

# Improved Network-based Identification of Protein Orthologs

Nir Yosef<sup>a,\*</sup>, Roded Sharan<sup>a</sup> and William Stafford Noble<sup>b</sup>

<sup>a</sup>School of Computer Science, Tel-Aviv University, Tel-Aviv 69978, Israel. <sup>b</sup>Department of Genome Sciences, Department of Computer Science and Engineering, University of Washington, Seattle, WA, USA.

\*To whom correspondence should be addressed.

## ABSTRACT

**Motivation:** Identifying protein orthologs is an important task that is receiving growing attention in the bioinformatics literature. Orthology detection provides a fundamental tool towards understanding protein evolution, predicting protein functions and interactions, aligning protein-protein interaction networks of different species and detecting conserved modules within these networks.

**Results:** Here, we present a novel diffusion-based framework that builds on the Rankprop algorithm for protein orthology detection and enhances it in several important ways. Specifically, we enhance the Rankprop algorithm to account for the presence of multiple paralogs, utilize protein-protein interactions, and consider multiple (>2) species in parallel. We comprehensively benchmarked our framework using a variety of training data sets and experimental settings. The results, based on the yeast, fly and human proteomes, show that the novel enhancements of Rankprop provide substantial improvements over its original formulation as well as over a number of state of the art methods for network-based orthology detection.

**Availability:** Data sets and source code are available upon request.

**Contact:** niryosef@post.tau.ac.il

## 1 INTRODUCTION

The notion that similar protein sequences imply similar protein functions has been traditionally employed to guide the identification of *orthologous* protein pairs [Brenner, 1999], *i.e.*, proteins in different species that evolved from a common ancestor. However, one problem with identifying orthologs by sequence similarity arises when the protein in question has similarity to not one but many paralogous proteins [Sjolander, 2004]. In these cases, every cross-species protein pair is technically orthologous, but it is still necessary to distinguish which protein pairs play functionally equivalent roles [Remm et al., 2001].

To date, one of the most successful paradigms for orthology identification uses a combination of local and global network properties. In this approach, the similarity of two proteins is determined by two factors: the similarity of their sequences (the local property) and the similarities among their neighbors in some network structure (the global property). One example is the Rankprop algorithm which identifies protein homologies by performing a diffusion operation in a protein sequence similarity network. Rankprop was shown to outperform iterative protein database search methods such as PSI-BLAST [Altschul et al., 1997], which considers only a limited proportion of protein similarities at a time (*i.e.*, focusing on the local property).

Subsequent studies based on these principles have extended the network structure used to quantify the global property by incorporating protein-protein interaction (PPI) information. Bandyopadhyay et al. [2006] used a Markov random field (MRF) formulation to derive probabilities for orthologies based on conserved PPI patterns. The ISORank algorithm [Singh et al., 2007] assigns similarity scores to pairs of proteins according to a random walk in the product graph of the two networks. These scores (or probabilities) then serve as a basis for orthology detection; in MRF, the derived probabilities were used to resolve ambiguous orthology predictions made by the Inparanoid algorithm [Remm et al., 2001]. In ISORank the scores were used to create a global alignment of the yeast and fly networks.

In this paper, we extend the Rankprop algorithm in three important ways and test the impact of these extensions on the accuracy of the obtained similarity scores in comparison to the original Rankprop formulation as well as to the two PPI-based approaches (MRF and ISORank). First, we modify Rankprop to include PPI information in addition to sequence similarity data. This variant, which we call *hybrid Rankprop*, combines the two types of information into a unified hybrid network and learns the weight of each of the factors in a supervised manner. Second, we present an improved version of Rankprop, termed *mutual Rankprop*, which explicitly accounts for the problem of multiple paralogs. Finally, we examine the utility of applying the algorithm to a network derived from more than two species.

Using the yeast and fly proteomes, we show that mutual Rankprop outperforms Rankprop, MRF and ISORank on the orthology detection task, and that incorporating the human network also improves the accuracy of the algorithm. The added value of including PPI information, however, remains unclear as we do not observe a significant improvement for hybrid Rankprop relative to the original Rankprop algorithm.

In addition to accurately identifying orthology relationships, the different Rankprop variants provide two other advantages compared with ISORank and MRF. First, the Rankprop methods can produce orthology predictions for any given protein. This is similar to ISORank but contrasts with the MRF, which is limited to proteins that participate in *at least one* conserved interaction, as defined in Section 3.4. For this reason, the MRF fails to make predictions for a substantial portion of the proteins in a given network.

Second, the Rankprop variants are efficient. The MRF method uses Gibbs sampling, the running time of which is difficult to characterize. ISORank's running time is  $O(E^k)$  where  $E$  is the number of edges in a network and  $k$  is the number of networks.

