BIOINFORMATICS ORIGINAL PAPER Vol. 22 no. 15 2006, pages 1871–1878
doi:10.1093/bioinformatics/btl187

Systems biology

Protein classification using probabilistic chain graphs and the Gene Ontology structure

Steven Carroll and Vladimir Pavlovic*

Department of Computer Science, Rutgers University, Piscataway, NJ 08854, USA

Received on February 25, 2006; revised on April 25, 2006; accepted on May 11, 2006
Advance Access publication May 16, 2006
Associate Editor: John Quackenbush

ABSTRACT

Motivation: Probabilistic graphical models have been developed in the past for the task of protein classification. In many cases, classifications obtained from the Gene Ontology have been used to validate these models. In this work we directly incorporate the structure of the Gene Ontology into the graphical representation for protein classification. We present a method in which each protein is represented by a replicate of the Gene Ontology structure, effectively modeling each protein in its own 'annotation space'. Proteins are also connected to one another according to different measures of functional similarity, after which belief propagation is run to make predictions at all ontology terms.

Results: The proposed method was evaluated on a set of 4879 proteins from the *Saccharomyces* Genome Database whose interactions were also recorded in the GRID project. Results indicate that direct utilization of the Gene Ontology improves predictive ability, outperforming traditional models that do not take advantage of dependencies among functional terms. Average increase in accuracy (precision) of positive and negative term predictions of 27.8% (2.0%) over three different similarity measures and three subontologies was observed.

Availability: C/C++/Perl implementation is available from authors upon request.

Contact: vladimir@cs.rutgers.edu

1 INTRODUCTION

Owing to the advent of high-throughput sequencing techniques, the complete sequences of several genomes are now known. However, biological function is still unknown for a large proportion of sequenced proteins. Moreover, a given protein may have more than one function, so many proteins that are known to be in some class may have as yet undiscovered functionalities.

Recently, belief networks have been utilized to infer protein functions over sets of partially annotated proteins (Deng *et al*., 2002, 2003, 2004; Letovski and Kasif, 2003). In these studies, protein–protein interaction data are used to define a Markov Random Field (MRF) topology over the full set of proteins. In these graphical models, a node represents each protein, and an interaction between two proteins is represented by an edge between the two nodes. Using partial knowledge of functional annotations of a subset of proteins, probabilistic inference on these models is used to elucidate other proteins' unknown functions.

Current graphical formulations consider only one protein functional category at a time, implying independence in annotations across multiple levels of protein description. However, functional categories are known to exhibit dependencies in a cellular context. In this work, we present a probabilistic graphical model framework that considers multiple functional categories (terms) in the Gene Ontology (GO) (Ashburner *et al*., 2000) *simultaneously*, making predictive use of the definitive and probabilistic relationships among the terms. This is possible owing to the well-defined structure of the Gene Ontology[1], a structured vocabulary of terms describing gene products. In our model, each protein is represented by its own ontology structure. Using this representation, information from annotated proteins is passed within the ontology structure as well as between neighboring proteins, leading to improved functional prediction on a set of functional terms. Furthermore, the use of negative as well as positive annotations to terms in the Gene Ontology gives our model a unique advantage over the previous studies.

2 METHODS

The Gene Ontology (GO) defines a set of terms to which any given protein may be annotated. GO representation entails a directed acyclic graph (DAG); in this graph the parent–child relationship among terms implies that the child term is either a special case of the parent term (the IS–A relationship) or describes a process or component that is part of the parent process/component (the PART-OF relationship). In either case, there is a clear directional dependency. Specifically, a protein positively annotated to a child term is, by definition, also positively annotated to the parent term(s), but not vice versa. As a logical consequence, a protein that is negatively annotated to a parent term is also negatively annotated to the child term(s). A negative annotation indicates that a protein has been experimentally verified not to be involved in a particular function.

The annotations of any single protein, both positive and negative, can be graphically represented by the entire GO graph. Thus a protein can be viewed in the annotation space defined by GO. For unannotated proteins, dependencies among different terms are unknown, but can be elucidated from similar dependencies within annotated proteins. In our model, therefore, each protein is represented by its own GO DAG ontology structure (or by a subontology of GO) and probabilistic dependencies among different functional terms, leading to a Bayesian Network representation (Pearl, 1988) of GO functional dependencies. Figure 1a shows a simple example of using a DAG to encode a subontology in GO.

*To whom correspondence should be addressed.

[1]Gene Ontology contains three types of terms: biological process, molecular function and cellular component. For clarity of explanation, we will refer only to function although any of the three types might apply.

© The Author 2006. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oxfordjournals.org

