Building Cellular Interaction Databases: Analysis, Synthesis, and Interfaces

Ananth Grama, Computer Sciences, Purdue University

In collaboration with

Mehmet Koyutürk (Purdue University),
Yohan Kim and Shankar Subramaniam (UC-San Diego)

Work supported by
National Institutes of Health and
National Science Foundation

Outline

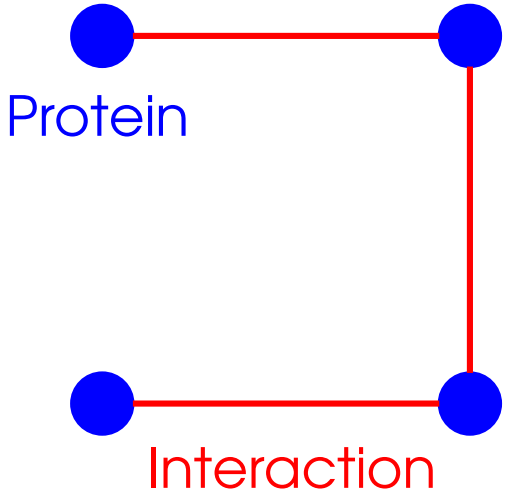
1. Biological Networks
   - Definitions, problems, applications
2. Current Work
   - Analyzing biological networks for conserved molecular interaction patterns
   - Alignment of protein interaction networks based on evolutionary models
   - Module identification based on phylogenetic profiles
3. Ongoing and Future Work
   - Projecting modules extracted from available networks to other genomes
   - Phylogenetic analysis at modular level
   - Constructing reference module maps
   - Building a fully functional interoperable signaling database

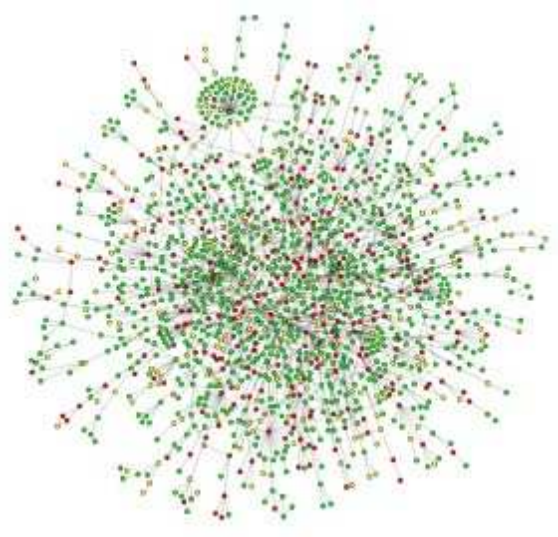
Biological Networks

- Interactions between biomolecules that drive cellular processes
  - Genes, proteins, enzymes, chemical compounds
  - Chemical transformation & energy generation, information transfer
  - Coarser level than sequences in life's complexity pyramid
- Experimental/inferred data in various forms
  - Protein-protein interaction (PPI) networks
  - Gene regulatory networks
  - Metabolic & signaling pathways
- What do we gain from analysis of cellular networks?
  - Modular analysis of cellular processes
  - Understanding evolutionary relationships at a higher level
  - Assigning functions to proteins through interaction information
  - Intelligent drug design: block protein, preserve pathway

Protein-Protein Interaction (PPI) Networks

- Interacting proteins can be discovered experimentally
  - Two-hybrid
  - Mass spectrometry
  - Tandem affinity purification (TAP)



