Genomics & Informatics Vol. 7(3) 141-147, September 2009

A New Approach to Find Orthologous Proteins Using Sequence and Protein-Protein Interaction Similarity

Min Kyung Kim[1], Young-Joo Seol[2], Hyun Seok Park[3], Seung-Hwan Jang[1], Hang-Cheol Shin[1] and Kwang-Hwi Cho[1]*

[1]Department of Bioinformatics and CAMDRC, Soongsil University, Seoul 156-743, Korea, [2]Research Planning & Information Div., National Institute of Agricultural Biotechnology, RDA, Suwon 441-707, Korea, [3]Department of Computer Science, Ewha Womans University, Seoul 120-750, Korea

Abstract

Developed proteome-scale ortholog and paralog prediction methods are mainly based on sequence similarity. However, it is known that even the closest BLAST hit often does not mean the closest neighbor. For this reason, we added conserved interaction information to find orthologs. We propose a genome-scale, automated ortholog prediction method, named OrthoInterBlast. The method is based on both sequence and interaction similarity. When we applied this method to fly and yeast, 17% of the ortholog candidates were different compared with the results of Inparanoid. By adding protein-protein interaction information, proteins that have low sequence similarity still can be selected as orthologs, which can not be easily detected by sequence homology alone.

Keywords: interolog, ortholog, protein-protein interaction

Introduction

Ortholog is a term of evolution, which means that proteins have originated from the same ancestor protein but exist in different species, in contrast to paralogs in the same species (Koonin, 2005). They have the nature of sequence and function similarity. Therefore, finding the ortholog of certain proteins is important for further analysis of the protein (Chervitz *et al*., 1998).

Automated proteome-scale ortholog identification methods are categorized into BLASTp-based, phylogeny-based, and evolutionary distance-based approaches (Chen *et al*., 2007). "Automated" means that there is no human interference, which frequently happens in the phylogenetic tree construction method. BLASTp-based methods have recently developed and mainly depend on sequence similarity. Phylogeny- and evolutionary distance-based methods have also been suggested to overcome the lack of evolutionary information of the BLASTp-based approach-for example, Orthostrapper (Hollich *et al*., 2002), and RIO (Zmasek and Eddy, 2002).

COG (Tatusov *et al*., 2001) is designed for finding prokaryote orthologs, and KOG (Koonin *et al*., 2004), Inparanoid (Remm *et al*., 2001), and OrthoMCL (Li *et al*., 2003) are for eukaryotes. The basic idea of COG is the use of the selection of a 3-way reciprocal best match and protein sequence comparisons that are conducted by BLASTp (Altschul *et al*., 1990). Once the proteins are selected as orthologs in COG, they are excluded in the candidate gene pool. Therefore, when an ortholog is defined, it can not be replaced by other reasonable candidates, even if they have higher scores. Because COG selects orthologs using domain information, it is difficult to apply it to eukaryotes, which are abundant in multi-domain proteins (Tatusov *et al*., 2001). For this reason, KOG has been developed for eukaryotes by the group who developed COG (Koonin *et al*., 2004). Inparanoid works by selecting orthologs by a reciprocal best hit, like COG. The method also considers the existence of paralogs. Therefore, Inparanoid defines ortholog relations that are sometimes one-to-many or many-to-many (Remm *et al*., 2001).

OrthoMCL (Li *et al*., 2003) is another kind of ortholog-finding system. This method can find orthologs among several species at the same time, which was impossible for other methods. OrthoMCL uses an all-against-all BLAST search and a clustering algorithm based on the Markov model to select an ortholog group.

However, these approaches above are basically using BLAST to check sequence similarity, and the proteins that have the best BLAST score are selected as the main orthologs. It has been known that proteins that have the best similarity scores in BLAST search are often not the closest relatives phylogenetically (Koski and Golding, 2001). This implies that there is a possibility that genuine orthologs can not be found using sequence similarity alone.

Not only can sequence similarity be used for ortholog finding but also structure and interaction data. Even though the sequence similarity is not discovered, the

*Corresponding author: E-mail chokh@ssu.ac.kr
Tel +82-2-820-0454, Fax +82-2-812-5762
Accepted 2 July 2009

proteins that have similar structures could conduct similar functions (Fribourg and Conti, 2003). Structural information, therefore, is an important key to classify the family and could be used to detect remote homologs. However, known protein structures are quite limited for use in genome-scale ortholog finding.

