Integrative Protein Function Transfer using Factor Graphs and Heterogeneous Data Sources

Antonina Mitrofanova
New York University
Computer Science department
antonina@cs.nyu.edu

Vladimir Pavlovic
Rutgers University
Computer Science department
vladimir@cs.rutgers.edu

Bud Mishra
New York University
Computer Science department
mishra@nyu.edu

Abstract

We propose a novel approach for predicting protein functions of an organism by coupling sequence homology and PPI data between two (or more) species with multi-functional Gene Ontology information into a single computational model. Instead of using a network of one organism in isolation, we join networks of different species by inter-species sequence homology links of sufficient similarity. As a consequence, the knowledge of a protein's function is acquired not only from one species' network alone, but also through homologous links to the networks of different species. We apply our method to two largest protein networks, Yeast (Saccharomyces cerevisiae) and Fly (Drosophila melanogaster). Our joint Fly-Yeast network displays statistically significant improvements in precision, accuracy, and false positive rate over networks that consider either of the sources in isolation, while retaining the computational efficiency of the simpler models.

1 Introduction

Proteins are the basis of life involved in many if not all biological processes, such as energy and RNA metabolism, translation initiation, enzymatic catalysis, and immune response. However, for a large portion of proteins, their biological function remains unknown or incomplete. Constructing efficient and reliable models for predicting protein functions remains the task of immense importance.

Recent modeling approaches, such as in [11], have shown that the predictive power of automated annotation systems rises significantly if they incorporate heterogeneous sources of data. This is particularly important as each type of data typically captures distinct aspects of cellular activity—PPI suggest a physical interaction between proteins, sequence similarity captures relationships on a level of orthologs (inter-species relationship) or paralogs (intra-species relationship), and gene ontology defines term-specific dependencies. One important source of information is, however, not typically used. Evolutionary relationships between species suggest that orthologous proteins of different species, which share high sequence similarity and whose functions have been established before speciation, are likely to share similar protein classifications.

The use of multi-species information can become particularly important as a number of modeling methods such as [11, 9, 4] rely on the computational power of networks to transfer the functional information from annotated to unannotated proteins. In such networks there may exist proteins with *no* edges connecting them to other proteins of their own species. For example, Fly's protein CG8793-PA has no edges of high sequence similarity to other proteins in its own Fly network, but it can be connected to the Yeast network through a high-similarity edge to the yeast YDR108W protein. Moreover, in a single species network, it is often the case that proteins are surrounded only by proteins with limited functional information. In such cases, using information from multiple species becomes crucial.

In this work, we design and evaluate a probabilistic approach which integrates multiple sources of information: PPIs, gene ontology, and intra as well as inter species sequence similarity. The approach incorporates our previous probabilistic graphical model with Gene Ontology [2] with information which describes evolutionary relationships between species. We demonstrate that this method can result in significant improvements in the accuracy of functional predictions using a probabilistic label-transfer paradigm. We apply our method to two largest protein networks of Yeast and Fly. The joint Fly-Yeast network outperforms networks that consider each source in isolation, while retaining the computational efficiency of the simpler models.

Our expanded Gene Ontology approach can also be interpreted as a special case of a new broader framework of "probabilistic graphical model checking" resembling classical model checking algorithms [6] implemented through message passing in a statistical graphical model. This con-

nection becomes explicit when a Gene subontology for a protein (Figure 1) is viewed as a family of properties encoded through logical propositions and connectives. These properties can be embedded and propagated in a general graphical structure with certain logical implications—all interpreted in a three-valued logic: True (positive), False (negative) and Unknown.

For specific species, our framework connects subontologies of all proteins by edges. In the language of model checking on graphical models, subontology network for each species can be viewed as an initial labeling of "possible worlds" with certain relationships/properties. By connecting networks of two different species we thus connect two neighboring "possible worlds" and try to gain some additional information from their distances (measured by orthology or PPI). Theoretically, if the two possible worlds are adjacent, they are expected to satisfy similar properties. Considering both "worlds" simultaneously will lead to algorithms with high fidelity and improved efficiency. As may be inferred from the preceding discussion, our approach suggests, for propositional and temporal logic, a potentially much broader range of applications including many non-biological problems.