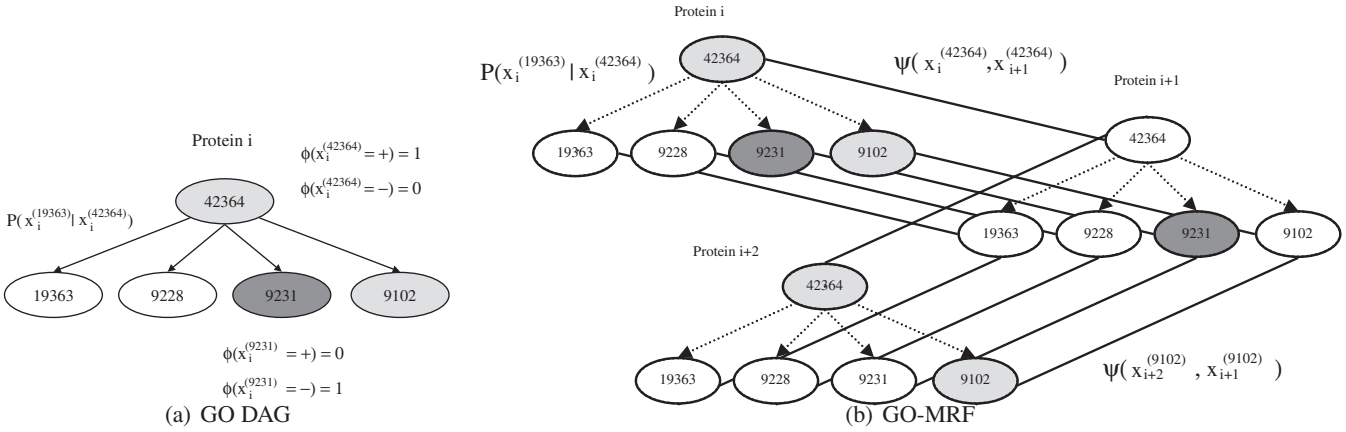

Fig. 1. (**a**) An ontology structure for a single protein. This hypothetical protein has been positively annotated to GO term 9102 and, therefore by definition, is also positively annotated to GO term 42364, the parent of 9102. The darker shading at term 9231 indicates that this protein has been negatively annotated to that term. The protein is unknown at the two unshaded terms. (**b**) A (simple) complete chain graph model with three proteins. Each protein is modeled by an ontology of size five, with different types of evidence present at each protein. Also shown are examples of the model functions P, ψ and ϕ, defined in Equation (1), corresponding to some model elements.

A common set of methods for inferring protein functions relies on the notion of similarity among proteins. Namely, similar proteins are more likely to share common functional aspects (terms in the GO notation) than proteins with less similarity. The notion of similarity may utilize primary and post-primary sequence homology (Liu and Rost, 2001, 2003; Pruess *et al*., 2003; Whisstock and Lesk, 2003), similarity in short signaling motifs, amino acid composition and expression data (Nakai and Horton, 1999; Drawid and Gerstein, 2000; Nair *et al*., 2003) implying subcellular localization, and protein–protein interactions (Galperin and Koonin, 2000; Valencia and Pazos, 2002), among others (Rost, 2003).

Similar to Deng *et al*. (2002, 2003, 2004) and Letovsky and Kasif (2003), we encode the ability of our model to transfer function among similar proteins using a probabilistic graphical representation of a Markov random field (MRF) (Geman and Geman, 1984); proteins are pairwise linked if they are considered to be similar rather than dissimilar. The notion of similarity, however, is associative, not directional as is the case with GO. Thus our complete model has two kinds of links: directed and undirected. Consequently, we define a chain graph model (Lauritzen, 1996), a hybrid between a Bayesian Network and a MRF. This is illustrated in Figure 1b.

If two proteins meet the criterion for similarity and are therefore linked by an undirected edge, then they are pairwise linked at each term of GO. Establishing links between proteins across all terms of GO enables the information about functional similarity to be potentially reinforced from multiple levels as well as multiple proteins.

3 ALGORITHM

3.1 Overview

In the proposed model, each protein is represented by a replicate of the GO (or a subontology of GO). Owing to the IS–A and PART–OF relationships between parent and child terms in GO, the relationships are directional, and are best modeled by directed edges with conditional probabilities of being in the child class given whether or not the protein is in the parent class(es). Thus, in isolation, each protein is modeled as a Bayesian Network (BN). Each protein's BN is then embedded in a larger network, including the BNs of other proteins. In this network similar proteins are linked by undirected edges representing the associative relationship implied by similarity. These undirected links are established pairwise at every term of GO between similar proteins.

Once the directed and undirected links of this network have been defined, as in Figure 1a and b, information is passed along the undirected links from annotated proteins to their neighbors, then to the neighbors' neighbors, and so forth, according to a function defined by protein similarity (the more similar two proteins are, the more this function influences them to have the same annotations). At the same time, information is passed within each protein's BN along the directed links, according to the conditional probabilistic relationships among different terms. This process continues until a state of convergence is reached (defined in Section 3.3). At convergence, the posterior probabilities of membership in the classes defined by GO are calculated at the target proteins and predictions are made based on those probabilities.

These main steps of our algorithm are outlined below:

Learning

1. Estimate parameters of Ontology Bayesian Network.
   a. Impose IS–A and PART–OF constraints.
   b. Estimate remaining parameters using GO data.
2. Estimate structure and parameters of Markov Random Field based on protein similarity measure(s).
   a. Estimate structure using a threshold on normalized pairwise similarity measure(s).
   b. Estimate parameters using normalized pairwise similarity scores.

Prediction

1. Infer posterior beliefs of terms of target proteins, given a set of known annotations of evidence proteins, using belief propagation in the chain graph.
2. Predict ontology term annotations of target proteins.

3.2 Learning

The learning of structure and parameters of the GO network and the MRF is achieved in two separate steps. In the GO phase a known GO structure is augmented with probabilistic parameters. The MRF modeling requires that both the structure and the parameters be estimated.

3.2.1 Ontology network learning Within each protein's GO DAG structure, we need to define the conditional probability distribution of all child terms given their parent terms,

$$P\left(x_i^{(c)} \mid Pa\left(x_i^{(c)}\right), \theta_{i,c}^{\mathrm{GO}}\right),$$

where $x_i^{(c)}$ denotes the positive or negative annotation of protein i to a particular term c in the GO[2] and $\theta_{i,c}^{\mathrm{GO}}$ are the distribution parameters. In our case, the distributions are modeled as the conditional probability tables and $\theta_{i,c}^{\mathrm{GO}}$ represent the entries in those tables.

