Pacific Symposium on Biocomputing 14: (2009)

PAIRWISE ALIGNMENT OF INTERACTION NETWORKS BY FAST IDENTIFICATION OF MAXIMAL CONSERVED PATTERNS*

WENHONG TIAN[1,2†], NAGIZA F. SAMATOVA[1,2]

[1]*Department of Computer and Mathematics Division , Oak Ridge National Laboratory, Oak Ridge, TN 37831 and* [2]*Computer Science Department, North Carolina State University, Raleigh, NC 27696, USA.*

A number of tools for the alignment of protein-protein interaction (PPI) networks have laid the foundation for PPI network analysis. They typically find conserved interaction patterns by various local or global search algorithms, and then validate the results using genome annotation. The improvement of the speed, scalability and accuracy of network alignment is still the target of ongoing research. In view of this, we introduce a connected-components based algorithm, called HopeMap for pairwise network alignment with the focus on fast identification of maximal conserved patterns across species. Observing that the number of true homologs across species is relatively small compared to the total number of proteins in all species, we start with highly homologous groups across species, find maximal conserved interaction patterns globally with a generic scoring system, and validate the results across multiple known functional annotations. The results are evaluated in terms of statistical enrichment of gene ontology (GO) terms and KEGG ortholog groups (KO) within conserved interaction patters. HopeMap is fast, with linear computational cost, accurate in terms of KO groups and GO terms specificity and sensitivity, and extensible to multiple network alignment.

1. Introduction

Protein-protein interactions (PPI) are of central importance for virtually every process in a living cell. Information about these interactions can improve our understanding of diseases and provide the basis for new therapeutic approaches. A fundamental goal in systems biology is to understand how proteins in the cell interact with each other. However, finding all protein interactions is both costly and labor-intensive. High-throughput experimental techniques (e.g., yeast two-hybrid, co-immunoprecipitation) are generating PPI networks data, which is being collected and stored in public databases, such as DIP [1] and SNDB [2].

* This research was supported by the "Exploratory Data Intensive Computing for Complex Biological Systems" project from U.S. Department of Energy (Office of Advanced Scientific Computing Research, Office of Science). The work of NFS was also sponsored by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, which is managed by UTBattelle for the LLC U.S. D.O.E. under contract no. DEAC05-00OR22725.
† Corresponding author: wenhong_tian@hotmail.com

Using PPI network data, integrated probability models are being developed to predict protein-protein interactions [3, 4]. Researchers [3-10] are trying to identify conserved patterns, such as ortholog groups and functionally similar pathways or protein complexes across species. The problem of identifying conserved network regions across species, known as the network alignment problem, is NP-hard [3-10]. This challenge attracts many researchers to find efficient heuristic solutions for the problem.

Most of previous network alignment tools focus on finding conserved functional patterns across species [3, 5, 6, 9] or on maximizing the overall number of matches between PPI networks globally [4]. The existing tools first find the conserved interaction regions by various search algorithms and then validate the results' accuracy against gene ontology (GO) [11] or KEGG ortholog (KO) groups [12] annotation.

In this paper, we explore a complementary approach to comparative analysis of PPI networks. Our focus is on fast identification of maximal conserved patterns across species with high accuracy validated against genome annotation. We propose an iterative connected-components-based algorithm with linear cost. This alignment algorithm combines information on HOmologs (genomic similarity), Physical interactions conservation and Equivalent functions (HOPE); we call our algorithm HopeMap**.**

1.1. Major Contributions

Our approach and focus are different from existing tools. The existing tools first find the conserved interaction regions by all kinds of search algorithms and then validate the results by gene annotations (e.g., GO, KO). Observing that the number of true homologs across species is relatively small compared to the total number of proteins in all species, we start with highly homologous clusters across species, find conserved interaction regions iteratively with a generic scoring system, and validate the results across multiple known functional annotations. Using known homolog clustering results, such as KO groups reduces the computational cost of finding sequence similar proteins by all kinds of searches in the existing tools. Applying connected-components based algorithm to find conserved regions across species assures a fast (if not the fastest) approach to identify maximal conserved patterns across species simultaneously; it is also a parameter-free clustering algorithm, unlike most of the existing tools that require setting up different sets of parameters for different alignments [8]. An iterative process can be applied to refine the identified regions. The generic scoring function is an open system; it currently combines

evolutionary evidence such as genomic, interaction and functional similarities; it can incorporate, if available and necessary, more features in the future.

