2009 Ohio Collaborative Conference on Bioinformatics

Comparative Analysis of Modularity in Biological Systems

Xin Li*[1], Sinan Erten*[1], Gurkan Bebek†‡, Mehmet Koyutürk*†, Jing Li*§[2]
Email: xxl41@case.edu, msinane@gmail.com, gurkan@case.edu, koyuturk@eecs.case.edu, jingli@eecs.cwru.edu
*Department of Electrical Engineering and Computer Science
†Case's Center for Proteomics and Bioinformatics
‡Case Comprehensive Cancer Center
§Department of Epidemiology and Biostatistics
Case Western Reserve University, Cleveland, OH

***Abstract*—In systems biology, comparative analysis of molecular interactions across diverse species indicates that conservation and divergence of networks can be used to understand functional evolution from a systems perspective. A key characteristic of these networks is their modularity, which contributes significantly to their robustness, as well as adaptability. In this paper, we investigate the evolution of modularity in biological networks through phylogenetic analysis of network modules. Namely, we develop a computational framework, which identifies modules in networks of diverse species independently and projects these modules into the networks of other species, with a view to capturing the evolutionary trajectories of functional modules. These trajectories can then be used to reconstruct modular phylogenies and whole-network phylogenies, or to enhance identification of functional modules. In the context of phylogeny reconstruction, our experiments on a comprehensive collection of simulated and real networks show that comparison of networks based on module trajectories is more informative than other measures of network similarity. These results demonstrate the key role of modularity in the functional evolution of biological systems and motivate further investigation of the evolution of functional modules.**

I. Introduction

As a fundamental concept, evolution has profound implications in a variety of applications in modern molecular biology; *e.g.*, functional annotation of DNA/protein sequences through comparative sequence analysis has become an important and integral part of biological sciences [1]. Accurate reconstruction of the evolutionary history of species, usually represented by a phylogenetic tree, is critical for the success of such applications. Phylogenetic analysis of molecular sequence data has drawn significant attention ever since protein/DNA sequences have become available [2], [3]. There exist many models (from the simplest Jukes-Cantor model to more complex General Time Reversible model), but all of them specify site evolutions at the DNA level for obvious reasons: structure constraints (secondary structures of RNAs and tertiary structures of proteins) are hard to model. Based on sequence evolution, different approaches have been developed to either explicitly use an evolutionary model (*e.g.*, Maximum Likelihood) or approximate one (*e.g.*, Maximum Parsimony) [4].

1) Comparative network analysis: Availability of high-throughput data that relates to the organization and dynamics of biological systems enables understanding of biological functions from a systems perspective [5]. An important source of data that pertains cellular organization and signaling is in the form of physical interactions between proteins, organized into genome scale protein-protein interaction (PPI) networks [6]. Comparison of recently available PPI networks that belong to diverse model organisms reveals that parts of extant molecular networks are conserved across diverse species [7]–[9]. Furthermore, it is observed that proteins that are organized into cohesive interaction patterns are more likely to be conserved [10]. Comparative network analysis is also shown to enhance the performance of computational approaches to basic problems in functional genomics, such as identification of orthologs across species and annotation of protein functions [11], [12]. Furthermore, recent studies show that incorporation of evolutionary models and knowledge enhances the performance of network alignment methods significantly [13]. Such findings demonstrate that network comparisons provide essential biological information beyond what is gleaned from the genome [14]. Consequently, phylogenetic analysis of network topology and function has the potential to provide key insights on the evolution of biological functions at a systems level [15].

2) Phylogenetic analysis of network modularity: In this paper, we propose a modularity based approach to phylogenetic network analysis. The proposed framework, MOPHY, is illustrated in Figure 1. Our approach differs fundamentally from existing approaches, in that we focus on the conservation and divergence of modular components, rather than one-to-one comparison of network topologies. In our framework, we first identify modular subgraphs in different networks independently. Then, we project these modules on networks of other species to understand the conservation and divergence of different modular processes in these networks. While projecting a module on different species, we rely on the conservation of network proximity between homologs (proteins with significant sequence similarity) of its constituent proteins in other networks. Consequently, by utilizing network information, our approach captures functional evolution beyond conservation

[1]contributed equally
[2]corresponding author

978-0-7695-3685-9/09 $25.00 © 2009 IEEE
DOI 10.1109/OCCBIO.2009.29

of sequences. Namely, network information is incorporated into the analysis by (i) considering network modules as "features" of each network and (ii) assessing the conservation of modularity in terms of the network proximity between proteins with conserved sequences.