Thus, this algorithm scales exponentially with the number of tested networks. In contrast, the running time of the Rankprop variants scales polynomially:  $O(k^3V^2 + k^2VE)$  with PPI information and  $O(k^3V^2)$  without using PPI information, where  $V$  is the number of nodes in the network. Furthermore, in many applications we are only interested in a subset of the proteins, *e.g.*, when looking for orthologs of a specific pathway or complex. The different Rankprop methods can infer orthologies for a specified subset of  $m$  proteins in a substantially shorter time ( $O(mk^2V + mkE)$  or  $O(mk^2V)$  with or without using PPI information, respectively). In practice, ISORank's exponential running time prevents the algorithm from considering more than two species at a time. Instead, it uses a pairwise incremental procedure in order to account for multiple networks [Singh et al., 2008]. The Rankprop variants, on the other hand, can be applied to multiple networks simultaneously. As our results show, this capability provides an important advantage, because the confidence of an inferred homology between a pair of proteins can be enhanced by the existence of common orthologs in other species.

## 2 ALGORITHMS

We start by reviewing the Rankprop algorithm and then describe the three novel enhancements of the basic algorithm.

### 2.1 Rankprop

The input to the Rankprop algorithm is a weighted network and a designated query node within that network. The nodes in the network correspond to proteins from two or more species. Edges connect pairs of proteins that share sequence similarity. The initial weights of the edges are set using a pairwise sequence comparison algorithm. In this work, we use BLAST  $E$ -values [Altschul et al., 1990], transformed so as to represent transition probabilities (setting the sum of weights of incoming edges to 1 for all nodes). Rankprop assigns scores to all of the nodes in the network by using a diffusion procedure across the weighted network. During diffusion, the query node is assigned a score of 1.0, and this score is continually pumped to the remaining nodes by means of the transition matrix. Upon termination, every protein is assigned a score, determined by the steady state of the diffusion process. A higher score implies a higher level of similarity.

In the next section we present a generalization of Rankprop to integrate protein-protein interaction data. We defer the formal description of the normalization procedures and the diffusion algorithm to that section.

### 2.2 Hybrid Rankprop

In the hybrid version of Rankprop, the edges of the weighted network encode *two* types of relations between proteins: protein-protein interaction and pairwise protein similarities. Specifically, edges between proteins of the same species correspond to protein-protein interactions, and edges connecting proteins from different species represent sequence similarity relations. The weights on the edges reflect the level of confidence in the protein-protein interaction, or the degree of sequence similarity.

In a preprocessing stage, we convert the weights in the graph to transition probabilities. First, we represent the input graph as

two separate matrices,  $W_{sim}$  and  $W_{ppi}$ . As a measure of sequence similarity we use the BLAST  $E$ -value, where  $W_{sim}[i, j]$  is the BLAST  $E$ -value assigned to protein  $i$  when querying with protein  $j$  (*i.e.*, the score is normalized by the length of sequence  $j$  and the number of proteins in the network). For PPI based similarity, we set  $W_{ppi}[i, j]$  to the complement  $1 - c$  of the confidence score  $c$  assigned to the interaction between proteins  $i$  and  $j$  (see Section 3.1). We use the complement in order to conform with the sequence similarity scores, where a lower score indicates a stronger signal.

Next we construct a weight matrix  $W$ , which encodes the hybrid network, by transforming the weights in the input matrix. The transformation is applied separately for PPI edges and for protein similarity edges in the following manner. For the PPI edges we define:

$$W_{ij} = -\log\left(\frac{W_{ppi}[i, j]}{MAX_{ppi} \cdot \sigma_{ppi}}\right), \quad (1)$$

where  $MAX_{ppi}$  is the highest weight (or the lowest confidence) assigned to protein pairs in the hybrid network. A similar formulation is used for sequence similarities:

$$W_{ij} = -\log\left(\frac{W_{sim}[i, j]}{MAX_{sim} \cdot \sigma_{sim}}\right), \quad (2)$$

where  $MAX_{sim}$  is the highest  $E$ -value assigned to protein pairs in the hybrid network. The logarithmic transfer functions in Equations 1 and 2 introduce two parameters,  $\sigma_{ppi}$  and  $\sigma_{sim}$ , for PPI and protein similarity edges, respectively. These parameters control the importance of highly scoring protein pairs compared to pairs with weaker links.

Finally, the matrix  $W$  is normalized such that for each node, the sum of weights of incoming edges is 1 (*i.e.*,  $\forall j : \sum_i W_{ij} = 1$ ). The normalization procedure introduces an additional parameter  $\rho$ , which determines the relative weight of sequence- and interaction-similarity edges. For each node, the sum of incoming sequence similarity edges is  $\frac{\rho}{1+\rho}$ , and the sum of incoming PPI edges is  $\frac{1}{1+\rho}$ . Note that although PPI edges are not originally directed, we treat the PPI edges as directed because the normalized weights depend only on one end-point—see discussion in Section 2.3. The parameter  $\rho$  allows us to reformulate the original Rankprop as a special case of the hybrid Rankprop by setting  $\rho = \infty$ .

