BIOINFORMATICS

Vol. 19 Suppl. 1 2003, pages i197–i204
DOI: 10.1093/bioinformatics/btg1026

Predicting protein function from protein/protein interaction data: a probabilistic approach

Stanley Letovsky and Simon Kasif*

Bioinformatics Program and Department of Biomedical Engineering, Boston University, 44 Cummington St., Boston, MA 02215, USA

Received on January 6, 2003; accepted on February 20, 2003

ABSTRACT

Motivation: The development of experimental methods for genome scale analysis of molecular interaction networks has made possible new approaches to inferring protein function. This paper describes a method of assigning functions based on a probabilistic analysis of graph neighborhoods in a protein-protein interaction network. The method exploits the fact that graph neighbors are more likely to share functions than nodes which are not neighbors. A binomial model of local neighbor function labeling probability is combined with a Markov random field propagation algorithm to assign function probabilities for proteins in the network.

Results: We applied the method to a protein-protein interaction dataset for the yeast *Saccharomyces cerevisiae* using the Gene Ontology (GO) terms as function labels. The method reconstructed known GO term assignments with high precision, and produced putative GO assignments to 320 proteins that currently lack GO annotation, which represents about 10% of the unlabeled proteins in *S. cerevisiae.*

Availability: Source code available upon request. Results available at http://genomics10.bu.edu/netmark.

Contact: sletovsky@aol.com

Keywords: protein–protein interaction, protein function prediction, gene ontology, Markov Random fields

INTRODUCTION

Since the first complete genome was sequenced in 1995, more than eighty microbial organisms and close to a dozen eukaryotic genomes have been sequenced. A critical problem in making sense of these genomes is the assignment of functional roles to newly discovered proteins. The primary tools for first pass function assignment, such as BLAST (Altschul *et al.*, 1990), are based on sequence similarity: they assign a function to a novel protein by propagating functional information from a similar protein of known function. This approach fails for the roughly 20-40% of proteins in newly sequenced genomes–many of them known only from *de novo* gene prediction–that do not have statistically significant sequence similarity to functionally annotated proteins. In addition, the transfer of functional assignment between proteins with low sequence identity (below 40%) is prone to significant error.

In recent years, high-throughput functional genomics techniques such as expression profiling and protein interaction mapping have generated new datasets that provide additional opportunities for inference of function. Newer computational methods for inferring protein function include analysis of gene fusion events (Enright *et al.*, 1999; Marcotte *et al.*, 1999; Yanai *et al.*, 2001); phylogenetically conserved linkage patterns (Overbeek, 1999; Yanai *et al.*, 2002; Zheng *et al.*, 2002) (sometimes called operon analysis); phylogenetic profiling, which looks at sharing of protein sets across organisms (Gaasterland and Ragan, 1998; Pellegrini *et al.*, 1999), and analysis of measurements of gene expression to identify genes that have similar expression patterns, which provides evidence of co-regulation and hence possible shared function (Ulanovsky *et al.*, 2002; Zhou *et al.*, 2002). Such methods have been greatly aided by the standardization of protein function descriptions in controlled vocabularies such as the GO hierarchies (Ashburner *et al.*, 2000), and by the production of carefully curated collections of protein annotations using those controlled vocabularies (Dwight *et al.*, 2002).

The method described here makes protein function predictions by analyzing networks of protein-protein interactions (PPI). We used a PPI dataset compiled and kindly provided to us by the GRID (Breitkreutz *et al.*, 2002) project. This dataset contains interactions from a number of published papers (Schwikowski *et al.*, 2000; Ito *et al.*, 2001a; Tong *et al.*, 2001; Gavin *et al.*, 2002; Ho *et al.*, 2002), as well as from the MIPS (Mewes *et al.*, 2002) and BIND (Bader *et al.*, 2001) databases. Evidence for the interactions was generated using a variety of methods, including the yeast two hybrid method (59% of interactions), affinity precipitation (34%), and synthetic lethality (5%). The yeast 2-hybrid method has been

*To whom correspondence should be addressed.

popular recently because it can be scaled up, but it can produce large numbers of both false positive and false negative interactions; any method which predicts function from these data must be robust in the presence of such errors. The problem of inferring function from PPI has previously been addressed by (Schwikowski *et al.*, 2000), in which a function was assigned to a protein based on the majority of functional labels of its interacting partners.

The current work attempts to provide a more robust probabilistic solution using a Markov Random Field (MRF) formalism. Markov Random Fields have been widely used in image analysis (Geman and Geman, 1984) for image restoration and segmentation. Our problem is reminiscent of segmentation, in that we wish to segment the protein-interaction network into subgraphs that share similar labels.

