Rutgers Computer Science Technical Report RU-DCS-TR587
December 2005

Protein Classification Using Probabilistic Chain Graphs and the Gene Ontology Structure

by

Steven Carroll and Vladimir Pavlovic
Department of Computer Science,
Rutgers University, Piscataway, NJ, 08854

ABSTRACT

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 amongst themselves according to different measures of functional similarity, after which belief propagation is run to make predictions at all ontology terms. Results indicate that such direct utilization of the Gene Ontology improves predictive ability, outperforming traditional models that do not take advantage of dependencies among functional terms.

1 Introduction

Due 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 [4, 5, 6, 12]. 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.

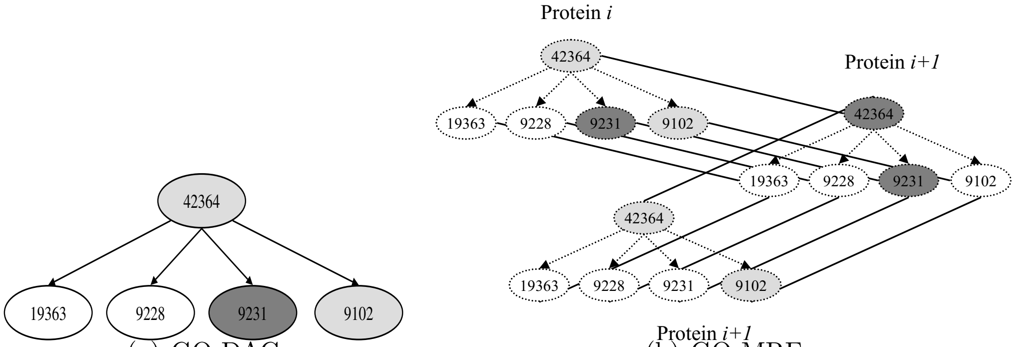
Current graphical formulations consider only one protein functional category at a time, implying independence in annotations across multiple levels of protein description. However, functional categories are known to exhibit dependencies in a cellular context. In this work, we present a probabilistic graphical model framework that considers multiple functional categories (terms) in the Gene Ontology [2] *simultaneously*, making predictive use of the definitive and probabilistic relationships among the terms. This is possible due 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. In this fashion 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 available in 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 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.

[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 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 on 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 molecular function terms, leading to a Bayesian Network representation [18] of GO functional dependencies. Figure 1(a) shows a simple example of using a DAG to encode a subontology in GO.



(a) GO DAG (b) GO-MRF

Figure 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 three, with different types of evidence present at each protein.

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) compared to the proteins with less similarity. The notion of similarity may utilize primary and post-primary sequence homology [13, 14, 19, 23], similarity in short signaling motifs, amino acid composition and expression data [15, 3, 16] implying subcellular localization, and protein-protein interactions [7, 22], among others [20].

Similar to [4, 5, 6, 12] we encode the ability of our model to transfer function among similar proteins using a probabilistic graphical representation of a Markov Random Field (MRF) [8]; 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 [11], a hybrid between a Bayesian Network and a MRF. This is illustrated in Figure 1(b).

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. Linking 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

The main steps of our algorithm are outlined below:

Learning

1. Estimate parameters of Ontology Bayesian Network.
2. Estimate structure and parameters of Markov Random Field based on protein similarity measure(s).

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.

3.1 Learning

Within each protein's GO DAG structure, we need to define the conditional probability distribution of all children given their parents,

$$P(x_c^{(i)}|a(x_c^{(i)}),\theta_{c,i}^{GO}),$$

where $x_c^{(i)}$ denotes the positive or negative annotation of protein i to a particular term c in the GO[2]. For some values of variables $x_c^{(i)}$, the conditional probabilities are constrained by definition of GO: for example, if a child term has one parent, then due to the IS-A and PART-OF relationships, the probability that the child term is negative , given that the parent term is negative, is one. If there are more parents than one, then any parent being negative immediately implies that the child is negative. Thus, the only conditional probabilities that need to be estimated are the probabilities of a protein being being positively vs. negatively annotated given that *all* parents are positive.