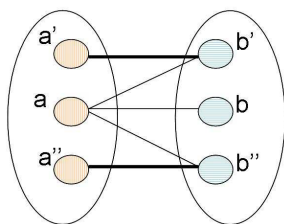
In the diffusion process, the query node continually pumps its score to the remaining nodes, this time by means of the weighted hybrid network. During the diffusion, a protein  $P$  pumps to its neighbors (either by homology or by PPI) at time  $t$  the linear combination of scores that  $P$  received from its neighbors at time  $t - 1$ , weighted by the strengths of the edges between them. The strength of the diffusion or the relative weighting between the local and global properties is controlled by an additional parameter  $\alpha$  (see Algorithm 1). For efficiency, the diffusion process is terminated after a fixed number of iterations, and the resulting diffusion values are used as an approximation to the ones we would obtain upon convergence. The output is a ranked list of putative homologs. Note that hybrid Rankprop can readily produce a second ranked list of putative interaction partners; however, in this study we concentrate only on homology detection. The algorithm is summarized as Algorithm 1.

**Algorithm 1 Pseudocode of the hybrid Rankprop algorithm.**

Given a set of proteins from two species, let  $W_{sim}$  and  $W_{ppi}$  denote the input weight matrices (encoding sequence similarity and protein interaction data, respectively),  $V$  denote the set of proteins in the hybrid network and  $q \in V$  denote the query protein. For each  $v \in V$  denote  $s(v)$  as the species to which the protein  $v$  belongs, and  $N(v)$  as the set of proteins adjacent to  $v$  in the hybrid network. In a preprocessing stage, the algorithm converts the input matrices into a single hybrid network using the procedure *makeTransitionMat*. The balancing parameter  $\rho$  determines the relative weight of sequence similarity and PPI edges, and the parameter  $\alpha$  controls the rate of the diffusion. The number of iterations is fixed to a constant  $L$ . Hybrid Rankprop produces a ranked list,  $rank_{hom}$  of candidate homologs of  $q$ .

```

1: procedure HYBRIDRANKPROP( $W_{sim}, W_{ppi}, q, s, \alpha, \rho$ )
2:    $W \leftarrow \text{makeTransitionMat}(W_{sim}, W_{ppi}, \rho)$ 
3:    $\gamma_q(0) = 1; \quad \forall i \neq q : \gamma_i(0) = 0$ 
4:   for  $t \leftarrow 1 \dots L$  do
5:     for  $i \in V$  do
6:        $\gamma_i(t+1) = \alpha \sum_{j \in N(i) \setminus q} W_{ji} \gamma_j(t) + \gamma_i(0)$ 
7:     end for
8:   end for
9:    $rank_{sim} \leftarrow \text{sort}(\{\gamma_i(t)\}_{S(i) \neq S(q)})$ 
10:  return  $rank_{sim}$ 
11: end procedure
    
```



**Fig. 1. An example for the utility of mutual scoring.** The two sets of nodes represent families of paralogs from two species. Thickness of edges reflects the magnitude of the respective Rankprop scores.

### 2.3 Mutual Rankprop

The ranking scheme applied by Rankprop is directional; the natural thing to ask then is whether it is also (at least to some extent) empirically symmetric. Note that the intrinsic lack of symmetry does not stem merely from the directionality of the BLAST scores but, more importantly, from the topology of the networks. Because the normalization of edge weights depends on the in-degree of the target node, two very different weights might be assigned to the same relation. The obvious example, presented in Figure 1, is in the case of duplication followed by divergence. In this case, a protein (denoted  $a$ ) in one network has multiple homology matches in the other network, representing a family of paralogs. Now assume that all the paralogs but one (denoted  $b$ ) have other clear homology matches (to paralogs of  $a$ ). Querying only from  $a$  will not be able to distinguish  $b$  from the rest, while querying from  $b$  and its paralogs will clarify that  $b$  is the most probable ortholog of  $a$ .

To exploit this asymmetry, we define the *mutual Rankprop* variant. This algorithm considers the top ranked candidates, applies a query from each of them, and then re-ranks them according to the corresponding scores assigned to the original query node. Formally, for a given query node  $q$  let  $a_1 \dots a_\mu$  be the top  $\mu$  proteins on the ranked list produced by Rankprop and  $s_1 \dots s_\mu$  be their scores (where  $\mu$  is a small constant, set here to 5). Mutual Rankprop calls Rankprop separately for every  $a_i$  as the query node. Let  $s_{q,i}$  be the score of  $q$  when querying from  $a_i$ . We rerank each of the  $a_i$ s based on a mutual score, which is defined as  $(s_i + s_{q,i})/2$ . Finally, the top  $\mu$  proteins are re-sorted according to this new score. In the following, we experiment with the mutual extension of both the original Rankprop and the hybrid Rankprop variants.

