

Phylogenetic Analysis of Molecular Interaction Networks¹

Mehmet Koyutürk

Case Western Reserve University
Electrical Engineering & Computer Science

¹Joint work with Sinan Erten, Xin Li, Gurkan Bebek, and Jing Li

Outline

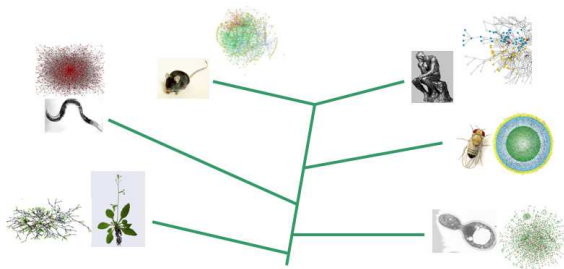
- 1 Background & Motivation
 - Why Network Phylogenetics?
- 2 Approach
 - Modularity Based Phylogenetic Analysis
 - Identifying Modular Network Components
 - Constructing Module Maps
 - Reconstructing Network Phylogenies
- 3 Results
 - Performance on Simulated Networks
 - Performance on Real Data
- 4 Conclusion
 - Acknowledgments

Outline

- 1 Background & Motivation
 - Why Network Phylogenetics?
- 2 Approach
 - Modularity Based Phylogenetic Analysis
 - Identifying Modular Network Components
 - Constructing Module Maps
 - Reconstructing Network Phylogenies
- 3 Results
 - Performance on Simulated Networks
 - Performance on Real Data
- 4 Conclusion
 - Acknowledgments

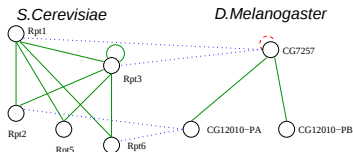
Large Scale Data on Diverse Biological Systems

- Protein-protein interactions (PPI): Which proteins bind to each other (possibly, to perform specific tasks together)?
 - Interacting proteins can be identified via high-throughput screening (e.g., Y2H, TAP)
- What can we learn from PPI data that belong to diverse species?



Comparative Network Analysis

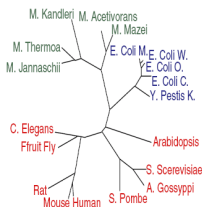
- Network Alignment: Identify "common" subgraphs in interaction networks of different species
 - These are likely to correspond to *modular* components of the cellular machinery
- Algorithms mostly focus on one-to-one matching of interactions
 - PathBlast (Kelley *et al.*, *PNAS*, 2003), NetworkBlast (Sharan *et al.*, *RECOMB*, 2004), MAWISH (Koyutürk *et al.*, *RECOMB*, 2005), MULE (Koyutürk *et al.*, *JCB*, 2006), Graemlin (Flannick *et al.*, *Genome Research*, 2007)



A conserved subgraph identified by MAWISH on yeast and fly networks: Proteasome regulatory particle

Network Evolution & Phylogenetics

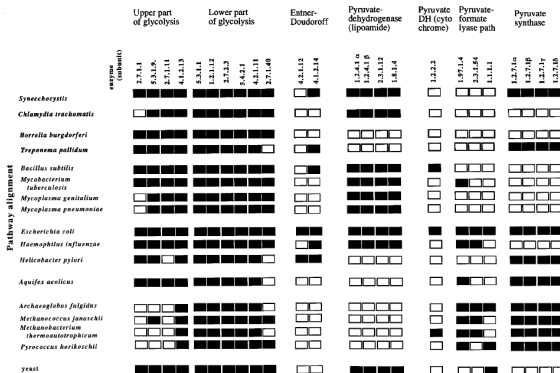
- Can we learn more about these networks by paying more attention to their evolutionary histories?
 - Incorporating evolutionary information enhances the performance of alignment algorithms (Hirsh & Sharan, 2007; Flannick *et al.*, *RECOMB*, 2008)
 - Network similarity based algorithms are quite successful in reconstructing evolutionary trees based on metabolic networks (Chor & Tuller, *RECOMB*, 2006)



Phylogenetic tree reconstructed by
Chor & Tuller's RDL-based algorithm

Modular Evolution

- What are the evolutionary mechanisms that create, preserve, and diversify modularity in biological systems?



