Probabilistic Paths for Protein Complex Inference

Hailiang Huang[1,2], Lan V. Zhang[3], Frederick P. Roth[3,4], and Joel S. Bader [1,2]

[1] Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21218
{hlhuang,joel.bader}@jhu.edu
[2] High-Throughput Biology Center, Johns Hopkins School of Medicine, Baltimore, MD 21205
[3] Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, and
[4] Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115
lanvzhang@gmail.com, fritz_roth@hms.harvard.edu

Abstract. Understanding how individual proteins are organized into complexes and pathways is a significant current challenge. We introduce new algorithms to infer protein complexes by combining seed proteins with a confidence-weighted network. Two new stochastic methods use averaging over a probabilistic ensemble of networks, and the new deterministic method provides a deterministic ranking of prospective complex members. We compare the performance of these algorithms with three existing algorithms. We test algorithm performance using three weighted graphs: a naïve Bayes estimate of the probability of a direct and stable protein-protein interaction; a logistic regression estimate of the probability of a direct or indirect interaction; and a decision tree estimate of whether two proteins exist within a common protein complex. The best-performing algorithms in these trials are the new stochastic methods. The deterministic algorithm is significantly faster, whereas the stochastic algorithms are less sensitive to the weighting scheme.

1 Introduction

The genome sequence of an organism provides a blueprint of its genes and proteins, but not the connections between these parts. Understanding how proteins are physically organized into complexes and pathways is increasingly based on observations from high-throughput experiments. Yeast has been the most widely used model for eukaryotic proteomics. High-throughput yeast two-hybrid screens have provided evidence for pair-wise links between proteins screens [1, 2]. Affinity purification followed by mass spectrometry identifies proteins that co-purify with a bait protein, suggesting shared membership in one or more protein complexes [3, 4].

Experimental interaction evidence can be unreliable due to high false-positive and false-negative rates [5, 6]. Experimental reports have included estimates of confidence based on multiple observations [1, 2, 7]. A more recent report of the fly protein interaction network included more sophisticated confidence metrics based on sequence analysis and network topology [8].

Here we consider three confidence-weighted networks derived from high-throughput data for yeast. The first, by Roth's group, is a naïve Bayes prediction (NB)

T. Ideker and V. Bafna (Eds.): Syst. Biol. and Comput. Proteomics Ws, LNBI 4532, pp. 14–28, 2007.

of the posterior probability w_{ij} that two proteins have a direct physical interaction conditioned on observed data [9],

$$\frac{w_{ij}}{1-w_{ij}} = \left[\prod\nolimits_{\tau} \frac{\Pr(x_{ij}^{\tau} \mid m_{ij})}{\Pr(x_{ij}^{\tau} \mid \bar{m}_{ij})}\right] \cdot \frac{\Pr(m)}{\Pr(\bar{m})}, \tag{1}$$

where τ labels the different types of experimental data, x_{ij}^{τ} is the experimental data of type τ relating to protein pair i and j, m_{ij} indicates that the proteins have a direct physical interaction, $\bar{m}_{ij}$ indicates that the proteins do not have a direct physical interaction, and $\Pr(m)$ is the prior probability that two arbitrary proteins have a direct physical interaction.

The second network, by Bader and coworkers, predicts the probability of a direct or indirect physical interaction using a logistic regression model (LR) [10],

$$w_{ij} / (1 - w_{ij}) = \exp(\Sigma_{\tau} \beta_{\tau} x_{ij}^{\tau} + \Sigma_{\tau,\tau'} \beta_{\tau,\tau'} x_{ij}^{\tau} x_{ij}^{\tau'} + \ldots) \cdot \Pr(m) / \Pr(\bar{m}). \tag{2}$$

The model parameters $\{\beta\}$ were estimated using a training set equally weighted for true-positives and false-positives, equivalent to using 1 for the prior likelihood ratio $\Pr(m)/\Pr(\bar{m})$. Although the logistic regression scores have been used as the posterior probability of a true interaction [11], they are overconfident to the extent that non-interacting protein pairs outnumber interacting pairs in the true interaction network. A one-parameter fit for $\Pr(m)/\Pr(\bar{m})$ similar to that used for the NB network would convert the LR confidence scores to probabilities.

The final network, again by Roth's group, used a decision tree to estimate probabilities of protein pairs being co-complexed (DT) [12]. Then the odds of being co-complexed are multiplied by an adjustable parameter to estimate the odds of a direct physical interaction. This single parameter may then be fit to optimize performance for a training set. Unlike the NB model, the LR and DT models have the benefit of explicitly modeling dependence between predictors.

Other groups have used related methods to infer confidence-weighted edges not observed in the high-throughput data [13-20]. Some such methods include inference of shared complex membership or common function, training on just one complex or function at a time [21]. While we restricted our attention to the NB, LR, and DT weighting schemes, the methods we describe are directly applicable to other weighting schemes as well. Thus, the starting point for the methods we describe is an undirected weighted graph, in which proteins are represented as vertices and edge weights in the range [0,1] represent the probability of a direct or indirect physical interaction between proteins.

