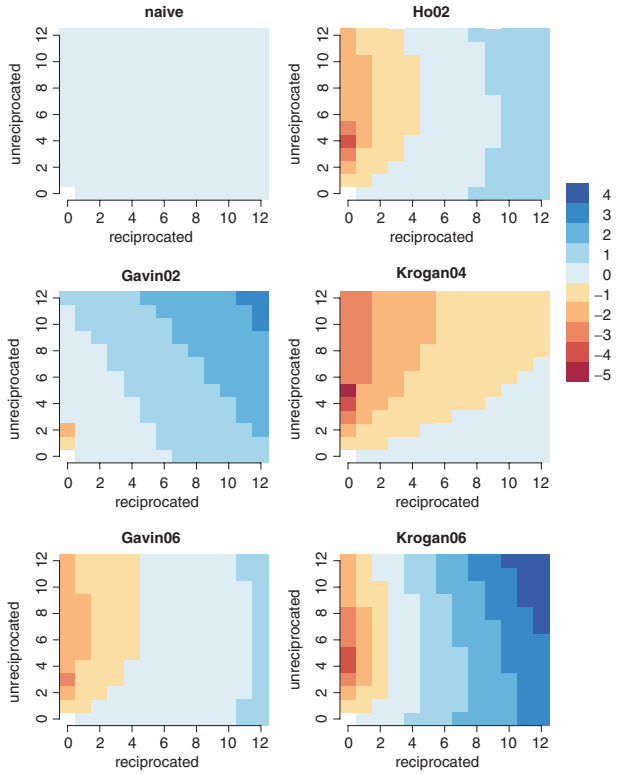

Fig. 4. Differences in degree estimates, $\hat{\Delta}_{\text{MLE}} - \hat{\Delta}_{\text{naïve}}$, under the stochastic error probability parameters in each of the AP-MS data sets. The gradation from orange to red marks negative differences ranging from -1 to -5. The gradation from light blue to dark blue marks non-negative differences ranging from 0 to 4. The pattern in under- and overestimation varies across experiments and depends on p_{TP} and the number of FPs per node. Differences are not plotted for $r = u = 0$ since degree estimation is not performed for nodes without any observed incident edges.

so the MLE estimates are consistently less than the naïve estimates for most of the reciprocated and unreciprocated pairs studied here. Second, Figure 4 illustrates that the largest disparities in node degree estimates tend to exist in the extremes of the observations, i.e. for large or small numbers of reciprocated and unreciprocated interactions.

In addition to between-experiment variability, Figures 3 and 4 also illustrate within-experiment variability. Observations within a data set that would be equivalent under a naïve paradigm are not when estimated via MLE. For example, in the Krogan06 data $r = 2$ and $u = 0$ yield $\hat{\Delta}_{\text{MLE}} = 2$, $r = 1$ and $u = 1$ yield $\hat{\Delta}_{\text{MLE}} = 1$, and $r = 0$ and $u = 2$ yield $\hat{\Delta}_{\text{MLE}} = 0$. While a node with two incident unreciprocated edges would naïvely be assumed to interact with two members of the viable prey population, in fact, the observed interactions could quite possibly be due to stochastic error and the protein may not contribute to the system under study at all. Further wet lab experiments would be required to confirm hypotheses along these lines, but modeling observed interactions according to reciprocity status and stochastic error probabilities points to the reliability of observed interactions and can foster well-informed experimental design and resource allocation.

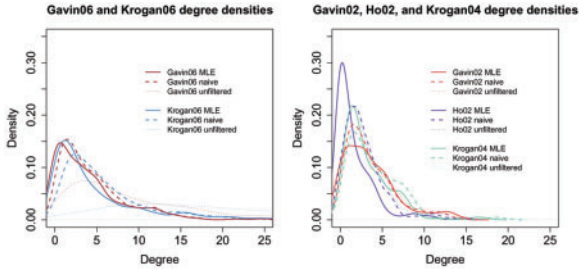

Fig. 5. Degree density estimates of $\hat{\Delta}_{\mathrm{MLE}}$ and $\hat{\Delta}_{\mathrm{naïve}}$, previous to and after filtering the VBP graph for baits prone to systematic bias.