Pathway alignment for glycolysis, Entner-Doudoroff pathway, and pyruvate processing

(Dandekar *et al.*, 1999)

Outline

1 Background & Motivation

- Why Network Phylogenetics?

2 Approach

- Modularity Based Phylogenetic Analysis
- Identifying Modular Network Components
- Constructing Module Maps
- Reconstructing Network Phylogenies

3 Results

- Performance on Simulated Networks
- Performance on Real Data

4 Conclusion

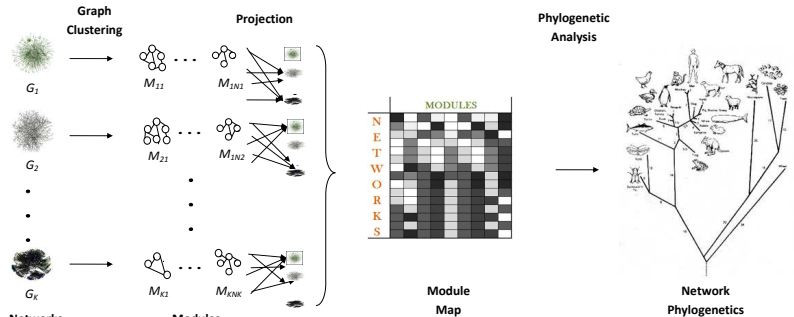
- Acknowledgments

Our Approach

- Instead of comparing interactions, compare emergent properties
 - We use conservation of modular network components as an indicator of evolutionary proximity
- Phylogeny reconstruction provides a testbed for hypotheses
 - Does network information provide information beyond what is gleaned from genome?
 - Does modularity say anything more about the evolutionary histories of these networks?
- Not another tool for phylogeny reconstruction!

MoPHY

- A framework for modularity based phylogenetic analysis
 - Reconstruction of whole-network and module phylogenies
 - Identification of module families

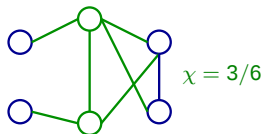


What is Modularity?

- *Functional module*: A group of molecules that perform a distinct function together (Hartwell, *Nature*, 1999)
 - Functional modules manifest themselves as highly connected subgraphs in the network (Rives & Galitski, *PNAS*, 2003)
- Many graph clustering algorithms have been developed to identify functional modules from interaction data
 - MCODE (Bader *et al.*, *BMC Bioinformatics*, 2003), MCL (Pereira-Leal *et al.*, *Proteins*, 2004), SIDES (Koyutürk *et al.*, *RECOMB*, 2006)
- Functional coherence is often used to verify modularity
 - VAMPIRE (Hsiao *et al.*, *NAR*, 2005), Ontologizer (Grossman *et al.*, *RECOMB*, 2006)

Network Connectivity

- Commonly used as an indicator of modularity
 - *Mutual Clustering Coefficient*: Fraction (significance) of shared interactions between two proteins (Goldberg & Roth, *PNAS*, 2003)
 - Significance can be computed using hypergeometric models

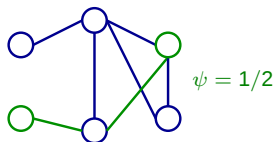


Mutual Clustering Coefficient

$$\chi(P_i, P_j) = \frac{|\{P_k \in V : P_i P_k \in E \vee P_j P_k \in E\}|}{|\{P_k \in V : P_i P_k \in E \wedge P_j P_k \in E\}|}$$