We investigate two general classes of algorithms that use confidence-weighted networks to infer protein complexes containing one or more seed proteins. First are deterministic algorithms, which directly calculate a threshold neighborhood around each seed protein, then identify proteins in the union of the neighborhoods as potential members of the complex. These algorithms include BESTPATH, published by Bader et al. as the SEEDY algorithm [22] and Shortest Path with Evidence (SPE), published previously by Roth's group [9] as a baseline for comparing improved algorithms. Here we report a new deterministic algorithm, SUMPATH, which attempts to combine information across multiple seeds.

The second class of algorithms generates a stochastic ensemble of networks using the edge weights as probabilistic measures that an edge taken from the high-throughput data is a true positive. This method was introduced by Roth's group in the PRONET algorithm [9], which requires the edge weights to refer to the probabilities of direct connections within a complex. Here we describe two related algorithms, PROPATH-ALG and PROPATH-EXP, designed to work well when edge weights also reflect the probability of indirect connections.

Although algorithms that are initialized with positive and negative seeds have been shown to be useful [21], the algorithms we describe use only positive seeds. Positive and negative seeds are particularly appropriate in the context of functional annotations using GO terms [23] or other ontologies in cases where terms in different lineages from the root are negatively correlated or mutually exclusive. The algorithms are also different from algorithms of finding complexes *de novo* [24, 25], which requires no seeds information.

Beyond introducing the new SUMPATH and PROPATH methods, the rationale of this report is to compare the abilities of each of these algorithms relative to recover well-annotated protein complexes when given partial information about these complexes. As in previous studies [9, 22], we use protein complexes from the MIPS catalog [26]. Furthermore, since the algorithms can be considered independently from the network confidence scores, we also compare performance as a function of the confidence score input. Because the PRONET algorithm was developed specifically for weights corresponding to direct connections, its performance is most fairly compared with other algorithms using the NB edge weights. Nevertheless, we provide results for all three networks using PRONET in the interests of completeness.

2 Methods

A summary of the algorithms is provided as Table 1. The input to each algorithm is a set of weighted edges $\{w_{ij}\}$ representing high-throughput interactions between proteins i and j, and a set of one or more seed proteins $\{s\}$. The output of each algorithm is a ranked list of other proteins in the network, where p_r is the protein with rank r in the list. Lower ranks correspond to greater probability that a protein is a member of a complex containing one or more of the seeds. In most of the algorithms, the ranks are calculated by first calculating a score S_i for each protein i, with higher scores corresponding to lower rank.

Each algorithm generates complex-membership scores differently based on the existence of one or more paths connecting seed proteins to other proteins in the network. For many proteins, no such path exists. These proteins are formally described as having distance = ∞ and/or score = 0 (the lowest possible value) and are appended to the end of the ranked list. We first describe the deterministic methods, Shortest Path with Evidence (SPE) [9], BESTPATH [22], and SUMPATH, then describe PRONET [9] and the probabilistic PROPATH algorithms.

Shortest Path with Evidence (SPE). The SPE method ignores the edge weights, treating each edge with any supporting evidence as having the same weight. The distance D_i of each protein in the network to the set of seeds is calculated as

$$D_i = \min_{s \in \text{seeds}} D_{is} \tag{3}$$

where D_{is} is the number of links in the shortest path connecting protein i to seed s, or $+\infty$ if no such path exists. Proteins are then ranked in decreasing order of D_i.

BESTPATH. The BESTPATH algorithm is identical to SEEDY, published earlier by Bader et al [22]. Here we term this algorithm BESTPATH to be more descriptive. With this algorithm, the weight of a path through proteins i_1, i_2, …, i_n is the product of edge weights $\prod_{k=1}^{n-1} w_{i_k i_{k+1}}$. The score of each protein is defined as

$$S_i = \max_{s \in \text{seeds}} S_{is} , \tag{4}$$

where S_{is} is the highest weighted path between protein i and seed s. These paths may be computed efficiently using standard algorithms for traversing weighted graphs. Our implementation uses a priority queue implemented through a max-heap.

SUMPATH. We developed the SUMPATH method in an attempt to improve BESTPATH by searching for multiple high-weight paths. SUMPATH is based on Ising models for spin lattices [27, 28]. Each protein is assigned a spin label, 1 (part of the complex) or –1 (not part of the complex). Weighted edges in the network are interpreted as couplings between spins [29], and the goal is to identify the set of labels $\{S_i\}$ that minimize an energy function $-\Sigma_{(ij)} S_i w_{ij} S_j - \Sigma_i \phi_i S_i$, where ϕ_i is an external field representing prior knowledge of the probability of each spin state. Approximations such as mean field theory [28] or belief propagation [30] can be applied to reduce the computational complexity, but are beyond the scope of this paper. Here we present a simplified method. In this method, each seed s is assigned a score $S_s = 1$ that remains fixed throughout the algorithm. The BESTPATH method is used to initialize the scores $S_i^{(0)}$ of the other proteins. Scores for iteration $q+1$ are obtained using the equations

$$T_i^{(q+1)} = \sum_j w_{ij} S_j^{(q)}, \quad Norm^{(q+1)} = \max_i T_i^{(q+1)} \text{ and } S_j^{(q+1)} = T_i^{(q+1)} / Norm^{(q+1)} \tag{5}$$

to update the scores from iteration q. The sum over j in the first equation includes seed proteins. Iterations proceed until convergence, with 8-10 iterations required for convergence according to the criterion $\max_i \left| S_i^{(q+1)} - S_i^{(q)} \right| < 0.001$. The converged scores are then output. The normalization is required to prevent scores from growing without bound and is performed for the entire network rather than separately for each connected component.