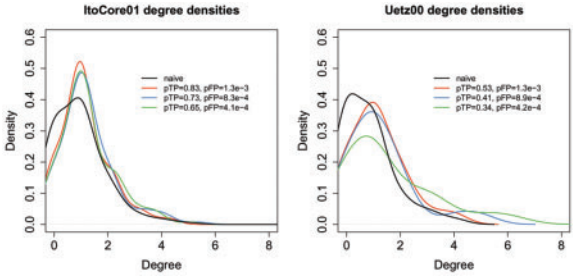

Fig. 6. Degree density estimates of $\hat{\Delta}_{\mathrm{MLE}}$ at the 25th, 50th and 75th percentiles from the family of solutions for p_{TP} and p_{FP} using the method of Chiang *et al.* (2007).

2.4.2 Global degree analysis Figure 5 contains density plots for estimates of AP-MS degree using $\hat{\Delta}_{\mathrm{naïve}}$ prior to filtering baits prone to systematic bias, $\hat{\Delta}_{\mathrm{naïve}}$ post-systematic-filtering, and $\hat{\Delta}_{\mathrm{MLE}}$. The left panel compares degree density estimates for the most recent and larger Gavin06 and Krogan06 AP-MS data sets, and the right panel compares density estimates for the three smaller Gavin02, Ho02 and Krogan04 data sets.

Degree density estimates for the Gavin06 and Krogan06 data suggest that both experiments were prone to systematic error, the latter more so than the former. Degree densities for $\hat{\Delta}_{\mathrm{naïve}}$ change drastically after the removal of biased baits with lower degrees being more prominent than in the raw unfiltered data sets. Even after removal of systematic error and modeling of stochastic error, the degree densities for $\hat{\Delta}_{\mathrm{MLE}}$ are not identical for the Gavin06 and Krogan06 data. Although both of these experiments were intended to be genome wide, experimental conditions differentially affected both bait and prey.

In contrast to the genome-wide surveys, the earlier, smaller data sets are far less subject to systematic error and in general have much lower degree. Although the graphs contained a similar number of nodes, the degree density for Ho02 weights low degree nodes more heavily than the degree density for the Gavin02 data. Krogan04 baits were selected based on involvement in RNA-processing and prefractionization by high-speed centrifugation was used as an intentional means of investigating smaller protein complexes. Relatively low degrees are expected due to the small complexes under investigation, but some connectivity is observed because of the functional commonality of the baits under study.

The interplay between total graph size and local node degree must be considered when drawing biological conclusions about graph data, particularly in the face of measurement error. For example, the Krogan06 data had the highest expected number of FPs per individual node (Table 2), but the global degree densities for Ho02 and Krogan04 actually experienced the greatest shrinkage toward zero after estimation with $\hat{\Delta}_{\mathrm{MLE}}$. Estimated degree for Krogan06 is in general much larger than Ho02 or Krogan04, hence the impact of modeling FPs on a local level plays out differently when the graphs are considered from a global point of view. In a second example, the number of nodes and global p_{TP} and p_{FP} estimates for Gavin02 are closest to those for Ho02. However, after modeling of stochastic error, the node degree distribution for Gavin02 in fact much more closely resembles that for the larger Gavin06 and Krogan06 data sets, indicating larger local degree estimates. For practical data analysis, both local degree estimates and global degree distributions provide complementary information about interactome graphs.

