Prediction of Protein Functions with Gene Ontology and Inter-Species Protein Homology Data

Antonina Mitrofanova, Vladimir Pavlovic, and Bud Mishra *Fellow, IEEE*

Abstract—Accurate computational prediction of protein functions increasingly relies on network-inspired models for the protein function transfer. This task can become challenging for proteins isolated in their own network or those with poor or uncharacterized neighborhoods. Here, we present a novel probabilistic chain-graph based approach for predicting protein functions that builds on connecting networks of two (or more) different species by links of high inter-species sequence homology. In this way, proteins are able to "exchange" functional information with their neighbors-homologs from a different species. The knowledge of inter-species relationships, such as the sequence homology, can become crucial in cases of limited information from other sources of data, including the protein-protein interactions or cellular locations of proteins. We further enhance our model to account for the Gene Ontology dependencies by linking multiple but related functional ontology categories within and across multiple species. The resulting networks are of significantly higher complexity than most traditional protein network models. We comprehensively benchmark our method by applying it to two largest protein networks, the Yeast and the Fly. The joint Fly-Yeast network provides substantial improvements in precision, accuracy, and false positive rate over networks that consider either of the sources in isolation. At the same time, the new model retains the computational efficiency similar to that of the simpler networks.

Index Terms—Biology and genetics, machine learning, bioinformatics (genome or protein) databases

1 INTRODUCTION

In protein-protein networks, each node represents a protein and edges between nodes represent different types of functional associations, such as protein-protein interactions, sequence similarity, co-expression patterns, and others. Majority of computational methods for protein classification rely on the property that close neighbors in a protein-protein network typically share a function [17], [15], [23], [21], [8], [6], [4], [20]. These methods assign the function (or functions) to a protein of interest based on the annotations of its neighbors. Such approaches have shown success in cases where proteins have multiple, mostly annotated neighbors. However, the methods display much less success on proteins with insufficient neighborhoods: those proteins isolated in their own network or the ones surrounded by poorly annotated neighbors.

In this work we propose a novel approach to protein function prediction, which overcomes these limitations and incorporates inter-species evolutionary information with multi-functional Gene Ontology (GO) dependencies. The fundamental conceptual innovation of our method is to connect protein-protein networks of two (or more) different, but related species, into a single computational model. Through the edges of high homology, proteins are able to expand their learning neighborhood and acquire additional functional information from their neighbors-homologs of a different species network.

Our new approach relies on a chain-graph probabilistic approach to integrate multiple sources of information: protein-protein interactions, multi-functional ontology information, intra-species sequence similarity, and inter-species homology which captures evolutionary relationships between species. In connecting networks, we rely on the fact that proteins of different species, which share high sequence similarity, are likely to share similar protein classification. In most cases such proteins, orthologs, had established functions before the speciation event. Thus, high similarity of sequences between species is likely to lead to shared functions. Even though the resulting large chain-graphs can suffer from increased time and space complexity of the models, compounded by the added complexity of the multi-species network, we show that the combined models often lead to efficient implementations and significant improvements in predictive accuracy not observed in isolated networks or other competing approaches.

The rest of the paper is organized as follows. In Section 2 we first present an overview of the closely related network approaches to protein function prediction. We then introduce, in Section 3, a chain-graph based probabilistic network model that combines both the GO structure and the information from protein-protein networks of multiple species. Section 4 demonstrates the effectiveness of the proposed approach when applied to large fly and yeast networks, at different granularities of the GO. We finally discuss the new results in Section 5

- *A. Mitrofanova is with the Department of Computer Science, Courant Institute of Mathematical Sciences, New York University, New York, NY, 10003.*
  E-mail: antonina@cs.nyu.edu
- *V. Pavlovic is with the Department of Computer Science, Rutgers University, Piscataway, NJ, 08854.*
  E-mail: vladimir@cs.rutgers.edu
- *B. Mishra is with the Department of Computer Science and Mathematics (Courant Institute of Mathematical Sciences) and Cell Biology (School of Medicine), New York University, New York, NY, 10003.*
  E-mail: mishra@nyu.edu

Manuscript received December 01, 2008; revised January 11, 2009.

and relate them to the performance of related state-of-the-art probabilistic network models.

The code (C/C++/Perl) and data files used in this work are available from http://research.rutgers.edu/~amitrofa/yeast_fly.html.

2 Related Work

Proteins are involved in many if not all biological processes, such as energy and RNA metabolism, signal transduction, and translation initiation. However, for a large portion of proteins, their biological function remains unknown or incomplete. Thus, constructing efficient and reliable models for predicting protein functions has thus become the task of immense importance.

A critical factor that impacts performance of network models is the choice of functional association between proteins. The most established methods for protein function prediction are based on sequence similarity (e.g., a BLAST score). 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, similarity in short signaling motifs, amino acid composition and expression data [18], [27], [22], [8], [20], [6].

Using PPI data to ascertain protein function within a network has been studied extensively. For example, methods in [17], [10], [11] 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 PPI networks should share a function [17], [15], [23].

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 [17], [10], [11]. 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.

Several related studies used various probabilistic frameworks to infer functions of proteins [25], [24], [26], [13], [19]. For example, the method in [26] used multiple Support Vector Machines for the classification of protein predictions using protein sequences of several organisms for training. GOtcha approach developed in [19] and method in [13] search for similar sequences, using the scoring scheme for GO annotations, based on degree of similarity of the original query and frequency of occurrence of GO in different sequences. Shin at el [24] proposed graph sharpening as a way to eliminate undesirable edges from sequence and 3D similarity graphs, and showed that graph sharpening together with data integration produced improvement in protein function prediction. Tsuda at el [25] proposed automated method to choose/weigh best networks (out of PPI, genetic interactions, protein complex, Pfam domain structure, gene expression networks) for each protein class, using Support Vector Machines.