3) Evaluating network comparison methods: In this study, we also take a novel approach to the calibration and validation of comparative network analysis methods. Based on theoretical models on the evolution of molecular interactions [19], [20], we simulate network evolution to generate networks with known underlying phylogeny. Then, we use our algorithms on the generated networks to reconstruct a phylogenetic tree. By comparing the reconstructed tree with the underlying tree, we evaluate the performance of different methods and assess the effect of various parameters on the accuracy of our methods. We also use simulated networks to evaluate the robustness of our methods against noise and missing data, by perturbing the simulated networks to mimic data collection processes. Extensive experiments on simulated data show that the proposed algorithm is extremely successful in reconstructing the underlying phylogenies and is quite robust to noise. We also use the proposed method, MOPHY, to reconstruct the phylogeny of seven species, for which reasonable interaction data is available. We show that our method constructs a phylogenetic tree that is in accordance with the phylogenetic relationships and evolutionary distances inferred by independent methods. Furthermore, we demonstrate that MOPHY outperforms existing phylogenetic network analysis methods.

4) Learning from phylogenetic network analysis: It should be noted that the application of the proposed framework extends well beyond phylogenetic tree reconstruction. The methods and results presented here rather constitute a step towards establishing modularity based analysis as a key approach in understanding the functional evolution of cellular organization. Indeed, our results show that conservation of modularity and network proximity is likely to provide useful insights into the evolutionary histories of networks, by providing statistical evidence for the following observations:

- Network modules are likely to be conserved more in evolutionarily closer species, in terms of the network proximity between the homologs of their constituent proteins (Figure 3).
- Conservation of network proximity is a better indicator of evolutionary relationships when modular network components are considered (as opposed to the proximity between arbitrary proteins) (Tables I and II).
- Modularity based analysis is quite robust to noise and missing data in terms of capturing evolutionary relationships, and therefore may be more promising in comparative analysis of extant protein-protein interaction networks, which are highly incomplete and susceptible to noise (Figure 2(c))).

These results motivate elaborate studies of modular evolution, including identification of module families and reconstruction of evolutionary trajectories for these module families, which in turn will be useful in constructing the "periodic table of systems biology" [5].

II. METHODS

Our modularity-driven approach to phylogenetic network analysis, MOPHY, which is illustrated in Figure 1, can be summarized as follows:

1) Considering each extant network individually, identify network modules that represent functional and topological properties of each network.
2) Project modules identified on each network to other extant networks, based on the conservation of functional and topological properties, to obtain a *module map* for each species. A module map can be thought of as a mathematical representation of the conservation of extant network modules in the corresponding species.
3) Using module maps, compare networks of diverse species to construct network phylogenies.
4) Using resulting network phylogenies, investigate the evolutionary histories of extant network modules to gain insights on the evolution of functional modularity.

III. RESULTS AND DISCUSSION

In this section, we evaluate the performance of MOPHY on simulated, as well as real data - in terms of (i) success in accurately reconstructing the underlying phylogeny, (ii) robustness to noise and missing data, and (iii) performance as compared to existing algorithms.

A. Results on Simulated Data

We test our method on synthetic networks generated by simulation of the evolutionary process. In our experiments, in order to keep the size of the networks at a realistic scale with sufficient variability, we set the average number of proteins in a network to 3000, with a standard deviation of $\sigma = 1000$. Here, average network size is kept relatively smaller as compared to that of extant networks for feasibility constraints, since these experiments are performed multiple times to assess statistical significance and the effect of varying parameters. Our results on extant networks show that the method also scales to larger networks and is applicable in practice. Using this configuration, we generate ten networks for each experiment. For all experiments, we generate five different instances, and for each performance figure, we report the average over these five instances. Note that, in these experiments, the interactions are not associated with reliability scores.