1.2. Related Work

The network alignment has been formulated formally [3-10]. While there are some variations among the tools, the key idea is similar: to combine gene sequence information and PPI network information to find conserved interaction regions across species. Each PPI network may be represented as an undirected graph $G = (V, E)$, where V is the set of nodes and E is the set of edges (G may be a weighted graph, i.e., a weight measure $w(e)$ may be associated with each edge e in E). A biological interpretation of network alignment is to find functional homologues across different organisms.

The general problem, known as the problem of finding common subgraphs across multiple networks, is NP-hard; heuristic methods combining sequence, interaction and functional similarities have been developed to tackle it. *A network alignment graph* can be built across two or more species based on protein sequence similarity, interaction conservation, and functional coherence. The nodes in the alignment graph represent sets of proteins, ideally one from each species, and edges for the conserved PPIs across the compared species. The heart of network alignment algorithms is to find conserved interaction regions across homologs among different species.

The history and some future directions for biological network alignment are reviewed in [13]. In general, network alignment may be classified as:

1) pairwise or multiple in terms of the number of networks aligned,
2) local or global in terms of the number of nodes simultaneously aligned,
3) divergence/duplication evolution, neighboring topology in PPI networks, and functional categories in MIPS [14] (GO, KO, etc.) in terms of the type of guided models.

Table 1 summarizes the classification of different network alignment tools. Note that no multiple alignment results are provided in this paper.

Alignment of protein-protein interaction networks went through three major generations. In the first generation, often called pair-wise alignment, conserved pathways/complexes between two species were indentified. PathBlast [6] is one of the pioneering works in this category. In the second generation, called multiple alignment, tools such as NetworkBlast [3], which extends PathBlast, aimed to align multiple networks and introduce a probability model for interactions. MaWish [9] focuses on divergence/duplication model guided by the evolution. Graemlin 1.0 [5] introduces an integration probability model to predict the interactions and, unlike NetworkBlast, have aligned more than three networks. Aforementioned tools are also called local alignment because their

Pacific Symposium on Biocomputing 14: (2009)

search algorithms for the conserved regions start from small local regions and then greedily expand [10]. Later, IsoRank [4] introduced the global alignment concept by adopting Google PageRank algorithm idea to network alignment. Comparing to global alignment, which finds global maximal matches across species, local alignment may just find local maximal results.

The third generation alignment tools, such as Graemlin 2.0 [8] and NetworkBLAST-M [5], are trying to improve both accuracy and speed. Previous tools such as NetworkBLAST-M are also called progressive alignment with possibly exponential representation of every set of potential orthologous proteins [10], which makes them slow and memory inefficient. The computational cost of Graemlin 2.0 is claimed to be linear with the number of PPIs in all the species.

Table 1. A summary of different network alignment tools.

Tools	local	global	pairwise	multiple	guided model
PathBLAST (PB)	×		×		evolution
NetworkBLAST(NB)	×			×	evolution
NetworkBlast-M (NM)	×			×	evolution
MaWish (MW)	×		×		duplication/ divergence
Graemlin 1.0 (Gr1)	×			×	evolution
Graemlin 2.0 (Gr2)		×		×	evolution, duplication.
Isorank (ISO)		×	×		evolution
Isorank-M (ISM)		×		×	evolution

As reported in Graemlin 2.0, which may be the only one to compare results of all tools against KO groups, the specificity of NetworkBLAST, Graemlin 1.0, MaWish, IsoRank and Graemlin 2.0 varies from species to species with the average accuracy of 42%, 53%, 57%, 70%, and 81%, respectively.