For some values of variables $x_i^{(c)}$, the conditional probabilities are constrained by definition of GO: e.g. if a child term has one parent, then owing to the IS–A and PART–OF relationships, the probability that the child term is negative, given that the parent term is negativ, is one. If there are more parents than one, then any parent being negative immediately implies that the child is negative,

$$P\left(x_i^{(c)} = -\mid Pa(x_i^{(c)}) = (\ldots, -, \ldots)\right) = 1,$$
$$P(x_i^{(c)} = +\mid Pa(x_i^{(c)}) = (\ldots, -, \ldots)) = 0.$$

Thus, the only conditional probabilities that need to be estimated are the probabilities of a protein being positively versus negatively annotated given that all parents are positive.

This is done by defining a binomial Bayesian network over GO, with Beta priors at each term, and estimating the free conditional binomial distribution parameters $\theta_{i,c}^{\mathrm{GO}}$ of GO DAG terms from training data. The priors are selected to express our indifference about the values of all relative frequencies (Neapolitan, 2004). Hence, the conditional probabilities in GO are estimated as

$$P\left(x_i^{(c)} = +\mid Pa\left(x_i^{(c)}\right) = (+, +, \ldots, +)\right) \sim$$
$$\#\left(x_i^{(c)} = +, Pa\left(x_i^{(c)}\right) = (+, +, \ldots, +)\right) + \beta(c)$$

$$P\left(x_i^{(c)} = -\mid Pa\left(x_i^{(c)}\right) = (+, +, \ldots, +)\right) \sim$$
$$\#\left(x_i^{(c)} = -, Pa\left(x_i^{(c)}\right) = (+, +, \ldots, +)\right) + \beta(c).$$

Here $\#(K)$ is the count of instances in the training data where condition K is satisfied and $\beta(c)$ is the Beta-prior pseudocount

$$\beta(c) = 2^{-d(c) - |\theta_{i,c}^{GO}| + 1},$$

where $d(c)$ denotes the depth of term c in GO and $|\theta_{i,c}^{\mathrm{GO}}|$ is the number of parameters in $\theta_{i,c}^{\mathrm{GO}}$ at the same term. For instance, all terms in Figure 1a have $|\theta_{i,c}^{\mathrm{GO}}| = 1$.

Without loss of modeling generality, we assume that the entire GO structure, along with the resulting conditional distributions, is identical at each protein in the set, $\theta_{i,c}^{\mathrm{GO}} = \theta_c^{\mathrm{GO}}, \forall i$. While estimating contextually different GO parameters for subsets of proteins would be desirable, in practice the lack of training data usually prevents one from obtaining sufficiently accurate parameter estimates.

3.2.2 MRF Learning In the case of the MRF for a set of proteins $\mathcal{I} = \{1, \ldots, N\}$, we need to estimate both the structure $\mathcal{G}^{\mathrm{MRF}}$, i.e. the set of undirected edges among proteins, and the potential functions, or pairwise compatibilities among interacting proteins terms,

$$\psi\left(x_i^{(c)}, x_j^{(c)} \mid \theta_{i,j,c}^{\mathrm{MRF}}\right),$$

$i, j \in \mathcal{I}, c \in \mathrm{GO}$. These potentials must capture the associative relationship defined by some notion of similarity between proteins i and j. For simplicity, we will assume that we have in hand a similarity measure $s_{i,j,c} \in [0, 1]$ at term c between two proteins. We then define pairwise potentials between corresponding terms of two proteins as:

$$\psi(+, +) = \psi(-, -) = s_{i,j,c} \quad \psi(+, -) = \psi(-, +) = 1 - s_{i,j,c}.$$

Defining potentials in this way, it is easy to see that the existence of an edge between two proteins with similarity <0.5, where one of the proteins is positive (negative) for some term, decreases the posterior probability that the other protein is positive (negative) for the same term. However, knowing that a protein is dissimilar to another protein does not suggest anything about annotation. For example, two dissimilar proteins may or may not be involved in the same biological process. Therefore, from a modeling perspective, it is not sensible to introduce edges between proteins with similarity <0.5.

Furthermore, in practice, most protein pairs across a set of proteins with diverse functions are characterized by low similarity $s_{i,j,c}$. Thus it is also not reasonable from a computational perspective to introduce edges for such pairs. For these reasons, in our model, no edge is created between two proteins with similarity $s_{i,j,c} < 0.5$.

Without loss of modeling generality, we assume that a single measure of similarity s between two proteins determines compatibility across all different functional terms c, i.e. that $\theta_{i,j,c}^{\mathrm{MRF}} = \theta_{i,j}^{\mathrm{MRF}}, \forall c$. While term-dependent compatibilities may in fact lead to better performance, obtaining such measures reliably for a large set of proteins is currently infeasible.

Therefore, given a measure of similarity s between pairs of proteins, we are able to determine both the structure $\mathcal{G}^{\mathrm{MRF}}$ and the parameters $\theta_{i,j}^{\mathrm{MRF}}$ of the MRF: we create undirected edges between two proteins at all GO terms if and only if their similarity is >0.5, and the parameters are defined by that similarity, as defined by the ψ function.

The combined GO + MRF model now defines a joint Gibbs distribution of functional term annotations over a set of proteins in the chain graph:

$$P\left(\{x_i^{(c)}\}_{c\in\mathrm{GO}, i\in\mathcal{I}}\right) = \frac{1}{Z}\prod_{c\in\mathrm{GO}}\prod_{i\in\mathcal{G}^{\mathrm{MRF}}}\phi\left(x_i^{(c)}\right)$$
$$\prod_{(i,j)\in\mathcal{G}^{\mathrm{MRF}}}\psi\left(x_i^{(c)}, x_j^{(c)} \mid \theta_{i,j}^{\mathrm{MRF}}\right)\prod_{i\in\mathcal{I}}P\left(x_i^{(c)} \mid Pa\left(x_i^{(c)}\right), \theta_c^{GO}\right), \quad (1)$$

where Z is the normalizing constant.