Network Proximity

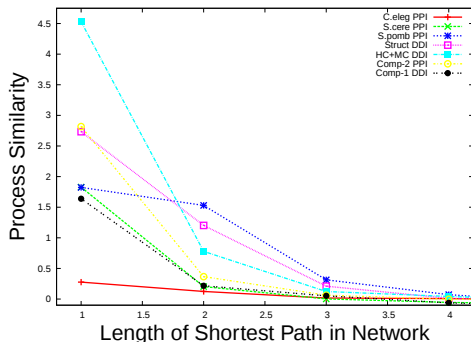
- A more flexible indicator of modularity
 - Data is incomplete: Indirect paths can account for missing interactions
 - Projection of modularity: Indirect paths may relax evolutionary pressure on preserving direct interactions



Proximity

$$\psi(P_i, P_j) = 1/\delta(P_i, P_j)$$

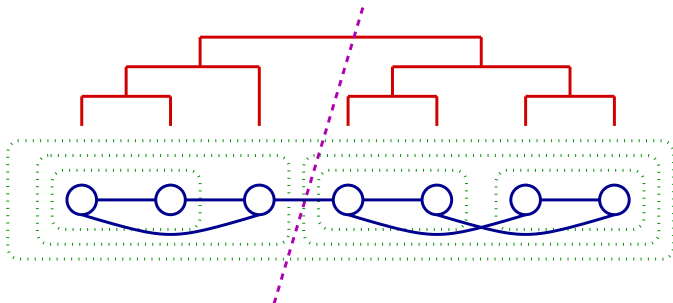
Network Proximity and Modularity



- Average process similarity with respect to network distance in a variety of PPI and DDI networks
 - Process similarity: Information content of the biological processes that describes all processes associated with the two proteins (Pandey *et al.*, *ECCB*, 2008)

Module Identification Algorithm

- Bottom-up complete linkage hierarchical clustering of proteins in the network
 - Primary similarity criterion: Proximity
 - Secondary similarity criterion: Mutual clustering coefficient
- Discrete modules are obtained by putting a threshold on proximity



Projection of Proximity

- Sequence similarity relates proteins in different networks
 - $\sigma(P_i, P_j) = 1 + 1/\log E_{ij}$, where E_{ij} denotes the BLAST E -value of the best bidirectional hit between proteins P_i and P_j
- The proximity between proteins $P_i, P_j \in V(G)$ with respect to network $G' \neq G$ is the aggregate proximity of their orthologs in G'

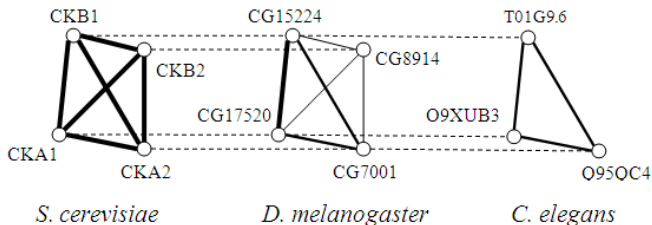
- $$\psi_{G'}(P_i, P_j) = \sum_{P_k, P_l \in V(G')} \hat{\sigma}_{G'}(P_i, P_k) \hat{\sigma}_{G'}(P_j, P_l) \psi(P_k, P_l)$$

- $$\text{Here, } \hat{\sigma}_{G'}(P_i, P_k) = \frac{\sigma(P_i, P_k)}{\sum_{P_l \in V(G')} \sigma(P_i, P_l)}$$

Projection of Modularity

- The modularity of a set $S \subseteq V(G)$ of proteins with respect to network G' is defined as the average proximity of all pairs of proteins in S with respect to G'

$$\mu_{G'}(S) = \frac{2 \sum_{P_i, P_j \in S} \psi_{G'}(P_i, P_j)}{|S|(|S| - 1)}$$