2.5 Y2H data results

Rather than select a particular pair of p_{TP} and p_{FP} parameters for estimating node degree via MLE, it is also of interest to explore a range of plausible error probability estimates and their effect on estimated node degree. Chiang *et al.* (2007) discuss a method of moments approach for estimating a family of solutions for p_{TP} and p_{FP}. After removing VBP proteins prone to systematic bias from the ItoCore01 (Ito *et al.*, 2001) and Uetz00 (Uetz *et al.*, 2000) data sets, we estimated $\hat{\Delta}_{MLE}$ using the 25th, 50th and 75th percentiles for $(p_{\mathrm{TP}}, p_{\mathrm{FP}})$ pairs from the family of solutions for these two parameters and compared the resultant degree densities to the naïve estimate in Figure 6. Interestingly, for the ItoCore01 data the three parameter pairs make very little difference in node degree density (left panel of Fig. 6). For small numbers of observed reciprocated and unreciprocated edges, the MLE estimates are equal for a wide range of error probability combinations. Degree estimate densities for the Uetz00 data are plotted in the right hand panel of Figure 6, and for these lower estimates of p_{TP}, we do see much more variation in resultant MLE density. For both Y2H data sets described here, the MLE estimates at all parameter combinations suggest that in general the naïve approach underestimates degree. We note that in the application of these methods to these two Y2H data sets, we assume the VBP paradigm as described in Section 3.2 is appropriate. These screens did use pooled tests and sampled a limited number of clones for analysis; were these parameters readily available, the bait-specific number of tests could easily be incorporated into analysis.

3 METHODS

3.1 Graph theory and notation

A graph can be represented as $G = (V, E)$ in which V is the set of nodes and E is the set of edges between the nodes. The nodes in V represent objects of interest, e.g. proteins, and $|V|$ denotes the size of V, e.g. the number of proteins. The edges in E represent relationships between nodes, e.g. complex co-membership. A graph in which all edges are not directed is called an *undirected* graph while any graph with one or more directed edges is a *directed* graph. For an undirected graph, *degree* of a node is the number of incident edges. For a directed graph,

the concept of degree is accommodated by considering the in-edges (in-degree) and out-edges (out-degree) separately. The true interactome graph is assumed to be an undirected graph. Data generated by AP-MS and Y2H technologies form a directed graph assay of the interactome with edges extending from baits to prey.

3.2 VBP graphs

Bait-prey technologies test for all relationships involving a set of pre-specified baits and all other prey expressed under the cellular conditions of the experiment. Experimental results are currently only reported for proteins that are observed in at least one interaction, leaving uncertainty as to which proteins were available at the time of interaction testing. Define V as the set of proteins reported as either bait or prey or both in the set of observed interactions. Let V_B denote the set of 'viable baits' defined as baits that detect at least one prey, and let V_P denote the set of 'viable prey' defined as the set of proteins detected by at least one bait. Note that $V_B \cup V_P = V$. The VBP graph is a subgraph of the original data induced by the set of nodes common to both V_B and V_P, or $V_{BP} = V_B \cap V_P$. While it is possible that all nodes in V could be contained in V_B and V_P, that is not the case for any of the five AP-MS data sets discussed here. Furthermore, even though for all five data sets $|V_B| < |V_P|$, it is not the case that $V_B \subset V_P$.

The distributions of unreciprocated in- and out-edges for nodes in the VBP graph are used to diagnose nodes prone to systematic bias using the Binomial method of Chiang *et al.* (2007). Systematically biased nodes are excluded from further analysis in this report.