### 2.4 Using more than two networks

The extension of Rankprop or hybrid Rankprop to more than two networks is straightforward. For example, in the latter variant, for three species we have a total of six networks (three PPI and three sequence similarity networks). The edge weights are transformed using Equations 1 and 2 and then normalized according to the parameter  $\rho$ .

In this case, the output with respect to a given query will be two orthology lists, one for each of the remaining species. It is easy to see that in a given hybrid network, every edge is visited a constant number of times (equal to the fixed number of iterations), so the running time of a single query on a hybrid graph encompassing  $k$  PPI networks, each with  $E$  edges and  $V$  nodes is  $O(kE + k^2Vh)$ , where  $h$  is the maximum number of sequence similarity links per protein. In our application  $h$  is limited to a constant value ( $h < 100$ ). Querying from all the nodes in the hybrid network therefore costs  $O(kV(k^2V + kE)) = O(k^3V^2 + k^2VE)$  time. The original version of Rankprop does not use PPI edges and, in our experiments, does not use sequence similarity edges within a single species; therefore, Rankprop's running time reduces to  $O(k^3V^2)$ .

### 2.5 Tuning the parameters

The parameters  $\sigma_{ppi}$ ,  $\sigma_{sim}$ ,  $\rho$  and  $\alpha$  enable us to control the operation of the Rankprop variants either by selective tuning of the weights of the two types of edges ( $\sigma_{ppi}$ ,  $\sigma_{sim}$ ), by determining the overall balance between their influences ( $\rho$ ), or by determining the rate of the diffusion ( $\alpha$ ).

We learn the values of these parameters from a labeled training set via cross-validation and grid search. In our implementation, we use five-fold cross-validation, and we consider all combinations of the different parameter values. We based our search on a series of three exponents:  $\Sigma = \{0, 2, 5\}$ . For the two tuning parameters  $\sigma_{ppi}$ , and  $\sigma_{sim}$ , we consider the values  $10^i$ ,  $i \in \Sigma$ . For the balance parameter  $\rho$  we consider the values  $5^{\pm i}$ ,  $i \in \Sigma$ . Finally, for the diffusion rate parameter,  $\alpha$ , we consider a low value of 0.3 and a higher value of 0.95. For the original Rankprop implementation, we set  $\rho$  to  $\infty$  and examine the different values only for  $\sigma_{sim}$  and  $\alpha$ .

## 3 EXPERIMENTAL SETUP

### 3.1 The analyzed networks

Initially, we apply our orthology detection scheme to the networks of *Saccharomyces cerevisiae* and *Drosophila melanogaster*. The

data set is identical to the one used by two previous orthology detection studies [Bandyopadhyay et al., 2006, Singh et al., 2007]), and downloaded from the online supplement of [Bandyopadhyay et al., 2006]. It contains protein sequences of 5878 yeast and 18,746 fly proteins from FlyBase and SGD [Crosby et al., 2007, Christie et al., 2004]. PPI information in the data set is from the Database of Interacting Proteins [Xenarios et al., 2000] and includes 14,319 and 20,720 interactions for yeast and fly respectively. The downloaded data set also provides confidence values to each PPI edge based on a logistic regression model [Bader et al., 2004]. Importantly, homology-based data were not used when determining these values. This is a crucial point because in the following experiments we will use the confidence values, in cross validation, to train and test orthology predictors.

In a second experiment, we also add a human network. This network contains 7915 protein sequences and 28,972 interactions collected from recently published papers [Rual et al., 2005, Stelzl et al., 2005] and from the HPRD database [Peri et al., 2003]. PPI confidence values for the human network were assigned using a logistic regression model similar to the one used for the yeast and fly networks.

Protein similarities ( $E$ -values) were computed using BLASTP [Altschul et al., 1997], using a threshold of  $E$ -value < 10.

### 3.2 Training Data

We consider two distinct gold standard sets of positive and negative orthology relations. The first set is based on the Inparanoid program [Remm et al., 2001]. These labels were used for validation in two previous studies [Bandyopadhyay et al., 2006, Singh et al., 2007]. Specifically, the positive cases are drawn from Inparanoid homology clusters that contain only one representative from each species (unambiguous orthology). We consider two methods for defining the negative set. The more stringent, which we term the *specific* negative set is composed of pairs of proteins that are best BLAST matches not assigned to the same Inparanoid cluster. The second definition for the negative set is the complement of the positive set. The positive and negative orthology data sets were provided by the authors of [Bandyopadhyay et al., 2006].

The second training set of positive and negative orthology relations is taken from the HomoloGene database [Wheeler et al., 2003]. Importantly, the homology detection procedure of HomoloGene uses both proteins and their matching DNA sequences and relies on a global optimization rather than local. In addition, HomoloGene considers synteny when applicable. The positive training set is composed of the HomoloGene orthologous pairs. As negatives we take the top five nonorthologous BLAST matches. As before, we also use a second definition for the negative set as the complement of the positive set.

