Global Network Alignment *

Lecturer: Nir Yosef Scribers: Rubi Boim and On Freund

Lecture 8, May 13, 2009

1 Introduction

Finding the common ancestries of different proteins across species is an important task. *Homologs* is the basic definition which denotes two genes that have descended from some common ancestral. For a more focused notation, we define *Paralogs* if the last evolutionary event separating the genes was duplication, and similarly we define *Orthologs* if the last evolutionary event separating the genes was speciation.

Understating the hierarchy of given genes is an important task. The most basic implementations of such knowledge can be understanding the evolutionary principles of the specific subject. Moreover, predicting/annotating the protein functions or interactions can be done with this information. Most of the methods of dealing with this problem are sequence-based models, thus sequence of proteins from different species was compared, in order to find a group of proteins that have the same homologs. Two such methods are COG (Clusters of Orthologous Groups) (Tatusov *et al.* [8]) and Inparanoid ([5]).

The COG approach defines orthologs using sets of proteins that contain reciprocal best BLAST matches across a minimum of three species. The Inparanoid approach is a sequence-based method of finding functional annotation. It uses clustering in order to derive orthology families, leaving some of the orthology relations ambiguous. For more information see [5].

1.1 Functional Orthology

Ambiguities in the functional annotation process arise when the protein in question has similarity to not one but many paralogous proteins, making it harder to distinguish which of these is the true ortholog that is, the protein that is directly inherited from a common ancestor. Especially in the genomes of mammals and other higher eukaryotes, large protein families are typically not the exception but the rule.

Moreover, the assignment of protein orthology depends largely on the evolutionary history. Protein families for which speciation predates gene duplication (out-paralogs) are particularly challenging. 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, i.e. are functional orthologs ([5]).

Conversely, when gene duplication predates speciation (out-paralogs), the family can often be subdivided into orthologous pairs which have higher sequence similarity to each other than to other members. However, evolutionary processes such as gene conversion serve to homogenize paralogous sequences over time, making these cases problematic as well. Even more complicated, protein function may be lost between distant organisms or conserved across multiple proteins within a single species.

Other approach for detecting orthology is by finding *inparalogs*. These are specific paralogs that were duplicated after the speciation and hence are orthologs to the other species genes. An example of this definition can be observed in Figure 1.

*Based Partially on a scribe by Ofer Lavi and Lev Ferdinskoif, 2006

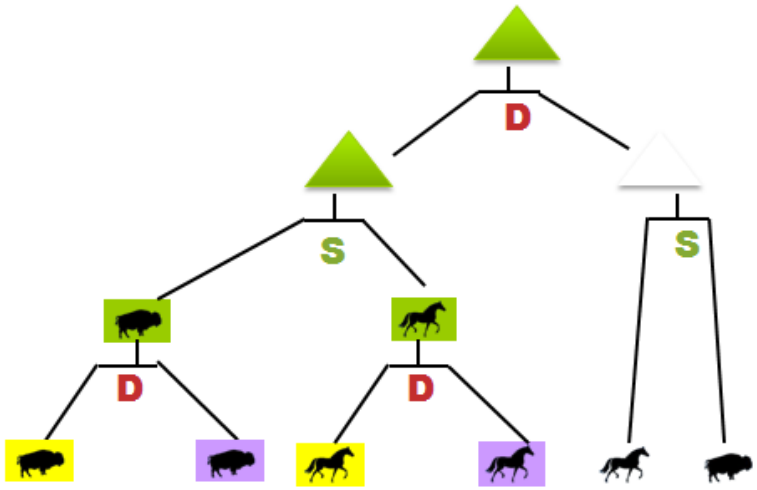


Figure 1: This figure shows an example of the inparalogs definition: paralogs that were duplicated after the speciation and hence are orthologs to the other species genes.

1.2 Global Approach for Orthology Prediction

Given two species A, and B, find a correspondence between their proteins which maps together proteins with similar sequences and enforces as much as possible the conservation of interactions. The matched proteins are predicted to be orthologs.

This problem can be seen as a problem in graph-theory: Given a pair of networks G and H, and a restriction on the mapping between the nodes of G and H, identify a $1-1$ mapping δ that preserves as many edges as possible. More formally:

$$\max_{\delta:V(G)\to V(H)} \#\{(v_1,v_2)\in E(G)\wedge(\delta(v1),\delta(v2))\in E(H)\}$$

Some related problems have been found to difficult. For example - the restricted injective homomorphism problem: Given a query graph Q, a reference graph H and a restriction on the mapping between the nodes of Q and H, is there an injective homomorphism that respects the homology constraints? Other variation is to find the best legal injective mapping (with the largest number of conserved edges). The difficult of this problem however varies among its parameters. Figure 2 depict the hardness changes according to the specific parameters. Another related problem is solving constrained maximum graph homomorphism. There is some work on solving this problem using linear programming [4].

The main purpose of this scribe is to present several *diffusion based techniques* for orthology prediction. All these methods share the basic concept: The similarity between a pair of proteins is determined by:

- The local property: The similarity of their sequences
- The global property: The similarity of their neighbors in some network structure

The input for these algorithms is usually a sequence-similarity matrix and PPI from 2 or more species. The output is a ranked list ranked list of protein pairs, ordered by their inferred similarity. These algorithm, however, differ as they try to improved the other ones drawbacks. First we will present the Markov Random Field [2]. We will then show the ISORank [6], and conclude with the Hybrid RankProp [1].

Mapping Restriction on Q	Mapping Restriction on H	Max degree Q	Max degree H	Hardness (decision)
2	--	--	--	Poly
*	1	--	2	Poly
3	2	1	2	NPC
3	1	3	4	NPC
2	1	3	2	Maximization is APX

Figure 2: This figure shows the differences in the difficulties of solving the problem of best legal injective mapping, according to the specific paramters.