2. Our Algorithm: HopeMap

Our algorithm, called HopeMap, can be described as follows (Figure 1 shows the five-step flow of HopeMap):

Step 1: Obtain and preprocess the PPI network data from PPI network databases, such as DIP and SNDB. Find all protein pairs that are interacting with each other in a species.

Step 2: Find highly similar protein sequences across the species. Use homolog clustering to identify homolog groups across different species based on all-versus-all BLAST scores or ortholog annotations. Existing tools, such as KO groups and INPARANOID [15] can be used for this purpose. Besides sequence similarity (homology), all the genes in the same KO group are assumed to be functionally similar. Once homolog groups are identified across the species, a

network alignment graph is built based on these groups. The nodes in the graph represent sets of proteins, ideally one from each species, in the same homolog group, and edges represent conserved protein-protein interactions across the compared species. One way of adding the edges between two node pairs (a_1,b_1) and (a_2,b_2) is when both (a_1,a_2) and (b_1,b_2) are directly interacting with each other in their corresponding PPI networks. Other rules for adding edges can be incorporated, such as those introduced in NetworkBLAST.

Step 3: Identify conserved protein interaction regions in the alignment graph. The major algorithm is based on strongly connected-components (clusters) in the alignment graph. A strongly connected component of a graph is a maximal set of vertices in which each vertex is reachable from another. The basic idea behind the connected-components approach is to use depth-first search. Figure 2 depicts a pseudocode for the connected-components algorithm.

Step 4: Score the identified clusters. For each cluster, *C*, our scoring function combines genomic similarity score, interaction conservation, and functional coherence. It is a normalized function with values from the [0, 1] interval so that it is convenient to compare the scores across different clusters. The scoring function of a cluster *C* is defined as:

$$Score(C) = w_1 S(C) + w_2 I(C) + w_3 F(C), \quad (1)$$

where *S(C)* is the sequence similarity score or the average confidence of homolog nodes in the cluster *C*, *I(C)* is the interaction conservation coefficient of
cluster *C*, and *F(C)* is the functional coherence score of cluster *C*; (w_1,w_2,w_3) is

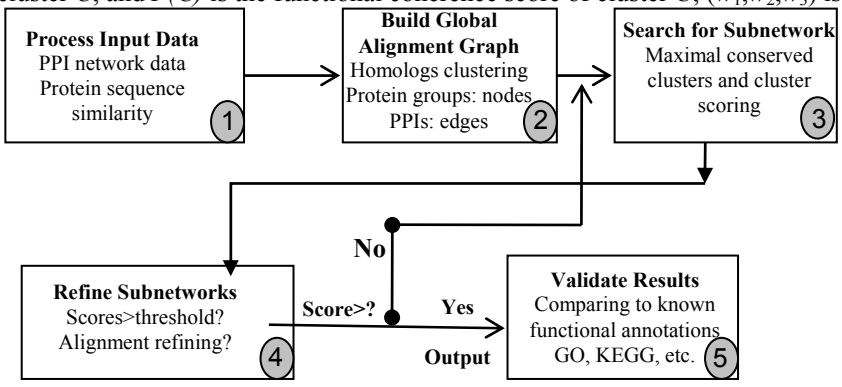

Figure 1. The HopeMap network alignment.

the corresponding weight coefficient and one-third for each is set as the default.