More recently, the approach of incorporating Gene Ontology structure into probabilistic graphical models [8] has shown promising results for predicting protein functions while outperforming approaches that do not take advantage of dependencies among different functional terms. The approach described in [8] considers multiple functional categories in the Gene Ontology (GO) simultaneously. In their model each protein is represented by its own annotation space - the GO structure. In this case, 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.

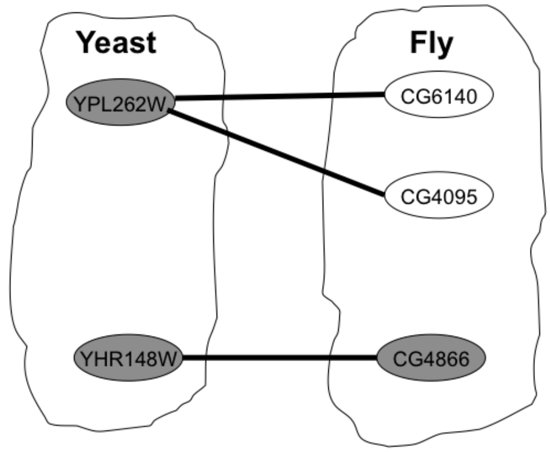


Fig. 1. Examples of proteins isolated in their own network, but connected to neighbors-homologs in the other species network.

Multiple approaches have proven that incorporating heterogeneous data to predict protein function can improve the overall predictive power of automated protein/gene annotation systems, as for example shown in [21], [4], [8]. Integrating multiple sources of information is particularly important as each type of data captures only one aspect of cellular activity—PPI data 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.

Many learning approaches rely on information available from neighbors in a protein network [21], [17], [4]. However, there may exist proteins with *no* edges connecting them to other proteins in their own networks, as demonstrated in Figure 1. For example, considering Yeast and Fly networks, yeast protein YPL262W has no edges of high sequence similarity to other proteins in its own yeast network, but it is connected to two fly proteins (CG6140-PA, CG4095-PA) through high similarity edges. On the other hand, fly protein CG4866-PA and yeast protein YHR148W do not share any sequence similarity with proteins in their own networks, but are connected through a highly homologous edge with each other. In a single species network it is often the case that proteins are surrounded only by proteins whose functional information is absent or very limited. In such cases, using information from multiple species becomes crucial: neighborhoods of many proteins are expanded by connecting them to proteins of high sequence

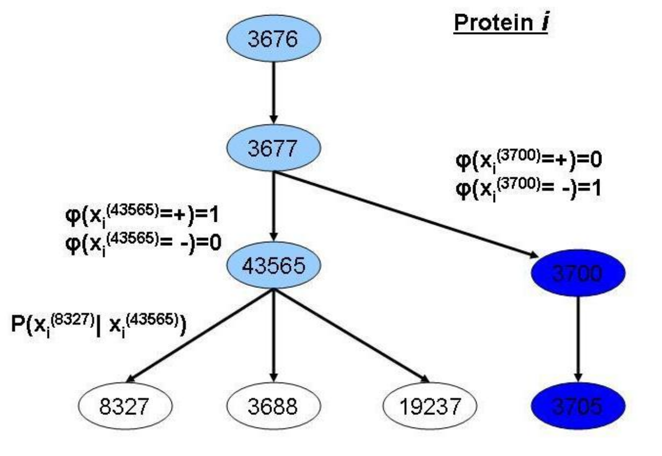


Fig. 2. The hypothetical protein is positively annotated (light blue color) to GO term 43565 and, thus, also positively annotated to its parent - GO term 3677 , and further up the tree to the parent's parent, term 3676. The term 3700, with the darker blue shade, indicates the negative annotation of the protein to this term. Its child, term 3705, inherits this negative annotation. The protein is unknown at the three unshaded (white) terms.

similarity in a different species' network. Through such multi-species networks sufficient information may be accumulated to improve the accuracy of protein functional prediction.

3 Methods

3.1 Single Species Network

In our work, we employ the idea of probabilistic chain graphs with incorporated Gene Ontology dependencies [8] to build protein network for each species (such as Yeast and Fly).

In this method, each protein is represented not by a single node, but by a replicate of a Gene Ontology (or subontology), as depicted in Figure 2. Gene Ontology (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 space corresponding to each of the functional terms in the Gene Ontology. The annotations can, in turn be, assigned positive, negative or unknown states.

Because the relationships between children and parents are directional, if a protein is positively annotated to a child, it is also, by definition, positively annotated to a parent. However, the reverse relationship does not hold. At the same time, if a protein is negatively annotated to a parent term, it will be negatively annotated to all the children terms.

From the above definition it becomes clear that the probability that the child term is negative, given that the parent term is negative, is one. In the presence of multiple parents, a negative state of any parent immediately yields a negative state for child. This step leaves the only probabilities that remain to be estimated as those that define the likelihood of a child being positively/negatively annotated when its parent is (or all parents are) positive.

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

By defining such probabilistic dependencies for the Gene Ontology terms (conditional probability distribution of all child terms given their parent terms), we create a Bayesian network (BN) representation for each protein, as represented in Figure 2.

We encode the ability of our model to transfer functions among similar proteins using a probabilistic graphical representation of a Markov Random Field (MRF) [12], similarly to [17], [10], [11]. In our work we consider two measures of similarity within each species network: sequence similarity determined through normalized BLAST scores and protein-protein interactions. The notion of similarity between proteins in this case is not directional, unlike the case of Gene Ontology.

For each measure of similarity we define a potential function, which corresponds to the probability of joint annotation of two proteins at a term, given that the proteins are similar. The sequence similarity-based potential for proteins i and j at term c is defined as

$$\begin{aligned}\psi(+,+) &= \psi(-,-) = s_{i,j,c}^{\text{within}}\\ \psi(+,-) &= \psi(-,+) = 1 - s_{i,j,c}^{\text{within}}\end{aligned}$$

where $s_{i,j,c}^{\text{within}}$ is a pairwise normalized BLAST score (we only consider normalized BLAST scores above 0.5). In this case, $s_{i,j,c}^{\text{within}} = s_{i,j}^{\text{within}}$ for all terms c.