PROPATH and PRONET Methods. PROPATH and PRONET are stochastic methods that require the generation of an ensemble of K replicate networks based on the edge weights. Each protein pair in each generated network receives a weight of either 0 or 1 based on a Bernoulli trial (i.e., a 'weighted coin flip') with probability w_{ij} that an edge between proteins i and j exists. Edges that are not included in the weighted network are assumed to have confidence 0 and never appear in a replicate network.

For each replicate network $k \in K$, the shortest path between protein i and seed s is denoted $D_{is}^{(k)}$, with $D_{is}^{(k)} = \infty$ if no path exists. These distances are calculated as with SPE, rather than BESTPATH, as the edge weights have already been taken into account in the generation of the replicate network. As with SPE, the distance to the closest seed is retained for each protein, $D_i^{(k)} = \min_s D_{is}^{(k)}$. If two proteins are in the same complex, we anticipate that multiple replicates in the ensemble will have a short path connecting the proteins. The mean distance over the ensemble, $K^{-1}\sum_k D_i^{(k)}$, is an inappropriate summary statistic because of the possibility that one of the replicates will generate an infinite distance.

The different PROPATH methods use distinct mathematical transforms to avoid this problem. Each transform maps infinite distance to zero score, and unit distance (the smallest possible distance for a protein that is not itself a seed) to unit score. The transforms we selected are

$$S_i^{(k)} = \begin{cases} I\left(D_i^{(k)} < \infty\right), & \text{PRONET} \\ \left(D_i^{(k)}\right)^{-\alpha}, & \text{PROPATH-ALG} \\ \exp\left(-\alpha D_i^{(k)} + \alpha\right), & \text{PROPATH-EXP} \end{cases}, \tag{6}$$

where $S_i^{(k)}$ is the transformed score of protein i in replicate network k, $I(\arg)$is an indicator function that is 1 for a true argument and 0 for a false argument, and α is a parameter defining the steepness of the decay of the algebraic or exponential transform. We have found that the PROPATH algorithms are insensitive to the exact value of α, with similar results for PROPATH-EXP for values of α up to 5 (a much faster decay; results not shown). For convenience, we used $\alpha = 1$ for PROPATH-ALG and PROPATH-EXP; performance may improve with an additional optimization over this single parameter.

In the PRONET algorithm, there is no distance-based decay. The existence of a path connecting a pair of vertices is converted to a 0/1 binary variable that is averaged over probabilistic networks. Formally, this is equivalent to taking the limit $\alpha \to 0$ in the PROPATH algorithms.

The final score of a protein is estimated as the average over replicates,

$$\hat{S}_i = K^{-1}\sum_k S_i^{(k)}. \tag{7}$$

The variance of $\hat{S}_i$ is bounded because $0 \le S_i^{(k)} \le 1$:

$$\mathrm{var}\left(\hat{S}_i\right) = K^{-1}\left(\left\langle S_i^2\right\rangle - \left\langle S_i\right\rangle^2\right) \le K^{-1}\left(\left\langle S_i\right\rangle - \left\langle S_i\right\rangle^2\right) \le \left(4K\right)^{-1}, \tag{8}$$

where the angle brackets refer to an average over a single replicate network. We used $K = 400$ to give a standard deviation no larger than 0.025. We checked that results had converged with respect to K.

Performance Metrics. We followed the same general procedure for each complex. First, we generated $N_{\mathrm{trial}} = 10$ random 50-50 splits of the complex into seed proteins and target proteins that were used as input to each algorithm. For complexes with an odd number of members, the seed group had one more member than the target group. The set of target proteins for trial t of complex c is denoted T_{ct}. The seeds were then used as input seeds for each of the algorithms, which returned lists of proteins ranked by decreasing likelihood of membership in the same complex as the seeds. Proteins used as seeds were omitted from the ranked list. The protein at rank r for trial t of complex c is denoted p_{ctr}. The indicator function $I\left(p_{ctr} \in T_{ct}\right)$ is 1 if this protein belongs to the target set and 0 otherwise.