Similarity scoring S(C)—node scoring: We find highly similar proteins across species by identifying close homologues first. For each node (one protein from each species) of cluster *C* in the alignment graph, we find its

Pacific Symposium on Biocomputing 14: (2009)

confidence score based on BLAST score or ortholog confidence score. Equation (2) is used for the similarity score of a cluster *C*:

```
DFS(G): V[G] is the set of all vertices in G
(1)      For each vertex u in V[G]
(2)      Do color [u]:=white //unvisited
(3)        Pi[u]:=null; Time:=0
(4)  For each vertex u in V[G]
(5)     Do if color[u]==white
(6)      then DFS-Visit(u)
DFS-Visit(u)
(7)    color[u] :=gray; time:= time+1; d[u]:=time
(8)    for each v in Adj[u] // explore u's neighbors
(9)      Do if color[v]==white
(10)       then {Pi[v]:=u;   DFS-Visit(v)}
(11)  color[u]:=black; f[u]:=time<-time+1
Strongly-connected-components (G)
(12)  call DFS(G) to compute f[u] for each vertex u;  compute G^T, transpose of G
(13)  call DFS(G^T), consider the vertices in decreasing order of f[u] in the main loop of
      DFS
(14)  output the vertices of each tree in DFS forest formed
```

Figure 2. Pseudocode for finding strongly-connected components

$$S(C) = (\sum_{k=1}^{k=|C|} s(k))/|C|, \quad (2)$$

where *s(k)* is the similarity (or ortholog) confidence score of node *k* in cluster *C*. For simplicity, the confidence score of a node in the alignment graph is set to 1, if the proteins in the node have the same known ortholog annotation, such as KO group.

Interaction conservation scoring I(C)—node and edge scoring: The conserved interactions are the number of edges connecting all nodes in the identified cluster of the alignment graph. *I(C)* is set to the portion of the direct interactions conserved in the clusters. *I(C)* is formally defined as:

$$I(C) = \frac{i(C)}{|C|(|C|-1)/2}, \quad (3)$$

where *i(C)* is the total number of conserved interactions (defined above), i.e. the total number of edges in cluster *C*, and $|C|(|C|-1)/2$ is a cliquishness measure of the cluster *C*.

Functional coherence scoring F(C)—node and edge scoring: Functional annotations, such as GO terms, complexes in MIPS [14], or pathways in KEGG can be used as functional coherence score of a cluster. We use the intersection over the union of the number of GO biological process terms covered in a cluster of a local species as *F(C):*

$$F(C) = \frac{Intersection(\text{GO process terms in } C)}{Union\ (\text{GO process terms in } C)} \quad (4)$$

The larger the *F(C)* is, the higher the significance of the cluster *C* in terms of the coherence of GO biological process terms.

Statistical significance assessment: To measure the statistical significance of the scoring functions, for each cluster (or ortholog pair), we randomly sample *N* clusters of the same size and compute the corresponding scores. Then, we find an empirical *P*-value of each cluster using the methods introduced in [16, 17]. Typically, the empirical *P*-value can be estimated as $P=(R+1)/(N+1)$, where *N* is the total number of random samples and *R* is the number of those samples that produce a test statistic greater than or equal to the value for the actual data. Finally, the scoring function can be formulated as follows:

$$Score(C) = w_1 \frac{R_{S(C)}+1}{N_{S(C)}+1} + w_2 \frac{R_{I(C)}+1}{N_{I(C)}+1} + w_3 \frac{R_{F(C)}+1}{N_{F(C)}+1}, \quad (5)$$

where $N_{I(C)}$ is the number of random sample clusters, $R_{I(C)}$ is the number of clusters that have the score values greater than or equal to *I(C)*, ($N_{F(C)}$, $R_{F(C)}$) are defined similarly for *F(C)*. For simplicity, one-third is set as default for the weight coefficient of each of the three functions. For *S(C),* which is based on a homologue node, we find the number of nodes that belongs to the same homologue groups (e.g., KO groups) as the $R_{S(C)}$ and the total possible nodes across all species as $N_{S(C)}$. Notice that the three parts in the scoring function (5) can be used jointly or independently.