Similarly, the PPI-based potential is defined in a term-specific way as shown below

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

where the quantities are estimated using relative frequency counts from the training data.

If both the similarity measure and the PPI occurred between a pair of proteins, the total potential ψ is defined as a product of the similarity-based potential and the PPI-based potential [8].

In the model, each protein i can have the evidential function ϕ at each term c, defined as follows. Let $x_i^{(c)}$ be the positive or negative annotation of a protein i to a particular term c. Then the evidential function models our knowledge of particular term annotations: a positively annotated protein at term c is modeled with $\phi(x_i^{(c)})$ defined as $\phi(+) = 1, \phi(-) = 0$. Similarly, when a protein is negatively annotated at c, the zero and one values are interchanged so that $\phi(+) = 0, \phi(-) = 1$. For proteins with no annotation the evidence ϕ is set to 0.5.

Our final model is embodied in a chain graph [16], a hybrid between a Bayesian Network (BN) and a MRF, see

$$P\left(\left\{x_i^{(c)}\right\}_{c\in GO, i\in\mathcal{I}_{\text{species 1}}\cup\mathcal{I}_{\text{species 2}}}\right) = \frac{1}{Z}\prod_{c\in GO}\prod_{i\in\mathcal{G}^{MRF(\text{species 1})}}\phi(x_i^{(c)})\prod_{i\in\mathcal{G}^{MRF(\text{species 2})}}\phi(x_i^{(c)})$$
$$\prod_{(i,j)\in\mathcal{G}^{MRF(\text{species 1})}}\psi_{within}(x_i^{(c)},x_j^{(c)}|\theta_{i,j}^{MRF(\text{species 1})})\prod_{(i,j)\in\mathcal{G}^{MRF(\text{species 2})}}\psi_{within}(x_i^{(c)},x_j^{(c)}|\theta_{i,j}^{MRF(\text{species 2})}) \tag{2}$$
$$\prod_{(i,j)\in\mathcal{G}^{MRF(\text{species 1}\cap\text{species 2})}}\psi_{between}(x_i^{(c)},x_j^{(c)}|\theta_{i,j}^{MRF(\text{species 1}\cap\text{species 2})})\prod_{i\in\mathcal{I}}P(x_i^{(c)}|Pa(x_i^{(c)}),\theta_c^{GO}),$$

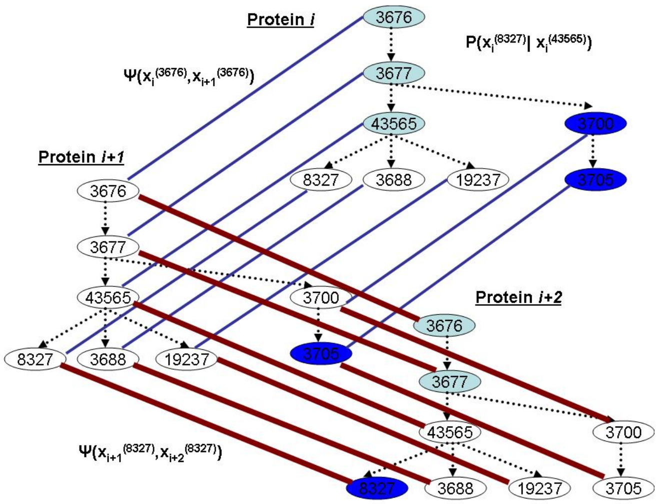


Fig. 3. A chain graph model with three proteins. Each protein is represented by GO subontology of size eight, with different annotations present at each protein. Some model elements, P and potential function ψ, are shown.

Figure 3. Operating all of the above parameters, the single-species model (of either Fly or Yeast, in our case) can now define a joint Gibbs distribution of functional term annotations over a set of proteins, as defined in Equation (1),where Z is the normalizing constant and $Pa(x_i^{(c)})$ is a parent (parents) of the GO term c in the protein x_i.

Once the network (chain graph) is built, the information is passed from annotated proteins through undirected links to their neighbors. At the same time the information flows within each protein's Bayesian network along the directed links, according to the conditional probabilistic relationships among different terms. In this fashion the annotation information is accumulated both via the similarity MRF and the ontology BN. For each term of a protein, a set of neighbors is defined by the local connectivity: for example, in the Figure 3 the neighbors of 3688 in the protein $i+1$ are $x_{i+1}^{(43565)}, x_i^{(3688)}, x_{i+2}^{(3688)}$.

The flow of information is modeled using a message-passing mechanism for chain graphs, similar to that described in [8]. Messages are passed until the state of convergence is reached; we define it as state at which all normalized messages change by less than 10^{-4} between successive iterations. We employ the "down" message-passing 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 their messages out.

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. We compare the beliefs, obtained thus, to a preselected threshold. Prediction decisions are based on 0.8 decision threshold, as suggested in [8], [17].

3.2 Multi-species network

Our next step is to join networks of two (or more) species by edges of high sequence similarity into one computational model. In particular, an edge is introduced between homologous proteins in two species if their normalized BLAST score is above 0.5 (the similarity is high). On the other hand, inter-species edges are not introduced when the score is below 0.5 (the similarity is low), since dissimilar proteins may or may not be involved in the same biological process. Moreover, most of the protein pairs would share some low similarity, which would obscure the network with potentially irrelevant low-similarity edges.