Alternatively, proteome-scale interaction data are available in *E. coli*, yeast, fly, worm, and human. It is known that proteins that have more interactors evolve more slowly (Fraser *et al*., 2002), interacting proteins co-evolve with their counterparts (Goh *et al*., 2000), and protein interfaces are more conserved than other surfaces (Caffrey *et al*., 2004). These findings are strong evidence of the relationship between evolution and protein-protein interaction data. The proteins that have conserved interactions are called "interologs" (Matthews *et al*., 2001; Yu *et al*., 2004). For this reason, we added a new criterion, "interaction data," to Inparanoid, which is a BLASTp-based ortholog finding system. The information of conserved protein-protein interactions is applied to identify functionally related proteins (Bandyopadhyay *et al*., 2006). However, they only analyze 121 cases of functional orthologs. In this paper, we suggest a genome-scale ortholog prediction system, OrthoInterBlast. In this method, not only sequence similarity but also protein-protein interaction data are used for finding orthologs across species. Comparison of the proteom-scale ortholog prediction system is shown in Table 1.

Methods

OrthoInterBlast predicts the orthologs by the following steps; (1) choosing ortholog candidates based on sequence similarity, (2) comparing the interacting partner of the ortholog candidates, (3) scoring the sequence and interaction similarity by graph alignment, and (4) deciding orthologs according to their score. OrthoInterBlast consists of three different modules: a sequence score module, interaction score module, and resolve module. The architecture of OrthoInterBlast is shown in Fig. 1.

Input data and Pre-processing

OrthoInterBlast requires 4 types of files as inputs: protein ID, protein sequence, sequence similarity, and interaction data. Protein ID file consists of a SwissProt or TrEMBL ID and its related information, such as the protein name and description (Boeckmann *et al*., 2003). Except for the protein ID, other information is not essential. They can be used as additional information for verification only. The information of protein sequences came from Inparanoid. We used 18,932 protein sequences of fly (*Drosophila melanogaster*) and 6706 protein sequences of yeast (*Saccharomyces cerevisiae*). The third one is the BLAST result of fly-fly, fly-yeast, yeast-fly, and yeast-yeast. In OrthoInterBlast, a proteome-scale BLAST search should be performed 4 times to find orthologs of the two species, because a change

Table 1. Comparison of the proteome-scale ortholog prediction systems

	OrthMCL	Inparanoid	OrthoInterBlast
BLAST version	WU-BLAST	NCBI-BLAST	BLASTp
BLAST Search	All-against-all	All-against-all	All-against-all
Similarity cutoff	p < 1e-5	Score ≥ 50 bits	Overlap > 50%
		Score ≥ 50 bits	Overlap > 50%[1]
Reciprocal best hits	p-value	Percent identity score	Percent identity score

[1]These options can be changed by the user in OrthoInterBlast.

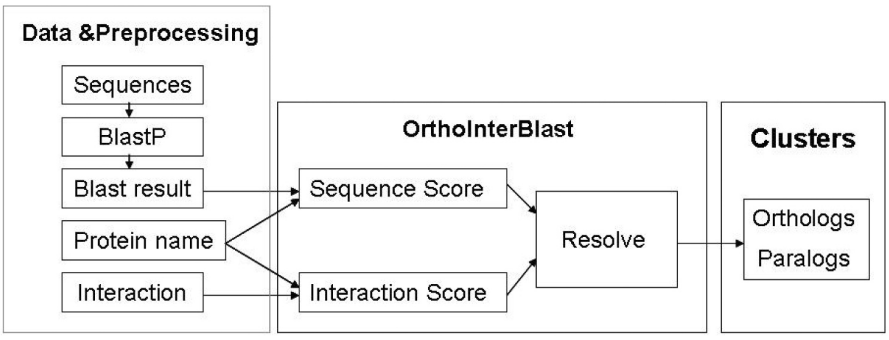

Fig. 1. System Architecture of OrthoInterBlast.

of query and target in a BLAST search sometimes gives different results. Therefore, we compared fly-yeast and yeast-fly separately and used the mean value as a similarity score (Li *et al*., 2003).