Step 5: Validate the results. Since our homolog groups are based on known KO annotation groups, we currently evaluate functional coherence of the identified local clusters in each species using gene ontology (GO). This is part of local alignment and refinement. To this end, GO termFinder [18] is used, which computes the empirical enrichment *P*-value and corrected values for multiple testing using the false discovery rate procedure. Similar to NetworkBLAST-M's way of measuring the results' specificity in terms of GO biological process terms enrichment, the percent of process-enriched clusters in each species is computed. The number of clusters with *P*-value below 0.05 is used as the sensitivity metric. Other known functional annotations, such as pathways or complexes enrichment in KEGG or MIPS, respectively, may also be used for validation.

After the fourth step, local alignment in each species can be iteratively applied to improve the cluster score, if necessary. If an identified cluster has the

score less than a threshold (e.g., 0.5), then HopeMap uses GO biological process terms enriched in each cluster as the indicator to keep or remove the cluster. To be more specific, we use the intersection over union (I/U) of GO biological process terms enriched in each cluster as an indicator to keep or remove the cluster. If the value of I/U is zero, we remove the cluster, otherwise, we keep it. Only two iterations were needed for convergence in the Analysis Results section.

The major step (Step 3) in HopeMap is based on identification of strongly connected-components in a graph. Its computational cost is linear with the number of nodes and edges in the alignment graph. All the other steps have linear cost with the input size. To the best of our knowledge, this is the fastest algorithm to identify maximal conserved patterns in the alignment graph.

3. Analysis Results

3.1. Input Data

Table 2 provides a summary of all the PPI networks analyzed in this paper. For GO terms, NetworkBLAST has higher specificity compared to other tools, as reported in NetworkBLAST-M. Therefore, we compare our results mainly against NetworkBLAST results for GO terms enrichment. For KO groups' enrichment, Graemlin 2.0 is reported to have the highest specificity and sensitivity, so we compare our results with Graemlin 2.0.

Table 2. The PPI networks analyzed in this paper.

Species (tax id, short name)	# proteins	#PPIs	Source
E.coli K12 (83333, eco)	4121	216426	SNDB[19]
S.typhimurium (99287,stm)	4239	94609	SNDB
C.crescentus (190650,ccr)	3365	40524	SNDB
S.cerevisiae (4932, sce)	4738	15417	DIP [18]
D.melanogaster (7227, dme)	7165	23484	DIP [18]

3.2. Alignment Results

The comparison of KO groups' enrichment against other tools is provided in Table 3. We provide specificity and sensitivity comparison in terms of KO groups for different global alignment tools (results for other alignment tools are also available in Graemlin 2.0). Specificity and sensitivity in terms of KO groups are introduced in Graemlin 2.0. In short, an equivalent class is defined as

correct if all protein members in the class are in the same KO group. The fraction of equivalence classes that are correct is denoted as C_{eq}, while the fraction of nodes that are in correct equivalence classes is C_{node}. C_{or} stands for the total number of nodes in correct equivalence classes and T_{ot} is the total number of equivalence classes with k species. (C_{eq}, C_{node}) and (C_{or}, T_{ot}) are used for measuring specificity and sensitivity, respectively. Other results but HM are from the original publication of Gr2, where Gr2 stands for Graemlin 2.0, ISO for Isorank, and HM for HopeMap. Since HopeMap uses KO groups for homologue clustering, it has higher specificity and sensitivity than the other tools in Table 3.

In Table 4, the results for NetworkBLAST (NB), 117 clusters with the node size larger than one, were obtained from the supplementary material of [2] using NB and INPARANOID ortholog groups. The results of HM1 were obtained using HopeMap on the same data from NetworkBLAST after adding edges to the alignment graph in a way similar to NetworkBlast. The results of HM-ko were obtained using KO groups as ortholog groups and no additional interactions being added to original PPI networks from DIP; the edges were added between node pairs (sce_1, dme_1) and (sce_2, dme_2) only when there were edges between both sce pairs (sce_1, sce_2) and dme pairs (dme_1, fly_2). NetworkBLAST used different techniques to add edges in the alignment graph; therefore, its total number of conserved regions is larger than for HopeMap. Table 4 shows that the specificity and sensitivity for GO terms enrichment by HopeMap are comparable to NetworkBLAST's, while HopeMap is simpler to use and is faster. Using KO groups in HopeMap improves the specificity. The range for cluster size in HM-ko is from 2 to 6, in NB is from 2 to 9, in NB-ko is from 4 to 6, and in HM1 is 2, 5, 7, and 50. HopeMap discovers more unique genes than NetworkBLAST, while keeping comparable specificities for GO terms' and KO groups' enrichment.