More formally, 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 $s_{i,j,c}^{\text{between}}$, a normalized pair-wise BLAST score. Consequently, the potential function for homologs between different species is defined as

$$\begin{aligned}\psi(+,+) &= \psi(-,-) = s_{i,j,c}^{\text{between}}\\ \psi(+,-) &= \psi(-,+) = 1 - s_{i,j,c}^{\text{between}}\end{aligned}$$

Similarly to a single-species model, we consider $s_{i,j,c}^{\text{between}} = s_{i,j}^{\text{between}}$ for all terms c of the Gene Ontology, as illustrated in Figure 3. While this assumption may be open to debate, it is shown to lead to improved annotation performance. Considering heterogeneous values of similarity $s_{i,j}^{\text{between}}$ at each term c may lead to additional improvements, at a cost of a more complex and demanding parameter estimation process.

The combined model for joint Fly-Yeast (referred to as species 1 and species 2) network now defines a joint Gibbs distribution of functional term annotations over a set of all proteins in the chain graph, detailed in Equation (2). Here, Z is the normalizing constant, ψ_{within} is similarity measure within one species network, $\psi_{between}$ is a similarity measure between the networks, $Pa(x_i^{(c)})$ is a parent (parents) of the GO term c in the protein x_i.

After the joint network is built, the belief propagation is used to make predictions at all ontology terms in both species. We consider a state of convergence and decision thresholds to be defined similarly to a single-species network.

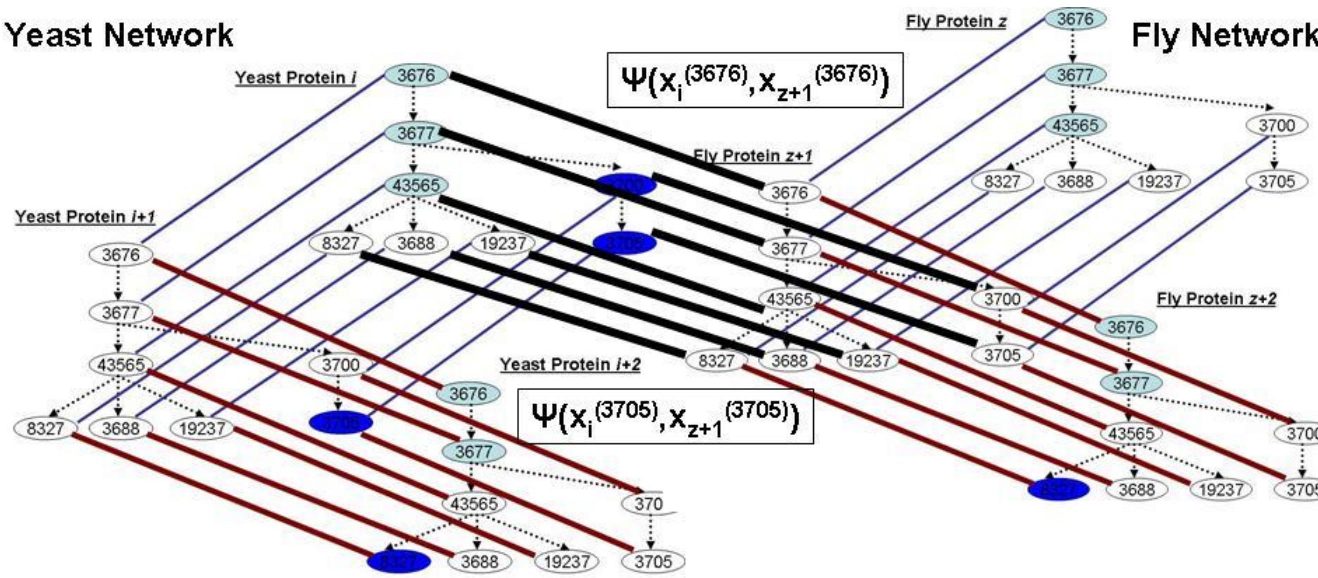


Fig. 4. Yeast and Fly networks joint by the similarity edges between Yeast's protein *i* and Fly's protein *z+1*. The edges between all GO terms of these proteins are in dark bold, with ψ shown.

Adding inter-species homology information into the learning model has unique advantages and shows significant improvements in protein function prediction. The model is specifically beneficial for proteins isolated in their own networks (having no interaction neighbors) or for proteins which are surrounded by poorly annotated neighbors. In a multi-species setting, the neighborhood of such proteins is expanded so that they can learn their functional annotations from their homologs in the different species.

4 Experiments and Results

4.1 Experiment design

We apply our method to two largest protein networks of Yeast and Fly as well as to a joint Yeast-Fly network. Predictive performance of our models is evaluated in a 5-cross validation setting. The test set consists of a random 20% of annotated proteins, that maintains the same proportion of negatively and positively annotated proteins as the remaining 80% of the data used for training the model. For each randomly chosen test protein, *all* of its annotations are left out—the Gene Ontology structure remains in place but the functions at all terms are now listed as unknown. In the case of a joint Fly-Yeast network, 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.

For each tested network, 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 are trained and evaluated on the same training/testing data.

For the measure of intra- and inter-species similarity we used normalized BLAST scores, defined as a BLAST score divided by self score of query (i.e. BLAST score of the homologue divided by the BLAST score of the protein against itself), ranging from 0 to 1. We obtained sequence and annotation data from Saccharomyces genome Database [3] for Yeast (February 2 and April 11, 2009 release) and FlyBase [1] for Fly (April 27, 2009 release). Protein-protein interaction data were obtained from BIOGRID [7] database (April 27, 2009 release). We considered only manual (higher quality) annotations, since computational predictions have been noted to present a conflicting evidence. To expand the applicability of our method, we considered *all* reported in the above sources Yeast and Fly proteins (as opposed to considering only proteins with specific evidence, such as protein-protein interactions). This resulted in a combined set of 12199 Fly and 6008 Yeast proteins that were used to construct our joint belief networks.