METHODS

We represent the evidence from PPI data using a graphical formalism called a *functional linkage graph* (Marcotte *et al.*, 1999; Yanai *et al.*, 2001), in which an edge (link) between two nodes (proteins) represents evidence that they might share the same function. The translation of PPI data into a functional linkage graph is straightforward: pairwise interactions become edges in the graph. If there are multiple pieces of evidence bearing on the same pairwise interaction they are combined into a single link. With each pairing of a protein i and a GO term t, we associate a Boolean random variable $L_{i,t}$ which is 1 if i is labeled with t, and 0 if it is not. This allows us to accommodate multiple labels for the same protein, which frequently arise, both as a result of the hierarchical nature of some controlled vocabularies, in which a label implies additional 'ancestral' labels, and also because proteins often carry out multiple functions.

The problem we then want to solve is to derive the marginal probability of a given protein taking a particular functional label given all the putative functional assignments to the other proteins in the graph. The Markov Random Field formulation provides a sound solution to this problem, subject to a conditional independence (Markov) assumption that states that probability distribution for the labeling of any node is conditionally independent of all other nodes given its neighbors. In a pairwise MRF, the label probability of a node is the product of node-specific (e.g. sensory) evidence about the node's state with pairwise joint probabilities with its neighbors (Yedidia *et al.*, 2001). In this application we have no node-specific evidence regarding the labeling of unlabeled nodes, so a node's label probability is entirely a function of its neighbors' states.

NEIGHBORHOOD FUNCTION

The MRF framework requires the specification of neighborhood functions that describe the dependence of the label probability of a node on the labels of its neighbors. Different types of neighborhood conditional probability functions can be used to model different types of local dependency structure. Our algorithm relies on the statistical property of *local density enrichment*: i.e. proteins with a particular label are more likely to have neighbors carrying that same label than proteins lacking the label. This property is not true for all terms, and randomizing the assignment of labels to proteins destroys it.

Figure 1 illustrates the variation in density enrichment across terms used in our example dataset. We will denote the pair of y-values associated with each term in this plot by p_1 (dark circles), corresponding to the probability that the target of an edge has a given label given that the source has this label, and p_0 (light triangles), corresponding to the probability that target protein has the label given that the source has some other label. The plot shows that for many terms there is a significantly enhanced probability of similar labels in the neighborhood of a labeled protein beyond what term frequency would predict. Our algorithm exploits this difference between p_0 and p_1 to make predictions.

We are interested in computing the probability that protein i has (or should have) label t, for all combinations of proteins and terms. We define our neighborhood function $p(L_{i,t})$ to be a function of N_i, the number of graph neighbors of i, and $k_{i,t}$, the number of those neighbors which are labeled with term t. We will denote this probability $p(L_{i,t} = 1|N_i, k_{i,t})$.[†] Applying Bayes' rule, and making an independence assumption[‡] we obtain:

$$p(L|N,k) = \frac{p(k|L,N) \cdot p(L)}{p(k|N)} \qquad (1)$$

where:

- $p(k|L,N)$is the probability of having k t-labeled neighbors out of N neighbors. If labels were randomly assigned to proteins we would expect $p(k|L,N)$ to follow a binomial[§] distribution. That is,

$$p(k|N) = B(N,k,f_t)$$

[†] Henceforth we will drop the subscripts except where needed for clarity, and for conciseness we will use L, $\overline{L}$ and $\overline{p}$ in place of $L=1$, $L=0$, and $(1-p)$ respectively.

[‡] $p(L|N) = p(L) \cdot p(N)$. This assumption states that the degree (# of neighbors) distribution of nodes labeled with t is not significantly different from the degree distribution overall. Although the former distribution may be poorly resolved for infrequent terms, the assumption is supported by visual inspection of the degree distribution of many terms.

[§] We could also use a hypergeometric distribution; the choice depends on whether we model the assignment as with or without replacement. Here we assume replacement and use binomial distributions; the difference between the two is small when the number of proteins is large.