### 3.3 Performance evaluation

We measure the quality of a given orthology predictor by comparing it with the gold standard and computing a receiver operating characteristic (ROC) curve [Hanley and McNeil, 1982]. Our quality metric is the area under this curve (the ROC score). In orthology prediction applications, we are primarily interested in the top-ranked predictions. To account for that, we measure the area under the curve up to the first 50 false positives (ROC<sub>50</sub>) [Gribskov and Robinson,

1996]. The ROC<sub>50</sub> scores are computed separately for each query, taking the complement of the positive set as the negative set.

The previous procedure is *relative*, in the sense that targets for a particular query are only ranked relative to one another. A more stringent quality metric requires that the scores produced for different queries lie on the same scale. We measure *absolute* quality by sorting together the outputs from multiple queries and computing a single ROC curve. The computation of the absolute quality is based on the specific negative sets.

### 3.4 Alignment graph for MRF

The MRF method is based on an *alignment* between the two given PPI networks. The nodes in the alignment graph represent pairs of proteins, one of each species. An edge between two alignment nodes  $(u, v)$  and  $(u', v')$  exists if  $u$  interacts with  $u'$  and the distance between  $v$  and  $v'$  is no more than 2 (or vice versa). We first use the original alignment graph used by [Bandyopadhyay et al., 2006], whose nodes are defined according to the orthology clusters of Inparanoid. However, a problem with this graph is that the vast majority of the proteins appear in only one alignment node and thus have only one candidate for orthology (in fact, this holds, by definition, for all positive pairs in the Inparanoid gold standard). Consequently, the relative performance of the algorithm could not be assessed. To circumvent this problem, we used a second alignment graph defining the alignment nodes according to the five top mutual BLAST matches of each protein. For a fair comparison, in this experiment we also limit the sequence similarity information available to Rankprop and ISORank to the five top mutual BLAST matches of each protein.

## 4 RESULTS

In the following we use the yeast and fly data to compare four Rankprop variants—the original Rankprop algorithm, mutual Rankprop, hybrid Rankprop, and mutual hybrid Rankprop. The variants are compared to each other and to the MRF method of [Bandyopadhyay et al., 2006], ISORank [Singh et al., 2007] and BLAST [Altschul et al., 1997]. In addition, we apply the original Rankprop and hybrid Rankprop to a three-species network (adding the human network), and we compare the quality of the resulting predictions to the two-species results. Because the ISORank, MRF and hybrid Rankprop are limited only to proteins that participate in the PPI network, we focus the computation on these cases only.

### 4.1 Performance evaluation on the Inparanoid training data

The evaluation of the Rankprop variants is done by cross validation. For each Rankprop variant and each cross validation iteration, we apply the algorithm with all possible parameter combinations (see Section 2.5). We then choose the best parameter set, using as our performance criterion either the relative or absolute ROC score on the training set. Finally, we apply the algorithm to the test set with the selected parameters. The MRF is also evaluated using cross validation as in [Bandyopadhyay et al., 2006]. For ISORank we do not use training and simply set the single parameter (the propagation rate  $\alpha$ ) to 0.6, its preferred value for this specific data set, according to [Singh et al., 2007]. Importantly, the MRF algorithm is applicable

only to proteins with conserved interactions. Therefore, for a fair comparison, we consider only cases for which this algorithm is applicable.

Figure 2A compares the relative performance of the different rankings using the Inparanoid positive and negative sets, where MRF is applied with the Inparanoid-based alignment graph. Evidently, the relative performance of the BLAST scores is extremely high. This stems from the fact that most of the positive pairs in the Inparanoid set are mutually best BLAST matches, and all the positive pairs are best BLAST matches in at least one direction. Because the ISORank scores are very similar to those of BLAST, ISORank performs almost as well as BLAST. For the Rankprop variants, the advantage of mutual approach is evident both for Rankprop and hybrid Rankprop. However, when considering only *non-trivial* test cases (inset), where true orthologs are not mutual best BLAST matches, we do not see any clear advantage for either of the variants.

Figure 2B depicts the absolute performance of the different rankings. Evidently, all four Rankprop variants outperform both MRF and ISORank. In addition, we see that the ability of the scores obtained from ISORank to separate true orthologies from false ones is very similar to that of the raw BLAST E-values. The MRF method performs worse than both. Focusing on the non-trivial test cases (inset), we see that MRF outperforms BLAST and ISORank and that all three are outperformed by the Rankprop variants.

We also tested the different methods using the Inparanoid gold standard where MRF is applied with the BLAST-based alignment graph. The results are qualitatively similar to the ones in Figure 3 (data not shown).

## 4.2 Performance comparison on the HomoloGene training data

In this experiment we use the HomoloGene positive and negative sets, and we use our alternative definition for the alignment graph of MRF. The results are displayed in Figure 3. Here we see that the mutual variants of Rankprop (with and without using PPI information) outperform their one sided counterparts both in the relative and absolute tests. In addition, the relative performance of mutual hybrid Rankprop is better than that of the mutual Rankprop, better than that of ISORank in the general case, and slightly worse than ISORank in the non trivial case.