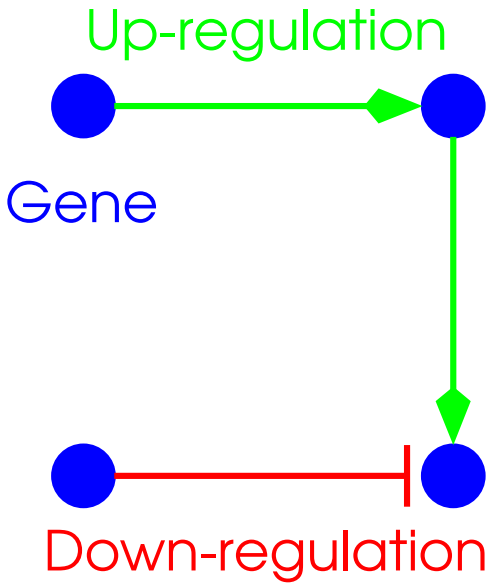


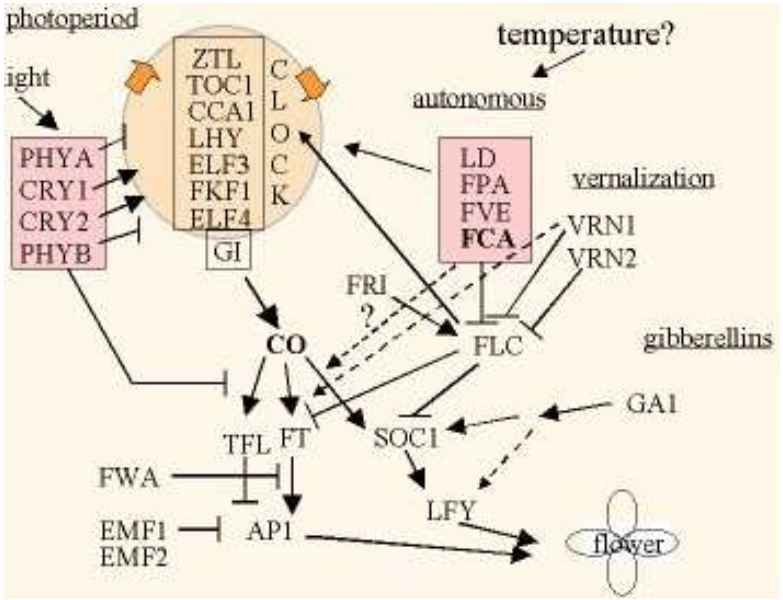
S. Cerevisiae protein interaction network

Source: Jeong et al. Nature 411: 41-42, 2001.

Gene Regulatory Networks

- Genes regulate each others' expression
  - A simple model: Boolean networks
  - Can be derived from gene expression data



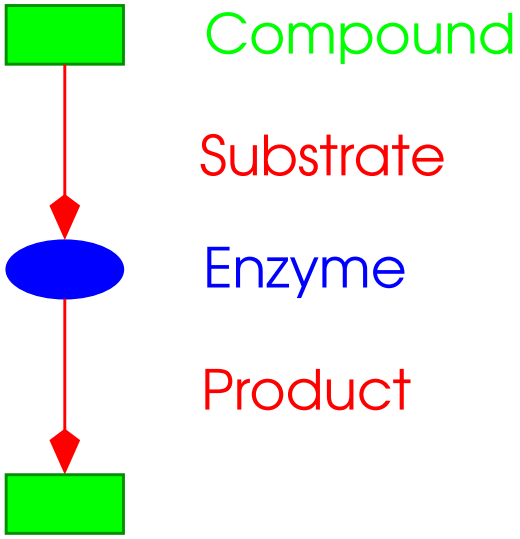


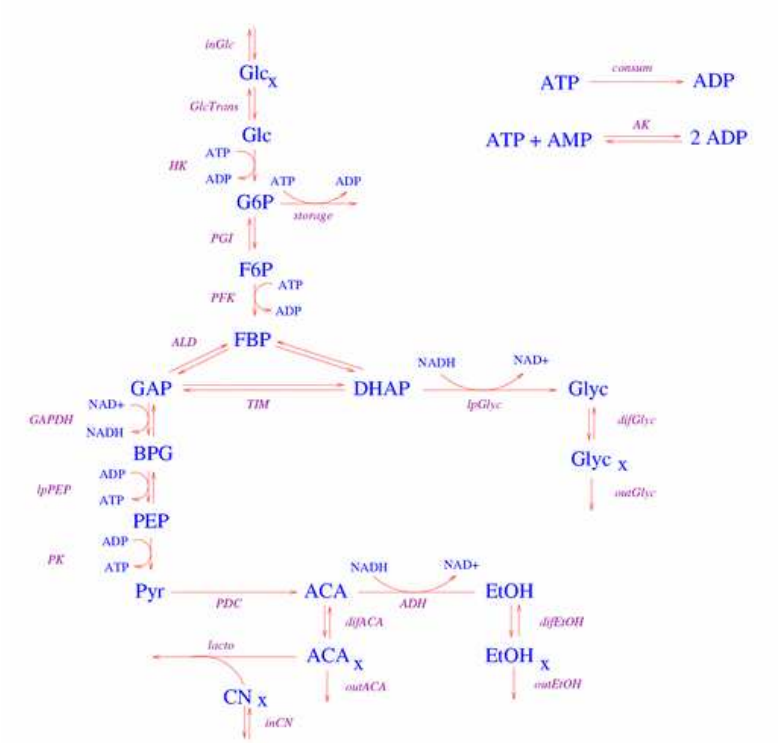
Genetic network that controls flowering time in *A. Thaliania*

Source: Blazquez et al. EMBO Reports 2: 1078-1082, 2001

Metabolic Pathways

- Chains of reactions that perform a particular metabolic function
  - Reactions are linked to each other through substrate-product relationships
  - Directed hypergraph/ graph models





Glycolysis pathway in *S. Cerevisiae*

Source: Hynne et al. Biophysical Chemistry, 94, 121-163, 2001.