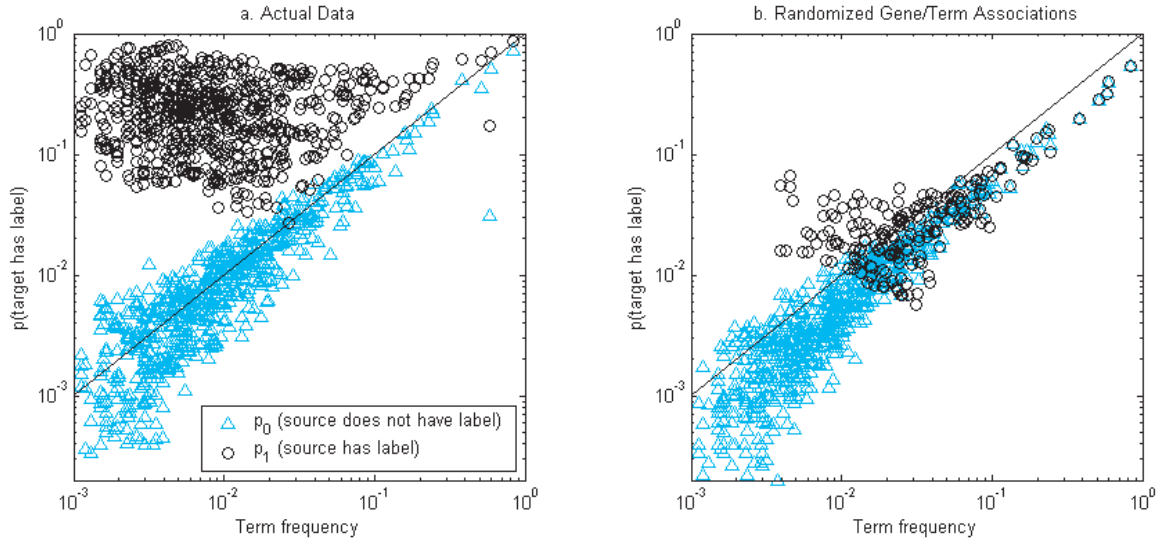

Fig. 1. For each term t we estimate two probabilities based on the set of graph edges $\{< i, j >\}$ where the labeling status of both i and j with respect to t is known. p_1 is the probability that the target node j is labeled with t given that source node i is also labeled with t; p_0 is the probability that j is labeled with t given that i is not. Both are shown as a function of term frequency on a log/log scale for a set of 689 terms with sufficient counts to estimate p_0 and p_1. (a) Actual data; (b) after shuffling rows (proteins) in the Protein By Term labeling matrix. (Shuffling rows preserves correlations between ISA-related terms). Values of p_0 are similar to the term frequency in both plots, suggesting that edges between labeled and unlabeled nodes occur at frequencies close to chance expectation. Values of p_1 are strikingly different, however; these are significantly higher in the actual data, while in the shuffled data they resemble p_0 values. This shows that the labeling structure of the actual graph is far from random. It also provides global evidence that there is significant information content in the PPI data which is correlated with known functions; i.e. the PPI data cannot be entirely noise.

where

$$B(N, k, p) = \binom{N}{k} p^k \overline{p}^{N-k}$$

and f_t is the frequency of occurrence of term t in the graph. If $p_o \neq p_1$ the label probability of a protein's neighbors will vary depending on its own label, and thus we expect neighbors of t and non-t proteins to have different conditional distributions:

$$p(k|\overline{L}, N) = B(N, k, p_0)$$

and

$$p(k|L, N) = B(N, k, p_1)$$

The latter term is used in the numerator of Equation (1).[¶]

- $p(L) = f$, the frequency of term t in the graph.
- $p(k|N)$ is the frequency-weighted average of the above two binomial terms:

$$p(k|N) = f \cdot p(k|L, N) + \overline{f} \cdot p(k|\overline{L}, N)$$

[¶] Note that we apply a conservative correction to our estimated values of p_0 and p_1 by adjusting them upward, and downward, respectively, to the edge of their 95% confidence intervals.

Putting all of these together gives the neighborhood function:

$$p(L|N, k) = \frac{f \cdot B(N, k, p_1)}{f \cdot B(N, k, p_1) + \overline{f} \cdot B(N, k, p_0)}$$

which can be rewritten as

$$p(L|N, k) = \frac{\lambda}{1 + \lambda}$$

where

$$\lambda = \frac{f \cdot B(N, k, p_1)}{\overline{f} \cdot B(N, k, p_0)}$$

is the likelihood ratio. This form makes it clear that the $\binom{N}{k}$ term in the binomial formula cancels out, giving

$$\lambda = \frac{f p_1^k \overline{p_1}^{N-k}}{\overline{f} p_0^k \overline{p_0}^{N-k}}$$

PROPAGATION OF PROBABILITIES

Since the label probability of a protein depends on its neighbors, which depend in turn on *their* neighbors, we would like a rigorous method of increasing our estimate

of the label probability of a protein if our estimate of its unlabeled neighbors' label probability increases. The Markov random field inference problem that corresponds to this is the compution of the marginal label probabilities of the unlabeled (hidden) nodes given some labeled (fully observable) nodes. While this problem is NP-hard in general even for grid topologies, a number of practical procedures exist that take advantage of the independence assumptions, including the junction tree algorithm, Gibbs sampling and belief propagation (Pearl, 1991; Yedidia *et al.*, 2001). Belief propagation is not guaranteed to converge on graphs with cycles, or to give even approximately correct results if does converge, although in practice it frequently does converge to approximately correct marginal probabilities. In this paper we use a simplified heuristic variant of belief propagation which is designed to ensure that adjacent nodes cannot mutually reinforce their estimated probabilities in a runaway fashion.