Summing the indicator function over ranks, trials, and complexes provides a quantitative assessment of algorithm performance by generating a receiver operating characteristic (ROC) curve. The order of summation was as follows. First, for each complex and trial, we calculated the numbers of true positives and false positives through rank r, $TP_{ct}\left(r\right)$ and $FP_{ct}\left(r\right)$, as

$$TP_{ct}\left(r\right) = \sum_{r'=1}^{r} I\left(p_{ctr'} \in T_{ct}\right) \text{ and } FP_{ct}\left(r\right) = r - TP_{ct}\left(r\right). \tag{9}$$

This makes the conservative assumption that the identity of each complex is correctly reported in the MIPS data. The true positive and false positives counts were then averaged over the trials for each complex,

$$TP_{c}\left(r\right) = \left(N_{\mathrm{trial}}\right)^{-1} \sum_{t=1}^{N_{\mathrm{trial}}} TP_{ct}\left(r\right) \text{ and } FP_{c}\left(r\right) = r - TP_{c}\left(r\right). \tag{10}$$

The counts were then converted to true-positive and false-positive rates for each complex,

$$tp_{c}\left(r\right) = TP_{c}\left(r\right) / \left|T_{c}\right|, \;\; fp_{c}\left(r\right) = FP_{c}\left(r\right) / \left|N_{\mathrm{tot}} - N_{c}\right|. \tag{11}$$

where $\left|T_c\right|$ is the cardinality of the target set for complex c, and $\left|N_{\mathrm{tot}} - N_{c}\right|$ is the number of proteins in the interaction network minus those that are also in the complex. Note that the maximum value for $tp_{c}\left(r\right)$ for large r is less than 1 if not every protein in the complex is in the protein interaction network. The maximum value of $fp_{c}\left(r\right)$ is 1, however. The overall true-positive and false-positive rates, averaged over complexes, are

$$tp(r) = C^{-1}\sum_{c=1}^{C} tp_c(r) \text{ and } fp(r) = C^{-1}\sum_{c=1}^{C} fp_c(r). \tag{12}$$

This procedure gives equal weight to each complex. The ROC curve is the parametric plot of $tp(r)$ vs. $fp(r)$.

As with microarray analysis, the false-discovery rate may be more informative than the false-positive rate because the maximum number of false-positives far outweighs the maximum number of true-positives. The false-discovery rate is defined as a function of r as

$$fd(r) = C^{-1}\sum_{c=1}^{C} \frac{FP_c(r)}{TP_c(r) + FP_c(r)}. \tag{13}$$

With ~4000 proteins in the network, the false discoveries begin to dominate the returned list of proteins when the false-positive rate is on the order of N_{tot}^{-1}, or ~ 10^{-3}.

The area under the ROC curve (AUC) provides a quantitative measure of performance, with higher AUC corresponding to better performance. Our focus is on the region of the ROC curve with few false-positives. Thus, rather than calculating the area under the entire curve, we calculate the area up to a false-positive rate typical of what would be used in practice. We normalize this area to return a value termed $\text{AUC}(fp)$ that increases with better recall,

$$\text{AUC}(fp) = (fp)^{-1}\int_{0}^{fp} d(fp')tp(fp'), \tag{14}$$

where the true-positive rate is considered to be a function of the false-positive rate. Results are provided for AUC(0.1%) and AUC(0.5%). The AUC for a complex (AUCc) is also calculated to measure the complex specific recovery performance,

$$\text{AUC}_c(fp_c) = (fp_c)^{-1}\int_{0}^{fp_c} d(fp')tp_c(fp').$$

3 Results

Algorithms for extracting protein complexes from confidence-weighted interaction data were tested by assessing their ability to extract a known complex based on partial knowledge of its components. As a gold standard of true complexes, we used C = 23 known complexes from MIPS [26]. These complexes include many of those used in the original reports of the PRONET and BESTPATH algorithms. In general, each algorithm returns a ranked list of possible complex members and, based on the known complex, calculates recovery rates as a function of proteins through rank r : the true-positive rate $tp(r)$ (the fraction of positive predictions that are correct); the false-discovery rate $fd(r)$ (the fraction of positive predictions that are incorrect); and the false-positive rate $fp(r)$ (the fraction of non-interacting pairs that are predicted positive). Performance is visualized by graphing $tp(r)$ vs. $fp(r)$ in the region of $0 < fp(r) < 5 \times 10^{-3}$, corresponding to ~20 false-positives, and the $tp(r)$ vs. $fd(r)$ graph in the full region of $0 < fd(r) < 1$. Quantitative measures such as normalized AUC (Area Under the Curve) and FP-50 (false-positive rate at 50% recall) provide a convenient summary metric for ranking the algorithms (Table 1). The AUC for each complex (AUCc) is calculated (Fig. 2).

Table 1. Summary of methods. For each network, each algorithm was ranked 1-6 in performance, 1 = best, 6 = worst. Superscripts in numbers stand for the ranking, and are also indicated by the background colors (Green = rank 1 or 2; Yellow = rank 3 or 4; Red = rank 5 or 6; ties are colored as the best rank). The ranks were averaged to give an overall measure of each algorithm's performance. [a]Normalized area under the curve (AUC) at a false-positive rate of 0.1%, in percentage scale. See Eq. [14] for the normalization. [b]Normalized AUC at a false-positive rate of 0.5%. [c]False-positive rate at 50% recall, in percentage scale.