Gene ontology structure was downloaded from the Gene Ontology database [2]. When reading Gene Ontology annotations, we consider two fundamental GO assumptions: GO hierarchy is expanded up for positively annotated proteins (if a proteins is positively annotated to a term, then it is also positively annotated to all of its parents/ancestors) and is expanded down for negatively annotated proteins (if a protein is negatively annotated to a term, then it is negatively annotated to all of its children/descendants). We construct a negative set relying on co-annotation (co-occurence) statistics of GO annotations in the data (further maintaining two fundamental GO assumptions). In particular, a protein is considered negatively annotated to a specific GO term if this term has never been observed to co-occur with a known function for this protein in the training data.

Our example of gene ontology was taken from molecular

	size	precis.	recall	accur.	FP rate	F1
8	Fly	100	99.78	99.86	0	99.89
	Fly, JN	100	99.78	99.86	0	99.89
12	Fly	99.00	99.40	99.25	0.90	99.23
	Fly, JN	99.36	99.33	99.37	0.60	99.34
16	Fly	98.44	98.93	98.75	1.25	98.68
	Fly, JN	99.20	98.25	98.95	0.625	98.72

TABLE 1
Average precision,recall,accuracy, false positive rate, and F1 over 10 runs for **Fly** species in isolated Fly and joint Fly-Yeast networks (percentage wise) for subontologies of various sizes. JN stands for joint Yeast-Fly network.

	size	precis.	recall	accur.	FP rate	F1
8	Yeast	89.52	97.66	91.13	29.32	93.41
	Yeast, JN	100	96.17	97.27	0	98.05
12	Yeast	94.98	97.05	95.27	7.24	95.96
	Yeast, JN	98.33	96.94	97.33	2.18	97.63
16	Yeast	95.06	96.31	95.54	4.9	95.64
	Yeast, JN	99.01	95.6	97.7	0.465	97.26

TABLE 2
Average precision,recall,accuracy, false positive rate, and F1 over 10 runs for **Yeast** species in isolated Yeast and joint Fly-Yeast networks (percentage wise) for subontologies of various sizes. JN stands for joint Yeast-Fly network.

function subtree of GO hierarchy [1], as depicted in Figure 2. As previously investigated in [17], [15], [8], [4], [20] among others, PPI networks have strong predictive power for molecular function categories of Gene Ontology, especially in combination with other sources of evidence (such as intra- and inter- species homology). Previously PPI and intra-species sequence homology *together* showed significant improvements in predicting molecular functions of proteins, as for example shown in [8], [20], [4]. Most importantly, the use of the proposed *inter-species homology* may render our computational method, a core concept of this work, broadly applicable to all three ontologies: molecular function, biological process and even the cellular component.

Our method can be applied to the entire gene ontology, at the expense of time and space complexity. However, in practice, biologists and clinicians are interested in *specific*, relatively small, subontologies, targeted in our study. For instance, vaccine and drug targets are usually the proteins that perform very specific functions, represented by the leaves of a specific Gene subontology.

583 Fly and 236 Yeast proteins are annotated to one or more terms of the selected subontology (among those 110 Fly and 29 Yeast proteins were assigned some negative annotations). Other proteins are unannotated to a given subontology and are used as intermediate points for information passage.

4.2 Results

For our model, we operate several performance measures, such as: precision, recall, accuracy, false positive rates, and F1 defined as: $recall = \frac{TP}{TP+FN}$, $precision = \frac{TP}{TP+FP}$, $accuracy = \frac{TP+TN}{TP+TN+FP+FN}$, $fpr = \frac{FP}{TN+FP}$, $F1 = \frac{2*precision*recall}{precision+recall}$, respectively.

The calculations are done separately for the Yeast network, the Fly network and the joint Fly-Yeast network. In the joint network, we separately calculate the performance of Fly and Yeast species and compare them to those in isolated networks.

In this work, we consider GO subontologies of different sizes. The main focus is on the GO subontology of size 8, similarly to our previous work in [5]. We expand our model to subontologies of bigger sizes: 150% the size of the original subontology (size 12) and 200% the size of the original ontology (size 16), shown in Figure 5. A typical run of the model with the 8-sized ontology on the joint Fly-Yeast network (on 3.6 GHz CPU with 8GB memory machine) takes approximately 28 minutes (with four iterations of message passing). In comparison, corresponding runs on individual species networks take 59 minutes for Fly and 35 minutes for Yeast.

While the difference in running times may at first appear to go against intuition, faster convergence rates in a Joint Network can be contributed to the presence of "denser" sources of evidence in networks of multiple species compared to that of the isolated runs.

Table 1 shows the average precision, recall, accuracy and false positive rate for Fly: in isolated Fly network, and in joint Fly-Yeast network, for subontologies of various sizes. Table 2 shows corresponding measures for Yeast.

The overall performance of Fly and Yeast networks is highly improved (compared to the results presented in our previous work [5]), which is most likely due to the more reliable sequence similarity scores, expanded protein coverage, and more general definition of a negative set.

The joint Fly-Yeast network significantly improves precision, accuracy, and FP rate while only slightly suffering from lowered recall, as shown in Table 1, for Fly, and Table 2, for Yeast. We stress the importance of F1 measure, a harmonic mean of precision and recall, and notice its consistent significant improvement in the joint network, even for larger subontologies. This indicates that despite the larger size and more complex structure, considering networks of multiple species jointly continues to offer important benefits to the prediction process.

4.3 Statistical analysis

Statistical analysis of significance of the aforementioned performance scores was done using the t-test and the Wilcoxon Signed-Ranks Test [9]. 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; and single Yeast network is compared

1. The original GO subontology covered eight terms: nucleic acid binding (3676), DNA binding (3677), sequence-specific DNA binding (43565), methyl-CpG binding (8327), DNA replication origin binding (3688), centromeric DNA binding (19237), transcription factor activity (3700), and RNA polymerase II transcription factor activity, enhancer binding (3705).