Our propagation algorithm is based on iterative application of equation 1, with k replaced by $E(k)$, the expected number of labeled nodes given their current estimated labeling probabilities, which is just the sum of those probabilities. Unlabeled nodes i are initialized to $p(L_{i,t}) = f_t$; probabilities of labeled nodes are clamped to the appropriate Boolean values throughout. In the first iteration, the estimated probabilities of unlabeled nodes are adjusted in parallel using Equation (1) to reflect their immediate neighborhood; in this step all unlabeled neighbors are still seen as having the label probability f_t. On the second iteration unlabeled nodes now see adjusted probabilities of their unlabeled neighbors, but those probabilities are based on initial values of *their* unlabeled neighbors. Only on the third iteration would influence propagate from unlabeled node i to its unlabeled neighbor j and back to i again, raising the possibility of invalid runaway self-reinforcement. To avoid this, we stop after the second iteration; apply a threshold (we use .8), and reclassify any unlabeled node whose labeling probability exceeds the threshold as now labeled. We then repeat the entire process, stopping when no further labeling occurs. This algorithm is applied to each term separately.

DATA SOURCES

The following datasets were used in our analysis:

Protein–Protein Interactions: The GRID dataset contained 20 985 distinct interactions catalogued between 13 607 distinct pairs of proteins. 4708 proteins participated in interactions. 4588 of these are in a single connected component, the second largest component has 4 proteins. 1442 unlabeled, connected proteins were potential labeling targets.

Yeast GO Labelings: 26 551 labelings of 6904 ORFs (including tRNAs and other nonprotein-coding ORFs) were taken from 12/1/02 version of SGD Yeast GO assignments. After merging on ORF name, the overlap between this ORF set and the GRID data consisted of 4692 proteins. 3267 of these had nontrivial labels (i.e. excluding 'unknown cellular compartment', 'unknown molecular function', 'unknown biological process') in at least one of the 3 GO hierarchies; however 2573 proteins were unlabeled in at least one of the three GO hierarchies and hence were candidates for labeling by our method. After expansion of protein labeling to include all ISA ancestors of each label in the GO hierarchy, 1951 GO terms were used as labels. The number of terms useful for labeling was further reduced by application of several filters. A term was excluded if it labeled exactly the same set of terms as a more specific term. A term was excluded if there were no links between proteins labeled with the term and other labeled proteins, so that the term-specific parameters p_0 and p_1 could not be estimated. A χ^2 test was used to verify that p_0 and p_1 were sufficiently different, by testing for non-independence of the 2×2 contingency table of source label versus target label, using a Bonferroni-corrected p-value of $0.001/T$, where T is the number of terms tested. There had to be edges between term-labeled and unlabeled[‖] proteins for propagation to operate. Finally, terms that occurred more than 300 times as known labels were eliminated; these high-frequency terms tended to be broad terms high in the GO hierarchy and of little predictive value, such as *metabolism* or *cell growth and maintenance*. These filters left 669 terms which were used in the analysis.

RESULTS

We implemented the above algorithm in MatLab and ran it on the above datasets, which took about 6 hours on a 1 GHz CPU. In order to generate predictions, the final inferred label probabilities must be thresholded at some cutoff. We determined a precision-optimizing cutoff for each term as part of our validation process, described below. The algorithm then produced 702 predictions for unlabeled proteins; 455 (65%) during the initialization phase and 247 during propagation. 404 of the predictions were ISA-minimal, i.e. not superterms of more specific terms predicted for the same protein. The full set of predictions is available on the web.

In order to assess the likely error rate in these predictions we first investigated the ability of the algorithm to reconstruct known labels. This is easier to do for the initialization step than for the propagation step, so we address these separately.

[‖] A protein was considered unlabeled with respect to a term if it did not contain any labels in the same GO hierarchy, other than ISA ancestors of the term.

In the initialization step we computed labeling probabilities for all proteins with respect to all terms. At a prediction threshold of $p(L) > 0.8$ overall precision was 85%, recall was 34%, and false positive rate was 0.15%. We also measured term-specific precisions as a function of a sliding prediction threshold. For every term we determined the maximum precision, and the threshold at which the precision first exceeded 80%. Terms which never attained 80% precision were culled from further consideration, leaving 228 terms with greater than 80% precision out of the 669 analyzed. For these the threshold value at which 80% precision was first attained was used as the prediction threshold.