2 Orthology Mapping take I - Markov Random Field (Bandyopadhyay *et al* '05)

The model consists of the following steps:

1. The protein interaction networks of two species are aligned by assigning proteins to putative orthology using the Inparanoid algorithm.
2. Networks are aligned into a merged graph representation (alignment graph).
3. Probabilistic inference is performed on the aligned networks to identify pairs of proteins, one from each species, that are likely to retain the same function based on conservation of their interacting partners.
4. A logistic function is used to compute the probability of functional orthology for a protein pair i given the states of functional orthology for its network neighbors.
5. The previous probability is updated for each pair over successive iterations of Gibbs sampling.

An overview of the method is seen in Figure 3.
More specifically, consider an alignment graph, G, with nodes representing sequence-similar protein pairs, and edges linking nodes (a, b) and $(a^{'}, b^{'})$ if one of $(a, a^{'})$ or $(b, b^{'})$ directly interacts, and the other interacts via a neighbor, which is directly connected to him (i.e. interaction of distance ≤ 2). Figure 4 illustrate this definition. An edge is *strongly conserved* if both its endpoints are true functional orthologs.

2.1 Conservation Index

The conservation index c of a node i (representing protein pair $(a, a^{'})$) is defined as twice its number of strongly conserved interactions, divided by its total number of interactions over both species:

$$c(i) = \frac{2d(i)}{d(a) + d(a^{'})} \tag{1}$$

where $d(i)$ denoted the number of strongly conserved links involving node i, and $d(a)$ and $d(a^{'})$ denots the degrees (number of interactions) of proteins a and $a^{'}$ in their respective networks.

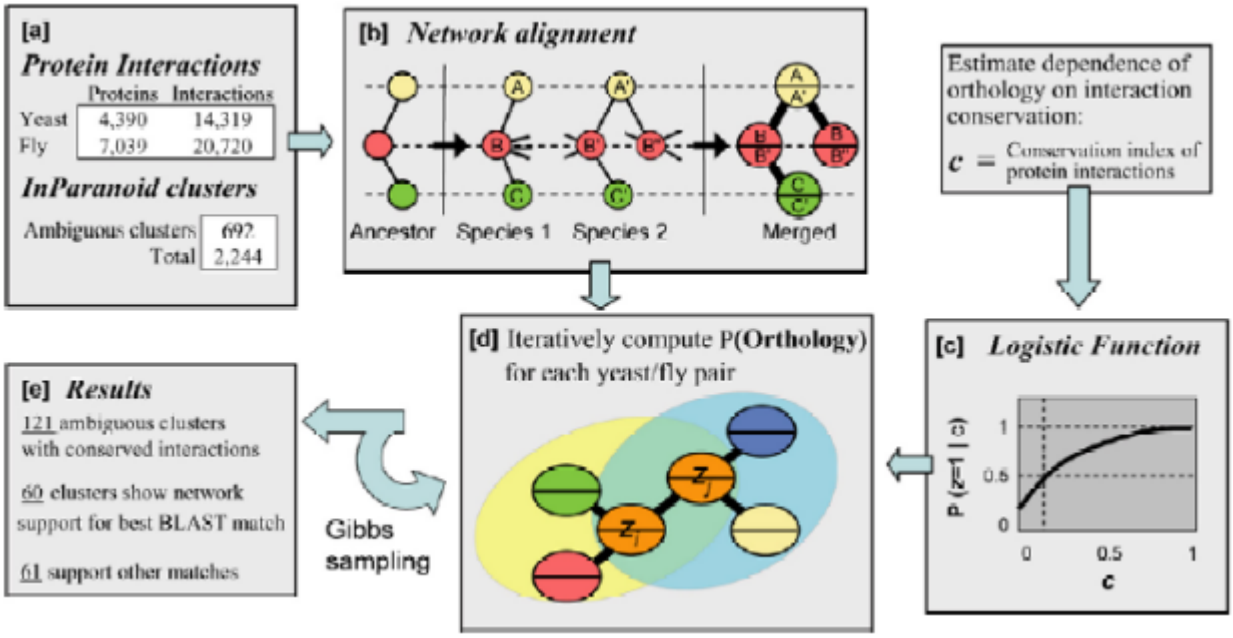


Figure 3: Source [2]. This figure shows an overview of the method, as an example on yeast/fly PPI networks. (a) PPI networks for yeast and fly are combined with clusters of orthologous yeast and fly protein sequences as determined by the Inparanoid algorithm. (b) Networks are aligned into a merged graph representation. (c) The logistic function is used to compute the probability of functional orthology for a protein pair i given the states of functional orthology for its network neighbors. (d) This probability is updated for each pair over successive iterations of Gibbs sampling. (e) The final probabilities confirm 60 of the best BLAST match pairings. The network supports a different hypothesis for 61 pairings.

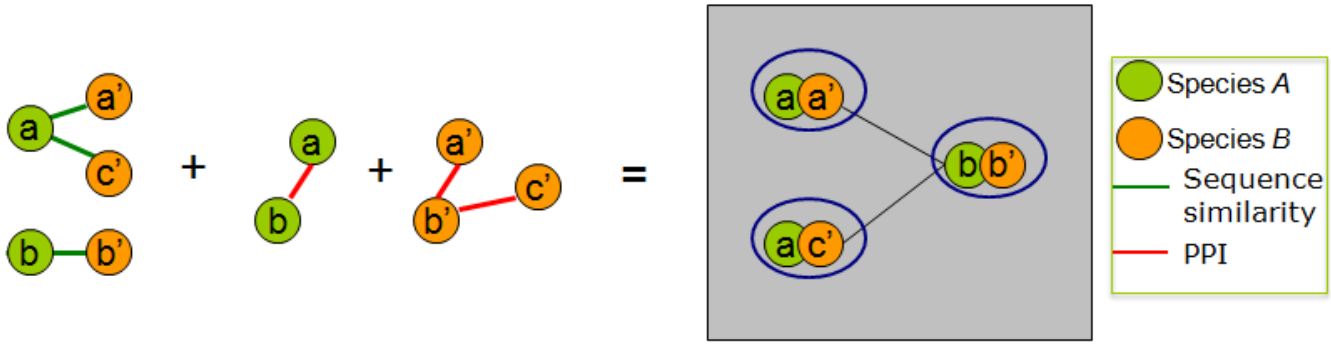


Figure 4: This figure shows an example of the alignment graph.