	Avg. Rank	AUC 0.1%(%)[a]			AUC 0.5%(%)[b]			FP-50(%)[c]			CPU Time (min)		
		NB	LR	DT	NB	LR	DT	NB	LR	DT	NB	LR	DT
PROPATH-EXP	2.25	10^1	17^1	20^1	19^1	34^1	34^1	7.3^1	1.2^2	1.0^2	8.1^5	240^5	2900^6
PROPATH-ALG	2.33	10^1	17^1	20^1	19^1	34^1	34^1	8.0^3	1.2^2	1.0^2	7.9^4	240^6	2700^5
BESTPATH	2.5	7.6^4	15^3	18^3	15^4	33^3	32^3	7.7^2	0.9^1	0.9^1	7.3^2	8.1^2	95^2
PRONET	3.8	10^1	0.03^5	18^3	19^1	0.18^5	31^4	9.5^5	27^6	1.6^4	7.6^3	230^4	2500^4
SPE	4.3	0.43^6	0.41^4	0.29^6	4.3^6	2.6^4	1.4^6	11^6	5.8^4	9.3^6	4.8^1	5.8^1	56^1
SUMPATH	4.8	1.3^5	0.009^6	4.1^5	12^5	0.092^6	8.4^5	8.1^4	22^5	5.8^5	77^6	42^3	300^3

NB network. We first compared algorithm performance for the confidence scores taken from NB [9] (Table 1 and Fig. 1A, B). The AUC (0.1%) and AUC (0.5%) measures show that PROPATH-EXP, PROPATH-ALG and PRONET have roughly equivalent performance in the region of stringent prediction, followed by BESTPATH. The SUMPATH algorithm has intermediate performance, and the SPE has the worst performance.

LR network. We then compared algorithm performance for confidence-weighted edges taken from LR [10]. The PROPATH-EXP and PROPATH-ALG algorithms perform the best and are comparable, followed closely by BESTPATH (Fig. 1C, D). These three algorithms dominate the other algorithms in this region of stringent prediction, returning ~40-50% of the target proteins.

DT network. The last edge weights we used are from DT [12]. This set of edge weights has a tunable parameter α. For each algorithm, we chose the value of α that maximized its AUC(0.5%). PROPATH-EXP and PROPATH-ALG have equivalent performance, followed closely by PRONET and BESTPATH. The remaining algorithms, SUMPATH and SPE, have the worst performance (Fig. 1E, F).

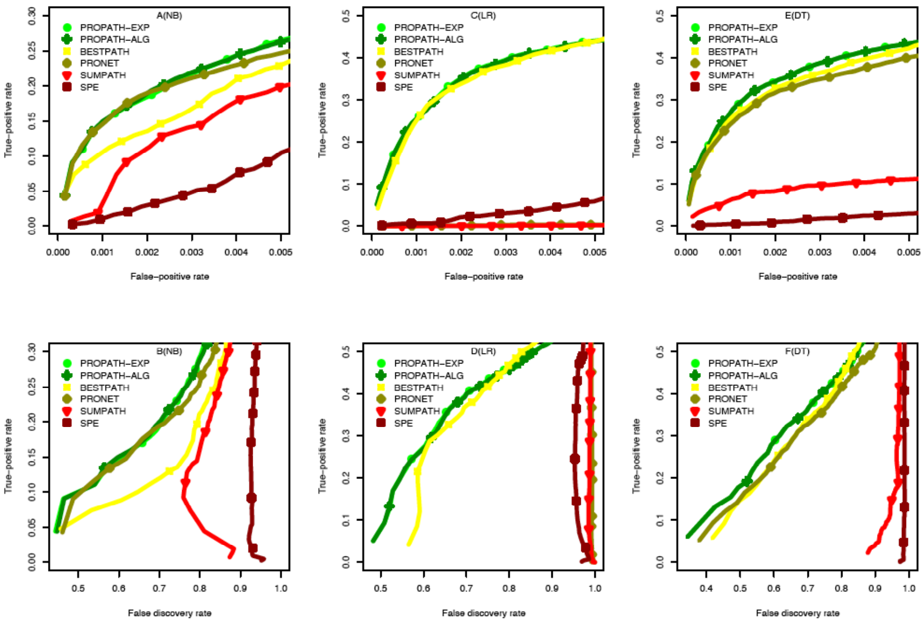

Fig. 1. Algorithm performance. Receiver operating characteristic (ROC) curves and false-discovery rates characterize the performance of algorithms to extract protein complexes from protein interaction networks. Fig. 1A and 1B are from edge weights using NB [9], Fig. 1C and 1D are from edge weights using LR [10] and Fig. 1E and 1F are from edge weights using DT [12].

Complex-specific recovery. We then investigated whether certain complexes are easier to recover than others. Given a set of network edges and a recovery algorithm, a one-sided Wilcoxon test was used to test the significance of the hypothesis that a particular complex had a higher than average AUC 0.5% compared to other complexes recovered using the same network edges and the same algorithm. A more complete description is provided in the Methods.