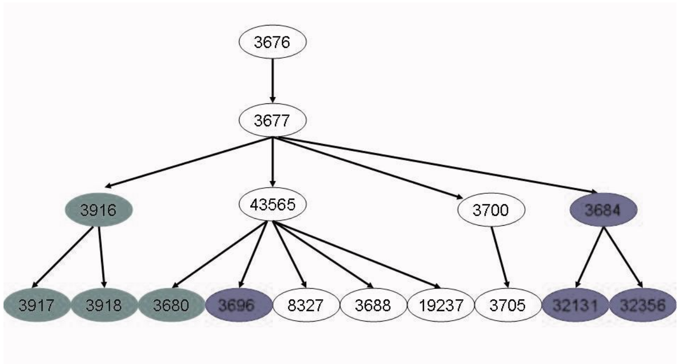


Fig. 5. Expanded subontologies of size 12 (added nodes are shown in gray) and 16 (added nodes are shown in black).

		precis.	recall	accur.	FP rate	F1
8	Fly, t-test	*	*	*	*	*
	Fly, WSR	*	*	*	*	*
	Yeast, t-test	$3 * 10^{-6}$	-	$2.1 * 10^{-5}$	$< 10^{-6}$	$4 * 10^{-6}$
	Yeast, WSR	0.0027	-	0.0046	0.0027	0.0029
12	Fly, t-test	0.22	-	0.35	0.20	0.35
	Fly, WSR	0.024	0.078	0.024	0.024	0.012
	Yeast, t-test	0.016	-	0.019	0.018	0.018
	Yeast, WSR	0.014	0.42	0.014	0.061	0.0053
16	Fly, t-test	0.11	-	0.33	0.10	0.47
	Fly, WSR	0.016	0.003	0.003	0.016	0.11
	Yeast, t-test	0.021	-	0.026	0.011	0.08
	Yeast, WSR	0.016	0.016	0.016	0.016	0.017

TABLE 3
p-statistics from t-test and Wilcoxon Signed-Ranks Test: p-values with respect to precision, recall, accuracy, false positive rate, and F1 as a measure of statistically significant improvements of a joint network performance, for subontologies of various sizes. "*" stands for "cannot be improved".

with the performance of the Yeast in the joint Fly-Yeast network. 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).

4.3.1 t-statistics per species

We present p-values calculated from t-statistics (degree of freedom= 9) to evaluate statistical significance of our findings in Table 3. We consider p-value to be statistically significant if it is less than 0.05. In general, Yeast shows more substantial improvements compared to Fly, which could indicate the higher quality of Fly data and better neighborhoods for the majority of Fly proteins.

4.3.2 Wilcoxon signed-ranks test

To remove the possible effects of outliers on the computed t-test statistics random samples can compensate for overall bad performance) we applied the Wilcoxon Signed-Ranks Test. Wilcoxon Signed-Ranks Test is a non-parametric alternative to the t-test, which assumes commensurability of differences in a qualitative way: greater differences count more. In many cases, this test is safer than the t-test since it does not assume a normal distribution.

Let $d_q = E_{c_q^1} - E_{c_q^2}$ be the difference between the performance scores of the approaches on the q-th out of the 10 random trials. Each difference is considered at its absolute value and the values are ranked. In the case of ties between differences, the average score among them is assigned. We use R^+ to denote the sum of ranks for the samples on which the Joint method outperforms the individual network approach; R^- is the sum of ranks when the individual methods "win":

$$R^+ = \sum_{d_q>0} rank(d_q) + \frac{1}{2}\sum_{d_q=0} rank(d_q)$$

$$R^- = \sum_{d_q<0} rank(d_q) + \frac{1}{2}\sum_{d_q=0} rank(d_q)$$

The z-statistic can be calculated as

$$z = \left(T - \frac{1}{4}N(N+1)\right) / \sqrt{\frac{1}{24}N(N+1)(2N+1)},$$

where $T = min(R^+, R^-)$. and$N = 10$ is the number of samples. With $\alpha = 0.05$, the null hypothesis will be rejected if $z < -1.96$. We calculate the corresponding p-values from the determined z-values.

The Wilcoxon test similarly confirms significant improvements in performances on the Joint network when compared to individual Yeast and Fly networks, as shown in Table 3. In fact, Wilcoxon test "catches" statistically significant improvements where t-test presents no evidence, such as for subontologies of size 12 and 16.

5 Comparative analysis

5.1 Gene Ontology vs single-term predictions

As a baseline test, we compare our methodology (with GO dependencies) to runs without GO in place, where the whole network of proteins is tested on a single ontology term (single protein function). As before, we perform 5-fold cross validation by choosing random 20% of annotated proteins as a testing set over 10 trials of the program. The results shown in Table 4 indicate the superiority of the network with built-in Gene Ontology over the single-term network even in the case of multiple species networks.

It is worth mentioning that the model with gene ontology in place makes a true positive prediction where the model without it commits a false negative error. This result is not surprising as there is only one term with one protein annotated to it. In general, similar to [8], 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.