2 Prior Work

One promising computational approach to protein function prediction utilizes the family of probabilistic graphical models, such as belief networks, to infer functions over sets of partially annotated proteins [9, 3, 4]. Using only a partial knowledge of functional annotations, probabilistic inference is employed to discover other proteins' unknown functions by passing on and accumulating uncertain information over large sets of associated proteins while taking into account different strengths of associations.

A critical factor that impacts performance of network models is the choice of functional association between proteins. The most established methods are based on sequence similarity using BLAST. A large set of methods relies on the fact that similar proteins are likely to share common functions, subcellular location, or protein-protein interactions (PPIs). Such similarity-based methods include sequence homology [10, 16, 12], similarity in short signaling motifs, amino acid composition and expression data [13, 15, 5]. Using PPI data to ascertain protein function within a network has been studied extensively. For example, methods in [9, 3, 4] used the PPI to define a Markov Random Field over the entire set of proteins. These methods are based on the notion that interacting neighbors in networks might also share a function [9, 7, 14].

More recently, the approach of incorporating Gene Ontology structure into probabilistic graphical models [2] has shown promising results for predicting protein functions. The approach considers multiple functional categories in the Gene Ontology (GO) simultaneously. In this model, each protein is represented by its own annotation space - the GO structure. The information is passed within the ontology structure as well as between neighboring proteins, leading to an added ability of the model to explain potentially uncertain single term predictions.

3 Methods

3.1 Single Species Model

We use the idea of probabilistic chain graphs with incorporated GO [2] to build protein network for each specie. In the model, each protein is represented not by a single node, but by a replicate of a Gene Ontology or subontology (see Figure 1). GO is a directed acyclic graph which describes a parent-children relationship among functional terms. The child term either *IS A* special case of the parent or is a *PART OF* the parent's process or its component. Every protein has its own annotation to each of the GO functional terms: it can be assigned one of three categorical values, namely, positive, negative or unknown.

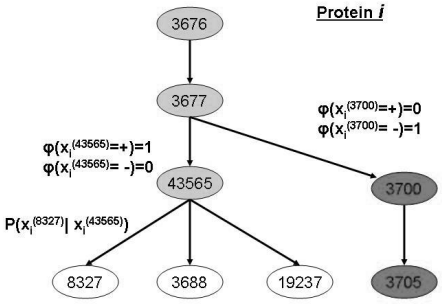

Figure 1. An ontology structure for a single (hypothetical) protein *i*: positive annotation (grey) to GO term 43565 and, thus, also to its parent - 3677 , and further up the tree to the parent's parent, term 3676. Darker shade indicates negative annotation (term 3700). Its child, term 3705, inherits this negative annotation. The protein is unknown at the three unshaded (white) terms.

The GO information is modeled as a Bayesian Network (BN), a directed graph where the child-parent relationships are defined in terms of conditional probability distributions. Proteins are then connected to each other by different measures of functional similarity (such as protein-protein interactions, sequence homology, etc) encoded in a Markov Random Field (MRF), an undirected probabilistic model. For each measure of similarity a potential function is defined, which corresponds to the probability of joint annotation of two proteins at a term given that the proteins are similar.

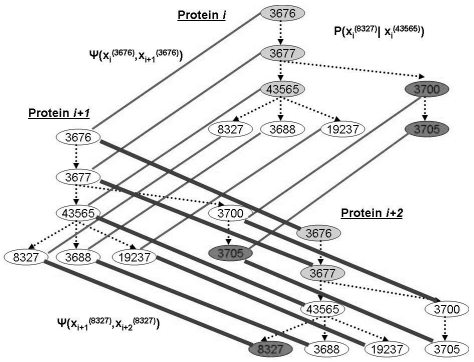

Figure 2. A chain graph model with three proteins. Each protein is represented by GO subontology of size eight, with different annotations at each protein. Some model elements, P and potential function ψ, are shown.

The similarity-based potential for proteins i and j (in a single network) at term c is defined as $\psi(+,+) = \psi(-,-) = s^{\text{within}}_{i,j,c}$ and $\psi(+,-) = \psi(-,+) = 1 - s^{\text{within}}_{i,j,c}$, for similarity measure $s_{i,j,c}$. For example, homology information is encoded as $s^{\text{within}}_{i,j,c} = 1 - p_{ij}$ where p_{ij} is a pairwise p-value determined by BLAST. The MRF and the BN are finally combined into a single graphical chain model [8], an example of which is shown in Figure 2. This model also includes the evidential function ϕ, as shown in Figure 1 that indicates the presence/absence of known annotations.