Module Maps

- Once modules are projected on all networks, each network is associated with a module map
 - A module map is a vector with each entry signifying the modularity of the corresponding module in the corresponding network
 - Modules identified in G_j : $\{S_1^j, S_2^j, \dots, S_{m_j}^j\}$
 - For all networks G_k , the module map of G_k with respect to G_j is given by

$$f_{ij} = [\mu_{G_i}(S_1^j) \mu_{G_i}(S_2^j) \dots \mu_{G_i}(S_{m_j}^j)]$$

- If there are K networks, then the complete module map of G_k is given by

$$f_i = [f_{i1} f_{i2} \dots f_{iK}]$$

Phylogeny Reconstruction Using Module Maps

- Estimate evolutionary distances based on module maps
 - $d(G_i, G_j) = 1 - \frac{f_i \cdot f_j}{||f_i|| ||f_j||}$
- Once evolutionary distances are available, reconstruct a phylogenetic tree using traditional algorithms
 - Neighbor Joining (Saitou & Nei, *Mol. Biol. Evol.*, 1987)
- Handling noise and missing data
 - Estimate the evolutionary distances between two networks based on only their module maps with respect to each other
 - $\hat{d}(G_i, G_j) = \min \left\{ 1 - \frac{f_{ii} \cdot f_{jj}}{||f_{ii}|| ||f_{jj}||}, 1 - \frac{f_{jj} \cdot f_{ij}}{||f_{jj}|| ||f_{ij}||} \right\}$

Outline

1 Background & Motivation

- Why Network Phylogenetics?

2 Approach

- Modularity Based Phylogenetic Analysis
- Identifying Modular Network Components
- Constructing Module Maps
- Reconstructing Network Phylogenies

3 Results

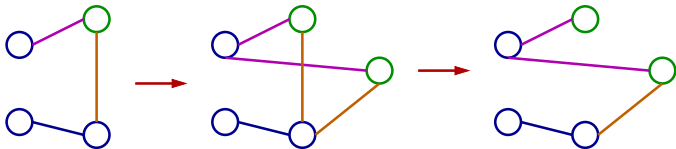
- Performance on Simulated Networks
- Performance on Real Data

4 Conclusion

- Acknowledgments

Simulation of Network Evolution

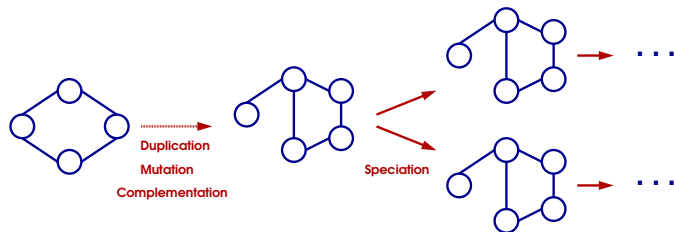
- Generate multiple networks based on theoretical models of network evolution
 - Duplication/Mutation/Complementation (DMC) model (Middendorf *et al.*, *PNAS*, 2005)



- Generalized models reconstruct the properties (degree distribution, clustering coefficient distribution) of extant networks (Bebek *et al.*, *Theo. Comp. Sci.*, 2006)

Evolving Multiple Networks

- Generate multiple networks based on the DMC model
 - Speciation: A network is duplicated, each copy evolves independently

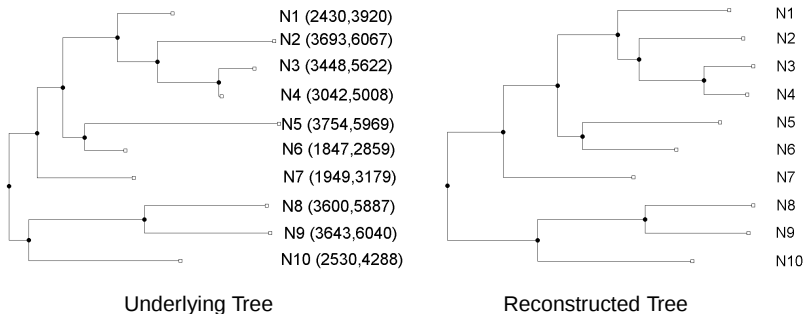


- Using the proposed framework, construct a phylogenetic tree based on the generated networks
 - Performance evaluation, comparison of different algorithms, adjustment of parameters

Performance Measures

- Comparison of the topologies of underlying and reconstructed phylogenetic trees
 - Symmetric Difference: Number of partitions that are on one tree and not the other (Robinson & Foulds, *Math. Biosci.*, 1981)
- Comparison of the actual and estimated evolutionary distances
 - Nodal Distance: Sum of distances of all node pairs in each tree (Bluis *et al.*, *IEEE BIBE*, 2003)
- Statistical significance
 - Random Method: Use random groups of proteins of the same size instead of identified modules
 - Compute p -values via t -test on repeated runs