Assessing precision of propagation is more difficult; we cannot reconstruct known labels unless we pretend that known proteins are unlabeled, in which case we change the information available to the algorithm, and hence the results. We therefore applied a jackknife procedure in which, for each term which generated predictions, we censored the labels of only 6 nodes of known label: 3 with the term, and 3 with a different term, each chosen at random. The number of perturbed nodes was kept small to minimize the disturbance to the results; the process was repeated 3 times for each term. From these data we estimate that the method had 98.6% precision and 21% recall using a threshold of p>.8; the false positive rate in that range is 0.3%. The actual error rate will be different because in practice the number of positive and negative nodes presented to the algorithm are not equal; the number of nodes which should not have a given label typically greatly exceed the number which should. After correcting for expected label frequency we estimated the prediction error rate at 71%, or 287 of the 404 ISA-minimal predictions are expected to be correct. This process corrects for multiple comparisons much more efficiently than Bonferroni.

Finally we assessed the plausibility of the predictions by direct examination. Looking only at predictions for proteins having a description line in SGD, we assigned each prediction a plausibility rating, as shown in Table 1. A rating of 2 is a Presumed True Positive; by which we mean that the assigned term was directly relevant to the description, often using some of the same keywords. This does not mean that the prediction is correct; for example if the protein description is *ubiquitin-like protein*, and the algorithm predicted the cellular location *nuclear ubiquitin-ligase complex*, that predicted location could be wrong, but it is highly relevant. A rating of −1, or Presumed False Positive, was assigned to predictions where no relationship could be established between the description and the assigned term, such as between the PDC6 with description *pyruvate decarboxylase isozyme* and the term *ubiquitin ligase complex;* as far as the authors admittedly shallow understanding of biology extends, these would seem to have nothing to do with each other. A rating of 0 (not shown) was applied if there was insufficient information to make a decision. Finally, and most interestingly, a rating of 1, for Plausible Prediction, was assigned to predictions if a search of PubMed retrieved one or more papers suggesting a relationship between keywords in the description and those in the term.

Examples of this latter category include assigning NHP1, an HMG1-box containing protein, to the GO molecular function category *chromatin binding* and the cellular component category *chromatin remodeling complex*. HGM1 proteins have been hypothesized (Wisniewski *et al.*, 1999) to have a role in chromatin structure. Another example is the SRO77 protein, which is described as a yeast homolog of the *Drosophila* tumor suppressor *lethal giant larvae* (see entry for l(2)gl in Flybase, The Flybase Consortium 2003) was assigned the molecular function *Motor*. (Asaba *et al.*, 2003) reported that some mammalian tumor suppressors interact with a kinesin-related motor; moreover the *Drosophila* homolog is described as interacting with TNFβ and myosin. MTH1 was predicted to be part of a transcription factor complex and to be a transcriptional regulator; its description is 'negative regulator of HXT gene expression'. UTP20, a U3 snoRNP protein, was predicted to be a structural constituent of the ribosome. A recent paper (Culver, 2002) finds that U3 SnoRNPs are attached to the 5′ ends of pre-rRNAs. UFO1, described as an F-box protein, was assigned to the cellular component *ubiquitin ligase complex*; SGD had already assigned it the molecular function *ubiquitin-protein ligase*. (Note that the latter assignment did not in any way contribute to the prediction, since the two terms are in different hierarchies.) BLM3 is described as being involved in protecting the cell against bleomycin damage; our algorithm assigned it the function *proteasome endopeptidase*. Ustrell et al. describe a link between disruption of a proteasome activator and bleomycin hypersensitivity. Asc1, described as a Gβ like protein, was assigned the function *N-acetyltransferase*. (Chetsawang *et al.*, 1999) reported that opioid receptors, a class of G-protein coupled receptors that include G-beta subunits, have a stimulatory effect on N-acetyltransferase. CYM1, described as a metalloprotease, was assigned the process pyruvate metabolism and the cellular component pyruvate dehydrogenase complex; (Opalka *et al.*, 2002) describe inhibition of pyruvate metabolism by matrix metalloproteinase inhibitors. These and other examples, along with references supporting the plausibility of the connection, are shown in Table 1.