5.2 Comparison with other methods

In this section we comprehensively compare our method to the most widely used group of techniques, such as in Nariai

networks		precision	recall	accuracy	FP rate
Fly	w/o GO	45.57	48.7	74.25	49.05
	GO	100	99.78	99.86	0
Fly | JN	w/o GO	49.5	53.78	54.94	32.13
	GO	100	99.78	99.86	0
Yeast	w/o GO	-	0	43.79	0
	GO	89.52	97.66	91.13	29.32
Yeast | JN	w/o GO	34.76	70.52	72.10	54.1
	GO	100	96.17	97.27	0
JN overall	w/o GO	44.36	59.63	60.9	39.81
	GO	100	98.70	98.98	0

TABLE 4
Comparison of results for the network with GO and without GO

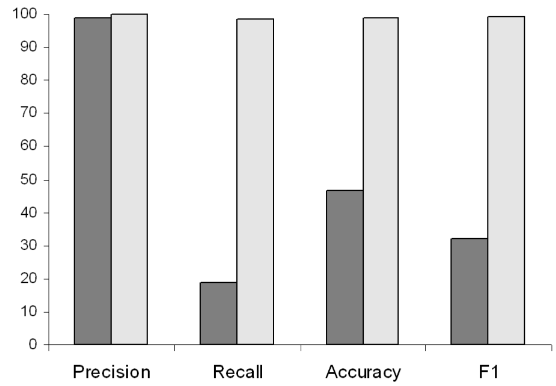


Fig. 7. Comparison of our method (light gray) to Nariai et. al. [21] (dark gray): overall performance of a joint network, %-wise precision, recall, accuracy, F1 rate.

et. al. [21], which are based on Bayesian probabilistic approaches. In such methodologies, proteins are embedded into protein-protein networks so that each protein is represented by a node and similarity measures between proteins (such as protein-protein interactions, sequence similarity, etc.) are represented by edges. In the model, each protein learns its functional annotation based on the number and character of his neighbors in the protein network, particularly the *total* number of neighbors and the number of *annotated* (to the GO term of interest) neighbors. This information is then embedded into a probabilistic Bayesian framework, which consequently assigns a probability to a protein of interest as positively or negatively annotated to a specific GO term [21]. Since fundamentals of Bayesian probabilistic approach are at the heart of the overwhelming majority of methods currently used for protein function prediction, we compare ourselves against this computational technique.

To achieve the most accurate comparative results, we use the same 10 training/testing sets as in our own experimental studies in a 5-fold cross validation setting. Similarly to our setting, both PPI and Sequence similarity (determined by normalized BLAST cores) are used to build protein interaction networks.

We present results as individual performance of Yeast and Fly species in the joint network (Figure 5), as well as the overall performance (Figure 6) in the joint network. The results indicate that our method performs better than the Bayesian probabilistic approach of Nariai et al [21] across F1, precision, recall, and accuracy scores, for all validation sets considered. We observed that the method of Narai et al. tends to produce higher rates of false negative predictions, resulting in lower recall rates. At the same time, the false negatives are correctly identified as positive using our approach. This improved performance can be attributed to the expanded neighborhood definition, endowed by the GO structure, which is not explicitly used in the approach of [21].

Interestingly, Fly species achieves precision of 100% and the false negative rate of 0 in the method of Nariai et. al. and for the subontology of size 8 in our proposed method. This may indicate that higher quality data was used to build the Fly protein network and that good learning neighborhood are induced for the majority of Fly proteins, which may not hold for the Yeast network. As for Yeast species, our method shows FP rate of 0 while the method of Nariai et. al. shows FP rate of 1.41%.

6 Conclusions

In this work we presented a novel approach that uses inter-species sequence homology to connect networks of two, and possibly more, species together with Gene Ontology dependencies in order to improve the predictive ability needed for protein classification. Joining the networks of two different species shows important advantages over runs on individual networks. While in single species networks 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 Gene Ontology 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 may produce improvements similar to the case of two species. The presence of multiple interacting networks may further enable integration of additional sources of evidence, thus contributing to increased accuracy of functional predictions.

References

[1] http://www.flybase.org/.
[2] http://www.geneontology.org/.
[3] http://www.yeastgenome.org/.
[4] Mitrofanova A., Kleinberg S., Carlton J., Kasif S., and Mishra B. Systems biology via redescription and ontologies (iii): Protein classification using malaria parasite's temporal transcriptomic profiles. *IEEE International Conference on Bioinformatics and Biomedicine*, pages 278–283, 2008.
[5] Mitrofanova A., Pavlovic V., and Mishra B. Integrative protein function transfer using factor graphs and heterogeneous data sources. *IEEE International Conference on Bioinformatics and Biomedicine*, pages 314–318, 2008.
[6] Engelhardt BE, Jordan MI, Muratore KE, and Brenner SE. Protein molecular function prediction by bayesian phylogenomics. *PLoS Comput Biol*, 1(5):e45, 2005.

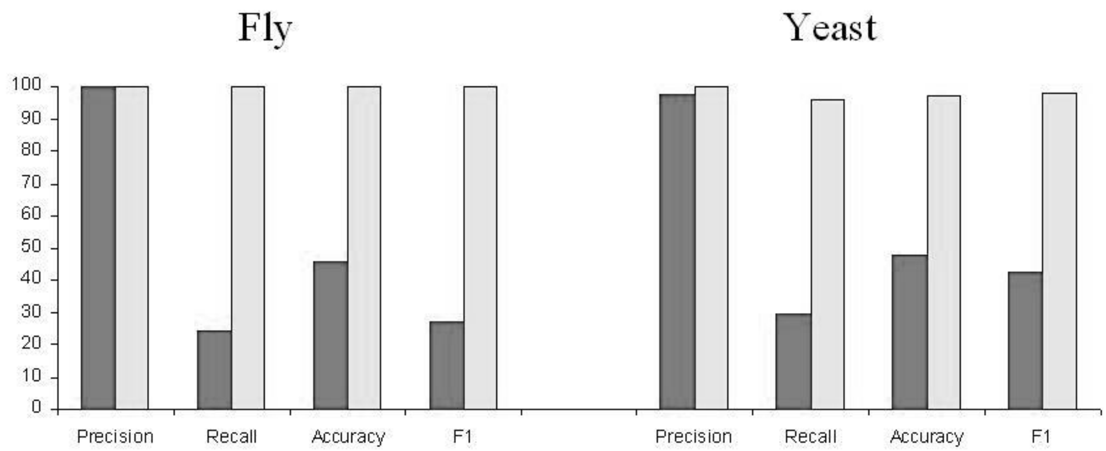


Fig. 6. Comparison of our method (light gray) to Nariai et. al. [21] (dark gray): performance of Fly and Yeast species in a joint Fly-Yeast network, %-wise precision, recall, accuracy, F1 rate.