Analysis of Biological Networks

- Evolution thinks in a modular fashion
  - Selective pressure on preserving interactions
  - Functional modules, protein complexes are highly conserved
- Computational methods for discovery and analysis of modules and complexes
  - Graph clustering: Functionally related entities are densely connected
  - Graph analysis: Common topological motifs, conserved interaction patterns reveal modularity
  - Graph alignment: Conservation/divergence of modules and pathways
  - Module maps: Canonical pathways across species
  - Phylogenetic analysis: Genes/proteins that belong to a common module are likely to have co-evolved

How do we detect conserved subgraphs?

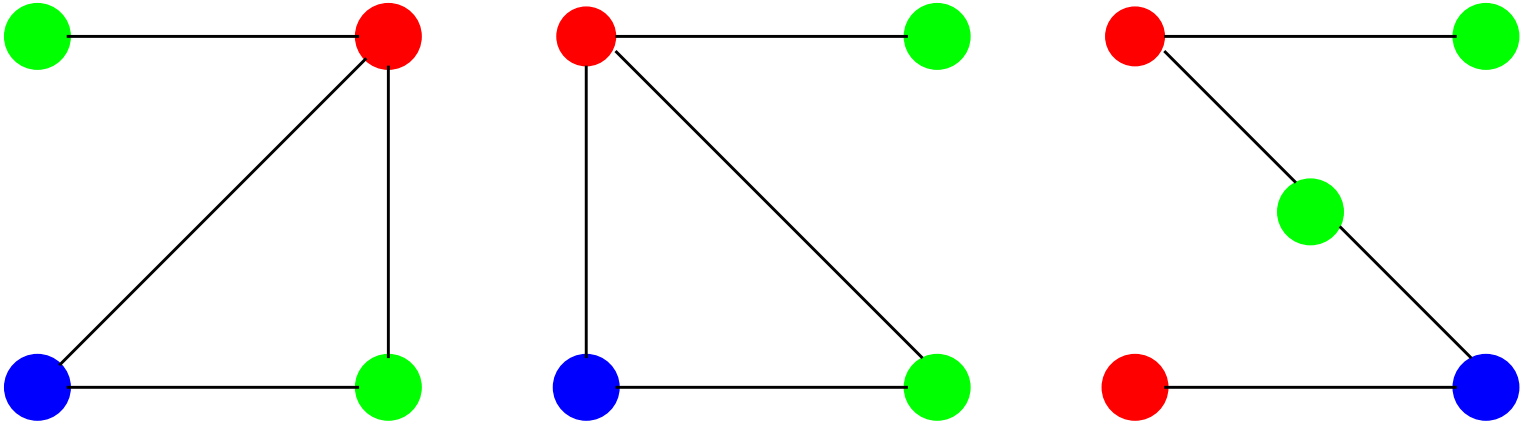
(Koyutürk, Grama, Szpankowski, ISMB04, Bioinformatics04)

- Given a collection of biological networks that belong to several organisms, discover sets of related interactions that frequently occur together
  - Protein interaction networks: Common interactions between orthologous proteins, possibly a conserved functional module
  - Metabolic pathways: Sub-pathways common to a group of organisms, may reveal functional conservation/divergence

- Earlier work focused on identifying common topological motifs
  - We can discover orthologous subgraphs by taking into account the identity of molecules
  - Contract orthologous proteins to relate networks between species
  - Contracting orthologs simplifies the computational problem as well

Graph Analysis: An Example

While mining PPI networks, orthologous proteins
are modeled by identically colored nodes