The evidential $\phi(x_i^{(c)})$ functions are defined to concur with whether and how a protein is annotated. If a protein is positively annotated, then

$$\phi(+) = 1, \phi(-) = 0, \quad (2)$$

[2]The positive + (negative −) annotation specifies that the protein in question performs (does not perform) the function specified by a particular term. For most terms of most proteins, the status is neither + nor −, but unknown, and only a small fraction of proteins have negative annotations.

whereas if a protein is negatively annotated (meaning that it has been experimentally verified to not be involved in a given process), the 0 and 1 are transposed. Finally, if a protein is unannotated at a given term, the ϕ function is set to a constant value of 0.5, indicating no bias toward being positive or negative.

3.3 Prediction

Given the model estimated in the learning phase, one can subsequently use it to elucidate functions of unannotated or partially annotated target proteins (annotated to a subset of GO terms). In view of the chain graph defined in Equation (1), functional prediction entails inference of state of the model's nonevidential variables

$$P\left(x_i^{(c)} \mid \{x_l^{(k)}\}_{(k,l)\in\text{evidence}}\right), \quad \forall (c,i) \in \text{target}.$$

Exact inference in this model is, however, not tractable. Nevertheless, an approximate estimate of the target terms' annotations can be obtained using generalized belief propagation in chain graphs (Yedidia *et al*., 2002). In this method the posterior probabilities are estimated by iteratively passing probabilistic messages among nodes in the chain graph. For instance, messages between different proteins connected at term c are updated as

$$m_{ij}\left(x_j^{(c)}\right) \leftarrow \sum_{x_i^{(c)}} \left[\phi_i\left(x_i^{(c)}\right)\psi_{ij}\left(x_i^{(c)}, x_j^{(c)}\right) \cdot \prod_{k\in N(i,c)\setminus\{j\}} m_{ki}\left(x_i^{(c)}\right)\right], \tag{3}$$

where m_{ij} represents the message from node i to node j about node j being in state $x_j^{(c)}$, ϕ and ψ are the evidence and potential functions, respectively. $N(i, c)$ represents the set of neighbors of node $x_i^{(c)}$, both across related terms in the ontology and similar proteins of the same term c. For example, for the term 9102 of protein $i + 1$ in Figure 1b the neighborhood would consist of nodes $N(i+1, 9102) = \{x_{i+1}^{(42364)}, x_i^{(9102)}, x_{i+2}^{(9102)}\}$.

At convergence, the posterior probability is estimated as the belief

$$b_i\left(x_i^{(c)}\right) \sim \phi_i\left(x_i^{(c)}\right) \prod_{k\in N(i,c)} m_{ki}\left(x_i^{(c)}\right). \tag{4}$$

Convergence is defined to be a state at which all normalized messages change by $<10^{-4}$ between successive iterations. Thus the belief is the normalized product of the local evidence ϕ and the messages coming in from neighbors[3].

A message passing schedule must be implemented in order to accurately estimate the posterior probability. We suggest two schedules, which we will refer to as 'down-up' and 'down', that have empirically shown good convergence properties. In the down-up schedule, messages are initiated from the annotated term nodes, sent to all of their neighbors, then to the neighbors of their neighbors, and so on, until all nodes have been sent messages out (the 'down'). Then the order is reversed (the 'up'). In the down schedule, the up iteration is skipped, and only successive downs are executed. Prediction of whether or not a protein performs the function corresponding to each term can then be achieved by comparing thus obtained beliefs to a fixed, preselected threshold.

[3]For a singly-connected graph, the belief has been shown to be exactly equal to the posterior probability.

4 EXPERIMENTS AND RESULTS

Sequence and annotation data were obtained from the *Saccharomyces* Genome Database (http://www.yeastgenome.org). We further restrict this set to the set of sequenced proteins (ORFs) whose protein–protein interaction data are available through the GRID project (Breikreutz *et al*., 2003); there were 4897 such proteins. The ontology structure was obtained from the Gene Ontology database.

To construct the MRF submodel, we tested two different measures of similarity. The first measure utilized the primary sequence homology, determined through BLAST scores. Instead of using the raw score, we defined s in this model in terms of the p-value returned by BLAST. We Blasted each sequence against the entire database, and defined s for a pair of proteins to be $(1 - p)$, where p is the pairwise p-value. While it is estimated that only 40–60% of certain aspects of all proteins' functions can be transferred using simple homology (Koonin, 2001), we selected it as a baseline measure for testing the model's predictive ability.

As the second measure we selected the protein–protein interactions available through GRID (Breikreutz, 2003). An edge was created between two nodes (terms) if and only if their corresponding proteins interact, and potentials were defined in a term-specific way:

$$\begin{aligned}
\psi(+,+) &= P(+,+\,|\,\text{interaction})\\
\psi(-,-) &= P(-,\pm-\,|\,\text{interaction})\\
\psi(+,-) &= P(+,-\,|\,\text{interaction})\\
\psi(-,+) &= P(-,+\,|\,\text{interaction}).
\end{aligned}$$

For example, for a given term, the potential between two interacting proteins for being positive at both proteins is defined as the probability that two proteins are positive for that term given that there is an interaction. This probability is estimated similar to the way the conditional probabilities within each protein's GO DAG structure are estimated: a beta prior distribution is assumed and the corresponding pseudo-counts are updated based on the data.

A third model-type was implemented in which both similarity and PPI data were used to define the network. An edge was created if either the similarity-based criterion was met or there was an interaction (or both), and an independence assumption was made: pairwise potential was defined as the product of the similarity-based potential and the PPI-based potential.