This is done by defining an augmented binomial Bayesian network over GO, with beta priors at each term and estimating the free conditional binomial distribution parameters $\theta_{c,i}$ of GO DAG terms from training data. Without loss of modeling generality, we assume that

[2] The positive + (negative -) annotation specifies that the protein in question has(does not have) 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.

the entire GO structure, along with the resulting conditional distributions, is identical at each protein in the set, $\theta_{c,i} = \theta_c, \forall i$. 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.

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

$$\psi(x_i^{(c)}, x_j^{(c)} | \theta_{c,i,j}^{MRF}),$$

$i, j \in \mathcal{I}, c \in GO$. 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 less than 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.

In practice, most protein pairs across a set of proteins with diverse functions are characterized by low similarity s. From computational as well as modeling perspective it is important to eliminate edges in graph $\mathcal{G}^{MRF}$ indicating the low similarity. In our model no edge is created between two proteins with similarity $s < 0.5$.

Without loss of model's generality, we assume that a single measure of similarity s between two proteins determines compatibility across all different functional terms c, $\theta_{i,j,c}^{MRF} = \theta_{i,j}^{MRF}$. While term-dependent compatibilities are favorable, obtaining such measures 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 $\mathcal{G}^{MRF}$ and the parameters $\theta_{i,j}^{MRF}$ of the MRF. The combined GO+MRF model now defines a joint Gibbs distribution of functional term annotations over a set of proteins in the chain graph:

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

3.2 Inference

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

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

Exact inference in this model is, however, not tractable. Nevertheless, an approximate estimate of the target terms' annotation can be obtained using generalized belief propagation in chain graphs [24]. A message passing schedule must be implemented in order to accomplish this. 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 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 a proteins most likely function for each term can then be achieved by comparing thus obtained beliefs to a fix, preselected threshold.

4 Experiments and Results

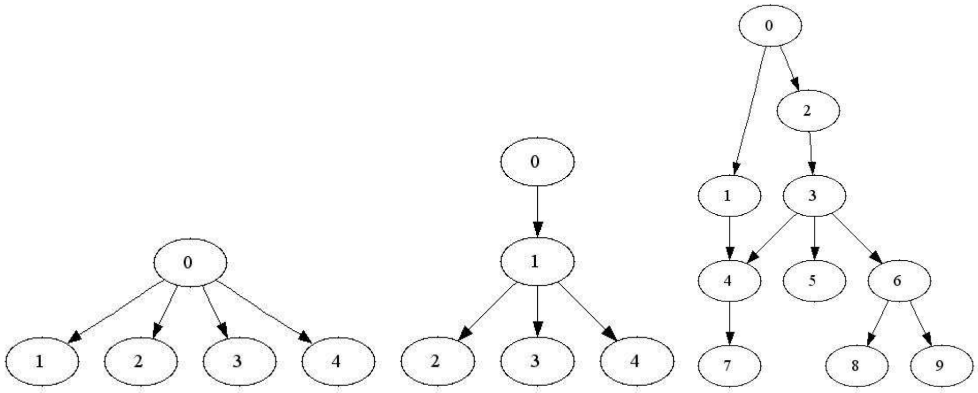
Sequence and annotation data were obtained from the *Saccharomyces* Genome Database [21]. We further restrict this set to the set of sequenced proteins (ORFs) whose protein-protein interaction data is available through the GRID project [1]; there were 4,897 such proteins. The ontology structure was obtained 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 PSI-BLAST scores. While it is estimated that only 40-60% of certain aspects of all proteins' functions can be transfered using simple homology [10], 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 [1].

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.

Results are presented of application of the method to several subontologies of the entire GO. Because only a small fraction of 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 ten 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. For instance, the first subset consists of several tens of proteins involved in functions of GO in Figure 2(a).

As a baseline test, we also implemented a model without GO, containing independent MRF for each term in the three subontologies. For each model, we calculated four measure



(a) Subontology 1 (b) Subontology 2 (c) Subontology 3

Figure 2: The structure of three GO subontologies to which the method was applied.

of performance: recall, false positive rate, accuracy, and precision:

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

where TP is the number of true positive predictions, TN the number of true negative predictions, FP the number of false positive predictions, FN the number of false negative predictions. Optimal prediction decisions are based on 0.8 decision threshold, similar to [12].