An example of the values of the conservation index in the yeast vs. fly experiment is shown in Figure 5(a). As we can observe, the $c(i)$ of the definite orthologs are indeed higher than the ambiguous ones and the random ones.

2.2 Probabilistic Model

The probabilistic model is based on the assumption that the probability of functional orthology for a pair of proteins is influenced by the probabilities of functional orthology for their network neighbors, which in turn depend on their network neighbors, and so on. This type of probabilistic model is known as a Markov random field (see [3]).

This model is specified by an undirected graph $G = (V, E)$ corresponding to a network alignment, and conditional probability distributions which relate the event that a given node represents a functionally orthologous pair with those events for its neighbors. A Markov random field model is specified in terms of potential functions on the cliques in the graph:

$$P(\overrightarrow{z}) = \frac{1}{Z} exp\{-U(\overrightarrow{z})\} \tag{2}$$

where $\overrightarrow{z}$ is some assignment to the states of all nodes in the graph, U is an "energy" function which integrates the potentials over all cliques in the graph, and Z is a normalizing constant. It is not necessary to compute the normalization constant, since all that is required are the conditional probabilities for each node given its neighbors (rather than the joint distribution). For computational efficiency, the common auto-logistic model ([3]) which assigns zero potential to cliques of size > 2 was used. Under this model, the energy takes the form:

$$U(\overrightarrow{z}) = -\sum_i \alpha_i z_i - \sum_{(i,j)\in E} \beta_{ij} z_i z_j \tag{3}$$

which, when substituted into the equation for $P(\overrightarrow{z})$ above, reduces to a logistic function.

Based on the initial observation that the functional orthology of a node is a function of its conservation index (well approximated by a logistic function see the next section), they set $\alpha_i = \alpha$, $\beta_{ji} = \beta_i = \frac{2\beta}{d(a_i + d(a'))}$ to obtain the following:

$$P(z_i|Z_{N(i)}) = \frac{1}{1 + exp\{-\alpha_i - \sum_{j\in N(i)} \beta_{ij} z_j\}} = \frac{1}{1 + exp\{-\alpha - \beta c(i)\}} \tag{4}$$

where $N(i)$ is the set of neighbors of node i, and $z_{N(i)}$ denotes the set of all z_j such that $j \in N(i)$. Note that α_i and β_{ij} could be set to accommodate other equations for conservation index, as long as they are linear in the number of strongly conserved neighbors $d(i)$.

2.3 Fitting the Logistic Function

In order to provide a set of training data for fitting the parameters α and β of the logistic function, a fraction (about a half) of the definite functional orthologs having at least one conserved interaction is chosen randomly, as positive examples, and their states are set to $z = 1$. Negative examples of "non-orthologs" are also generated by randomly selecting a fraction (about a half) of the proteins in one species and pairing each with its best BLAST E-value matching protein in the other species not in the same cluster; their states are set to $z = 0$ (ideally, the negative training data would consist of orthologs that are not functional orthologs, but few such examples exist). Parameters α and β are optimized by maximizing the product of $P(z_i|z_{N(i)})$ over all positive and $(1 - P(z_i|z_{N(i)}))$ over all negative training data using the method of conjugate gradients. The logistic function obtained for yeast vs. fly is shown in Figure 5(b).