1) Evaluating performance: Comparison of phylogenetic trees: In order to quantify the performance of a tree reconstruction method, it is necessary to compare the reconstructed tree with the underlying tree based on a sound measure of similarity between two phylogenetic trees. For this purpose, we investigate the similarity between the two phylogenetic trees by Nodal distance [33]. Nodal distance takes into account the branch lengths and computes a similarity metric by comparing the sum of distances of every node pair in each tree. We use this method to evaluate the performance of algorithms in capturing the evolutionary distances between different networks.

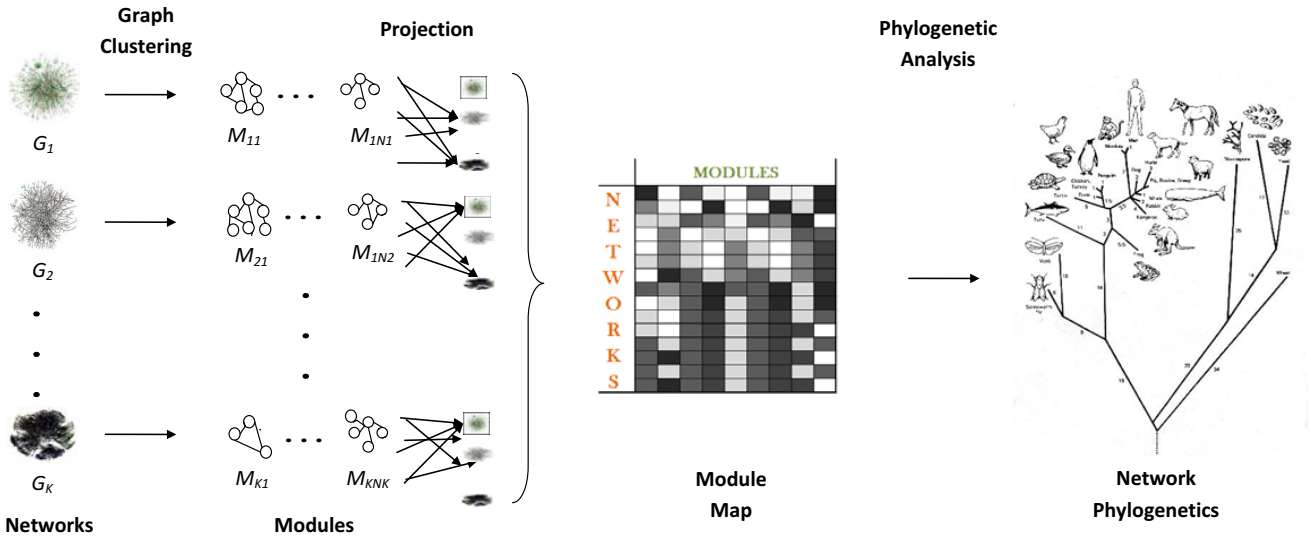

Fig. 1. Modularity Based Phylogenetic Analysis of Molecular Interaction Networks.

TABLE I
COMPARISON OF PERFORMANCES OF MOPHY WITH USING RANDOM PROTEIN MODULES, USING PROTEIN SIMILARITIES ONLY AND RDL.

METHOD	Run 1	Run 2	Run 3	Run 4	Run 5	Avg.
MOPHY	5.28	5.85	8.82	5.82	8.15	6.79
RDL	14.40	15.50	19.80	13.12	14.83	15.53
Random Modules	14.81	9.33	11.54	8.11	12.6	11.29
Protein Similarity	11.72	11.56	13.96	11.22	10.75	11.84

For the simulated networks, we compare the performance of MOPHY with the random module method, RDL and a method that only uses protein similarity in the networks. Nodal distance for five simulation instances as well as the average values of these five runs are shown in the table. For MOPHY, the result used is the best performance achieved with coverage 0.60 and diameter 2, by using the most specific modules. Similarly for the random module method, the best performance is achieved for the instance with coverage 0.40 and diameter 4, with the most comprehensive modules. As clearly seen, MOPHY outperforms the other methods in terms of capturing the evolutionary distances between species.