Table 3. Specificity and sensitivity comparison in terms of KO groups.

	eco/stm				**eco/ccr**				**sce/dme**			
Tools	**C_{eq}**	**C_{node}**	**C_{or}**	**T_{ot}**	**C_{eq}**	**C_{node}**	**C_{or}**	**T_{ot}**	**C_{eq}**	**C_{node}**	**C_{or}**	**T_{ot}**
NB	0.8	0.45	457	1016	0.78	0.5	346	697	0.39	0.14	43	306
ISO	0.9	0.91	2026		0.65	0.65	1014		0.63	0.63	534	
Gr2	0.96	0.96	2024		0.78	0.78	1012		0.73	0.73	637	
HM	1.00	1.00	2159	3151	1.00	1.00	1061	1365	1.00	1.00	768	1664

Table 4. Comparison of NetworkBLAST and HopeMap for Yeast/Fly.

Methods	Spec. (%) GO	Spec. (%) KO	#GO enriched	#conserved regions	#unique genes
sce (NB)	94.87	N/A	67	117	348
dme (NB)	84.62	N/A	62	117	256
sce (HM1)	98.73	N/A	65	79	1645
dme (HM1)	78.48	N/A	46	79	1913
sce (NB-ko)	100	100	9	9	34
dme (NB-ko)	100	100	8	9	27
sce (HM-ko)	100	100	24	26	747
dme(HM-ko)	92.31	100	24	26	753

Table 5. Specificity and sensitivity results for eco/stm and eco/ccr.

Species	Spec. (%) GO	Spec. (%) KO	#GO enrich.	#conserved regions	# unique genes
eco (eco/stm) HM	100	100	49	58	2085
stm (eco/stm) HM	96.55	100	46	58	2183
eco (eco/ccr) HM	95.24	100	37	42	1069
ccr (eco/stm) HM	90.48	100	31	42	1138
eco (eco/stm) NB	100	100	267	309	657
stm (eco/stm) NB	100	100	252	309	654
eco (eco/ccr) NB	100	100	34	39	108
ccr (eco/ccr) NB	100	100	38	39	100

In Table 5, we use the PPI network data from SNDB for two organismal pairs, eco/stm and eco/ccr. The interaction probability above 0.5 was set as the cut-off. The total number of conserved regions only included those sizes that are larger than two for eco/stm, while regions with the size larger than one were included for eco/ccr. The values for cluster size in eco/stm for HM are 3, 12, 14, 20, 22, 35, 97, 177, and 280, while the cluster sizes in eco/ccr for HM are 2, 3, 4, 5, 6, and 8; NB-ko has cluster size range between 4 and 15. HopeMap discovers more unique genes than NetworkBLAST does, while keeping comparable specificities for GO terms' and KO groups' enrichment.

4. Conclusion

Based on genome similarity across different species, interactome conservations and functional coherence, we developed a pairwise network alignment tool, called HopeMap, to improve the speed, accuracy and generality of the alignment.

HopeMap is fast; it is linear in terms of the number of nodes and edges in the alignment graph. Our results show that HopeMap has specificity and sensitivity as comparable with the existing best-performing tools. Specifically, in terms of GO terms' enrichment, HopeMap performs comparably with NetworkBLAST, and HopeMap has higher specificity and sensitivity in terms of KO groups' enrichment than the other tools. Our scoring system is generic, and the main algorithm is parameter-free. HopeMap is also extensible to multiple network alignment.