We found that the best-performing algorithms (PROPATH-EXP, PROPATH-ALG, and BESTPATH) consistently recovered four complexes with a higher than average AUC 0.5% regardless of the network edges used: the PROTEASOME, HISTONEAC, HISTONEDEAC and NUCLEARPORE. One reason for better-than-average recovery of these specific complexes may be the number of proteins contained in these gold-standard examples, 36, 17, 4, and 24 respectively. These are less than the mean number of proteins across all complexes, 45.8. A possible interpretation is that these four represent distinct single complexes. Other gold-standard complexes may in fact comprise a number of more loosely coupled sub-complexes that are more difficult to recover as single cohesive units. Such sub-complexes might also be expected to have more interactions outside the gold-standard complex, which would reduce the AUC. Furthermore, signaling pathways might also be expected to be more loosely coupled and not recovered as well.

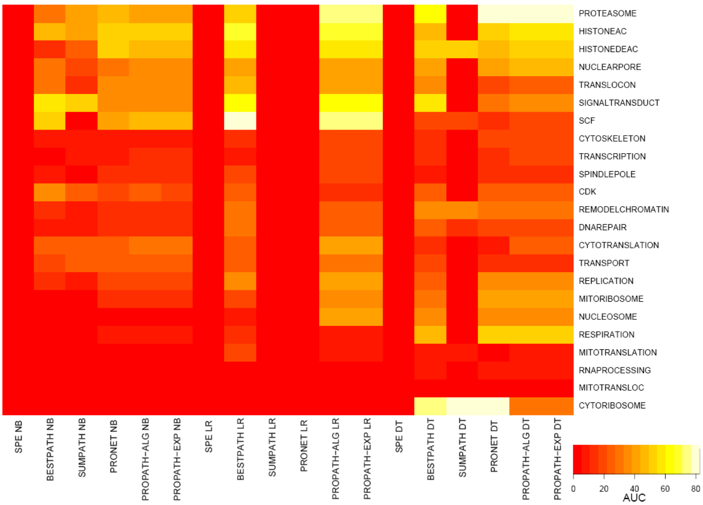

Fig. 2. Complex-specific performance plot identifies complexes that are better extracted using our methods. AUC 0.5% for each complex is visualized using the color key on the bottom right corner. Complexes have been reordered to show clusters of similar performance.Lighter colors indicate better performance.

Recovery performance may be visualized using a color-coded display of true-positives and nominal false-positives predicted by an algorithm (Fig. 3). We focus on a specific complex, histone acetyltransferase (HAC), which has 17 members. We generated random 50-50 splits of the complex into seed proteins (E_t) and target proteins. The seed proteins were then used as input to PROPATH-EXP with LR to generate a list of proteins ranked by decreasing likelihood of their memberships in HAC. We kept the first half of the ranked list, $P_t = \left\{ p_{rt}, r \le \frac{N_{\text{tot}} - N_c/2}{2} \right\}$, excluding the seeds. N_{tot} is the total number of proteins in the list and N_{c} is the number of proteins in the complex. The number of times protein p has been used as seed is $N_{sp} = \sum_{t=1}^{N_{\text{trial}}} I\left(p \in E_t\right)$, where $I\left(p \in E_t\right)$ is the indicator function, $I\left(p \in E_t\right) = \begin{cases} 1 & p \in E_t \\ 0 & p \notin E_t \end{cases}$. The maximum possible recovery count for protein p is $N_{\text{trial}} - N_{sp}$, and the recovery rate for protein p is $R_p = \sum_{\text{t=1}}^{\text{N}_{\text{trial}}} I\left(p \in P_t\right) \Big/ \left(N_{\text{trial}} - N_{\text{sp}}\right)$. We defined three categories of recovered proteins:

$$\{p \in \text{HAC}\} \cap \{p, \text{with } R_p \geq 0.5\}, \text{ High recovery rate true-positive protein}$$

$$\{p \in \text{HAC}\} \cap \{p, \text{with } R_p < 0.5\}, \text{ Low recovery rate true-positive protein}$$

$$\{p \notin \text{HAC}\} \cap \{p, \text{with } R_p \geq 0.5\}, \text{ High recovery rate false-positive protein}$$

In the graph, we have 9 out of 17 HAC proteins recovered with $R \geq 0.5$ and 2 false positive proteins with $R \geq 0.5$. Despite not being included in the MIPS catalogue for HAC, these two proteins, SGF29 and SGF73, are annotated in SGD as probable subunits of the SAGA HAC.

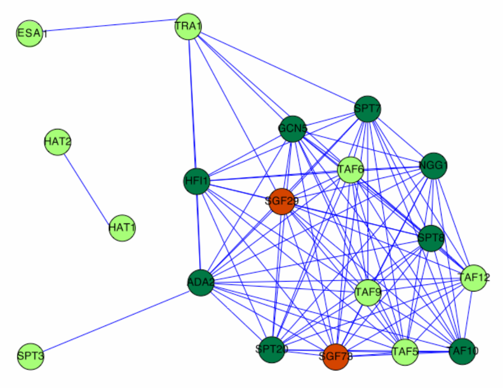

Fig. 3. Complex recovery graph. Histone acetyltransferase complex recovery graph shows the rate of proteins being recovered. Dark green nodes indicate high recovery rate true-positive proteins. Light green nodes indicate low recovery rate true-positive proteins. Red nodes indicate high recovery rate false-positive proteins.