Reconstructing Known Phylogeny



- MoPHY reconstructs the topology of the underlying phylogeny perfectly
- Nodal distance: 5.12 ($p < 0.002$)

Effect of Parameters

Most Specific Modules

	Diameter								
	2			3			4		
Coverage	MoPHY	Random	$p <$	MoPHY	Random	$p <$	MoPHY	Random	$p <$
20%	1.6**	11.2	0.004	1.6**	11.2	0.004	1.6**	12.0	0.003
40%	1.6**	12.0	0.001	1.6**	10.8	0.009	1.6**	12.0	0.001
60%	1.6**	11.2	0.002	1.6**	11.6	0.004	1.6**	11.6	0.002

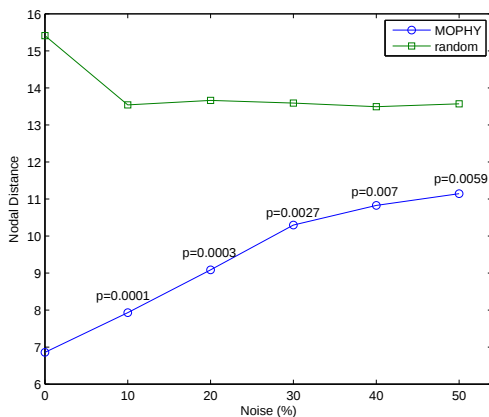
Most Comprehensive Modules

	Diameter								
	2			3			4		
Coverage	MoPHY	Random	$p <$	MoPHY	Random	$p <$	MoPHY	Random	$p <$
20%	2.4**	11.6	0.005	2.8**	10.8	0.009	4.4*	10.8	0.012
40%	2.8**	12.0	0.004	2.8**	10.8	0.003	4.4*	10.4	0.018
60%	1.6**	10.8	0.006	2.4**	11.2	0.003	3.6**	10.0	0.005

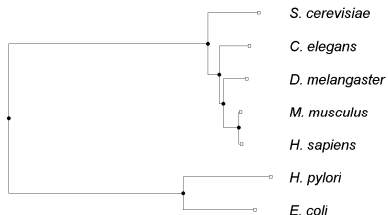
- Performance difference between MoPHY and random method is consistently significant
 - Conservation of proximity between modular groups of proteins captures evolutionary histories better than that of arbitrary proteins

Effect of Noise

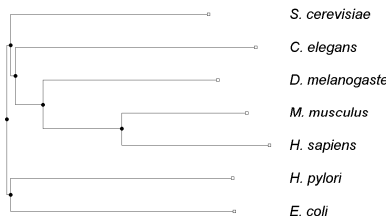
- Once networks are generated, we add noise to each network by randomly swapping interactions



Comparison to Existing Algorithms



Sequence-based Phylogenetic Tree



MoPHY



RDL (Chor & Tuller, 2006)

Outline

1 Background & Motivation

- Why Network Phylogenetics?

2 Approach

- Modularity Based Phylogenetic Analysis
- Identifying Modular Network Components
- Constructing Module Maps
- Reconstructing Network Phylogenies

3 Results

- Performance on Simulated Networks
- Performance on Real Data

4 Conclusion

- Acknowledgments

Remarks & Ongoing Work

- Success in phylogenetic tree reconstruction can be seen as proof of concept
- We have worked on "rows" of the module map so far
 - What do we learn if we work on columns?
- How can we trace evolutionary histories of modules?
 - Find modules that map to each other, co-evolve, or complement each other
 - Evolutionary trajectories can be used to identify hierarchical module families and construct module ontologies

Thanks...

- CWRU
 - Sinan Erten, Xin Li, Gurkan Bebek, Jing Li
- Purdue
 - Ananth Grama
- UC-San Diego
 - Shankar Subramaniam