We also tested the different methods using the HomoloGene positive and negative sets where MRF is applied with the Inparanoid-based alignment graph. The results are qualitatively similar to the previous experiment in Figure 2 (data not shown).

## 4.3 Disambiguating Inparanoid orthology predictions

Inference based on sequence similarity alone, using the Inparanoid program, is often insufficient to determine orthology relations. In such cases, we obtain orthology clusters containing a number of paralogs from each species, where the actual mapping of functional orthologs is unknown. Bandyopadhyay et al. [2006] used the scores obtained by the MRF method to determine the most probable orthologs and resolve these ambiguities based on PPI information. In a similar manner, we used the different variants of our method to disambiguate the same clusters, assigning the top ranked protein as the putative ortholog of the query protein. We then

**Table 1. Overlap in resolving ambiguous Inparanoid orthology predictions.** Given two methods  $a$  and  $b$ , the table presents the overlap index  $\frac{O_{ab}}{T_{ab}}$  where  $O_{ab}$  is the number of yeast proteins that belong to an ambiguous Inparanoid cluster (*i.e.*, a cluster with more than one yeast protein or more than one fly protein) and were assigned the same ortholog by  $a$  and  $b$ .  $T_{ab}$  is the total number of such proteins that were applicable by both  $a$  and  $b$ . Analyzed methods include BLAST, hybrid Rankprop (HRP), mutual hybrid Rankprop (MHRP), Rankprop (RP), mutual Rankprop (MRP) and the markov random field (MRF) method.

	HRP	MHRP	RP	MRP	MRF
BLAST	0.45	0.61	0.48	0.63	0.64
HRP	–	0.72	0.95	0.70	0.40
MHRP	–	–	0.71	0.92	0.53
RP	–	–	–	0.72	0.41
MRP	–	–	–	–	0.53

compared our results to each other, to the results of Bandyopadhyay et al. [2006] and to annotations from HomoloGene.

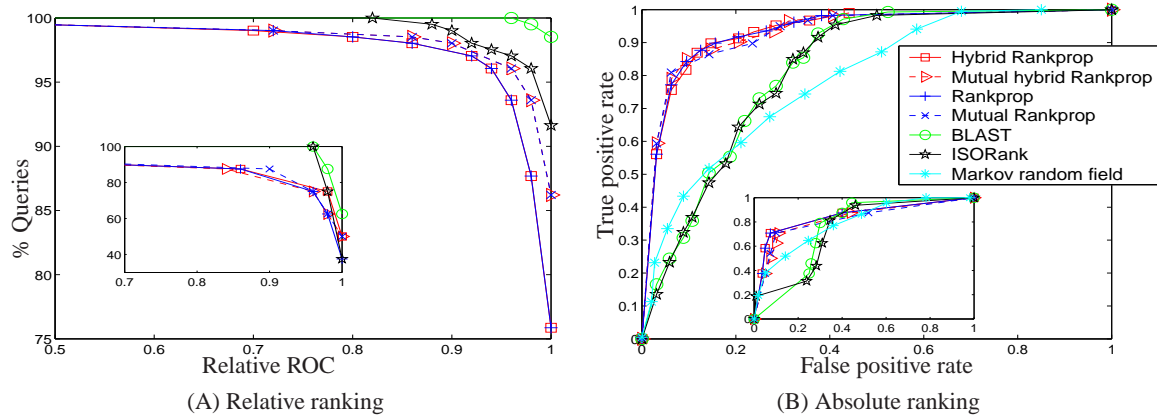
Overall, the Rankprop variants can resolve a much larger percentage of the ambiguous Inparanoid predictions than MRF can. Out of 1238 yeast proteins in the ambiguous clusters, the MRF method was applicable only to 146, whereas the Rankprop variants are applicable to 703, excluding proteins with no PPI information.

As a gold standard for this disambiguation task, we used orthology annotations from the HomoloGene database. We identified an ortholog in the HomoloGene database for 189 of the 1238 ambiguous Inparanoid predictions. Comparing the various Rankprop variants to one another shows that the mutual Rankprop variants strongly outperform the single-sided variants. On these 189 examples, the mutual Rankprop variants agrees with HomoloGene in 137 (72%) and 139 (73%) of the cases, with and without PPI data, respectively. By comparison, the one-sided Rankprop variants match HomoloGene in 104 (55%) and 106 (56%) of the cases.