Graph database

Subgraphs with frequency 3

Extending Frequent Itemset Mining to Graph Analysis

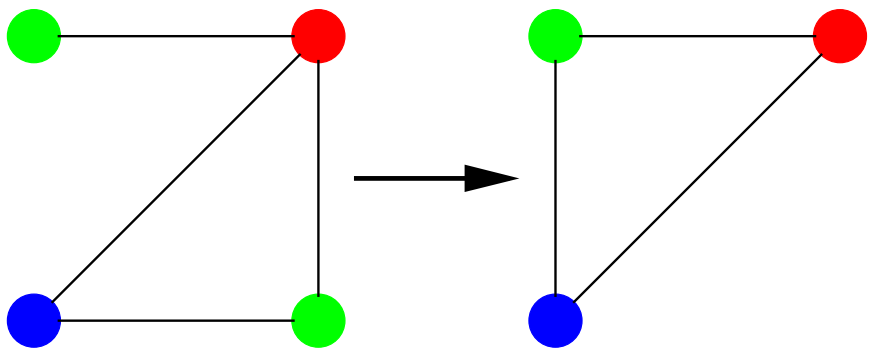
- Given a set of transactions, find sets of items that are frequent in these transactions
- Extensively studied in data mining literature
- Algorithms exploit downward closure property
  - A set is frequent only if all of its subsets are frequent
  - Generate itemsets from small to large, pruning supersets of infrequent sets
- Can be generalized to mining graphs
  - transaction $\longrightarrow$ graph
  - item $\longrightarrow$ node, edge
  - itemset $\longrightarrow$ subgraph
- However, the graph analysis problem is considerably more difficult!

Analyzing Graphs: Challenges

- Subgraph Isomorphism
  - For counting frequencies, it is necessary to check whether a given graph is a subgraph of another one
  - NP-complete
- Canonical labeling
  - To avoid redundancy while generating subgraphs, canonical labeling of graphs is necessary
  - Equivalent to subgraph isomorphism
- Connectivity
  - Patterns of interest are generally connected, so it is necessary to generate only connected subgraphs
- Existing algorithms mainly focus on minimizing redundancy and mining & extending simple substructures
  - AGM, FSG, gSpan, SPIM, CLOSEGRAPH