The flow of information is modeled using a message-passing mechanism for chain graphs. Messages are passed until the state of convergence is reached. At that point, posterior probabilities of membership in the classes defined by GO are calculated at the target proteins. The predictions are made by comparing those probabilities with a fixed threshold (0.8, as suggested in [9]). See [2] for a detailed description of this model.

3.2 Connecting the networks

In this work, we use inter-species sequence homology information to connect the chain graphs of multiple but related species. During the MRF building stage, we combine the individually constructed networks of two species, Yeast and Fly, through sequence similarity edges.

An edge is introduced between corresponding terms of two species based on similarity measured using BLAST scores (p-value below 0.5, similarly to [2]). In a two-species setting, we define a similarity measure between protein i in Yeast network and protein j in Fly network, at term c, as $s^{\text{between}}_{i,j,c} = 1 - p_{ij}$, where p_{ij} is the pairwise p-value. Edges are not introduced when the similarity is less than 0.5 (p-value above 0.5), since dissimilar proteins may or may not be involved in the same biological process. This similarity measure then translates into the potential function ψ in a manner analogous to the similar potential within one species: $\psi(+,+) = \psi(-,-) = s^{\text{between}}_{i,j,c}$ and $\psi(+,-) = \psi(-,+) = 1 - s^{\text{between}}_{i,j,c}$.

Similar to the single-specie model, we connect two proteins at all GO terms so that $s^{\text{between}}_{i,j,c} = s^{\text{between}}_{i,j}$ for all terms c. While using same potential for all terms may not be optimal, it was shown to improve the annotation performance. Heterogeneous values of similarity $s^{\text{between}}_{i,j}$ at each term c may lead to additional improvements, but also a more complex and demanding parameter estimation process.

This model directly generalizes to scenarios with multiple species and types of associations. Even though chain graphs can suffer from increased time and space in the multi-species networks, they are amenable to distributed implementations and often lead to significant improvements in predictive accuracy not observed in other approaches.

3.3 Protein classification

When predicting multiple protein functions, it is important to elucidate both the "negative" as well "positive" annotations for the proteins of interest. This task is rarely undertaken in practice, in part due to the lack of data and the accompanying computational methods.

Our choice of the GO subontology was driven by the task of predicting both types of annotations. The chosen subontology contains terms with negative as well as positive GO annotations for both Yeast and Fly. The subontology is depicted in Figure 1 and consists of eight terms: nucleic acid binding (3676), DNA binding (3677), sequence-specific DNA binding (43565), methyl-CpG binding (8327), DNA replication of origin binding (3688), centromeric DNA binding (19237), transcription factor activity (3700), and RNA polymerase II transcription factor activity, enhancer binding (3705). However, only a small fraction of proteins contains negative annotations. One reason for this asymmetry is the need for comprehensive tests in order to ensure that a certain protein *cannot* perform a specific function.

The leaves in the GO subontology represent the leaves in the entire GO structure implying very specific functional terms. To perform one of such functions, a protein should have specific binding motifs and configurations, suggesting that it cannot be involved in more than one function. In particular, if a protein is positively assigned to a certain GO term, we assume that it is negatively annotated to all of its siblings. 1256 out of 7260 Fly proteins and 503 out of 5298 Yeast proteins are positively annotated to one or more terms of the used subontology. After we assign possible negative annotations, there are 305 Fly and 91 Yeast proteins with at least one negative annotation. Other proteins are unannotated and are used as information conduits.

Our method can be applied to the entire GO, at the ex-

pense of time and space complexity. However, *specific*, relatively small, subontologies can be of particular interests to biochemists. For instance, vaccine targets are usually the proteins with very particular functions, represented by specific subontologies.

4 Results and Discussion

4.1 Experiments

Our experiments focused on inferring functional annotations in a combined Yeast-Fly network. The GO structure was obtained from the Gene Ontology database. We expand GO hierarchy up for positively annotated proteins and down for negatively annotated proteins. Saccharomyces genome Database for Yeast and FlyBase for Fly were used as the sources of the sequence and annotation data. The PPI data were obtained from GRID [1]. This resulted in a combined set of 7260 Fly and 5298 Yeast proteins that were used to construct the joint belief networks.