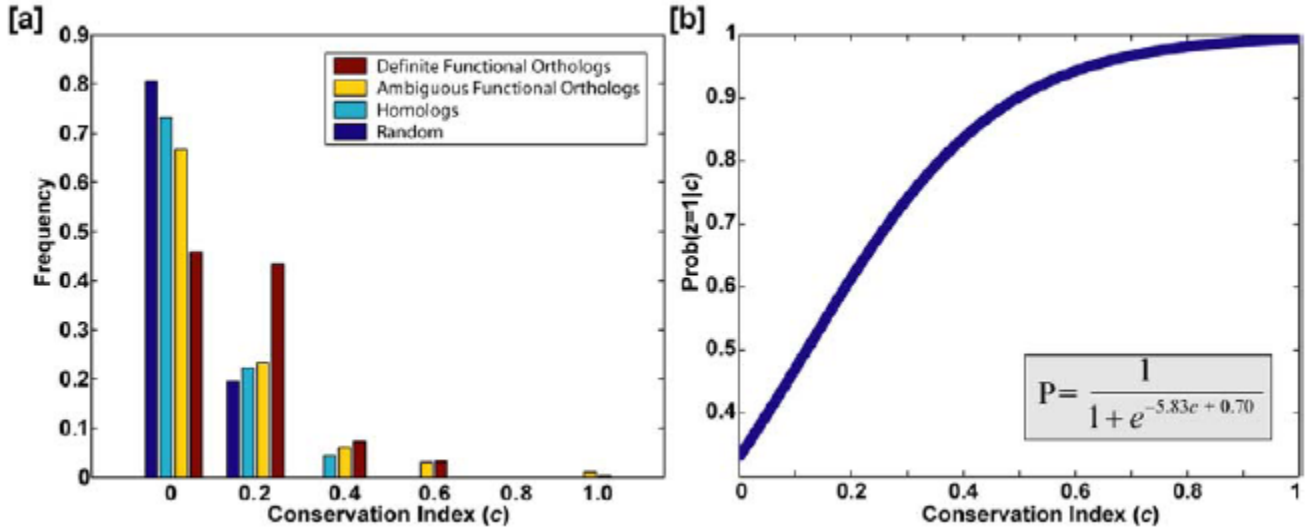


Figure 5: Source [2]. This figure shows two graphs: (a) Network neighborhood conservation for definite orthologs versus other yeast/fly protein pairs. The distribution of the conservation index is shown for definite functional orthologs (sole members of an Inparanoid group); ambiguous functional orthologs (in a group with multiple members); homologs (different groups but similar sequences); and random protein pairs. Definite functional orthologs show a shift towards higher conservation of protein interactions between the yeast and fly protein networks. Mean c=0.1512, 0.1171, 0.0870, 0.0615 for definite functional orthologs, ambiguous functional orthologs, homologs and random pairs respectively. (b) Logistic function relating conservation index to probability of functional orthology. Logistic regression was performed using the "definite functional ortholog" and "homolog" pairs as positive vs. negative training data, respectively. The resulting function is shown.

Species	No. of proteins	No. of interactions
yeast	4,389	14,319
fly	7,038	20,720

Table 1: Source [2]. Number of proteins and interactions used in the experiment.

2.4 Orthology Inference

In order to identify the functional orthologs the above model was used to estimate the final posterior probabilities $P(z_i)$ using the method of Gibbs sampling ([7]). In this approach, nodes representing ambiguous functional orthologs are each assigned a temporary state $z = 0$ or $z = 1$, initially at random. At each iteration, a node i is sampled (with replacement) and its value of z_i is updated given the states of its neighbors, $z_{N(i)}$. The new value of z_i is set to 0 or 1 with probability $P(z_i|z_{N(i)})$. Over all iterations, the nodes designated as definite functional orthologs and "non-orthologs" are forced to states of 1 and 0, respectively.

2.5 Experimental Results

The method was applied on yeast and fly, and PPI data was downloaded from DIP ([9]). The statistics for the PPI networks of yeast and fly are shown in Table 1.

A total of 2,244 clusters were generated, covering 2,834 proteins in yeast and 3,881 proteins in fly. Of these, 1,552 clusters contained only a single yeast and fly protein pair and were assumed to represent unambiguous or "definite" functional orthologs. The remaining 692 clusters contained multiple proteins from yeast and/or fly, leaving the functional orthologs ambiguous.

To determine the extent to which proteins and their functional orthologs had conserved protein interactions, the network neighborhoods of definite functional orthologs were examined and compared to the neighborhoods of less related protein pairs (Figure 5). As a measure of local network conservation, the conservation index of each protein pair was computed as proportional to the fraction of interactions that were conserved across the two species.

For example, in Figure 3(b) the orthologous pairing $B/B^{'}$ has a higher conservation index $(4/9)$ than the alternative pairing $B/B^{''}$ $(2/9)$. Figure 5(a) shows the set of conservation indices for definite functional orthologs versus those of ambiguous functional orthologs, non-orthologous homologs (best cross-species BLAST matches not assigned to the same Inparanoid group), and random pairs of proteins chosen independent of sequence similarity.

The set of definite functional orthologs had the highest occurrence of conserved interactions. Moreover, the mean conservation index was related to the stringency of the pairing: definite functional orthologs tended to have higher conservation indices than ambiguous functional orthologs, ambiguous functional orthologs higher indices than homologs, and homologs higher indices than random protein pairs.

Beyond the mean conservation index, there were also significant differences among the four distributions. These findings confirm that yeast/fly proteins classified as definite functional orthologs are more likely to have equivalent functional roles in the protein network and, conversely, that conserved network context could be used to help discriminate functional orthology from general sequence similarity.

They applied their approach to resolve ambiguous functional orthology relationships in the yeast and fly protein networks. Of the 692 ambiguous Inparanoid clusters, 121 contained protein pairs for which at least one pair had conserved interactions between networks. Application of their Gibbs sampling procedure yielded estimates of probability of functional orthology for each protein pair in these 121 ambiguous clusters. In 60 of these clusters, the highest probability was assigned to the protein pair that was also the most sequence-similar via BLAST. These cases reinforced the intuition that the best sequence matches are also the most functionally similar.

The remaining 61 clusters showed the opposite behavior, i.e., the highest probability pair was not the most sequence similar pair. Of these 61 cases, 15 were supported by two or more conserved interactions. Because the yeast and fly networks are incomplete (i.e., they contain false negatives because of the "noisy" data of the PPI networks), in some of these cases they could not rule out the possibility that conserved interactions with the best BLAST matches have been missed.

A complete listing of the results can be found on their website (http://bioinf.ucsd.edu/ sbandyop/GR). For some examples of the clusters found see Figure 6.

2.5.1 Validation

One approach to validate their results would be to compare them against databases of functional annotations (such as GO). However, such databases are based directly on sequence similarity, thus they are lack the specificity to discriminate among subtle functional differences across large gene families. Therefore, cross-validation was used in order to test the ability of their approach to reclassify protein pairs in the definite functional ortholog set (positive test data), against the non-orthologs homolog set (negative test data).

In each cross-validation trial, 1% of these assignments were hidden (declassified) and monitored during Gibbs sampling to obtain probabilities of functional orthology for positive and negative examples. Reclassification was judged successful if the probability of functional orthology exceeded a particular cutoff value. These statistics were compiled over 100 trials. Figure 7(a) charts cross-validation performance over a range of probability cutoffs.

At a probability cutoff of 0.5, they observed a 50% true positive rate and a 15% false positive rate. This shows marked improvement over a random predictor where we would expect to see the same true positive rate as false positive rate.