DISCUSSION

The performance of the algorithm was surprisingly good at reconstructing known labels; however, more experience

Table 1. Sample Predictions. A rating of 2 is a presumed true positive, 1 is plausible prediction, −1 is a presumed false positive. Highlighted predictions were generated in the propagation phase. Type is *f* for functional role, *p* for biological process, and *c* for cellular component

Rating	ORF	Gene	Description	Term	Type
2	YBL004W	UTP20	U3 snoRNP protein	RNA binding	f
2	YCL010C	SGF29	Probable 29kKDa Subunit of SAGA histone acetyltransferase complex	transcription regulator	f
2	YDR139C	RUB1	ubiquitin-like protein	nuclear ubiquitin ligase complex	c
2	YGL145W	TIP20	transport protein that interacts with Sec20p; required for protein transport from the endoplasmic reticulum to the golgi apparatus	intracellular transporter	f
2	YGR232W	NAS6	26S proteasome interacting protein	proteasome endopeptidase	f
2	YJL069C	UTP18	U3 snoRNP protein, U3 snoRNA associated protein	RNA binding	f
2	YKR068C	BET3	transport protein particle (TRAPP) component	intracellular transporter	f
2	YNL110C	NOP15	ribosome biogenesis	RNA binding	f
2	YLR306W	UBC12	ubiquitin-conjugating enzyme	nuclear ubiquitin ligase complex	c
2	YPL151C	PRP46	pre-mRNA splicing factor	RNA binding	f
2	YPR066W	UBA3	ubiquitin-like protein activating enzyme	nuclear ubiquitin ligase complex	c
2	YPR101W	SNT309	protein complex component associated with the splicing factor Prp19p	RNA binding	f
1	YBL004W	UTP20	U3 snoRNP protein	structural constituent of ribosome (Culver 2002)	f
1	YBL106C	SRO77	yeast homolog of the *Drosophila* tumor suppressor, lethal giant larvae	Motor (Asaba *et al.*, 2003)	f
1	YDL002C	NHP10	HMG1-box containing protein	chromatin remodeling complex (Wisniewski *et al.*, 1999)	c
1	YDR091C	RLI1	ATP-binding cassette (ABC) superfamily nontransporter group (putative)	translation initiation factor	f
1	YDR277C	MTH1	Msn3p homolog (61% identical)	transcription factor complex	c
1	YDR277C	MTH1	Msn3p homolog (61% identical)	transcriptional activator	f
1	YDR430C	CYM1	Metalloprotease	pyruvate dehydrogenase complex (Opalka *et al.*, 2002)	c
1	YDR430C	CYM1	Metalloprotease	pyruvate metabolism (Opalka *et al.*, 2002)	p
1	YER173W	RAD24	cell cycle exonuclease (putative)	DNA clamp loader	f
1	YFL007W	BLM3	involved in protecting the cell against bleomycin damage	proteasome endopeptidase (Ustrell *et al.*, 2002)	f
1	YJL044C	GYP6	GTPase activating protein (GAP) for Ypt6	TRAPP	c
1	YML088W	UFO1	F-box protein	nuclear ubiquitin ligase complex	c
1	YMR116C	ASC1	G-beta like protein	N-acetyltransferase (Chetsawang *et al.*, 1999)	f
1	YOR243C	PUS7	pseudouridylate U2 snRNA at position 35	structural constituent of ribosome	f
1	YDL002C	NHP10	HMG1-box containing protein	chromatin binding	f
−1	YBL066C	SEF1	transcription factor (putative)	ribonuclease P	f
−1	YBR272C	HSM3	MutS family (putative)	proteasome endopeptidase	f
−1	YDR400W	URH1	uridine nucleosidase (uridine ribohydrolase); EC 3.2.2.3	signalosome complex	c
−1	YGR087C	PDC6	pyruvate decarboxylase isozyme	ubiquitin ligase complex	c
−1	YJL057C	IKS1	serine/threonine kinase (putative)	nuclear pore	c
−1	YJR082C	EAF6	Subunit of the NuA4 complex	intracellular protein transport	p
−1	YNL135C	FPR1	peptidyl-prolyl cis-trans isomerase (PPIase)	negative regulation of transcription from Pol II promoter, mitotic	p
−1	YOL044W	PEX15	44 kDa phosphorylated integral peroxisomal membrane protein	Nuclease	f
−1	YOR279C	RFM1	DNA-binding protein	dolichyl-diphospho-oligosaccharide-protein glycosyltransferase	f
−1	YOR279C	RFM1	DNA-binding protein	N-linked glycosylation	p
−1	YBR188C	NTC20	splicing factor	transcriptional activator	f
−1	YDL049C	KNH1	KRE9 homolog	structural constituent of ribosome	f
−1	YKR068C	BET3	transport protein particle (TRAPP) component	cation transporter	f
−1	YLL036C	PRP19	RNA splicing factor	cation transporter	f
−1	YLL036C	PRP19	RNA splicing factor	transcriptional activator	f
−1	YLR037C	DAN2	putative cell wall protein	RNA binding	f
−1	YLR117C	CLF1	pre-mRNA splicing factor	cation transporter	f
−1	YLR117C	CLF1	pre-mRNA splicing factor	transcriptional activator	f
−1	YML077W	BET5	TRAPP 18kDa component	cation transporter	f
−1	YMR213W	CEF1	protein complex component associated with the splicing factor Prp19p	cation transporter	f

will be needed to assess its success rate for novel predictions.