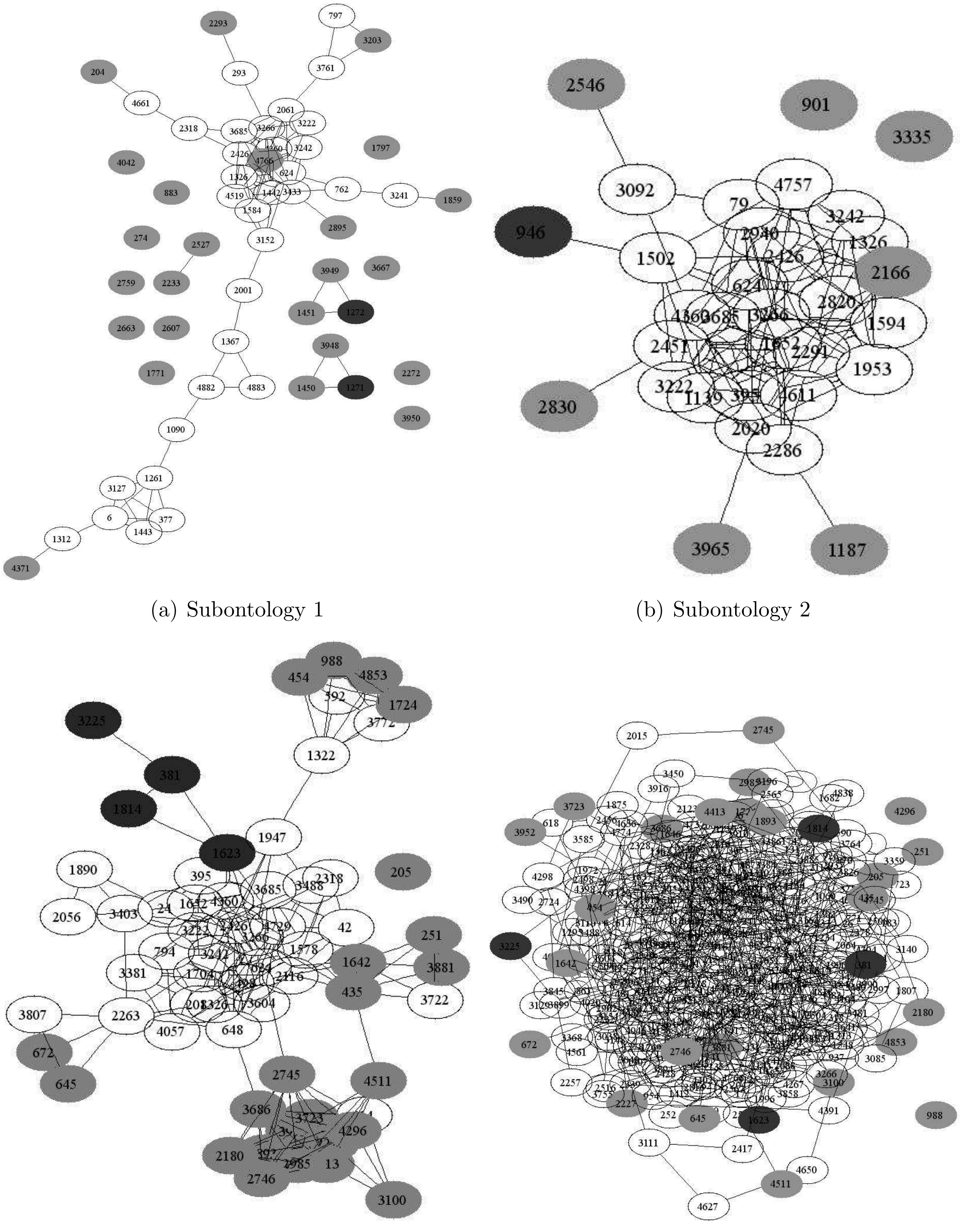
A typical run of the model with the ontology, with 4,897 proteins and five ontology nodes per protein, required approximately 275 iterations of message-passing, which took about two hours on a 2.80 GHz CPU. The corresponding runs on networks defined by separate terms for the MRF model took a total of approximately 40 minutes.

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. Interestingly, the two models perform equally well on negative predictions. 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) was 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 this 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 subontology of Figure 2(a) occur because each of the two negatively annotated proteins is directly connected to two positively annotated proteins, but the negative proteins

Table 1: Performance measures for the various types of models. 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. Subscript t indicates results combined for a given model type over the three different ontologies, yielding overall estimates of performance for each model type.