Leave-one-out cross-validation was used to test performance of the method; in this case, the annotations of an entire protein were left out and predictions made at all terms. These predictions were then compared against the actual annotations. The ontology structure is assumed for all proteins in the network, including annotated proteins, the one whose annotations are left out, and all other unannotated proteins. This structure applies to all proteins because by definition of the Gene Ontology, it represents a set of terms to which any given protein may be annotated, whether the protein actually has already been annotated or not. In other words, the structure represents a set of functions which any given protein may perform, even if a protein's functionality is as yet unknown.

Results are presented of application of the method to several subontologies of the entire GO. Because only a small fraction of

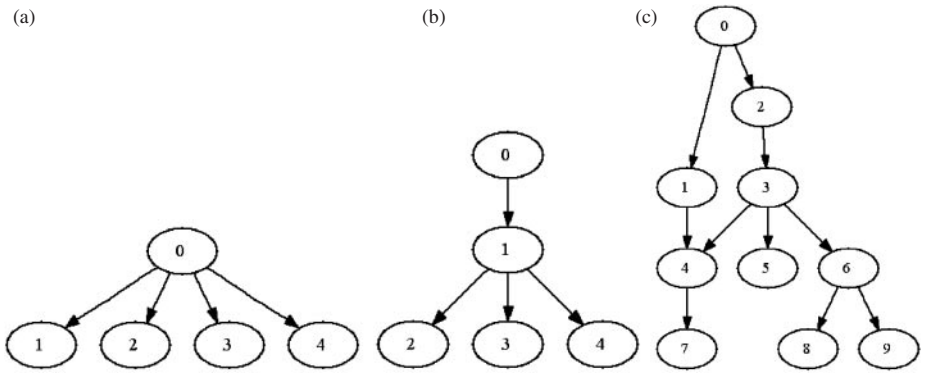

Fig. 2. The structure of three GO subontologies to which the method was applied. (**a**) Subontology 1. (**b**) Subontology 2. (**c**) Subontology 3.

proteins are negatively annotated to some term, subontologies were chosen to contain terms to which there exist negative annotations. The three ontology structures to which the method was applied are shown in Figure 2. The leaves in these subontologies are leaves in the entire GO structure (highly specific terms). This implies that relatively few proteins, between 10 and several hundred, are annotated to these terms. We consider the subsets of proteins involved in the three subontologies as three different sets on which different models are evaluated. Specifically, there were 26 proteins positively or negatively annotated to one or more terms in Subontology 1 (Fig. 2a), 8 proteins annotated to terms in Subontology 2 (Fig. 2b) and 292 proteins annotated to terms in Subontology 3 (Fig. 2c). In each case, the annotated proteins comprised a subset of the 4897 proteins under consideration, and all 4897 proteins were in the network. Most proteins in the network were unannotated in these subontologies and served as intermediate points through which information was passed.

The annotated proteins for each of these networks can be identified by the terms in the subontologies. Subontology 1 contains the following terms: 'water-soluble vitamin biosynthesis' is the root term, and its children are 'biotin biosynthesis', 'pyridine nucleotide biosynthesis', 'riboflavin biosynthesis' and 'thiamin biosynthesis'. The root term of Subontology 2 is 'secretion', its child is 'protein secretion', and the three low-level terms are 'cytokine secretion', 'immunoglobulin secretion' and 'regulation of protein secretion'. For Subontology 3, the root term is 'transporter activity', term 1 is 'amine transporter activity', term 2 is 'organic acid transporter activity', term 3 is 'carboxylic acid transporter activity', term 4 is 'amino acid transporter activity', term 5 is 'dicarboxylic acid transporter activity', term 6 is 'monocarboxylic acid transporter activity', term 7 is 'amino acid-polyamine transporter activity', term 8 is 'acetate transporter activity' and term 9 is 'allantoate transporter activity'.

As a baseline test, we also implemented a model without GO, containing an independent MRF for each term in the three subontologies. For each model, we calculated four measures of performance: recall, false positive rate, accuracy and precision:

$$\text{recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \qquad \text{fpr} = \frac{\text{FP}}{\text{TN}+\text{FP}},$$
$$\text{accuracy} = \frac{\text{TP}+\text{TN}}{\text{TP}+\text{TN}+\text{FP}+\text{FN}}, \quad \text{precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{5}$$

where TP is the number of true positive predictions, TN the number of true negative predictions, FP the number of false positive predictions, and FN the number of false negative predictions. Prediction decisions are based on 0.8 decision threshold, similar to Letovsky and Kasif (2003).

A typical run of the model with the ontology, with 4897 proteins and five ontology nodes per protein, required ~275 iterations of message-passing, which took about 2 h on a 2.80 GHz CPU. The corresponding runs on networks defined by separate terms for the MRF model took a total of ~40 min.

Results for each of the three subontologies and the various types of models are shown in Table 1. Overall, one concludes that for the homology-based models, the model with the ontology always outperforms the simple model without the ontology in terms of positive predictions. We observed an average relative increase in accuracy of 58.2% over the three different subontologies (a, b and c in Fig. 2). Interestingly, the two models perform equally well on negative predictions. This is reflected in the average increase in precision of 2.0%. Therefore, the model with the ontology has higher recall, precision and accuracy for all three ontology structures, and equivalent false positive rate to that of the simple model.