The design of any label propagation algorithm must address a number of issues. For example, nodes (proteins) may have multiple labels; does one want to assume that the presence of one label is evidence for the absence of others? Our approach treats every term as a separate binary Markov Random Field, which allows multiple labels to be inferred for the same protein. We do, however, assume that if a protein has a label, we can treat it as a negative for all other labels in the same hierarchy except sub- and superterms, the assumption being that its function was well enough characterized that absence of other labels can be interpreted as evidence against them. We are currently experimenting with a generalization of the system described here which will allow a labeling probability $p(L_{i,t})$ to influence the probability $p(L_{j,u})$ of another term u across an edge $\langle i, j \rangle$. This will allow terms to be preferentially adjacent to other terms representing upstream or downstream pathways; we have in fact identified such correlations in the data.

A related issue is whether an algorithm can handle the sort of hierarchical controlled vocabularies which have been popularized by the GO consortium. We chose to work with the ISA-transitive closure of the yeast GO labeling, that is, if a protein was labeled with term u, we also labeled it with all terms v where v is an ancestor of u in the GO ISA-hierarchy. Consequently very broad, general terms had very high frequencies, and if a term did not have enough density in the graph to support label propagation, it may be that one of its broader terms did.

Since terms have different frequencies, the likelihood of having labeled terms varies with the term frequency. Algorithms that use simple heuristics such as majority voting by neighbors (Schwikowski *et al.*, 2000) or thresholding the fraction of neighbors with a given label are vulnerable to being either too conservative or too liberal in their predictions as the term frequency varies. On the other hand we make the assumption that a term's frequency among labeled proteins will be predictive of its frequency among unlabeled proteins. This may not be the case, e.g. if the term describes a well-studied pathway most of its members may already be labeled, and conversely for a poorly studied one. Where the assumption of equal frequencies is wrong our predictions may run astray. Other biases in the data, such as nonrandom sampling of interactions or ascertainment biases such as are described in (Bader and Hogue, 2002), may also lead to errors.

Another issue is the fact that not all terms are equally predictive. Our algorithm allows each term's predictiveness to be determined separately. We hope that by explicitly taking frequency and predictiveness into account we will achieve greater robustness in the face of 'edge noise' in the PPI data.

Our approach raises some interesting general questions about the relationship between function labels and network representations of the biology. Should there be some sort of structural constraint on a functional label in a graph, and if so what? Perhaps in an ideal wiring diagram, terms would label connected subgraphs[**]. This is unlikely to be the case for many GO terms, including many that failed our precision filter, such as *chaperone, signal transducer*, and *protein folding*. In the present study we are taking functional labels defined by historical and often poorly articulated criteria and assessing their coherence with respect to an independent PPI dataset. In the future, as the biological networks are resolved with greater precision, it may make sense to define functions explicitly as connected subgraphs of the network.

This paper presents a particular choice of neighborhood function based on the binomial distribution, but the question of which functions are optimal for different types of biological evidence is an important topic for further investigation. We believe the MRF framework will be general enough to support a variety of different neighborhood functions, and that different neighborhood functions may be appropriate for different types of evidence. In the future we plan to explore the application of the MRF framework to different types of evidence beyond PPI, such as gene expression, sequence similarity, biochemical flux networks, and protein-DNA interactions; to incorporate degrees of belief in the evidence, such as edge probabilities reflecting quality of the PPI data; and to compare different propagation algorithms such as BP and Gibbs sampling to our current approach.

ACKNOWLEDGEMENTS

Simon Kasif was supported in part by NSF (KDI grant). The authors thank T.M. Murali and Ulas Karaoz for numerous useful discussions related to analysis of protein-protein networks. The MathWorks generously provided Stan Letovsky with MatLab software.

REFERENCES

Altschul,S.F., Gish,W. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Asaba,N., Hanada,T. *et al.* (2003) Direct interaction with a Kinesin-related motor mediates transport of mammalian discs large tumor suppressor homologue in epithelial cells. *J. Biol. Chem.*, **278**, 8395–8400.

[**]For our algorithm to be applicable to a term we require $p_1 >> p_0$; this turns out to be neither necessary nor sufficient to ensure that there are few connected components in the term subgraph. For example, a graph in which every labeled protein occurred only in connected components of 2, both labeled, would have very high p_1/p_0 but potentially many t-components. In practice, however, most terms with fewer than expected connected$t$-components do have $p_1 >> p_0$, although the reverse is not always true.