3.3 Multinomial model for node degree

Let R_T and U_T be random variables representing the number of reciprocated and unreciprocated observations for edges incident on the node of interest that exist in truth. Given Δ and the sensitivity of the technology, p_{TP}, R_T and U_T are jointly distributed according to a Multinomial model, specifically

$$\Pr(R_T = r_T, U_T = u_T | \Delta, p_{TP}) = \frac{\Delta!}{r_T! u_T! (\Delta - r_T - u_T)!} \times p_{T2}^{r_T} p_{T1}^{u_T} p_{T0}^{(\Delta - r_T - u_T)}, \quad (1)$$

where $p_{T2} = p_{TP}^2$, $p_{T1} = 2p_{TP}(1 - p_{TP})$, and $p_{T0} = (1 - p_{TP})^2$. These probabilities arise when an individual edge is subjected to two independent assays, each with probability p_{TP} of producing a TP result. Under stochastic error, the direction of the unreciprocated edges is uninformative. An FN or FP observation can be made in either direction with equal probability. Since bait-prey technologies are not perfectly sensitive, $0 < p_{TP} < 1$.

Both reciprocated and unreciprocated FP edges may also arise, and these are also jointly distributed according to a Multinomial model. Let R_F and U_F be random variables representing the number of reciprocated and unreciprocated observations for edges incident on the node of interest that *do not* exist in truth. Given Δ and the specificity of the technology, $1 - p_{FP}$, R_F and U_F are jointly distributed according to

$$\Pr(R_F = r_F, U_F = u_F | \Delta, p_{FP}) = \frac{(|V_{BP}| - \Delta)!}{r_F!\, u_F!\, ((|V_{BP}| - \Delta) - r_F - u_F)!} \times p_{F2}^{r_F} p_{F1}^{u_F} p_{F0}^{((|V_{BP}| - \Delta) - r_F - u_F)}, \quad (2)$$

where $p_{F2} = p_{FP}^2$, $p_{F1} = 2p_{FP}(1 - p_{FP})$, $p_{F0} = (1 - p_{FP})^2$, and $|V_{BP}|$ is the total number of nodes in the VBP graph. Since bait-prey technologies are not perfectly specific, $0 < p_{FP} < 1$.

Data arising from bait-prey technologies do not afford direct observation of R_T, U_T, R_F and U_F for a node. Rather, the observations are generated by $R = R_T + R_F$ and $U = U_T + U_F$ where R and U are random variables representing the total number (both TPs and FPs) of reciprocated and unreciprocated edges incident upon that node, respectively. The joint distribution of R and U is the convolution of the two distributions in (1) and (2), with a truncation factor that accounts for the fact that the VBP graph includes only nodes for which at least one in- and out-edge is observed. In particular, the convolution is divided by $1 - p_{T0}^{\Delta} p_{F0}^{(|V_{BP}| - \Delta)}$.

The current practice naïve estimate of degree is $\hat{\Delta}_{naïve} = r + u$. The maximum likelihood estimate of Δ, $\hat{\Delta}_{MLE}$, is that which maximizes the joint probability of observing r and u given estimates of p_{TP} and p_{FP}.

3.4 Gold standard complex co-membership edges and estimation of p_{TP}

Candidate complexes for the AP-MS gold standard data sets were obtained from both GO and MIPS. For the GO repository, the Cellular Component Ontology was searched to identify terms with the entire word *complex* or the suffixes *ase* or *some*. We performed manual curation to exclude the following terms that were not protein complexes but were rather subcellular locations with descriptions containing the words *chromosome*, *endosome*, *chitosome*, or *kinetochore*: GO:0000794, GO:0000780, GO:0000781, GO:0000784, GO:0000778, GO:0000942, GO:0031902, GO:0045009, GO:0000776. We also manually checked that all selected GO terms used the word *complex* as a noun. For the MIPS repository, the *S. cerevisiae* genome database was parsed under the *complex* catalog. We included all terms containing the word *complex* in the description but excluded any containing the word *complexes* so as to avoid miscellaneous collections of multiple complexes. After collection and curation, the set of protein complexes was then merged to form one aggregate set of 335 complexes after deleting redundant protein complexes. The complete set is available in the `ScISIC` data set in the R package *ScISI* (version 1.9.7).