If we conduct a BLAST search whenever OrthoInterBlast is running, the speed of the system will be decreased. Therefore, we conducted BLASTp prior to the main search and saved the result as an input file. The conditions for the BLASTp search were as follows: Grey zone is 0 bits, Score cutoff is 50 bits, In-paralogs, paralogs defined by Inparanoid, with confidence less than 0.05, Sequence overlap cutoff is 0.05, Group merging cutoff is 0.05, and Scoring matrix is BLOSUM62 (Li *et al*., 2003).

The result of BLASTp between different species, such as fly-yeast and yeast-fly, is used to find ortholog candidates. For example, *a′* and *a″* show the similarity of *a* and *a′*, and *a″* is the ortholog candidate in OrthoInterBlast. Then, the information is analyzed further by using the interaction score module for finding the main ortholog. The result of BLAST between the same species, such as fly-fly and yeast-yeast, will be used to find the paralog.

The last input data are the information of interacting partners. The information originates from the DIP (Database of Interacting Proteins) (Xenarios *et al*., 2002). The numbers of data used in OrthoInterBlast that are the overlap of both databases (DIP and Inparanoid) are 12,868 interactions among 4398 proteins and 18,197 interactions among 6628 proteins for yeast and fly, respectively (Table 2).

OrthoInterBlast

OrthoInterBlast consists of *sequence, interaction*, and *resolve* modules, as shown in Fig. 1. First, the sequence module selects ortholog candidates according to their sequence similarity. The threshold of sequence similarity score was set to the cutoff value. The current default value is 50 bit homology, as used in Inparanoid, and each candidate also has the sequence similarity score.

Table 2. The numbers of proteins and interactions used in OrthoInterBlast

	Organism	
	Fly	Yeast
No. of proteins from DIP	7,052	4,749
No. of proteins from Inparanoid	18,931	6,705
No. of interactions from DIP	20,789	15,131
No. of proteins from OrthoInterBlast	6,628	4,398
No. of interactions from OrthoInterBlast	18,197	12,868

Then, the candidates enter the *interaction* module. OrthoInterBlast performs a local graph alignment only for the candidates that have a sequence score above the cutoff value. The process of the local graph alignment is as follows: (1) If *A*'s ortholog candidate is *a*, it generates new node *Aa*. (2) A's interacting partner is *B*, *D*, and *F* in yeast. Its ortholog candidates are *b*, *d*, and *f*, which are interacting with *a* in fly. Then, the new nodes *Bb*, *Dd*, and *Ff* are generated by direct interaction with *Aa* in local graph alignment. (3) *c* protein is related to *a* directly, but *C* protein is related to *A* through *B* (the distance between *C* and *A* is 2). This situation between *Aa* and *Cc* is defined as a "gap" in graph theory. The relation between *Bb* and *Cc* is the same. (4) The "mismatch" stands for the case that has the same distance through a different bridge node, as shown in the relation between *Aa* and *Ee* (Fig. 2).

Local graph alignment is similar to sequence alignment. "Direct interaction" in the graph is the same as "match" in the sequence alignment. "Gap" and "mismatch" are used terms in sequence alignment, too. These concepts are applied in PathBlast (Kelley *et al*., 2004) at the first time, which is aligning two different protein interaction networks to find conserved pathways. OrthoInterBlast follows the same graph alignment rule to find new ortholog candidates from the interaction network.

After the construction of a global network for each ortholog candidate, we were able to get the score of interaction similarity. We calculated the interaction similarity scores of each ortholog candidate according to the sum of the direct, gap, and mismatch interaction numbers with their weights. The weights were 10, 7,

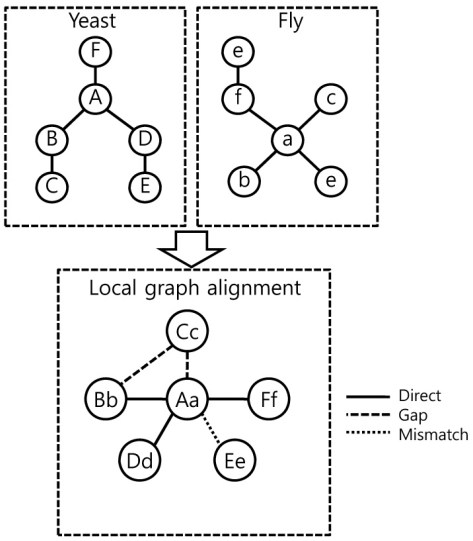

Fig. 2. Construction of local graph alignment.

and 4 for the direct, gap, and mismatch interaction, respectively, which are used in the PathBlast scoring function. In case of Fig. 2, 3 direct interactions, 2 gaps, and 1 mismatch were found, so the total sum will be 48 ($3\times10+2\times7+1\times4=48$). Then, the total sums are divided by the average number of interacting protein partners.