Acknowledgments

The authors would like to thank NetworkBLAST and NetworkBLAST-M team, Prof. Roded Sharan and Maxim Kalaev et al. as well as the Graemlin team (Jason Flannick et al.) from Stanford University for sharing the datasets and helpful communication.

5. References

1. DIP, *http://dip.doe-mbi.ucla.edu/*.
2. SNDB, *Stanford Network Database (SNDB), ftp://ftp-networks.stanford.edu/pub/graemlin_nets*.
3. Sharan, R., S. Suthram, R.M. Kelley, T. Kuhn, S. McCuine, P. Uetz, T. Sittler, R.M. Karp, and T. Ideker, *Conserved patterns of protein interaction in multiple species.* Proc Natl Acad Sci U S A, 2005. **102**(6): p. 1974-9.
4. Singh, R., J. Xu, and B. Berger, *Global alignment of multiple protein interaction networks.* Pac Symp Biocomput, 2008: p. 303-14.
5. Flannick, J., A. Novak, B.S. Srinivasan, H.H. McAdams, and S. Batzoglou, *Graemlin: general and robust alignment of multiple large interaction networks.* Genome Res, 2006. **16**(9): p. 1169-81.
6. Kelley, B.P., R. Sharan, R.M. Karp, T. Sittler, D.E. Root, B.R. Stockwell, and T. Ideker, *Conserved pathways within bacteria and yeast as revealed by global protein network alignment.* Proc Natl Acad Sci U S A, 2003. **100**(20): p. 11394-9.
7. Matthews, L.R., P. Vaglio, J. Reboul, H. Ge, B.P. Davis, J. Garrels, S. Vincent, and M. Vidal, *Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs".* Genome Res, 2001. **11**(12): p. 2120-6.
8. Flannick, J., A. Novak, C.B. Dol, B.S. Srinivasan, and S. Batzoglou. *Automatic Parameter Learning for Multiple Network Alignment*. in *RECOMB*. 2008.
9. Koyuturk, M., Y. Kim, U. Topkara, S. Subramaniam, W. Szpankowski, and A. Grama, *Pairwise alignment of protein interaction networks.* J Comput Biol, 2006. **13**(2): p. 182-99.

Pacific Symposium on Biocomputing 14: (2009)

10. Kalaev, M., V. Bafna, and R. Sharan. *Fast and Accurate Alignment of Multiple Protein Networks*. in *RECOMB*. 2008.
11. Ashburner, M., C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, *Gene ontology: tool for the unification of biology. The Gene Ontology Consortium.* Nat Genet, 2000. **25**(1): p. 25-9.
12. Kanehisa, M. and S. Goto, *KEGG: kyoto encyclopedia of genes and genomes.* Nucleic Acids Res, 2000. **28**(1): p. 27-30.
13. Sharan, R. and T. Ideker, *Modeling cellular machinery through biological network comparison.* Nat Biotechnol, 2006. **24**(4): p. 427-33.
14. Mewes, H.W., C. Amid, R. Arnold, D. Frishman, U. Guldener, G. Mannhaupt, M. Munsterkotter, P. Pagel, N. Strack, V. Stumpflen, J. Warfsmann, and A. Ruepp, *MIPS: analysis and annotation of proteins from whole genomes.* Nucleic Acids Res, 2004. **32**(Database issue): p. D41-4.
15. Remm, M., C.E. Storm, and E.L. Sonnhammer, *Automatic clustering of orthologs and in-paralogs from pairwise species comparisons.* J Mol Biol, 2001. **314**(5): p. 1041-52.
16. North, B.V., D. Curtis, and P.C. Sham, *A Note on the Calculation of Empirical P values from Monte Carlo Procedures.* Am. J. Hum. Genet., 2002. **71**: p. 439-441.
17. Zhang, B., B.H. Park, T. Karpinets, and N.F. Samatova, *From pull-down data to protein interaction networks and complexes with biological relevance.* Bioinformatics, 2008. **24**(7): p. 979-86.
18. GOtermFinder, *http://go.princeton.edu/cgi-bin/GOTermFinder*.