Declassifying 1% of the known functional orthologous and non-orthologous pairs reduces the amount of information available to the algorithm and, thus, can reduce its predictive ability. Therefore, the cross–validation analysis was repeated at varying percentages of declassification of positive and negative data (ranging from 1% to 100%) (Figure 7(b)). For instance, changing the amount of declassification of available training data from 1% to 25% reduced the maximum precision from 83% to 75%. Further declassification yielded more marked reductions in precision and recall.

2.5.2 Summary

Although MRF is a very useful algorithm, it suffers from a few disadvantages that more modern algorithms try to overcome. The first major disadvantage is its low applicabillity - only 11% of the cases have a conserved interaction. Consequently, MRF is probably not applicable for multiple species. The second major disadvantage is that the algorithm does not allow control in serveral aspects. It cannot be parameterized to tune the contribution of the local factor against the global factor, nor can it be parametrized to tune the contribution of sequence similarity against PPI data.

3 Orthology Mapping take II - ISORank (Singh *et al* '07, '08)

In order to overcome the disadvantages of MRF, the ISORank algorithm looks at the problem from a slightly different angle, and constrcuts an input graph in which all possible protein pairs are represented, each pair is a node, and the edges of the graph are not necessarily interlogs. This creates a network product as can be seen in Figure 8.

The main idea being that a and b are a good match if their sequences align and their respective neighbors are a good match with each other.
The algorithm associates a functional similarity score R_{ij} with each possible match between every pair i, j

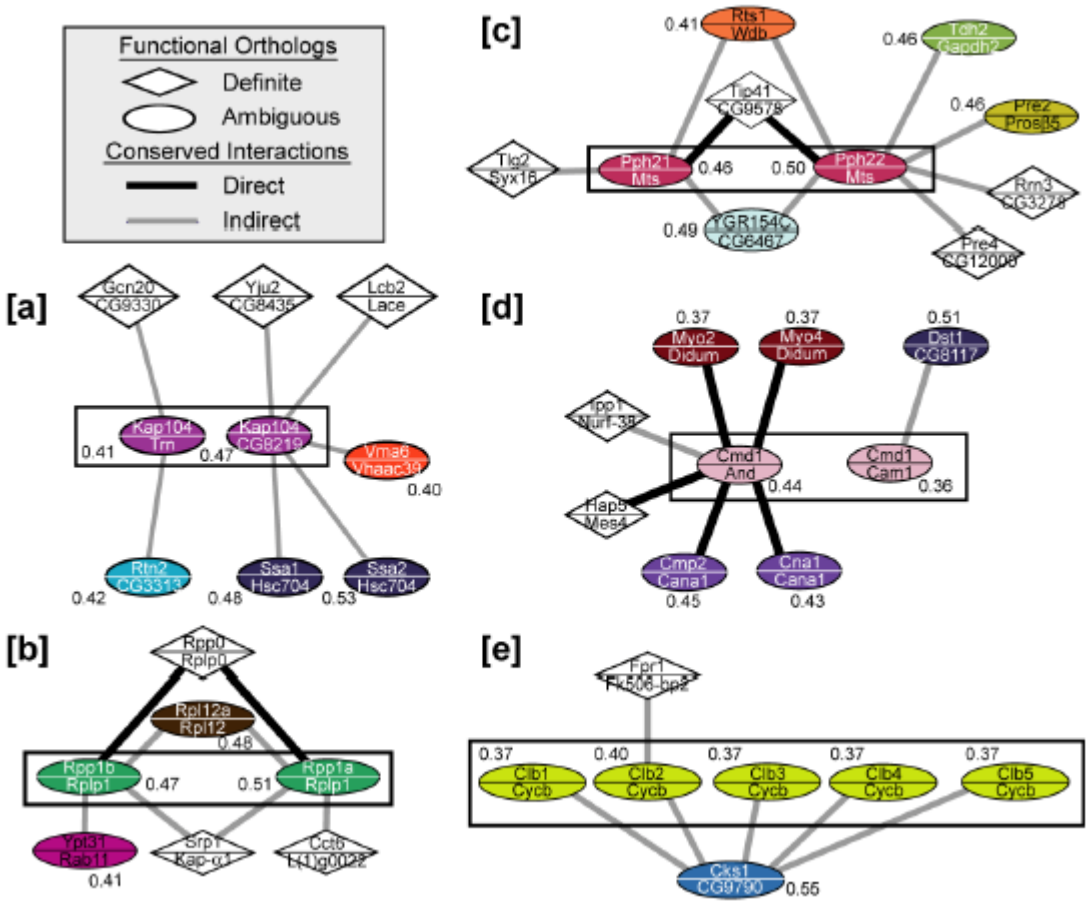


Figure 6: Source [2]. Example orthologs resolved by network conservation. Each node represents a putative functional match between a yeast/fly protein pair (with names shown above/below the line,respectively). Links between nodes denote conserved interactions (thick black, direct interactions in both species; thin gray, indirect interaction in one of the species). Diamond vs. oval nodes represent definite vs. ambiguous functional orthologs. Oval nodes of the same color represent ambiguous protein pairs belonging to the same Inparanoid cluster. The mean probability of functional orthology is given next to each ambiguous pair. (a)-(e) show examples of clusters that were disambiguated by conserved network information;the cluster resolved in each panel is outlined by a black rectangle.