In order to better understand the performance of the model, neighborhoods of evidence (i.e. annotated proteins) were calculated for the homology-based model. This neighborhood is constructed in the following manner: for every pair of annotated proteins the shortest path between them was calculated based on the similarity measure. Every such shortest path is included in the neighborhood of evidence; there is generally a great deal of overlap among the shortest paths, so the total number of nodes in the neighborhood is visually tractable. While the similarity-based distance measure should not be expected to correlate perfectly with what occurs in this type of belief network, the resulting neighborhood graphs have proven to be quite elucidating. The graphs illustrating this property are shown in Figure 3. The false positives in the case of the subontology of Figure 2a occur because each of the two negatively annotated proteins is directly connected to two positively annotated proteins, but the negative proteins are not connected to each other. Hence no pathway for information exists between negative proteins, as depicted in Figure 3a. The second subontology of Figure 2b has recall 100% and false positive rate 25% due to a single negatively annotated protein (negatively annotated to four terms). Finally, for the ontology of Figure 2c, all predictions were correct, i.e. recall,

Table 1. Performance measures for the various types of models

	Sequence homology		PPI		Sequence homology + PPI	
	MRF	MRF + GO	MRF	MRF + GO	MRF	MRF + GO
REC_a	0.436	1.00	0.957	1.00	0.974	1.00
FPR_a	1.00	1.00	1.00	1.00	1.00	1.00
$PREC_a$	0.944	0.975	0.957	0.959	0.974	0.975
ACC_a	0.425	0.975	0.918	0.959	0.95	0.975
REC_b	0.714	1.00	0.714	1.00	0.857	1.00
FPR_b	0.250	0.250	1.00	1.00	1.00	1.00
$PREC_b$	0.909	0.933	0.714	0.778	0.750	0.778
ACC_b	0.722	0.944	0.556	0.778	0.667	0.778
REC_c	0.870	1.00	0.896	1.00	0.990	1.00
FPR_c	0.00	0.00	0.75	1.00	0.00	0.00
$PREC_c$	1.00	1.00	0.975	0.970	1.00	1.00
ACC_c	0.874	1.00	0.877	0.970	0.990	1.00

The three basic model types are homology-based network, PPI-based network and combined homology-PPI.
For each of these basic types, there are two subtypes, MRF without ontology and MRF + GO. REC stands for recall, FPR stands for false positive rate, PREC represents precision, and ACC is accuracy. Subscripts *a*, *b*, and *c* correspond to subsets of proteins derived from terms in the three different ontologies of Figure 2.

precision and accuracy were all 100%, and the false positive rate was 0%. Viewing Figure 3c, one can see that the positively annotated proteins tend to cluster, as do the negatively annotated proteins. This is an ideal situation in terms of prediction. Furthermore, since this is the largest ontology, with the most general root term out of the three ontologies, it has the highest number of annotated proteins and total predictions, both measuring in the hundreds.

In most cases, the model with the ontology makes a true positive prediction where the model without the ontology makes a false negative prediction because there is a term with only one protein annotated to it. In the simpler model, this results in that protein being isolated from evidence. However, in the model with the ontology, the protein is often connected to proteins annotated to other terms, and belief propagates from those terms to that protein's GO DAG structure, down to the term in question, leading to a more accurate prediction.

In the case of the PPI-based model, the model with the ontology again ubiquitously outperforms the simple model in terms of positive predictions, again attaining 100% recall for all three ontology structures. However, the more complex model also exhibits a 100% false positive rate, i.e. all negative annotations were predicted positive, possibly due to high clustering of positive proteins and isolation of the negative ones, as illustrated in Figure 3d. This fact is not surprising, since knowing that two proteins interact should tell us something about the functions they perform, not the functions they do not perform. As a result, PPI-based models are not generally expected to yield high accuracy when predicting GO functions. In particular, PPI is not expected to be a good predictor of negative annotations. In our case, overall accuracy was generally high for the PPI-based models because there were very few negative annotations in the data. Overall, the inclusion of ontology resulted in relative increases in accuracy (precision) of 18.3% (2.9%) over the traditional methods.

Finally, considering the model that accounts for both homology and PPI, again recall is always 100% for the complex model and lower for the simple model. The two types of models performed equally well (or equally poorly, depending on the ontology) on negative predictions. For the simple models of all three ontologies, using similarity and PPI in conjunction improved recall over using either type of potential function alone. However, such an improvement over the similarity-based model with the ontology incorporated was not possible, as recall was already 100%. This resulted in relative increases in accuracy (precision) of 6.8% (1.3%).

Similar conclusions carry over for different settings of the decision threshold (ROC analysis). Our experimental results, using homology-based similarity, yield area-under-curve (AUC) estimates at 0.192, 0.256 for MRF and MRF+GO of Figure 2a, 0.679, 0.804 for MRF and MRF+GO of Figure 2b and 1.000, 1.000 for MRF and MRF+GO of Figure 2c. For the models corresponding to Figure 2c, although AUC was equivalent (with a perfect score of 1.000 in both cases), there was substantially more separation between beliefs at positive versus negative terms for the MRF+GO model than for the simple MRF model, indicating that the MRF+GO model more strongly separated the positives from the negatives. Again, inclusion of ontology seems to be a crucial aspect for improving functional predictions.

It is important to note that negative annotations at present are very rare in the data. For this reason, the estimates of the false-positive rate may not be reliable, and may be improved as these datasets become richer in negative annotations.

5 CONCLUSIONS

In this work, we presented a method that integrates the GO structure into a protein classification framework based on similarity-induced function transfer. Incorporation of the ontology structure, along with the dependencies among its functional terms, improves performance over traditional model that considers each term separately. In this context we also showed that particular choices of similarity functions can lead to improved positive as well as negative predictions at ontology terms.