2) Comparison of performance with other methods: We compare the performance of MOPHY in reconstructing the correct phylonegy to that of three alternate methods

- **RDL:** An existing method for network-based phylogeny reconstruction, which uses relative description length to assess the similarity between networks [15].
- **Random Modules:** This method implements a similar algorithm with MOPHY, but it uses random groups of proteins as modules. These random modules are selected in a way that they reflect the modules incorporated by MOPHY in terms of their quantity and size distribution. We use this method as a reference method to assess the contribution of the information on modularity in reconstructing the correct phylogeny.
- **Only Protein Similarity:** This method incorporates only the similarities between proteins to reconstruct a phylogenetic tree. Namely, we still compute feature vectors for each network, but each entry of the feature vector represents the conservation (score of best sequence similarity match) of a single protein. The purpose of using this method as a reference is to assess the contribution of the use of network information (proximity and modularity) in reconstructing the correct phylogeny.

The comparison of the performances of these methods over five different instances, obtained through simulation of network evolution, is shown in Table I. As seen on the table, MOPHY performs drastically better than any of the three alternate methods in terms of minimizing the nodal distance between the correct evolutionary history and the reconstructed evolutionary history. Furthermore, the Random Modules method performs clearly better than RDL, suggesting that incorporation of network proximity, i.e., aggregation of interactions, is more useful than incorporation of network topology, i.e., incorporation of single interactions, in capturing the similarity of networks. However, comparison of the performances of Random Modules and Only Protein Similarity suggests that, when modularity is not considered, incorporation of network information provides marginal improvement.

Finally, to evaluate the performance of MOPHY statistically, we evaluate the statistical significance of its performance with respect to the Random Modules method. The performance difference between MOPHY and Random Modules can be thought of as an indicator of the usefulness of relying on conservation of modular network structures as opposed to arbitrary (groups of) proteins and their interactions. We quantify the statistical significance of the performance difference between MOPHY and Random Modules based on Student's t-test to compare the means of two populations. The p-value for an experiment gives the probability that an algorithm that incorporates sequence conservation and network proximity, but not modularity can achieve as good as MOPHY solely based on chance.

3) Design parameters and module selection: In MOPHY, the module identification process can be tuned by adjusting several parameters: (i) The threshold on proximity adjusts the trade-off between the tightness and comprehensiveness of modules (higher threshold on proximity results in smaller

and more tightly coupled modules). Since the interactions in the simulated networks are unweighted, we use *diameter*, *i.e.*, the maximum distance between two proteins in a module, to represent the proximity threshold. (ii) As multiple modules are identified in each network, using all modules in phylogeny reconstruction may lead to problems associated with high-dimensionality. Therefore, we investigate the effect of network coverage provided by the modules considered, where *coverage* is defined as the percentage of proteins included in the selected modules. (iii) In order to understand which modules are more informative, we consider two different module selection strategies: *most specific*, *i.e.*, the set of smallest (with size ≥ 3) modules for a given coverage or *most comprehensive*, *i.e.*, the set of largest modules for a given coverage.

4) Performance of MOPHY *for different parameters:* Detailed statistics on the comparison of underlying and reconstructed phylogenetic trees for a sample instance are shown in Tables II, and in Figure 2. As seen in Table II, for any configuration of parameters, the accuracy of the phylogenetic tree reconstructed by MOPHY is highly significant. In general, more specific (smaller) modules appear to be more informative. Indeed, as seen in Table II, when evolutionary distances are considered, the performance with more comprehensive (larger) modules is not statistically significant. Furthermore, performance degrades with increasing diameter (less proximity), suggesting that conservation of tightly coupled modules is more informative in reconstructing evolutionary histories. The effect of coverage on performance is shown in Figure 2(a) and (b). When more specific modules are used, the effect of coverage on performance is marginal. This indicates that careful selection of a concise set of small, tightly coupled modules may be adequate to reconstruct network phylogenies accurately. Finally, it is interesting to note that the randomized method performs better with large clusters, which is probably due to the increased likelihood that a random group of proteins will contain an informative subset of proteins.

5) Robustness against noise and missing data: Currently available PPI data is likely to be highly noisy and incomplete. Hence, we evaluate the robustness of MOPHY against random noise and incompleteness of data. For this purpose, after generating the networks via simulation of network evolution, we randomly perturb the resulting networks by repeatedly swapping randomly selected interactions. The behavior of the performance of MOPHY with respect to noise rate (percentage of interactions that are swapped) is shown in Figure 2(c). This experiment is performed for $diameter = 3$, $coverage = 60\%$. As seen in the figure, although the accuracy of MOPHY decreases with noise as expected, the performance difference between MOPHY and the randomized method is significant even at the presence of 50% noise. This observation suggests that MOPHY can be used to extract meaningful information on evolutionary histories of networks even when the networks are highly noisy and incomplete. Moreover, note that the performance of the randomized method is not effected by noise, and the performance of MOPHY becomes equivalent to that of the randomized method at the presence 100% noise (*i.e.*, random edge swapping is repeated for a sufficiently large number of iterations). These results indicate that the biological signals captured by MOPHY depend on network topology and the use of network proximity and modularity provide significant information on conservation of function that is beyond sequence similarity.