To ensure that both PPI and the homology measures are available for MRF potential esitmation on all proteins we restricted the study to the data with available PPI information. Predictive performance of our models is evaluated in a cross-validation setting. The test set consists of a random 20% of annotated proteins, with the same proportion of negatively and positively annotated proteins as the remaining 80% for training the model. For each randomly chosen test protein, its GO structure remains in place but *all* of its annotations are left out and are listed as unknown. In the case of the joint Fly-Yeast network (JN), we eliminate annotations of 20% of annotated proteins from *each* network. In the testing phase, upon convergence of the message-passing process, predictions at terms whose annotations were left out are tested against the known eliminated annotations.

We conduct a total of ten experimental rounds using the random splitting process. In each round, we compared results of runs on single networks (without joining) to that of the joint network. Individual and joint networks were trained and evaluated on the same training/testing data.

A typical run of the model with GO on the JN took approximately 28 min (4 iterations of message passing). Corresponding individual network runs took 59 min for Fly and 35 min for Yeast. Faster convergence rates in JN can be contributed to the "denser" sources of evidence in networks of multiple species compared to that of the isolated runs.

4.2 Results

For our model, we calculate five measures of performance: $recall = \frac{TP}{TP+FN}$, $precision = \frac{TP}{TP+FP}$, $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, $FPr = \frac{FP}{TN+FP}$, where negatives are as defined in 3.3.

Table 1. Comparison of average statistics (%) in joint and individual networks, over 10 runs.

networks	precision	recall	accuracy	FP rate
Fly	97.94	98.41	97.62	3.83
Fly | JN	98.71	97.98	97.87	2.35
Yeast	94.82	93.48	91.74	12.86
Yeast | JN	97.56	96.58	95.82	6.20
JN, overall	98.49	97.76	97.54	2.88

The calculations are done separately for the Yeast network, the Fly network and the joint Fly-Yeast network. In the joint network, we first calculate the overall performance (ignoring the differentiation of species), and then the performance of Fly and Yeast in the joint network separately.

Table 1 shows the average precision, recall, accuracy and FP rate for four cases: Fly network, Fly in Fly-Yeast network, Yeast network, and Yeast in Fly-Yeast network. The Fly-Yeast JN shows a clear improvement in all of the above measures. Most importantly, it significantly decreases the FP rate for both Fly and Yeast, compared to their isolated networks. In particular, FP rate for Fly decreases by 48%, and for Yeast by 52 %. For Fly, the increase in precision is 1%, in accuracy is 0.4%; for Yeast, the increase in precision is 3%, in recall is 3.3%, and in accuracy is 4.5 %. Fly does not show improvements for recall in the JN.

4.3 Statistical analysis

Statistical analysis of significance of the aforementioned performance scores was done using the t-test. The tests were conducted separately for each species and each performance measure: single Fly network is compared with the performance on the Fly in the joint Fly-Yeast network; similarly for Yeast. For comparison to be sound, the evaluations on single and joint networks were done using the same random samples (splits for testing and training sets).

The joint Fly-Yeast network shows significant improvement in performances for both Fly and Yeast ($p - value < 0.05$), as seen in Table 2(degree of freedom $= 9$). For example, for Fly the joint Fly-Yeast network shows a significant improvements compared to the Fly network alone, with respect to precision (p=0.0056)and false positive rate (p=0.0082). At the same time, for Yeast the joint network shows a significant improvement for all four measures: precision (p=0.0162), recall (p=0.0096), accuracy (p=0.0093), and false positive rate (p=0.0132).