Beyond recall performance, CPU performance may also be a criterion for selecting an algorithm. Timings are provided for a Perl implementation of each algorithm (FreeBSD 5.2.1, 3.0Ghz Pentium-4 CPU, 1GB memory). The deterministic algorithms SPE and BESTPATH are approximately 3 to 5 times faster than the probabilistic PROPATH algorithms. A naïve expectation is that the running time would scale as the number of probabilistic replicates sampled for the PROPATH algorithms; the difference is likely due to initialization overhead common to the probabilistic and deterministic algorithms. The SUMPATH algorithm, although deterministic, requires iterations for convergence. Thus, it is much slower than the other deterministic algorithms by about three times. The same algorithms implemented in C run an order of magnitude faster or more than those implemented in Perl, depending on the size of the network, but the relative timings of the algorithms are similar.

4 Discussion and Conclusion

We have introduced novel algorithms for predicting additional members of a protein complex based on knowledge of a subset of known members and access to a database of confidence-weighted protein-protein interactions. These algorithms have been tested against one another, and with related algorithms described previously in the literature. Important future work is to benchmark these algorithms against other methods that predict process-specific networks [31] or model the dynamical structure of protein complexes [32, 33]. Such comparisons will require standardized data sets and performance criteria [34].

The best-performing algorithms overall, PROPATH-EXP and PROPATH-ALG, share two distinctive characteristics. First, they rely on probabilistic sampling of protein interaction networks based on the confidence weights. Second, they use a distance measure, rather than the mere existence of a path, to rank potential complex members. A deterministic algorithm that performs almost as well in this test, BESTPATH, uses a greedy approach to identify the single path with greatest probability, but does not explicitly consider the length of a path. We attempted to improve the performance of BESTPATH by incorporating multiple paths. The resulting SUMPATH algorithm performed worse, however. The BESTPATH algorithm has an additional speed advantage over all other algorithms tested, excepting the poorly performing SPE method, which ignores confidence weights.

An important conclusion of this work is that algorithms may be sensitive to the meaning of an edge, in particular whether it represents a direct physical interaction or a more general functional association (such as co-membership in a protein complex). The PRONET algorithm, which was developed specifically for inference based on a network of direct interactions, indeed performs less well beyond its intended range. Other algorithms, including BESTPATH and PROPATH, appear more robust to the inclusion of indirect interaction edges. Quantitative measures of performance can depend on the examples used for testing; we find that some complexes are consistently recovered better than others regardless of algorithm or network edges.

While BESTPATH performed nearly as well as PROPATH-EXP and PROPATH-ALG in this test, we anticipate that the performance of BESTPATH will degrade in networks with many interaction edges having a weight close to 1, which should happen increasingly often as individual interactions are experimentally validated. As the number of high-weight edges increases, the BESTPATH algorithm will necessarily return an increasingly large fraction of proteins in the network. In this regime, however, the probabilistic PROPATH-EXP and PROPATH-ALG algorithms that explicitly consider the length of a high-confidence path should continue to give good performance.

Acknowledgements. JSB acknowledges support from NIH/NIGMS 1R01GM067761-01, NIH/NCRR U54RR020839, NSF 0546446, and the Whitaker Foundation. FPR acknowledges support from the Keck Foundation and NIH grants HG0017115, HG003224, & HL81341. LVZ was supported by Fu, Ryan, and AAUW Fellowships.

References

1. Uetz, P., Giot, L., Cagney, G., Mansfield, T.A., Judson, R.S., Knight, J.R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., Rothberg, J.M.: A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae. Nature 403, 623–627 (2000)
2. Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., Sakaki, Y.: A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proc Natl Acad Sci. USA 98, 4569–4574 (2001)
3. Gavin, A.C., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J.M., Michon, A.M., Cruciat, C.M., Remor, M., Hofert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M.A., Copley, R.R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., Superti-Furga, G.: Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature 415, 141–147 (2002)
4. Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S.L., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A.R., Sassi, H., Nielsen, P.A., Rasmussen, K.J., Andersen, J.R., Johansen, L.E., Hansen, L.H., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B.D., Matthiesen, J., Hendrickson, R.C., Gleeson, F., Pawson, T., Moran, M.F., Durocher, D., Mann, M., Hogue, C.W., Figeys, D., Tyers, M.: Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415, 180–183 (2002)
5. Deane, C.M., Salwinski, L., Xenarios, I., Eisenberg, D.: Protein interactions: two methods for assessment of the reliability of high throughput observations. Mol. Cell Proteomics 1, 349–356 (2002)
6. von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S.G., Fields, S., Bork, P.: Comparative assessment of large-scale data sets of protein-protein interactions. Nature 417, 399–403 (2002)
7. Li, S., Armstrong, C.M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.O., Han, J.D., Chesneau, A., Hao, T., Goldberg, D.S., Li, N., Martinez, M., Rual, J.F., Lamesch, P., Xu, L., Tewari, M., Wong, S.L., Zhang, L.V., Berriz, G.F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H.W., Elewa, A., Baumgartner, B., Rose, D.J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S.E., Saxton, W.M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K.C., Harper, J.W., Cusick, M.E., Roth, F.P., Hill, D.E., Vidal, M.: A map of the interactome network of the metazoan C. elegans. Science 303, 540–543 (2004)
8. Giot, L., Bader, J.S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y.L., Ooi, C.E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C.A., Finley, R.L., White Jr., K.P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R.A., McKenna, M.P., Chant, J., Rothberg, J.M.: A Protein Interaction Map of Drosophila melanogaster. Science 302, 1727–1736 (2003)