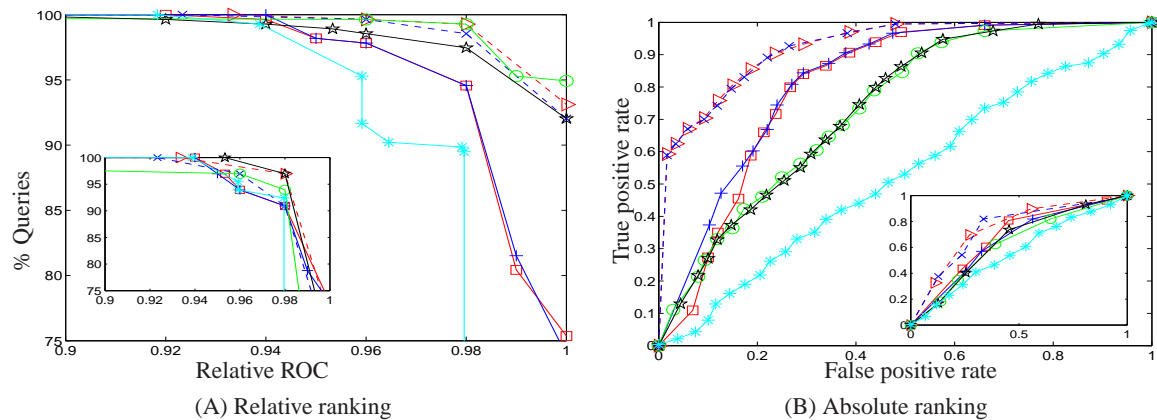
Among the same set of 189 proteins with HomoloGene annotations, MRF is only applicable to 51. For these proteins MRF agrees with HomoloGene in 36 (70%) of the cases, whereas mutual Rankprop agrees with HomoloGene in 35 (68%) or 37 (72%) cases, depending upon whether PPI information is used or not. The overlaps among the results from MRF and from the different Rankprop variants are summarized in Table 1.

An example for an ambiguous orthology prediction by Inparanoid is yeast Ubiquitin UBI4 which marks various proteins for selective degradation via the ubiquitin-26S proteasome system [Ozsolak et al., 1987]. Along with UBI4 the respective Inparanoid cluster contains two fly paralogs—Ubiquitin-63E and Ubiquitin-5E, where the latter has a slightly better sequence similarity with UBI4. The true ortholog according to all the Rankprop variants as well as MRF, however, is the former. This result is further supported by the HomoloGene database.

Another example is the kinetochore protein Skp1, which participates in multiple protein complexes, including the SCF ubiquitin ligase complex, the centromeric DNA binding CBF3 complex, and the RAVE complex that regulates assembly of the V-ATPase [Seol et al., 2001]. The Inparanoid cluster of Skp1 contains a number of paralogous fly proteins. Among those, the ortholog



**Fig. 2. Homology detection benchmark using the Inparanoid gold standard** (A) Relative performance: the figure plots the percentage of queries (y-axis) for which the associated  $ROC_{50}$  score is greater than a given threshold (x-axis). (B) Absolute performance: an ROC curve is displayed for each of the predictors. Seven methods are shown, including the four Rankprop variants, BLAST, ISORank and MRF. The insets present only *non-trivial* cases where true orthologs are not mutual best BLAST matches.



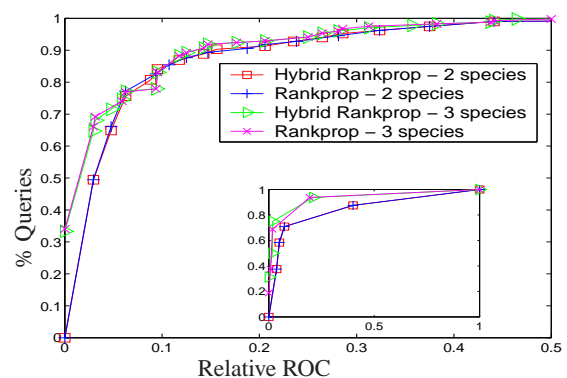
**Fig. 3. Homology detection benchmark using the HomoloGene gold standard.** Methods, labels and insets are as in Figure 2.

predicted by MRF is the Skpa centromeric DNA binding protein. However, the correct ortholog according to HomoloGene and the mutual Rankprop variants is Skpb, a paralog of Skpa (which has a slightly lower sequence similarity with the yeast's Skp1). The reason for this discrepancy is that Skpb does not have any known conserved interactions, and is excluded from the MRF analysis.

These two examples demonstrate that network based methods can produce accurate orthology predictions, which are not necessarily in line with the best BLAST matches. Additionally, we see that the limited applicability of MRF may harm its accuracy by excluding prominent candidates from the analysis, a limitation which does not hold for the Rankprop variants.

#### 4.4 Orthology detection based on three networks

To examine the utility of employing more than two networks in orthology detection, we repeated the above experiments using the human network in addition to those of yeast and fly. We applied the Rankprop and hybrid Rankprop variants as described above and tested the accuracy of the scores assigned to the yeast-fly protein pairs. The basic assumption is that the scores of true



**Fig. 4. Homology detection benchmark using three species** Absolute performance scores are shown for the Inparanoid gold standard. The inset presents only *non-trivial* cases as in Figure 2.

orthologous pairs from yeast and fly will increase when also

accounting for their common orthologs in human. The resulting absolute performance with the Inparanoid gold standard (Figure 4) shows a clear improvement in accuracy when adding the human network. We also examined the effect on the relative performance and observed a smaller yet evident improvement when using the additional network (data not shown). Similar results were also obtained with the HomoloGene dataset (data not shown).