B. Results on Extant PPI Networks

We test our method on the available PPI networks from seven diverse species. The PPI data is obtained from the Database of Interacting Proteins (DIP) [34]. These networks include those of *D. melanogaster* (7471 proteins, 22656 interactions), *S. cerevisiae* (4968 proteins, 17286 interactions), *E. coli* (1848 proteins, 5930 interactions), *C. elegans* (2646 proteins, 3977 interactions, *H. sapiens* (1334 proteins, 1539 interactions), *H. pylori* (710 proteins, 1359 interactions), and *M. musculus* (414 proteins, 337 interactions). Although the network sizes in this database vary dramatically for different species, MOPHY can effectively deal with such incompleteness by considering modules from each pair of species.

To reconstruct the phylogeny of these seven networks via MOPHY, we use the most specific modules that contain at least three proteins and set the coverage to 50%. As in our experiments on simulated data, we compare MOPHY with three alternate methods; (i) RDL, (ii) using only protein similarities and (iii) using random modules. For reference, we also consider the phylogenetic tree that is reconstructed based on sequenced genomes [35], which is shown in Figure 3(a). The phylogenetic trees reconstructed based on the seven PPI networks by MOPHY, RDL, using only protein similarities and using random modules are shown in Figures 3 (b), (c), (d) and (e) respectively. Unlike other methods, the tree reconstructed by MOPHY complies well with common knowledge on the underlying phylogeny of these seven diverse species and is also consistent with the whole genome based phylogeny. As seen in Figure 3, network-based distance measures tend to overestimate evolutionary distances between extant species. Therefore, methods for normalizing the estimated distances between networks are necessary.

Incidentally, these results also provide evidence supporting the *Coelomata* topology in the *Coelomata vs. Ecdysozoa* debate regarding the evolutionary relationship between nematodes, arthropods, and vertebrates, which has also been supported recently through rigorous analysis of the conservation patterns in intron positions [36]. It is worth to note that, due to limited availability of data, PPI networks differ significantly in size from one species to another. This actually introduces a lot of artificial variation between networks, which might, on a common graph measure, overwhelm desired biological signals. Indeed, as seen in Figure 3(c), RDL is significantly effected by the variability in data availability; it assigns mouse PPI network to the same clade with prokaryotic networks, presumably because the interaction data for this species is quite limited. On the other hand, by focusing on the signals harbored by some more informative modules, we avoid the interference of this global difference among networks.

TABLE II
PERFORMANCE OF MOPHY IN CAPTURING THE TOPOLOGY OF UNDERLYING PHYLOGENY FOR SIMULATED NETWORKS.

Most Specific Modules

	Diameter								
	2			3			4		
Coverage	MOPHY	Random	*p*-value	MOPHY	Random	*p*-value	MOPHY	Random	*p*-value
20%	6.87^{**}	16.40	0.0020	6.84^{**}	15.85	0.0017	6.97^{**}	15.78	0.0019
40%	6.81^{**}	16.14	0.0017	6.86^{**}	15.85	0.0017	7.01^{**}	15.53	0.0029
60%	6.79^{**}	15.85	0.0017	6.86^{**}	15.41	0.0016	7.02^{**}	15.25	0.0026

Most Comprehensive Modules

	Diameter								
	2			3			4		
Coverage	MOPHY	Random	*p*-value	MOPHY	Random	*p*-value	MOPHY	Random	*p*-value
20%	8.89	11.72	0.2283	9.67	11.76	0.3512	10.83	11.82	0.6277
40%	7.62^{*}	13.12	0.0277	8.44	11.61	0.2029	9.63	11.29	0.4922
60%	6.70^{**}	14.93	0.0018	7.92	12.90	0.0529	8.96	11.51	0.3263

(b) Most Comprehensive Modules

Performance of MOPHY in capturing the underlying phylogeny for simulated networks. For each parameter setting, the Nodal distance between the underlying tree and the tree reconstructed by MOPHY/randomized method is shown. Reported values are averages over five runs. p-values indicate the statistical significance of the performance difference between MOPHY and the randomized method. **: $p < 0.01$, *: $p < 0.05$.