9. Asthana, S., King, O.D., Gibbons, F.D., Roth, F.P.: Predicting Protein Complex Membership Using Probabilistic Network Reliability. Genome Res. 14, 1170–1175 (2004)
10. Bader, J.S., Chaudhuri, A., Rothberg, J.M., Chant, J.: Gaining confidence in high-throughput protein interaction networks. Nat. Biotechnol. 22, 78–85 (2004)
11. Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M., Ideker, T.: Conserved patterns of protein interaction in multiple species. Proc Natl Acad Sci. USA 102, 1974–1979 (2005)
12. Zhang, L.V., Wong, S.L., King, O.D., Roth, F.P.: Predicting co-complexed protein pairs using genomic and proteomic data integration. BMC Bioinformatics 5, 38 (2004)
13. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O., Eisenberg, D.: A combined algorithm for genome-wide prediction of protein function. Nature 402, 83–86 (1999)
14. Goldberg, D.S., Roth, F.P.: Assessing experimentally derived interactions in a small world. Proc Natl Acad Sci. USA 100, 4372–4376 (2003)
15. Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N.J., Chung, S., Emili, A., Snyder, M., Greenblatt, J.F., Gerstein, M.: A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science 302, 449–453 (2003)
16. Troyanskaya, O.G., Dolinski, K., Owen, A.B., Altman, R.B., Botstein, D.: A Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). Proc Natl Acad Sci. USA 100, 8348–8353 (2003)
17. von Mering, C., Huynen, M., Jaeggi, D., Schmidt, S., Bork, P., Snel, B.: STRING: a database of predicted functional associations between proteins. Nucleic Acids Res. 31, 258–261 (2003)
18. Walhout, A.J., Sordella, R., Lu, X., Hartley, J.L., Temple, G.F., Brasch, M.A., Thierry-Mieg, N., Vidal, M.: Protein interaction mapping in C elegans using proteins involved in vulval development. Science 287, 116–122 (2000)
19. von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B., Bork, P.: STRING 7–recent developments in the integration and prediction of protein interactions. Nucleic Acids Res. 35, D358–362 (2007)
20. Lee, I., Date, S.V., Adai, A.T., Marcotte, E.M.: A probabilistic functional network of yeast genes. Science 306, 1555–1558 (2004)
21. Letovsky, S., Kasif, S.: Predicting protein function from protein/protein interaction data: a probabilistic approach. Bioinformatics 19, 197–204 (2003)
22. Bader, J.S.: Greedily building protein networks with confidence. Bioinformatics 19, 1869–1874 (2003)
23. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M., Sherlock, G.: Gene ontology: tool for the unification of biology. The. Gene Ontology Consortium. Nat. Genet. 25, 25–29 (2000)
24. Bader, G.D., Hogue, C.W.: An automated method for finding molecular complexes in large protein interaction networks. BMC Bioinformatics 4, 2 (2003)
25. Spirin, V., Mirny, L.A.: Protein complexes and functional modules in molecular networks. Proc Natl Acad Sci. USA 100, 12123–12128 (2003)
26. Mewes, H.W., Frishman, D., Guldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Munsterkotter, M., Rudd, S., Weil, B.: MIPS: a database for genomes and protein sequences. Nucleic Acids Res. 30, 31–34 (2002)
27. Wu, F.Y.: The Potts model. Reviews of Modern Physics 54, 235–268 (1982)
28. Chandler, D.: Introduction to modern statistical mechanics. Oxford University Press, New York (1987)

29. Vazquez, A., Flammini, A., Maritan, A., Vespignani, A.: Global protein function prediction from protein-protein interaction networks. Nat. Biotechnol. 21, 697–700 (2003)
30. Leone, M., Pagnani, A.: Predicting protein functions with message passing algorithms. Bioinformatics 21, 239–247 (2005)
31. Myers, C.L., Robson, D., Wible, A., Hibbs, M.A., Chiriac, C., Theesfeld, C.L., Dolinski, K., Troyanskaya, O.G.: Discovery of biological networks from diverse functional genomic data. Genome Biol. 6, R114 (2005)
32. Scholtens, D., Gentleman, R.: Making sense of high-throughput protein-protein interaction data. Stat Appl Genet Mol Biol. 3 Article39 (2004)
33. Scholtens, D., Vidal, M., Gentleman, R.: Local modeling of global interactome networks. Bioinformatics 21, 3548–3557 (2005)
34. Califano, A., Stolovitzky, G.: DREAM Project. http://magnet.c2b2.columbia.edu/news/DREAMInitiative.pdf