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

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

[9] J Demsar. Statistical comparison of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[10] M Deng, T Chen, and F Sun. An integrated probabilistic model for functional prediction of proteins. In *Proceedings of the Seventh InternationalConference on ComputationalMolecularBiology (RECOMB)*, pages 95–103, 2003.

[11] 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.

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

[13] T Hawkins, S Luban, and S Kihara. Enhanced automated function prediction using distantly related sequences and contextual association by pfp. *Protein Sci*, 15:1550–1556, 2006.

[14] Edmund M. Clarke Jr, Orna Grumberg, , and Doron A. Peled. *Model Checking*. The MIT Press, 1999.

[15] 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.

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

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

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

[19] DM Martin, M Berriman, and GJ Barton. Gotcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinformatics*, 5:178, 2004.

[20] Yosef N., Sharan R., and Stafford Noble W. Improved network-based identification of protein orthologs. *Bioinformatics*, 24(16):i200–i206, 2008.

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

[22] 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.

[23] B Schwikowski, P Uetz, and Fields. S. A network of protein-protein interactions in yeast. *Nat Biotechnol*, 18:1257–1261, 2000.

[24] H Shin, AM Lisewski, and O Lichtarge. Graph sharpening plus graph integration: a synergy that improves protein functional classification. *Bioinformatics*, 23:3217–3224, 2007.

[25] K Tsuda, HJ Shin, and B Scholkopf. Fast protein classification with multiple networks. *Bioinformatics*, 21:ii59–ii65, 2005.

[26] A Vinayagam, R Konig, J Moormann, R Schubert, F ad Eils, K-H Glatting, and Suhai S. Applying support vector machines for gene ontology based gene function prediction. *BMC Bioinformatics*, 5:116, 2004.

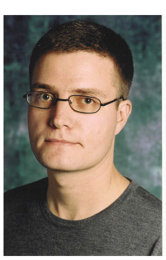
[27] J Whisstock and A Lesk. Prediction of protein function from protein sequence and structure. *Quarterly Review of Biophysics*, 36:307–340, 2003.

APPENDIX

A Model Checking Interpretation. Our expanded Gene Ontology approach can also be interpreted as a special case of a new broader framework of "probabilistic graphical model checking." The framework resembles classical model checking algorithms [14] implemented through message passing in a statistical graphical model. This connection becomes explicit when a Gene subontology for a protein (Figure 2) is viewed as a family of properties encoded through logical propositions and connectives. Also modal operators and quantifiers may be added, if further generalizations are desired. 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 example, in the currently used Gene subontology, the positive information about a child implied positive information about a parent; and negative information about a parent implied negative information about child. Additionally, we define a probability for a child being positive/negative given that a parent is positive, which defines a probabilistic framework for the model. Thus, if we view our graphical model as not strictly related to a GO subontology, but to a more general framework such as this, we can define any set of properties on the elements of this graphical structure, introduce time frames, or imply hierarchical relationships for this graph. Once we define relationships/properties, we can then propagate these properties in the entire model (which in our application, corresponds to message passing).

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. Our approach suggests, for propositional and temporal logic, a potentially much broader range of applications including many non-biological problems.

Antonina Mitrofanova Antonina Mitrofanova did this work when she was pursuing her PhD in Computer Science at New York University, Courant Institute of Mathematical Sciences, under the supervision of Prof. Bud Mishra. She received her PhD in September 2009 and currently is a PostDoctoral Computing Innovation Fellow at Columbia University. Before coming to NYC, Antonina spent four years studying medicine at the National Medical University in Kiev, Ukraine. Her research interests include probabilistic models and algorithms applied to protein and gene networks as well as transcriptional and post-transcriptional networks in cancer. Antonina received several prestigious awards and scholarships, e.g., PostDoctoral CIFellow award, Sandra Bleistein Prize, NSF Graduate Research Fellowship Honorable Mention List, CRA-W distributed mentor project award, Stewart M. Monchik Memorial Scholarship in Computer and Information Science, and Jack Wolfe Award in Computer and Information Science.



Vladimir Pavlovic Vladimir Pavlovic is an Associate Professor in the Computer Science Department at Rutgers University. He received the PhD in electrical engineering from the University of Illinois in Urbana-Champaign in 1999. From 1999 until 2001 he was a member of research staff at the Cambridge Research Laboratory, Cambridge, MA. Before joining Rutgers in 2002, he held a research professor position in the Bioinformatics Program at Boston University. Vladimir's research interests include probabilistic system modeling, time-series analysis, statistical computer vision and bioinformatics. He has published over 80 peer-reviewed papers in major computer vision, machine learning and pattern recognition journals and conferences.

Bud Mishra Prof. Bhubaneswar (Bud) Mishra is a professor of computer science and mathematics at NYU's Courant Institute of Mathematical Sciences, professor of human genetics at Mt. Sinai School of Medicine, and a professor of cell biology at NYU School of Medicine. Prof. Mishra has a degree in Physics from Utkal University, in Electronics and Communication Engineering from IIT, Kharagpur, and MS and PhD degrees in Computer Science from Carnegie-Mellon University. He is editor of Molecular Cancer Therapeutics, AMRX (Applied Mathematics Research Exchange), Nanotechnology, Science and Applications, and Transactions on Systems Biology, and author of a textbook on algorithmic algebra and more than two hundred archived publications. He is an ACM fellow and a NYSTAR Distinguished Professor (2001). He also holds adjunct professorship at Tata Institute of Fundamental Research in Mumbai, India. From 2001-04, he was a professor at the Watson School of Biological Sciences, Cold Spring Harbor Lab.