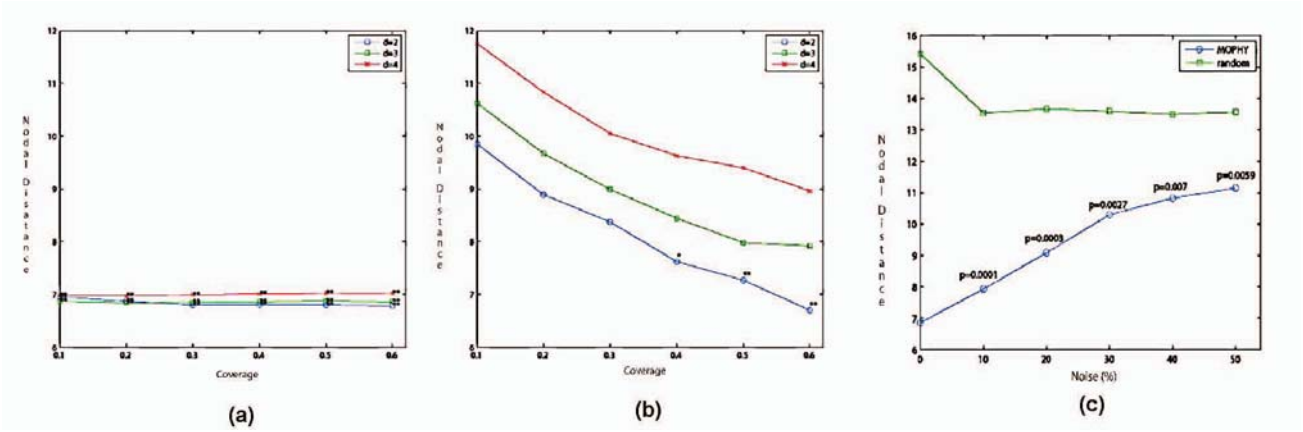

Fig. 2. Effect of Coverage and Noise on Performance

(a), (b): Performance of MOPHY in capturing the underlying evolutionary distances between simulated networks with respect to coverage (fraction of modules that are used in phylogeny reconstruction). (a) Most specific modules, (b) most comprehensive modules. (c): The effect of noise on the performance of MOPHY. Even when the data is perturbed with 50% noise, MOPHY's accuracy in reconstructing the phylogeny is statistically significant.

IV. CONCLUSIONS

In this paper, we propose a phylogenetic framework for analyzing modularity in protein-protein interaction networks. Our approach is motivated by the premise that biomolecular interactions and their modularity are likely to provide direct functional information on the evolution of biological systems. We also develop a method based on the simulation of network evolution to evaluate phylogenetic tree reconstruction methods. Comprehensive experimental results on simulated, as well as real data show that our algorithm is highly successful in reconstructing the underlying phylogenies based on PPI networks, is quite robust to noise, and performs significantly better than existing network-based phylogeny reconstruction algorithms on available protein-protein interaction data. These results demonstrate the promise of modularity-based approaches in comparative network analysis and motivate the study of the evolution of network modularity within a phylogenetic framework.

REFERENCES

[1] Altschul SF, Madden TL, Schffer AA, J Zhang ZZ, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs**. *Nucleic Acids Research* 1997, **25**(17):3389–3402.

[2] Gascuel O: *Mathematics of Evolution and Phylogeny*. Oxford: Oxford University Press 2005.

[3] Nei M, Kumar S: *Molecular Evolution and Phylogenetics*. Oxford: Oxford University Press 2003.

[4] Linder C, Warnow T: **An Overview of Phylogeny Reconstruction**. In *Handbook of Computational Molecular Biology*. Edited by Aluru S, CRC Press 2005.

[5] Kitano H: **Systems biology: a brief overview**. *Science* 2002, **295**(5560):1662–1664.