Only dependencies directly implied by the GO structure were exploited in this study. It is possible that other dependencies exist among terms, and if so, that their incorporation could also

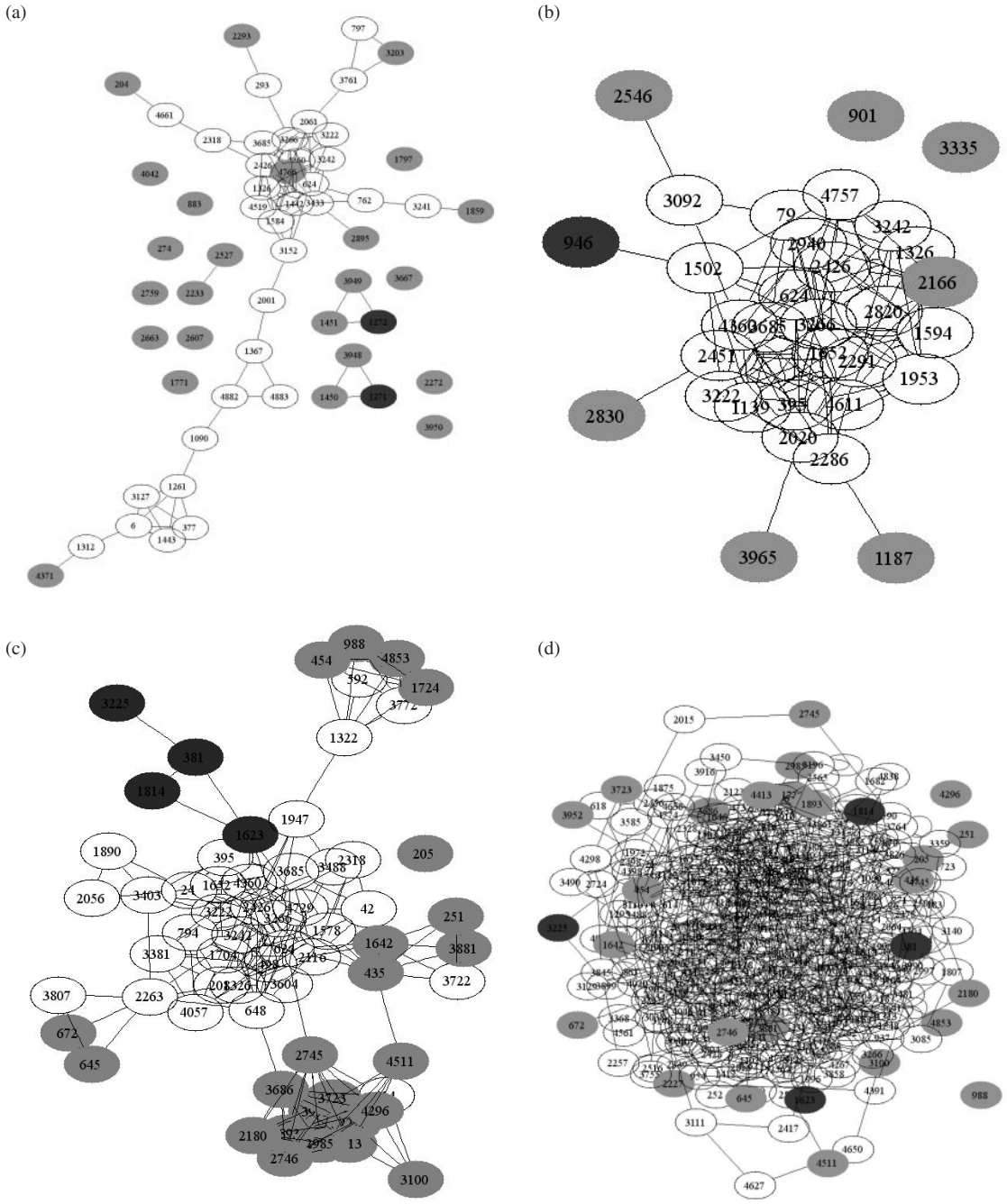

Fig. 3. A neighborhood of evidence for the annotated proteins corresponding to the three different ontologies and homology-based similarity together with a neighborhood obtained using the PPI similarity. The ontology structure is omitted for visual clarity, each protein is represented by a single node. Proteins with any negative annotations are shaded dark grey proteins with positive annotations are light grey and unannotated proteins are clear. For the PPI network the positively annotated proteins often cluster, but the negative ones do not. (**a**) Subontology 1. (**b**) Subontology 2. (**c**) Subontology 3. (**d**) PPI.

improve results. Additional improvement in predictive performance could also be achieved by combining different types of protein correlative evidence, such as sequence similarity, PPI data, phylogeny, etc. Finally, our current approach relies on models of GO whose parameters are identical over the set of proteins in question. A similar assumption is made for MRF models that are not term specific. Relaxing these two assumptions may lead to more realistic models. However, accurately estimating parameters of such models may be infeasible given current sizes of functionally annotated datasets.

At present, computational issues preclude an application of the model to the entire Gene Ontology, which contains over 19 000