Ashburner,M., Ball,C.A. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.

Bader,G.D., Donaldson,I. *et al.* (2001) BIND—the Biomolecular Interaction Network Database. *Nucleic Acids Res.*, **29**, 242–245.

Bader,G.D. and Hogue,C.W. (2002) Analyzing yeast protein–protein interaction data obtained from different sources. *Nat. Biotechnol.*, **20**, 991–997.

Breitkreutz,B.J., Stark,C. *et al.* (2002) The GRID: the General Repository for Interaction Datasets. *Genome Biol.*, **3**).

Chetsawang,B., Casalotti,S.O. *et al.* (1999) Gene expressions of opioid receptors and G-proteins in pineal glands. *Biochem. Biophys. Res. Commun.*, **262**, 775–780.

Culver,G.M. (2002) Sno-capped: 5′ ends of preribosomal RNAs are decorated with a U3 SnoRNP. *Chem. Biol.*, **9**, 777–779.

Dwight,S.S., Harris,M.A. *et al.* (2002) Saccharomyces Genome Database (SGD) provides secondary gene annotation using the Gene Ontology (GO). *Nucleic Acids Res.*, **30**, 69–72.

Enright,A.J., Iliopoulos,I. *et al.* (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402**, 86–90.

Gaasterland,T. and Ragan,M.A. (1998) Microbial genescapes: phyletic and functional patterns of ORF distribution among prokaryotes. *Microb. Comp. Genomics*, **3**, 199–217.

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

Geman,S. and Geman,D. (1984) Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

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

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

Ito,T., Chiba,T. *et al.* (2001b) Exploring the protein interactome using comprehensive two-hybrid projects. *Trends Biotechnol.*, **19 (Suppl 10)**, S23–S27.

Marcotte,E.M., Pellegrini,M. *et al.* (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.

Mewes,H.W., Frishman,D. *et al.* (2002) MIPS: a database for genomes and protein sequences. *Nucleic Acids Res.*, **30**, 31–34.

Opalka,J.R., Gellerich,F.N. *et al.* (2002) Effect of the new matrix metalloproteinase inhibitor RO-28-2653 on mitochondrial function. *Biochem. Pharmacol.*, **63**, 725–732.

Overbeek,R., Fonstein,M. *et al.* (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci. USA*, **96**, 2896–2901.

Pearl,J. (1991) *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pellegrini,M., Marcotte,E.M. *et al.* (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl Acad. Sci. USA*, **96**, 4285–4288.

Schwikowski,B., Uetz,P. *et al.* (2000) A network of protein–protein interactions in yeast. *Nat. Biotechnol.*, **18**, 1257–1261.

The Flybase Consortium (2003) The FlyBase database of the *Drosophila* genome projects and community literature. *Nucleic Acids Res.*, **31**, 172–175.

The Gene Ontology Consortium (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.

Tong,A.H., Evangelista,M. *et al.* (2001) Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science*, **294**, 2364–2368.

Ulanovsky,H., Ron,S. *et al.* (2002) *ProToGO—Evaluating biological features for a set of proteins using GO annotations*, GO Users Meeting, Hinxton, UK.

Ustrell,V., Hoffman,L. *et al.* (2002) PA200, a nuclear proteasome activator involved in DNA repair. *Embo J.*, **21**, 3516–3525.

Wisniewski,J.R., Krohn,N.M. *et al.* (1999) HMG1 proteins from evolutionary distant organisms distort B-DNA conformation in similar way. *Biochim. Biophys. Acta*, **1447**, 25–34.

Yanai,I., Derti,A. *et al.* (2001) Genes linked by fusion events are generally of the same functional category: a systematic analysis of 30 microbial genomes. *Proc. Natl Acad. Sci USA*, **98**, 7940–7945.

Yanai,I., Mellor,J.C. *et al.* (2002) Identifying functional links between genes using conserved chromosomal proximity. *Trends Genet.*, **18**, 176–179.

Yedidia,J., Freeman,W.T. *et al.* (2001) *Understanding Belief Propagation and its Generalizations*, International Joint Conference on Artificial Intelligence (IJCAI)..

Zheng,Y., Roberts,R.J. *et al.* (2002) Genomic functional annotation using co-evolution profiles of gene clusters. *Genome Biol.*, **3**, RESEARCH0060.

Zhou,X., Kao,M.C. *et al.* (2002) Transitive functional annotation by shortest-path analysis of gene expression data. *Proc Natl Acad. Sci. USA*, **99**, 12783–12788.