[6] Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P, Qureshi-Emili A, Li Y, Godwin B, Conover D, Kalbfleisch T, Vijayadamodar G, Yang M, Johnston M, Fields S, Rothberg JM: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae**. *Nature* 2000, **403**(6770):623–627.

[7] Flannick J, Novak A, Srinivasan BS, McAdams HH, Batzoglou S: **Graemlin: general and robust alignment of multiple large interaction networks.** *Genome Research* 2006, **16**(9):1169–1181.

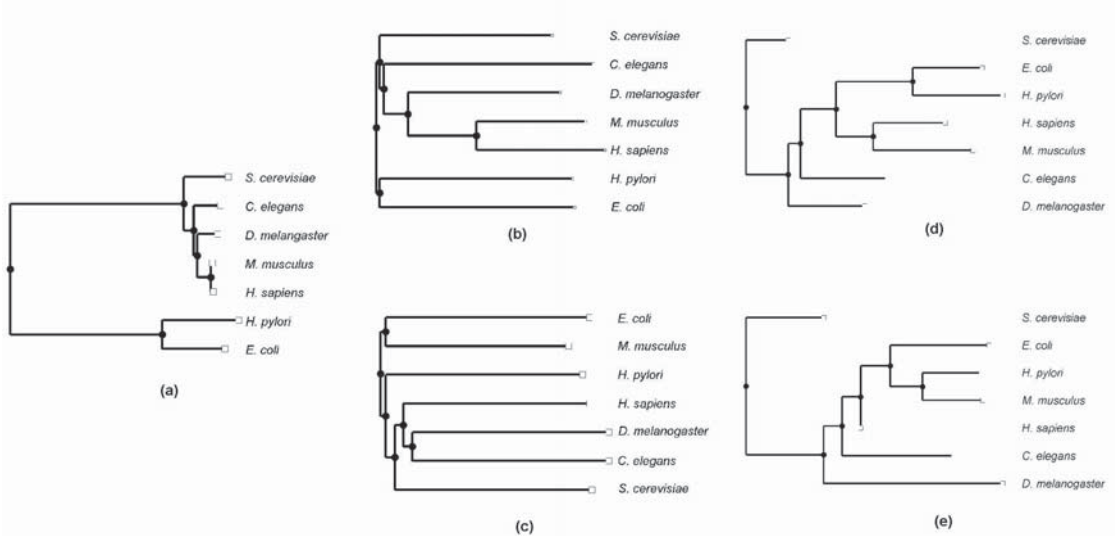

Fig. 3. Comparison of the performances of MOPHY and other methods in reconstructing the phylogenetic tree of seven PPI networks.

(a) Tree based on genome sequences [35], (b) tree reconstructed by MOPHY, (c) tree reconstructed by RDL [15], (d) tree reconstructed by using protein similarities only, (e) tree reconstructed by using random groups of proteins as modules.

[8] Koyutürk M, Kim Y, Subramaniam S, Szpankowski W, Grama A: **Detecting conserved interaction patterns in biological networks**. *Journal of Computational Biology* 2006, **13**(7):1299–1322.

[9] Sharan R, Ideker T: **Modeling cellular machinery through biological network comparison**. *Nature Biotechnology* 2006, **24**(4):427–433.

[10] Wuchty S, Oltvai ZN, Barabási AL: **Evolutionary conservation of motif constituents in the yeast protein interaction network**. *Nature Genetics* 2003, **35**(2):176–179.

[11] Bandyopadhyay S, Sharan R, Ideker T: **Systematic identification of functional orthologs based on protein network comparison**. *Genome Res.* 2006, **16**(3):428–435, [http://dx.doi.org/10.1101\%2Fgr.4526006].

[12] Singh R, Xu J, Berger B: **Global alignment of multiple protein interaction networks**. In *13th Pacific Symposium on Biocomputing (PSB'08)*, *Volume 13* 2008:303–314.

[13] Flannick J, Novak A, Do CB, Srinivasan BS, Batzoglou S: **Automatic Parameter Learning for Multiple Network Alignment**. In *RECOMB* 2008.

[14] Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, Sittler T, Karp RM, Ideker T: **Conserved patterns of protein interaction in multiple species.** *Proc Natl Acad Sci U S A* 2005, **102**(6):1974–1979.

[15] Chor B, Tuller T: **Biological Networks: Comparison, Conservation, and Evolutionary Trees**. In *RECOMB* 2006:30–44.