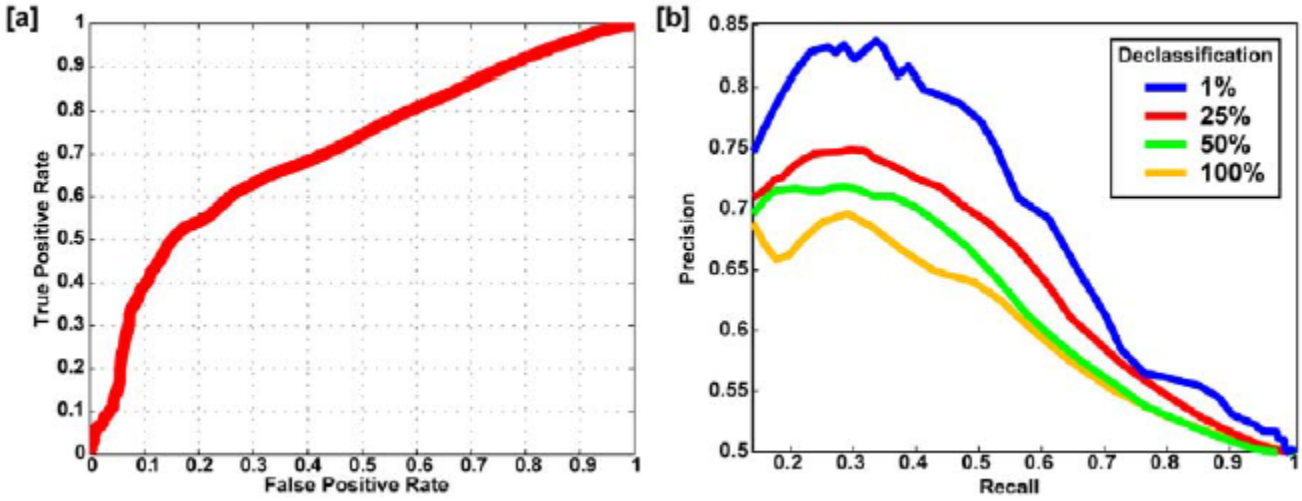


Figure 7: Source [2]. (a) The Receiver Operating Characteristic (ROC) curve shows the true positive rate (percent of true data predicted correctly as positive) vs. the false positive rate (percent of false data predicted incorrectly, i.e. as positives) of the method. (b) Dependence of predictions on number of available training examples. Percent recall (true positive rate) vs. precision (percent of positive predictions that were correct) is plotted as the probability cutoff ranges from [0-1]. Different color plots correspond to different percents of declassification of training examples.

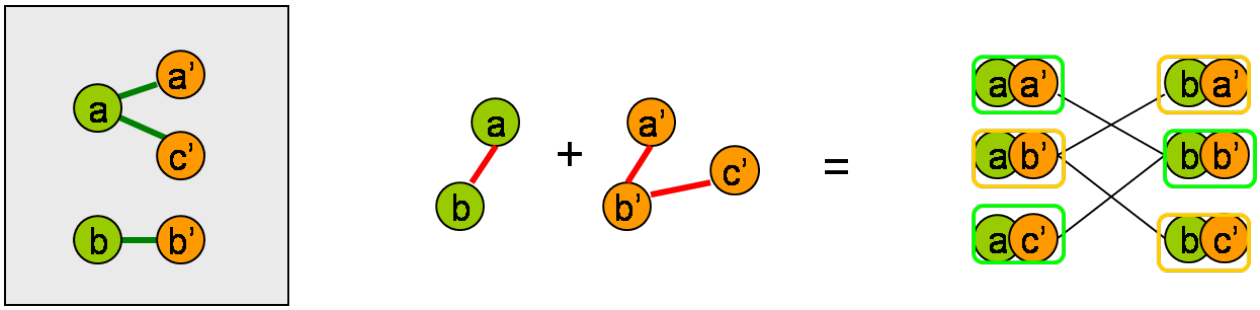
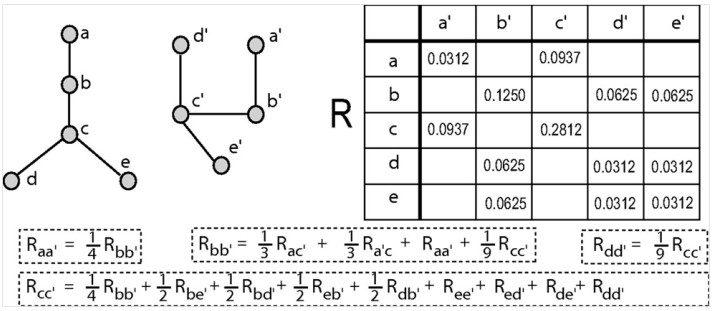


Figure 8: In ISORank's input graph every node represents a pair and all possible protein pairs are considered.



	a'	b'	c'	d'	e'
a	0.0312		0.0937		
b		0.1250		0.0625	0.0625
c	0.0937		0.2812		
d		0.0625		0.0312	0.0312
e		0.0625		0.0312	0.0312

of nodes from the two networks. R is found by solving an eigenvalue problem, which is based on Google's PageRank formula:

$$R_{ij} = \alpha \sum_{u \in N(i)} \sum_{v \in N(j)} \frac{w(i,u)w(j,v)}{\sum_{r \in N(u), q \in N(v)} w(r,u)w(q,v)} R_{uv} + (1-\alpha)E_{ij}$$

Where $N(x)$ are the neighbors of x, w is the weight function, α is a parameter balancing the contribution of the global factor and local factor.
The problem explicitly models that tradeoff between the twin objectives of high network overlap (the global factor, represented by the left term of the equation) and high sequence similarity (the local factor, represented by the right term of the equation) between matching nodes. The problem allows to "play" with this tradeoff, by tuning the value of α.
The score R_{ij} for any match (i,j) must be equal to the total support provided to it by each of the $|N(i)N(j)|$ possible matches between the neighbors of i and j. By these equations a system of constrains is created, and the neighborhood score is computed in a recursive fashion.