Finally, the sum of the calculated sequence and interaction similarity scores enters the *resolve* module as an input. Then, the *resolve* module defines orthologs and paralogs from the candidates by using their interaction and sequence similarity scores. The final decision of ortholog and paralog is made by the rules used in Inparanoid (Remm *et al.*, 2001). The major difference from Inparanoid is that the score in OrthoInterBlast contains interaction similarity scores.

Results

OrthoInterBlast suggests 1922 ortholog groups between fly and yeast. Part of the results is shown in Fig. 3, and the full results are available in the supplementary materials. For further analysis, we compared the results with Inparanoid (Fig. 4) and summarized them in Table 3. An identical cluster should have the same ortholog and paralog pair. A non-identical cluster can be classified to mismatch and match groups. *Match* means that its cluster has the same ortholog protein but has a different paralog protein(s). *Mismatch* has a different ortholog protein pair but includes it as a paralog protein.

To find out the relative contributions of two pieces of information, sequence and interaction information, 2 sets of the ratio between sequence and interaction scores were tested (50:50, and 20:80). As the portion of the interaction score increased, the identical cluster with the result of Inparanoid decreased from 83% for 50:50 to 78% for 20:80. The number of clusters found with OrthoInterBlast decreased slightly compared to that of Inparanoid. Because proteins that are used in OrthoInterBlast should have the protein interaction information as well, the proteins without any interaction information were eliminated prior to the OrthoInterBlast calculation (As shown in Table 2).

For the case of 50:50, the number of identical clusters is 1622 out of 1963 (83%). The 17% difference has been made by introducing the interaction score into OrthoInterBlast compared to Inparanoid. The trend for

========== **0 cluster** ==========	
RIR2_DROME : 1.0	**RIR2_YEAST** : 1.0
DESC : *Ribonucleoside diphosphate reductase small subunit (CG8975-PA)*	DESC : *ribonucleoside-diphosphate reductase small chain*
Total Score : 68.35858585858585	Total Score : 68.35858585858585
seq Length : 393.0	seq Length : 399.0
Seq Score : 54.7979797979798%	Seq Score : 54.7979797979798%
Inter Score : 99.99999999999997%	Inter Score : 99.99999999999997%
Target protein: RIR2_YEAST	Target protein: RIR2_DROME
---interaction pairs-----	---interaction pairs-----
Q9VIT3:*CG10364-PA open reading frame*	ALG5_YEAST:*dolichyl-phosphate beta-glucosyltransferase*
Q9V8M7:*CG15080-PA open reading frame*	RIR4_YEAST:*ribonucleoside-diphosphate reductase chain RNR4*
Q9V996:*CG18584-PA open reading frame*	CYAA_YEAST:*adenylate cyclase*
Q9VCC9:*CG6147-PA open reading frame*	DBP8_YEAST:*helicase homolog*
Q8SX59:*CG9083-PA open reading frame*	RA10_YEAST:*RAD10 protein*
----------	COPP_YEAST:*coatomer complex beta' chain*
	SIR3_YEAST:*regulatory protein SIR3*
	PP11_YEAST:*phosphoprotein phosphatase SIT4*
	SMK1_YEAST:*protein kinase SMK1*
	TEM1_YEAST:*GTP-binding protein TEM1*
	WTM2_YEAST:*transcription modulator WTM2*
	YAH3_YEAST:*FUN16 protein*
	Q04177:*hypothetical protein YDR398w*
	ASF1_YEAST:*ASF1 protein*
	P89501:*nonhistone chromosomal protein NHP6B*

Fig. 3. Snapshot of OrthoInterBlast result. The ortholog pairs are shown in the first line of each cluster number. Protein function description, score, and interaction pair list are supplied as additional information.