[16] Vespignani A: **Evolution thinks modular**. *Nature Genetics* 2003, **35**(2):118–119.

[17] Yamada T, Goto S, Kanehisa M: **Extraction of phylogenetic network modules from prokayrote metabolic pathways.** *Genome Inform Ser Workshop Genome Inform* 2004, **15**:249–258, [http://view.ncbi.nlm.nih.gov/pubmed/15712127].

[18] Titz B, Schlesner M, Uetz P: **What do we learn from high-throughput protein interaction data?** *Expert Review of Proteomics* 2004, **1**:111–121.

[19] Middendorf M, Ziv E, Wiggins CH: **From The Cover: Inferring network mechanisms: The Drosophila melanogaster protein interaction network**. *Proceedings of the National Academy of Sciences* 2005, **102**(9):3192–3197, [http://www.pnas.org/cgi/content/abstract/102/9/3192].

[20] Pastor-Satorras R, Smith E, Sole RV: **Evolving protein interaction networks through gene duplication.** *J Theor Biol* 2003, **222**(2):199–210.

[21] Lee I, Date SV, Adai AT, Marcotte EM: **A probabilistic functional network of yeast genes**. *Science* 2004, **306**(5701):1555–1558.

[22] Spirin V, Mirny LA: **Protein complexes and functional modules in molecular networks**. *PNAS* 2003, **100**(21):12123–12128.

[23] Bader GD, Hogue CW: **An automated method for finding molecular complexes in large protein interaction networks**. *BMC Bioinformatics* 2003, **4**(2).

[24] Bebek G, Yang J: **PathFinder: mining signal transduction pathway segments from protein-protein interaction networks**. *BMC Bioinformatics* 2007, **8**:335+.

[25] Pandey J, Koyutürk M, Subramaniam S, Grama A: **Functional coherence in domain interaction networks**. *Bioinformatics Suppl.on ECCB'08* 2008, **24**(16):i28–i34.

[26] Sharan R, Ulitsky I, Shamir R: **Network-based prediction of protein function**. *Mol Syst Biol* 2007, **3**.

[27] Kelley BP, Sharan R, Karp RM, Sittler T, Root DE, Stockwell BR, Ideker T: **Conserved pathways within bacteria and yeast as revealed by global protein network alignment**. *PNAS* 2003, **100**(20):11394–11399.

[28] Goldberg D, Roth F: **Assessing experimentally derived interactions in a small world**. *Proceedings of the National Academy of Sciences* 2003, **100**(8):4372–4376.

[29] Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4**(4):406–425.

[30] Vazquez A, Flammini A, Maritan A, Vespignani A: **Modeling of protein interaction networks**. *Complexus* 2003, **1**:38.

[31] Bebek G, Berenbrink P, Cooper C, Friedetzky T, Nadeau J, Sahinalp S: **Improved Duplication Models for Proteome Network Evolution**. In *Systems Biology and Regulatory Genomics*, *Volume 4023/2006 of* Lecture Notes in Computer Science 2006:119–137.

[32] Robinson DF, Foulds LR: **Comparison of phylogenetic trees**. *Mathematical Biosciences* 1981, **53**(1-2):131–147, [http://dx.doi.org/10.1016\%2F0025-5564\%2881\%2990043-2].

[33] Bluis J, Shin DG, Shin DG: **Nodal distance algorithm: calculating a phylogenetic tree comparison metric**. *Bioinformatics and Bioengineering, 2003. Proceedings. Third IEEE Symposium on* 2003, :87–94.

[34] Xenarios I, Fernandez E, Salwinski L, Duan XJ, Thompson MJ, Marcotte EM, Eisenberg D: **DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions.** *Nucleic Acids Res* 2002, **30**:303–305, [http://dip.doe-mbi.ucla.edu].

[35] Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward Automatic Reconstruction of a Highly Resolved Tree of Life**. *Science* 2006, **311**(5765):1283–1287, [http://www.sciencemag.org/cgi/content/abstract/311/5765/1283].

[36] Zheng J, Rogozin IB, Koonin EV, Przytycka TM: **Support for the Coelomata Clade of Animals from a Rigorous Analysis of the Pattern of Intron Conservation**. *Molecular Biology and Evolution* 2007, **24**(11):2583–2592.