The propagation equations can then be solved by defining a matrix A such that:

$$A[i,j][u,v] = \begin{cases} \frac{w(i,u)w(j,v)}{\sum_{r \in N(u)} w(r,u) \sum_{q \in N(v)} w(q,v)} & \text{if } (i,u) \in E_1, (j,v) \in E_2 \\ 0, & \text{otherwise} \end{cases}$$

And re-writing the recursive formula as matrix multiplication:

$$R = \alpha AR + (1-\alpha)E, 0 \leq \alpha \leq 1, \; or$$

$$R = (\alpha A + (1-\alpha)E1^T)R$$

The desired solution R is the first eigenvector of the matrix $(\alpha A + (1-\alpha)E1^T)$.
Since the matrix is stochastic, we have an efficient solution using the power method:

- Start from a "random" R such that $||R|| = 1$.
- 1. Apply $R \leftarrow AR$
  2. Normalize R

The ISORank algorithm also has two major disadvantages. Firstly its running time grows exponentially with the number of networks, making it almost unfeasible to handle more than two species. Secondly, it does not allow control of sequence similarity vs. PPI in the same fashion that is allows control of the local and global factors. When we look at Hybrid RankProp we will see how such control can be acheived.

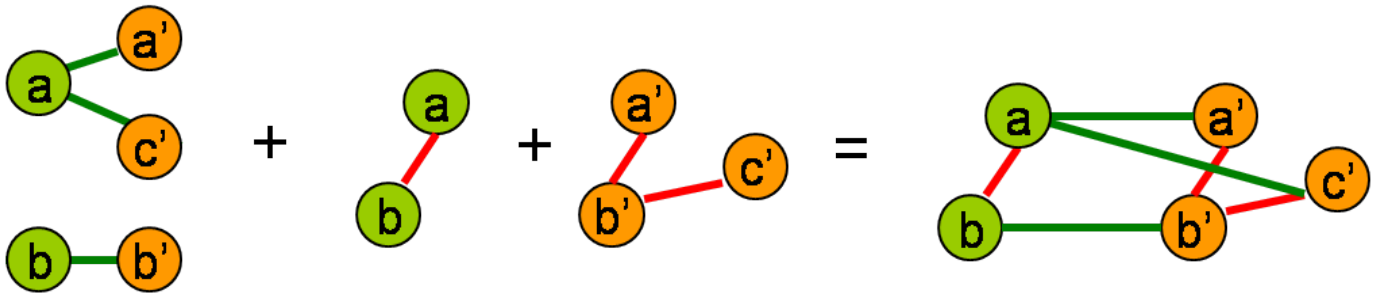


Figure 9: Creation of a Hybrid RankProp input graph. The graph is created by combining sequence similarities (left term) and PPI (right term).

4 Orthology Mapping take III - Hybrid RankProp (Yosef *et al* '08)

The next step in Orthology Mapping, the Hybrid RankProp, was motivated by the following needs:

- High Applicability, compared with that of MRF
- Efficiency (allowing more than two species to particapate in the mapping)
- Account for multiple hits (inparalogs)
- Supply control in two dimensions:
  - Local vs. Global (as in ISORank)
  - Sequence similarity vs. PPI

The graph structure in Hybrid RankProp is based on network concatenation. Two types of edges exists in the graph, sequence similarities (between species) and PPI (within species).

The main idea is that a and b are a good match if their sequences align and their respective neighbors (either by PPI or by sequence similarity) are a good match with each other.
Unlike ISORank, Hybrid RankProp does not handle all pairs in a single run, but rather takes a "query node" as an input, and outputs a list of genes from the second species ordered by their similarity to the query node. It does so by a diffusion procedure: First, the query node (q) is assigned a score of 1.0. This score is continually pumped to the remaining nodes by means of the network, and upon termination, every protein is assigned a score determined by the steady state of the diffusion process.

$$\delta_i(0) = 0, \delta_q 0 = 1$$

$$\delta_i(t+1) = W_{qi} + \alpha \left(\frac{\rho}{1+\rho} \sum_{j \in N_{ppi}(i) \setminus q} W_{ji}\delta_j(t) + \frac{1}{1+\rho} \sum_{j \in N_{hom}(i) \setminus q} W_{ji}\delta_j(t) \right)$$

Where N_{ppi} and N_{hom} stand for neighbors from the PPI network and neighbors from the sequence homology networks respectively.α controls the tradeoff between the local and the global factors, and ρ controls the tradeoff between sequence similarity and PPI.
The algorithm has several configurable parameters. Firstly, the different normalization for PPI and homology scores

$$W[i,j] = -log\left(\frac{W_{ppi}[i,j]}{max_{ppi}\sigma_{ppi}}\right) \; W[i,j] = -log\left(\frac{W_{hom}[i,j]}{max_{hom}\sigma_{hom}}\right)$$

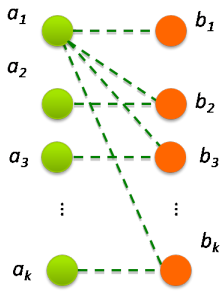
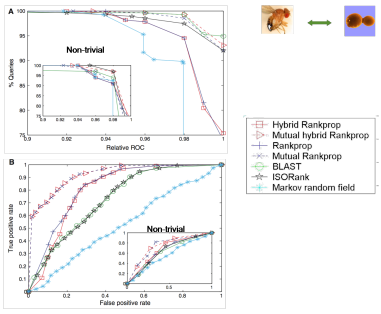
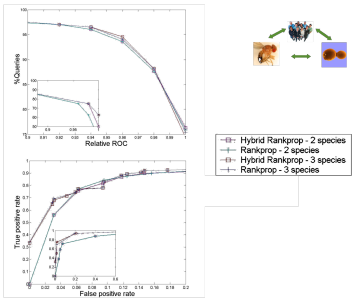


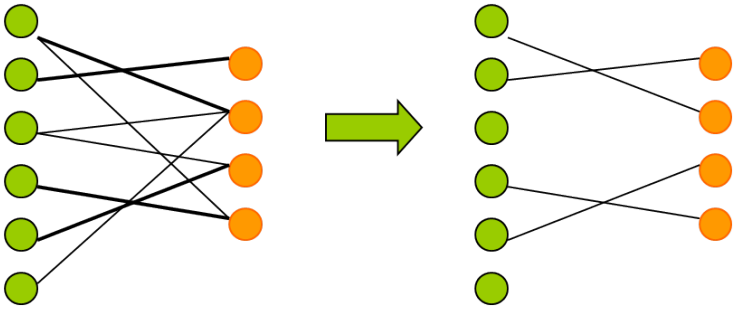
Figure 10: In the figure we see that all the paralogs but one (b_1) have strong matches (with paralogs of the query node a_1). Querying from a_1 will not distinguish b_1 from the rest, while querying from b_1 and its paralogs will clarify that it is the most likely ortholog of a_1.