For each data set, let Δ_{GS} represent the total number of edges in the complex co-membership graph induced by the gold standard data set. Reciprocated and unreciprocated observations of this set of edges that we believe to exist in truth are then used to estimate p_{TP}. Equation (1) is maximized for p_{TP} after replacing Δ by Δ_{GS} and letting R_T and U_T represent the number reciprocated and unreciprocated observations on the gold standard set, respectively. Truncation of observations is not a concern in this case since it is possible for none of the gold standard edges incident on any unique node to be observed. Specifically, the maximum-likelihood estimate for p_{TP} is $\hat{p}_{TP} = (2r_T + u_T)/(2\Delta_{GS})$. Variance for this estimator can be directly calculated to be $p_{TP}(1 - p_{TP})/(2\Delta_{GS})$, and is estimated by plugging in $\hat{p}_{TP}$.

3.5 Method of moments estimator for p_{FP}

Chiang *et al.* (2007) describe a method of moments approach for the problem of estimating p_{FP}. We briefly describe their method here. Let $|V_{BP}|$ be the number of nodes in the VBP graph, then the largest number of possible distinct interacting protein pairs is $\binom{|V_{BP}|}{2}$. Let Δ_{VBP} be the true number of unique interacting pairs and $\Delta_{VBP}^c = \binom{|V_{BP}|}{2} - \Delta_{VBP}$ the number of non-interacting protein pairs. The expected number of reciprocally adjacent pairs R_{VBP} and unreciprocally adjacent pairs U_{VBP} in the VBP graph are:

$$E[R_{VBP}] = \Delta_{VBP}\, p_{TP}^2 + \Delta_{VBP}^c\, p_{FP}^2 \quad (3)$$

$$E[U_{VBP}] = \Delta_{VBP}\, 2\, p_{TP}\, (1 - p_{TP}) + \Delta_{VBP}^c\, 2\, p_{FP}\, (1 - p_{FP}) \quad (4)$$

These two independent equations in three parameters $\{(p_{TP}, p_{FP}, \Delta_{VBP})\}$ generate a family of solutions in which a value of any one of these parameters determines unique solutions for the other two (Chiang *et al.*, 2007).

4 DISCUSSION

Experimental data from bait-prey technologies are prone to sampling bias, FP and FN observations. While these issues are widely recognized, little has been done to statistically

model these errors when estimating global and/or local graph statistics. Straightforward application of likelihood techniques can be used to estimate node degree with more accuracy than the current practice naïve estimate. Simulation studies demonstrate the accuracy of the MLE approach even under misspecification of stochastic error probabilities, and depictions of graphs with known node degree distributions show the extent to which MLE degree estimation can be used to more closely resemble the true distribution. Node degree itself is useful for understanding the range of influence of an individual protein in the cell, and is a helpful contributor for elucidating the identities of a protein's interactors given the possibilities reported in a data set. Accurate estimation of node degree, and many other graph parameters, cannot be overlooked.

Our MLE approach limits analysis to the VBP graph and in this report we compare it to the naïve estimator also used on the VBP graph. In practice, naïve summation is generally applied to the graph including all baits and prey-only proteins, regardless of the severe imbalance in the number of tested edges incident on each. Under perfect sensitivity and specificity, the number of detected interactions for a bait would equal its true degree since all incident edges are tested. On the other hand, for prey-only proteins, only the subset of edges connected to bait proteins is tested. The naïve sum approach generally treats the remaining untested edges as though they are tested but not observed. Hence, even under perfect sensitivity and specificity for tested edges, naïve summation actually only yields a lower bound for the true degree for prey-only proteins. Estimation of degree for nodes outside the VBP set could proceed quite naturally using the Hypergeometric distribution under the assumption that the VBP nodes are a random sample from the population of proteins. If the VBP nodes are not a random sample then it is not, in general, possible to make specific statements about the probability that a prey is observed. The AP-MS and Y2H data sets discussed here give no evidence that the set of VBP nodes is a representative random sample of proteins in the cell, hence extrapolation of the results to untested portions of the interactome using either the MLE or naïve degree estimation approach would be largely irrelevant.