## 5 DISCUSSION

We have presented three novel extensions for the Rankprop algorithm—the hybrid Rankprop which includes PPI information, the mutual Rankprop which was designed to account for multiple paralogous candidates and an application of Rankprop (and hybrid Rankprop) to three species concomitantly. We have demonstrated that the Rankprop algorithm and its novel variants provide improved scoring methodologies compared to several state-of-the-art methods. We also showed that, in the majority of cases, both the mutual variant and the addition of a third network improve upon the original Rankprop algorithm in both relative and absolute performance.

With regards to the utility of PPI data, the picture is not as clear. Evidently, diffusion based only on sequence similarity performs just as well as the diffusion with PPI information, and is often better than the PPI based approaches MRF and ISORank. A possible explanation for this observation might be that the majority of pairs in both positive sets are mutually best matches with very significant *E*-values, which makes them easy for detection by sequence similarity. This phenomenon is illustrated by the exceptionally high BLAST scores in the relative performance estimations. An alternative explanation might be that yeast and fly are just too distant to apply methods that are based on interaction conservation.

The probabilities obtained by the MRF algorithm were used by [Bandyopadhyay et al., 2006] to resolve Inparanoid clusters that had a few paralogs from each species. In Section 4.3 we compared the utility of the Rankprop variants in this task to that of MRF. We showed that the mutual variants of Rankprop compare favorably to MRF as they were applicable to roughly five times more proteins and achieved a similar accuracy.

The ISORank algorithm uses the obtained similarity scores to construct a global alignment of the investigated PPI networks by seeking the best one-to-one orthology assignment. Naturally, such an assignment should not necessarily fit the best matches on the ranked lists. Hence, a prerequisite for a successful construction of a global alignment is for the scores assigned to target proteins to be well calibrated from one query to the next. This property is reflected by our absolute performance measure, which points to a clear advantage of the Rankprop variants over ISORank. In this regard, a natural next step for our work would be to use the ranking scores to construct an alignment graph.

## REFERENCES

- S. F. Altschul et al. A basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.
- S. F. Altschul et al. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, 25:3389–3402, 1997.
- J. S. Bader, A. Chaudhuri, J. M. Rothberg, and J. Chant. Gaining confidence in high-throughput protein interaction networks. *Nature Biotechnology*, 22(1):78–85, 2004.
- S. Bandyopadhyay, R. Sharan, and T. Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Research*, 16:426–35, 2006.
- S. E. Brenner. Errors in genome annotation. *Trends in Genetics*, 15: 132–133, 1999.
- K. R. Christie et al. *Saccharomyces* Genome Database (SGD) provides tools to identify and analyze sequences from *saccharomyces cerevisiae* and related sequences from other organisms. *Nucleic Acids Research*, 32:D311–4, 2004.
- M. A. Crosby et al. FlyBase: Genomes by the dozen. *Nucleic Acids Research*, 35(Database issue):D486–491, 2007.
- M. Gribskov and N. L. Robinson. Use of receiver operating characteristic (ROC) analysis to evaluate sequence matching. *Computers and Chemistry*, 20(1):25–33, 1996.
- J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143:29–36, 1982.
- E. Ozsolak, D. Finley, M. J. Solomon, and A. Varshavsky. The yeast ubiquitin genes: a family of natural gene fusions. *EMBO J.*, 6(5): 1429–1439, 1987.
- S. Peri et al. Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, 13:2363–71, 2003.
- M. Remm, C. E. Storm, and E. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314:1041–1052, 2001.
- J. F. Rual et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature*, 437:1173–8, 2005.
- J. H. Seol, A. Shevchenko, and R. J. Deshaies. Skp1 forms multiple protein complexes, including rve, a regulator of v- atpase assembly. *Nat Cell Biol.*, 3(4):384–391, 2001.
- R. Singh, J. Xu, and B. Berger. Pairwise global alignment of protein interaction networks by matching neighborhood topology. *RECOMB07*, pages 16–31, 2007.
- R. Singh, J. Xu, and B. Berger. Global alignment of multiple protein interaction networks. *Pacific Symposium on Biocomputing*, 13: 303–314, 2008.
- K. Sjolander. Phylogenomic inference of protein molecular function: Advances and challenges. *Bioinformatics*, 20(2): 170–179, 2004.
- U. Stelzl et al. A human protein-protein interaction network: a resource for annotating the proteome. *Cell*, 122:830–2, 2005.
- D. L. Wheeler et al. Database resources of the national center for biotechnology. *Nucleic Acids Res*, 31(1):28–33, 2003.
- I. Xenarios et al. DIP: the Database of Interacting Proteins. *Nucleic Acids Research*, 28(1):289–291, 2000.