Additionally, the algorithm support a configurable diffusion rate α that allows control over Local vs. global in a similar fashion to ISORank. Finally, the algorithm allows tuning of ρ - the relative weight of PPI vs. homology edges. An extra enhancement to the algorithm, titled *Mutual RankProp*, accounts for multiple matches, by utilizing "back queries". Let the query protein a have multiple homology matches $b_1, ..., b_k$, representing a paralogous set. We query back from the paralogous set and report the mean activation score. This removes the algorithm's dependance on the order of species. An example is shown in Figure lec09:fig:Mutual RankProp

As mentioned ealier, one of Hybrid RankProp's goal is the ability to support more than two species, or to propogate on 3 (or more) networks concomitantly. The idea behind this ability is that scores of true orthologs will increase when also accounting for their common orthologs on a third species. The runnning time of the algorithm is quadratic in the number of networks (as opposed to ISORank which is exponential). In order to test Hybrid RankProp's characteristics, training data was taken from InParanoid Homologene. We use two criteria to evaluate the performance, while focusing on non-trivial cases, of course: The relative criterion measures how far down the list is the true ortholog located. Each query is considered separately, using the permissive negative set, and looking at the distributtion of ROC_{50} (considering only the 50 top ranking proteins); the absolute criterion tests if the scores are comparable across queries and if true matches tend to get higher scores than false matches. To this end, all queries are considered toghether, and the stringent negative set was used. The activation scores from all queries were sorted together, and a single ROC curve computed.





5 Constructing a global network alignment

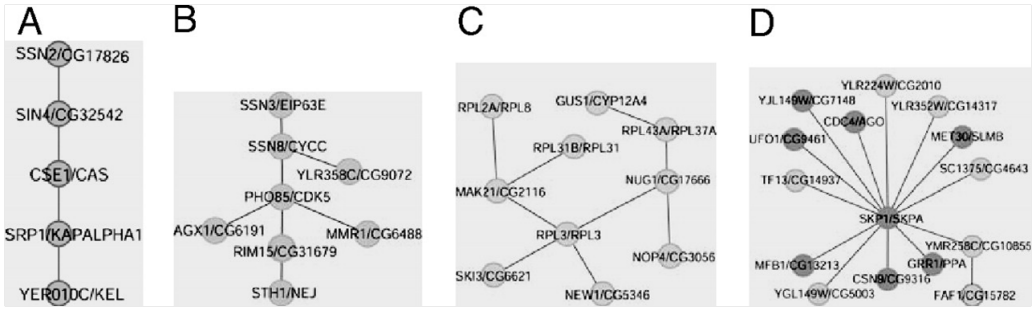
One to one mapping extracts a set of high-scoring, mutually consistent matches by means of an Orthology Mapping algorithm. This translates to finding a maximum weight matching in a bipartite graph, which can be solved efficiently.

$$max_{\delta:V(G)\to V(H)} \sum_{v\in V(G)} R_{v,\delta v}$$

In this example, te results of ISORank (represented by R) are utilized.

Constructing a one to one mapping on yeast and fly using ISORank scores conserves about 1400 edges (out of about $25,000$). The conserved components reveal functional modules that are encriched in proteins involved in various biological processes. The conserved functions range from various signaling cascades to core cellular functions like ribosoma synthesis and function DNA transcription and translation, cell divsion, etc... The conservation of network regions annotated with core cellular functions is an expected outcome as evolution tends to keep them in place.

With many to many mapping we note that the previous formulation ignore issues like gene duplication because more than one match per protein is possible. Therefore, we need to find a partition of the proteins



into sets of orthologs. An idea in this direction is noting that each gene in the set has high pairwise R scores with the rest of the genes in the set, while there are no genes outisde the set with this property.

References

[1] Nir Yosef abd Roded Sharan and William Stafford Noble. Improved network-based identification of protein orthologs.

[2] S. Bandyopadhyay, R. Sharan, and T. Ideker. Systematic identification of functional orthologs based on protein network comparison. *Genome Research*, 16(3):428–435, 2006.

[3] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society*, B:192–236, 1974.

[4] Klau et al. 2009.

[5] 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.

[6] Rohit Singh, Jinbo Xu, and Bonnie Berger. Global alignment of multiple protein interaction networks with application to functional orthology detection. *Proceedings of the National Academy of Sciences*, 35:12763–12768, 2008.

[7] A. Smith and G. Roberts. Bayesian computation via the gibbs sampler and related markov chain monte carlo methods. *Journal of the Royal Statistical Society*, B55:3–23, 1993.

[8] R.L. Tatusov, M.Y. Galperin, A., Darren, A. Natale, and E.V. Koonin. The cog database:a tool for genome–scale analysis of protein functions and evolution. *Nucleic Acids Research*, 28(1):3336, 2000.

[9] I. Xenarios, L. Salwinski, X.J. Duan, P. Higney, S.M. Kim, and D. Eisenberg. DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. *Nucleic Acids Research*, 30:303–305, 2002.