	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
REC_t	0.827	1.00	0.896	1.00	0.984	1.00
FPR_t	0.167	0.167	0.833	1.00	0.294	0.294
PREC_t	0.992	0.993	0.964	0.962	0.989	0.989
ACC_t	0.828	0.994	0.869	0.962	0.974	0.989



(a) Subontology 1

(b) Subontology 2

(c) Subontology 3

(d) PPI

Figure 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.

are not connected to each other. Hence no pathway for information exists between negative proteins as depicted in Figure 3(a). The second subontology of Figure 2(b) has recall 100% and false positive rate 25% due to a single negatively annotated protein. Finally, for the ontology of Figure 2(c), all predictions were correct, i.e. recall, precision and accuracy were all 100%, and the false positive rate was 0%. Viewing 3(c), 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.

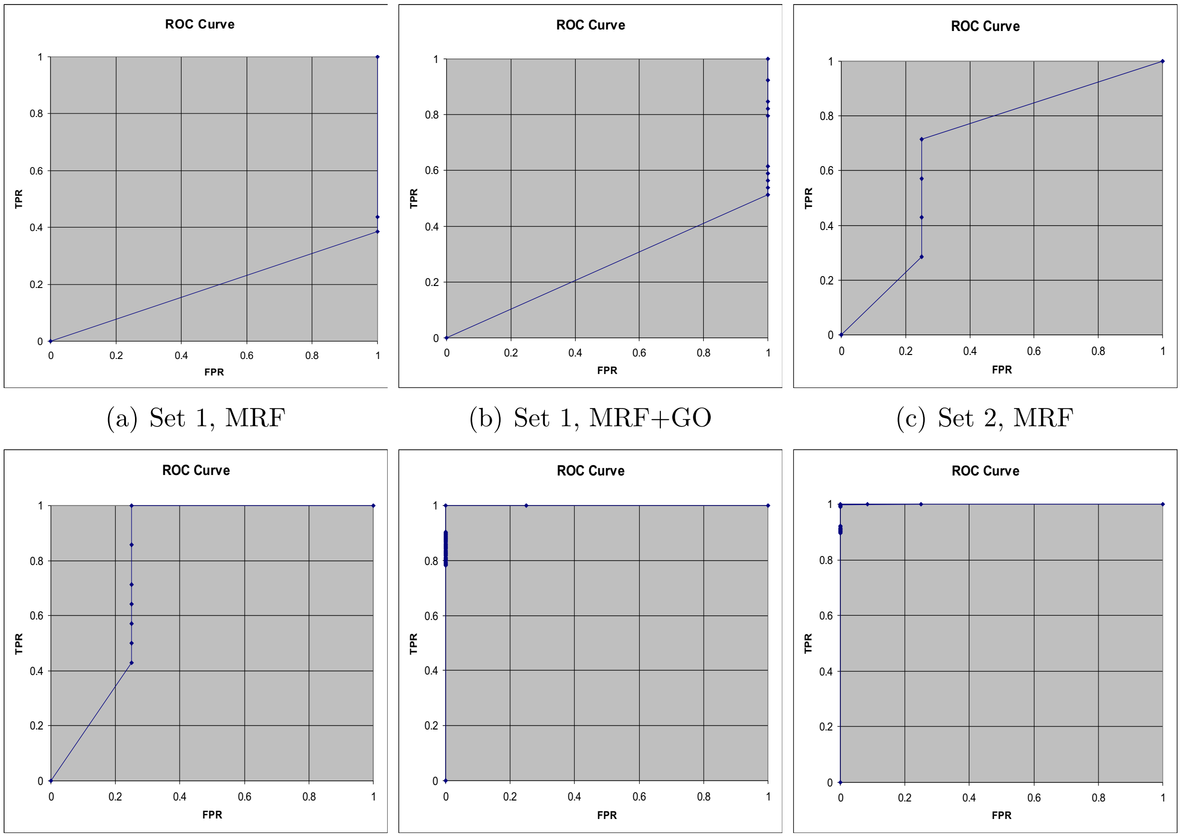
In most cases, the model with the ontology makes a true positive prediction where the model without the ontology makes a false negative prediction because there is a term with only one protein annotated to it. In the simpler model, this results in that protein being isolated from evidence. However, in 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 models, 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 annotated negative predictions were predicted positive possibly due to high clustering of positive proteins and isolation of the negative ones, as illustrated in Figure 3(d). 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.