Focusing on the VBP graph for data analysis does therefore demand careful interpretation of the term *degree*. Degree in this context represents the number of interactions between viable baits that were also viable prey, and is therefore representative of both bait selection and the set of constitutive proteins under the experimental conditions. For intended genome-wide experiments such as Gavin06 and Krogan06 or for smaller-scale experiments with baits targeting specific cellular processes such as Krogan04, interpretation is more straightforward. Experiments between the two extremes of genome-wide assay and those targeting specific cellular functions, e.g. Gavin02, Ho02, ItoCore01 and Uetz00 make the interpretation of VBP degree slightly more difficult. That said, degree density estimates for these experiments do point to different characteristics of the detected interactions. While AP-MS and Y2H data are used here to demonstrate MLE degree estimation, the technique is equally applicable and similar conclusions can be made for data generated by other bait-prey technologies for which the VBP paradigm is appropriate.

One limitation of the MLE estimation method in its current formulation is that the same estimate of degree is made for all nodes with the same numbers of observed unreciprocated and reciprocated edges within a data set. In reality, protein-specific covariates, e.g. PFAM domains, likely contribute to variability in the observed data. As further relationships between these types of protein characteristics and node degree are uncovered, they can easily be accommodated into the MLE paradigm by introducing protein-specific FP and FN error probabilities.

Classic statistical techniques are readily applicable to graph feature analysis, and in this specific example yield accurate estimates of node degree. This work represents a first step toward rigorous handling of experimental noise in bait-prey data when estimating graph statistics. Likelihood methods are also appropriate for estimation of other features of *interactome* behavior, and a move beyond naïve summary statistics of observed data is both needed and warranted. Provided data are reported with bait-prey distinctions, sampling, systematic errors and stochastic errors can be easily addressed. Such approaches promise greatly increased accuracy in estimation of global and local statistics, as well as more holistic model development.

ACKNOWLEDGEMENTS

We acknowledge funding through the Human Frontiers Science Program Research Grant RGP0022/2005 to W.H. and R.G., also in support of T.C.

Conflict of Interest: none declared.

REFERENCES

Barabási,A.-L. and Albert,R. (1999) Emergence of scaling in random networks. *Science*, **286**, 509–512.

Chiang,T. *et al.* (2007) Coverage and error models of protein-protein interaction data by directed graph analysis. *Genome Biol.*, **8**, R186.

Collins,S. *et al.* (2007) Toward a comprehensive atlas of the physical interactome of *Saccharomyces cerevisiae*. *Mol. Cell. Proteomics*, **6**, 439–450.

Deng,M. *et al.* (2003) Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac. Symp. Biocomput.*, **8**, 140–151.

D'haeseleer,P. and Church,G. (2004) Estimating and improving protein interaction error rates. In *Proceedings of the IEEE Computational Systems Bioinformatics Conference August 16-19 2004*. IEEE Computer Society, California, pp. 216–223.

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

Gavin,A.C. *et al.* (2006) Proteome survery reveals modularity of the yeast cell machinery. *Nature*, **440**, 631–636.

Hart,G.T. *et al.* (2006) How complete are current yeast and human protein-interaction networks? *Genome Biol.*, **7**, 120.

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

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

Krogan,N. *et al.* (2004) High-definition macromolecular composition of yeast RNA-processing complexes. *Mol. Cell*, **13**, 225–239.

Krogan,N. *et al.* (2006) Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature*, **440**, 637–643.

Li,L. *et al.* (2006) Towards a theory of scale-free graphs: Definition, properties, and implications. *Internet Mathe.*, **2**, 4.

Mewes,H. *et al.* (2004) MIPS: analysis and annotations of proteins from whole genomes. *Nucleic Acids Res.*, **32**, D41–D44.

The Gene Ontology Consortium (2000) Gene Ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

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