identical			
NAME:Q9VBU7 ID:1904 Score:1.993007 Confidence:1.0	NAME:Q9VBU7 ID:1390 Score:114.0 Confidence:1.0	NAME:YRB1_YEAST ID:1904 Score:1.993007 Confidence:1.0	NAME:YRB1_YEAST ID:1390 Score:114.0 Confidence:1.0
--Match-- : OrthoInterBlast cluster:1903 Inparanoid cluster:1495			
NAME:Q8IH66 ID:1903 Score:2.3028786 Confidence:1.0 NAME:Q9VC62 ID:1903 Score:2.1389241 Confidence:1.0 NAME:Q8MRR8 ID:1903 Score:1.9977762 Confidence:1.0 NAME:Q9V4C4 ID:1903 Score:1.4836112 Confidence:0.3518163	NAME:Q8IH66 ID:1495 Score:101.0 Confidence:1.0 NAME:Q9VC62 ID:1495 Score:101.0 Confidence:1.0	NAME:YMH2_YEAST ID:1903 Score:2.1389241 Confidence:1.0 NAME:YNI7_YEAST ID:1903 Score:2.3028786 Confidence:1.0 NAME:Q12466 ID:1903 Score:1.9977762 Confidence:1.0	NAME:YMH2_YEAST ID:1495 Score:101.0 Confidence:1.0 NAME:YNI7_YEAST ID:1495 Score:101.0

Fig. 4. Comparison of ortholog groups identified by OrthoInterBlast with those with Inparanoid (Each column indicates OrthoInterBlast Fly, Inparanoid fly, OrthoInterBlast yeast, Inparanoid yeast from left to right. This is a snapshot of OrthoInterBlast output).

20:80 is similar to that of 50:50 (Table 3).

We tried to verify the results in several ways. First, the EC number was used to evaluate the quality of the predicted orthologs with OrthoInterBlast. However, EC numbers are mostly based on sequence similarity and the groups in EC gene numbering system that are larger than the clusters in OrthoInterBlast and Inparanoid. Most of the EC numbers of the members in the clusters from both methods are the same, so it does not discriminate the superiority of our systems to Inparanoid. Second, structural similarity among orthologs has been considered. Unfortunately, structural information for yeast and fly are not enough to verify the result. As the structural information is increased, the results will be verified in the future.

Most of the methods so far have been developed based on sequence similarity, so it has a genuine limitation for finding remote homology that has very low sequence similarity. Even though it was not successful to find a proper way of validating our method, this method can overcome the genuine limitation of a sequence- only-based method, and the ability of the method could be improved as the interaction data increased. Also, global graph alignment without using sequence similarity would improve the predictability of OrthoInterBlast.

Table 3. Comparison of ortholog groups identified by OrthoInterBlast with Inparanoid

50:50[1]	No. of fly proteins	No. of yeast proteins
Clusters from OrthoInterBlast	3,398	2,368
Clusters from Inparanoid	3,792	2,473
Match	104	40
Mismatch	284	12

20:80[1]	No. of fly proteins	No. of yeast proteins
Clusters from OrthoInterBlast	3,185	2,239
Clusters from Inparanoid	3,792	2,473
Match	98	38
Mismatch	247	16

Ratio	OrthoInterBlast	Identical clusters with Inparanoid	Percent of identical clusters[2]
50:50[1]	1,922	1,622	83%
20:80[1]	1,814	1,533	78%

[1]Ratio between the sequence similarity and interaction data used, [2]Compared to Inparanoid.

Discussion

One advantage of OrthoInterBlast is its expansibility to other groups of species when genomic-scale interaction data are available. We also applied it to *E. coli*, yeast, fly, C. *elegans*, mouse, and humans using OrthoInterBlast. The tendencies are also shown to have the same trend with fly-yeast (you can find these results at ebio.ssu.ac.kr/OIB～).

The difference between Ideker's methods (Bandyopadhyay *et al*., 2006) and OrthoInterBlast is the usage of sequence information. Although sequence similarity decides the first ortholog candidate in Ideker's system and OrthoInterBlast, sequence score also contributes to the final decision of the ortholog in OrthoInterBlast. It is because we want to reduce the effect of false positive interaction data (Li *et al*., 2006).