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%.

Similar conclusions carry over for different setting of the decision threshold (ROC analysis). Our experimental results, using homology-based similarity, yield area-under-curve (AOC) estimates at 0.192, 0.256 for MRF and MRF+GO of Figure 2(a), 0.679, 0.804 for MRF and MRF+GO of Figure 2(b) and 1.000, 1.000 for MRF and MRF+GO of Figure 2(c). 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 data sets become richer in negative annotations.



(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

Figure 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.

5 Conclusions

Incorporation of ontology structure, along with the dependencies among its functional terms, improves performance over a more naive model that considers each term separately. Furthermore, defining potential functions based on similarity allows for correct positive and negative predictions, whereas using protein-protein interaction data does not yield correct negative predictions.

Only dependencies directly implied by the Gene Ontology structure were exploited in this study. It is possible that other dependencies exist among terms, and if so, that their incorporation could also 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. Similar assumption is made for MRF models that are not term specific. Relaxing the 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 17,000 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

[1] Breitkreutz, BJ., Stark, C., Tyers M. (2003) The GRID: the General Repository for Interaction Datasets. *Genome Biology* 4(3):R23

[2] Ashburner,M., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M. et al. (2000). Geneontology: tool for the unification of biology. The gene ontology consortium. *Nat.Gen.*, 25, 25-29.

[3] Drawid, A.and Gerstein, M. (2000). A Bayesian system integrating expression data withsequence patterns for localizing proteins: comprehensive application to theyeast genome. *J. Mol. Biol.*, 301,1059-1075.

[4] Deng,M., Zhang,K., Mehta,S., Chen,T. and Sun,F. (2002) Prediction of protein function using protein–protein interaction data. *Proceedings of the First IEEE Computer Society Bioinformatics Conference (CSB2002)*, 197–206.

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

[6] Deng M., Tu, Z., Sun, F. and Chen, T. (2004) Mapping gene ontology to proteins based on protein–protein interaction data. ***Bioinformatics,*** *Vol. 20 no. 6,* 895–902.

[7] Galperin, M.Y. and Koonin, E. V. (2000). Who's your neighbor? New computationalapproaches for functional genomics. Nat. Biotechnol., 18, 609-613.

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

[9] Iyer, L. M.,Aravind, L., Bork, P., Hofmann, K., Mushegian, A. R. et al. (2001). Quod eratdemonstrandum? The mystery of experimental validation of apparently erroneous-computational analyses of protein sequences. Genome Biol., 2, RESEARCH0051.

[10] Koonin, E. V.(2001). Computational genomics. *Curr. Biol.*, 11, R155-8.

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

[12] Letovsky, S. and Kasif, S**.** (2003). Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics Vol. 19 Suppl. 1 2003*, i197-i204.

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

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

[15] Nakai, K. and Horton, P. (1999). PSORT: a program for detecting sorting signals in-proteins and predicting their subcellular localization. *TIBS*, 24, 34-6.

[16] Nair, R.,Carter, P. and Rost, B. (2003). NLSdb: database of nuclear localizationsignals. Nucl. Acids Res.,31, 397-399.

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

[18] Pearl, J. (1988). Probabilistic Reasoning in Intelligent Systems : Networks of Plausible Inference. Morgan Kaufmann.

[19] Pruess, M.,Fleischmann, W., Kanapin, A., Karavidopoulou, Y., Kersey, P. et al. (2003). TheProteome Analysis database: a tool for the in silico analysis of wholeproteomes. *Nucl. Acids Res.*, 31, 414-417.

[20] Rost, B. , Liu, J. , Nair, R., Wrzeszczynski, K. and Ofran, Y. (2003). Automatic prediction of protein function. Cellular Molecular Life Sciences, 60, 2637-2650.

[21] http://www.yeastgenome.org

[22] Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of proteininteractions. *Curr. Opin. Str. Biol.*, 12, 368-373.

[23] James C. Whisstock and Arthur M. Lesk (2003). Prediction of protein function from protein sequence and structure. Quarterly Review of Biophysics, 36, 307-340.

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