4.4 GO vs single-term predictions

As a baseline test, we apply our method to networks without GO in place, similarly to [2], where the whole net-

Table 2. T-test p-values for precision, recall, accuracy, and FP rate.

	precis.	recall	accur.	FP rate
Fly, t-test	0.0056	-	0.2523	0.0082
Yeast, t-test	0.0162	0.0096	0.0093	0.0132

work of proteins is tested on a single ontology term. As before, in ten trials, we choose at random 20% of the network as a testing set and learn the parameters on the remaining 80%. The results shown in Table 3 indicate the superiority of the network with built-in Gene Ontology over the single-term network even in the case of multiple species networks

Table 3. Comparison of results for the network with GO and without GO

networks		precision	recall	accuracy	FP rate
Fly	w/o GO	89.97	98.37	88.67	98.57
	GO	97.94	98.41	97.62	3.83
Fly | JN	w/o GO	90.51	96.80	87.97	91.56
	GO	98.72	97.98	97.87	2.35
Yeast	w/o GO	-	0	42.62	0
	GO	94.83	93.48	91.74	12.86
Yeast | JN	w/o GO	57.38	1	57.38	1
	GO	97.56	96.58	95.82	6.20
JN overall	w/o GO	86.90	97.04	84.75	94.35
	GO	98.49	97.76	97.54	2.88

The model with GO makes a TP prediction, where the model without it commits a FN error. This result is not surprising as there is only one term with one protein annotated to it. In general, similar to [2], incorporating the ontology structure, along with the dependencies among its functional terms, considerably improves performance over that of traditional models that consider each term in isolation.

5 Conclusions

In this work we presented a new approach that uses inter-species information and the GO to simultaneously consider multiple functional categories connected in networks of two (or more) species in order to improve the predictive ability for protein classification. We show statistically significant improvements in performance of the joint model over the prediction runs on isolated species/category networks.

While in single species proteins may exist that have no annotated partners, they have the potential to acquire annotated interacting partners-homologs in a two-species setting. Additional benefits emerge for species with poorly defined protein functions and/or protein interactions. The use of the GO enables simultaneous consideration of multiple but related functional categories, opening information paths for further improvements to the model's predictive ability.

Our method readily extends to multiple species settings, and is likely to produce similar improvements. The presence of multiple interacting networks may further enable integration of additional sources of evidence, thus contributing to increased accuracy in functional predictions.

References

[1] B. Breitkreutz, C. Stark, and M. Tyers. The grid: the general repository for interaction datasets. *Genome Biology*, 4(3):R23, 2003.

[2] S. Carroll and V. Pavlovic. Protein classification using probabilistic chain graphs and the gene ontology structure. *Bioinformatics*, 22(15):1871–1878, 2006.

[3] M. Deng, T. Chen, and F. Sun. An integrated probabilistic model for functional prediction of proteins. In *RECOMB*, pages 95–103, 2003.

[4] M. Deng, Z. Tu, F. Sun, and T. Chen. Mapping gene ontology to proteins based on protein-protein interaction data. *Bioinformatics*, 20(6):895–902, 2004.

[5] M. Y. Galperin and E. V. Koonin. Who's your neighbor? new computational approaches for functional genomics. *Nat. Biotechnol*, 18:609–613, 2000.

[6] E. M. C. Jr, O. Grumberg, , and D. A. Peled. *Model Checking*. The MIT Press, 1999.

[7] U. Karaoz, T. Murali, S. Letovsky, Y. Zheng, C. Ding, and et al. Whole-genome annotation by using evidence integration in functional-linkage networks. *Proc Natl Acad Sci*, 101:2888–2893, 2004.

[8] S. L. Lauritzen. *Graphical Models*. Oxford University Press, New York, 1996.

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

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

[11] N. Nariai, E. Kolaczyk, and S. Kasif. Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS ONE*, 2(3), 2007.

[12] M. Pruess, W. Fleischmann, A. Kanapin, Y. Karavidopoulou, P. Kersey, and et al. The proteome analysis database: a tool for the in silico analysis of whole proteomes. *Nucl. Acids Res*, 31:414–417, 2003.

[13] B. Rost, J. Liu, R. Nair, K. Wrzeszczynski, and Y. Ofran. Automatic prediction of protein function. *CMLS*, 60:2637–2650, 2003.

[14] B. Schwikowski, P. Uetz, and F. S. A network of protein-protein interactions in yeast. *Nat Biotechnol*, 18:1257–61, 2000.

[15] A. Valencia and F. Pazos. Computational methods for the prediction of protein interactions. *Curr. Opin. Str.Biol*, 12:368–373, 2002.

[16] J. Whisstock and A. Lesk. Prediction of protein function from protein sequence and structure. *Quarterly Review of Biophysics*, 36:307–340, 2003.