We suggest a new system for finding ortholog proteins, OrthoInterBlast, based on sequence similarity and graph alignment. The result is compared to that of Inparanoid, and 17% of clusters are different between both methods. For verification, comparing the EC number shows as good a result as Inparanoid. Structural verification was not successful, because structural information for yeast and fly are very limited. However, including interaction data is very useful to find orthologs that have very low sequence similarity (remote homologous protein). As the quantity of interaction data increases, OrthoInterBlast can be a more powerful tool for functional ortholog finding.

Acknowledgements

This work was supported by a Korea Research Foundation Grant (KRF-2005-005-J01101).

References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. *J. Mol. Biol.* 215, 403-410.

Bandyopadhyay, S., Sharan, R., and Ideker, T. (2006). Systematic identification of functional orthologs based on protein network comparison. *Genome Res*. 16, 428-435.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., and Phan, I. (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucl. Acids Res*. 31, 365-370.

Caffrey, D.R., Somaroo, S., Hughes, J.D., Mintseris, J., and Huang, E.S. (2004). Are protein-protein interfaces more conserved in sequence than the rest of the protein surface? *Protein Sci*. 13, 190-202.

Chen, F., Mackey, A.J., Vermunt, J.K., and Roos, D.S. (2007). Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS One* 2, e383.

Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A., Koonin, E.V., Dwight, S. S., Harris, M.A., Dolinski, K., Mohr, S., and Smith, T. (1998). Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 282, 2022-2028.

Fraser, H.B., Hirsh, A.E., Steinmetz, L.M., Scharfe, C., and Feldman, M.W. (2002). Evolutionary rate in the protein interaction network. *Science* 296, 750-752.

Fribourg, S., Conti, E. (2003). Structural similarity in the absence of sequence homology of the messenger RNA export factors Mtr2 and p15. *EMBO Rep*. 4, 699-703.

Goh, C.S., Bogan, A.A., Joachimiak, M., Walther, D., and Cohen, F.E. (2000). Co-evolution of proteins with their interaction partners. *J. Mol. Biol*. 299, 283-293.

Hollich, V., Storm, C.E., and Sonnhammer, E.L. (2002). OrthoGUI: graphical presentation of Orthostrapper results. *Bioinformatics* 18, 1272-1273.

Kelley, B.P., Yuan, B., Lewitter, F., Sharan, R., Stockwell, B.R., and Ideker, T. (2004). PathBLAST: a tool for alignment of protein interaction networks. *Nucl. Acids Res*. 32, W83-W88.

Koonin, E.V., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Krylov, D.M., Makarova, K.S., Mazumder, R., Mekhedov, S.L., Nikolskaya, A.N., and Rao, B.S. (2004). A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol*. 5, R7.

Koonin, E.V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet*. 39, 309-338.

Koski, L.B., and Golding, G.B. (2001). The closest BLAST hit is often not the nearest neighbor. *J. Mol. Evol*. 52, 540-542.

Li, L., Stoeckert, C.J.Jr., and Roos, D.S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Res*. 13, 2178-2189.

Li, D., LI, J., Ouyang, S., Wang, J., Wu, S., Wan, P., Zhu, Y., Xu, X., and He, F. (2006). Protein interaction networks of saccharomyces cerevisiae, caenorhabditis elegans and drosophila melanogaster: large-scale organization and robustness. *Proteomics* 6, 456-461.

Matthews, L.R., Vaglio, P., Reboul, J., Ge, H., Davis, B.P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of Potential Interaction Networks Using Sequence-Based Searches for Conserved Protein-Protein Interactions or "Interologs". *Genome Res*. 11, 2120-2126.

Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol*. 314, 1041-1052.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. (2001). The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucl. Acids Res*. 29, 22-28.

Xenarios, I., Salwinski, L., Duan, X.J., Higney, P., Kim, S.M., and Eisenberg, D. (2002). DIP, the Database of Interacting Proteins: a research tool for studying cellular net-

works of protein interactions. *Nucl. Acids Res*. 30, 303-305.

Yu, H., Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M., and Gerstein, M. (2004). Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*. 14, 1107-1118.

Zmasek, C.M., and Eddy, S.R. (2002). RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics* 3, 14.