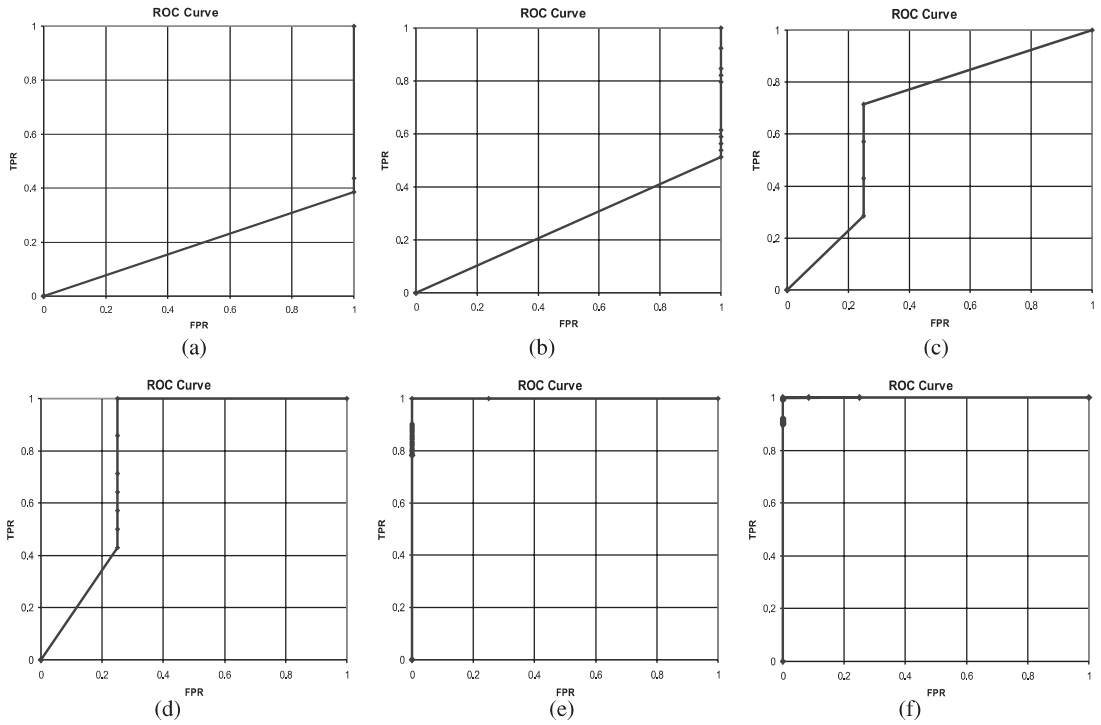

Fig. 4. ROC graphs for models evaluated on three subsets of proteins corresponding to terms involved in the three ontologies in Figure 2. Evaluated on the three datasets are both MRF and MRF + GO models, constructed using homology-based similarity. (**a**) Set 1, MRF. (**b**) Set 1, MRF + GO. (**c**) Set 2, MRF, (**d**) Set 2, MRF + GO. (**e**) Set 3, MRF. (**f**) Set 3, MRF + GO.

terms, with a substantial number of proteins. Instead, one could apply the model to high level, general terms in the ontology, resulting in candidate root terms for the next round of application, and so on, until the most specific predictions possible are made.

REFERENCES

Breitkreutz,B.J. *et al*. (2003) The GRID: the general repository for interaction datasets. *Gen. Biol*., **4**, R23.

Ashburner,M. *et al*. (2000) Geneontology: tool for the unification of biology. The gene ontology consortium. *Nat. Gen*., **25**, 25–29.

Drawid,A. and Gerstein,M. (2000) A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J. Mol. Biol*., **301**, 1059–1075.

Deng,M., Zhang,K., Mehta,S., Chen,T. and Sun,F. (2002) Prediction of protein function using protein–protein interaction data. In *Proceedings of the First IEEE Computer Society Bioinformatics Conference (CSB2002)*. Standford University, Palo Alto, CA, pp. 197–206.

Deng,M., Chen,T. and Sun,F. (2003) An integrated probabilistic model for functional prediction of proteins. In *Proceedings of the Seventh International Conference on Computational Molecular Biology (RECOMB2003),* Berlin, Germany, pp. 95–103.

Deng,M. *et al*. (2004) Mapping gene ontology to proteins based on protein–protein interaction data. *Bioinformatics*, **20**, 895–902.

Galperin,M.Y. and Koonin,E.V. (2000) Who's your neighbor? New computational approaches for functional genomics *Nat. Biotechnol*., **18**, 609–613.

Geman,S. and Geman,D. (1984) Stochastic relaxation,Gibbs distribution and the Bayesian restoration of images. *IEEE Trans. Pattern Anal. Mach. Intell*., **6**, 721–741.

Iyer,L.M. *et al*. (2001) Quod erat demonstrandum? The mystery of experimental validation of apparently erroneous computational analyses of protein sequences. *Genome Biol*., **2**, RESEARCH0051.

Koonin,E.V. (2001) Computational genomics. *Curr. Biol*., **11**, R155–R1558.

Lauritzen,S.L. (1996) *Graphical Models*. Oxford University Press, New York, NY.

Letovsky,S. and Kasif,S. (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, **19** (Suppl), i197–i204.

Liu,J. and Rost,B. (2001) Comparing function and structure between entire proteomes. *Prot. Sci*., **10**, 1970–1979.

Liu,J. and Rost,B. (2003) CHOP proteins into structural domain-like fragments. *J. Mol. Biol*., (submitted 2003-03-25).

Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *TIBS*, **24**, 34–36.

Nair,R., Carter,P. and Rost,B. (2003) NLSdb: database of nuclear localization signals. *Nucleic Acids Res*., **31**, 397–399.

Neapolitan,R.E. (2004) *Learning Bayesian Networks*. Prentice-Hall, Upper Saddle River, NJ.

Pearl,J. (1988) *Probabilistic reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann.

Pruess,M. *et al*. (2003) The Proteome Analysis database: a tool for the *in silico* analysis of whole proteomes. *Nucleic Acids Res*., **31**, 414–417.

Rost,B. (2003) Automatic prediction of protein function. *Cellular Molecular Life Sciences*, **60**, 2637–2650.

Valencia,A. and Pazos,F. (2002) Computational methods for the prediction of protein interactions. *Curr. Opin. Struct. Biol*., **12**, 368–373.

Whisstock,J.C. and Lesk,A.M. (2003) Prediction of protein function from protein sequence and structure. *Quart. Rev. Biophys*., **36**, 307–340.

Yedidia,J.S., Freeman,W.T. and Weiss,Y. Understanding belief propagation and its generalizations. Technical Report TR2001-22, Mitsubishi Electric Research Labs, January 2002.