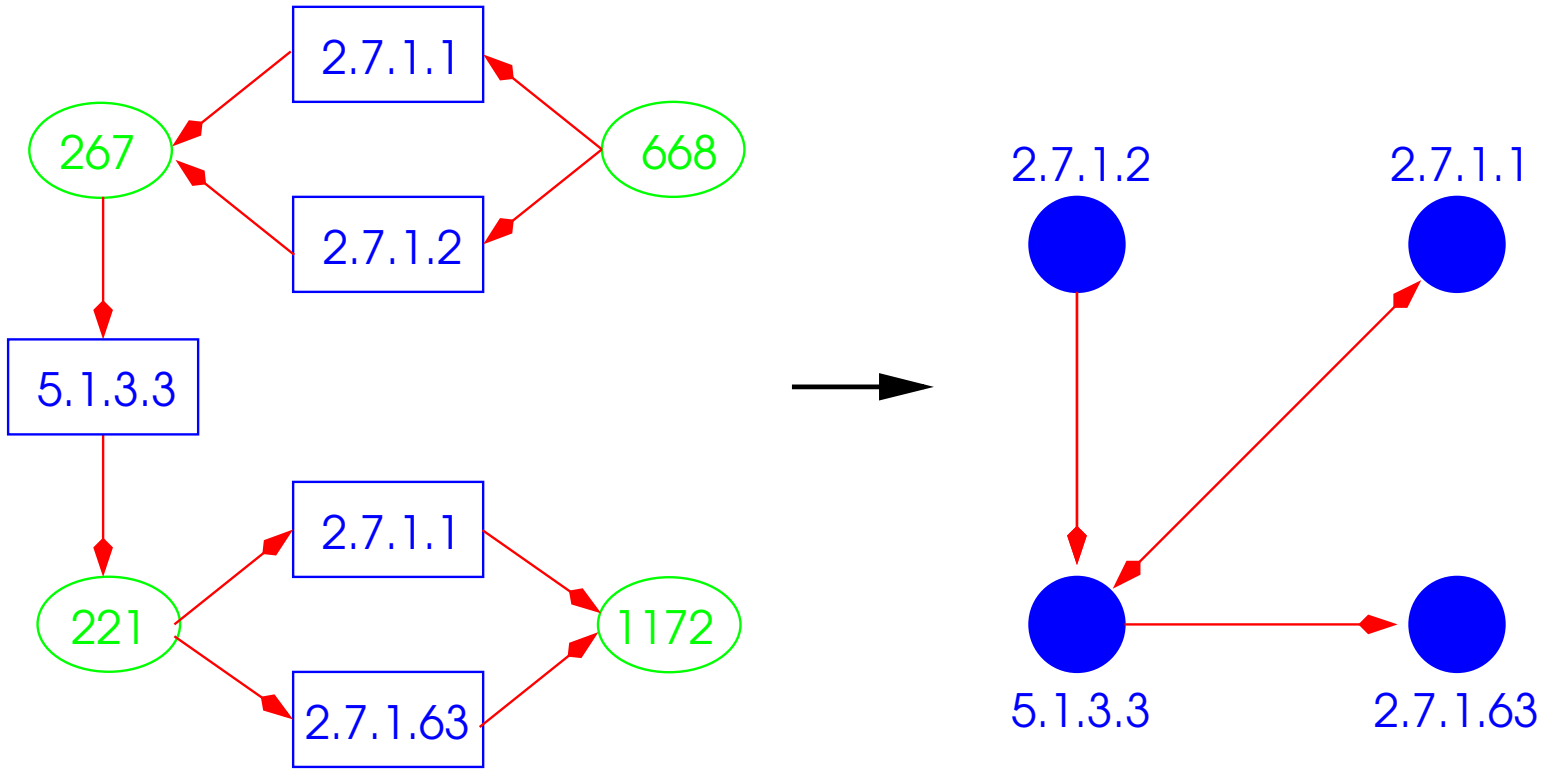
Contracting Orthologous Nodes

- Contract orthologous nodes (proteins, enzymes) into a single node
- No subgraph isomorphism
  - Graphs are uniquely identified by their edge sets
- Frequent subgraphs are preserved $\Rightarrow$ No information loss
  - Subgraphs that are frequent in general graphs are also frequent in their ortholog-contracted representation
- Discovered frequent subgraphs are still biologically interpretable!
  - Interaction between proteins becomes interaction between protein families



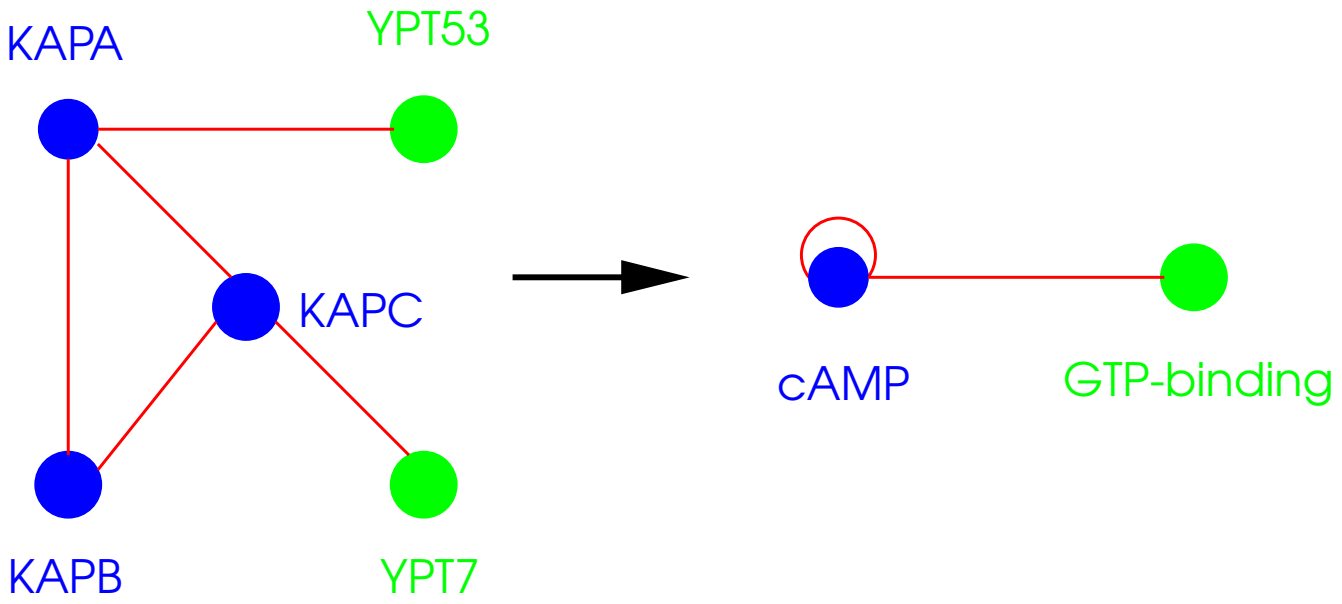
Node Contraction in Metabolic Pathways

- Enzyme-contracted directed graph model
  - Nodes represent enzymes
  - Global labeling by enzyme nomenclature (EC numbers)
  - A directed edge from one enzyme to the other implies that the second consumes a product of the first



Node Contraction in Protein Interaction Networks

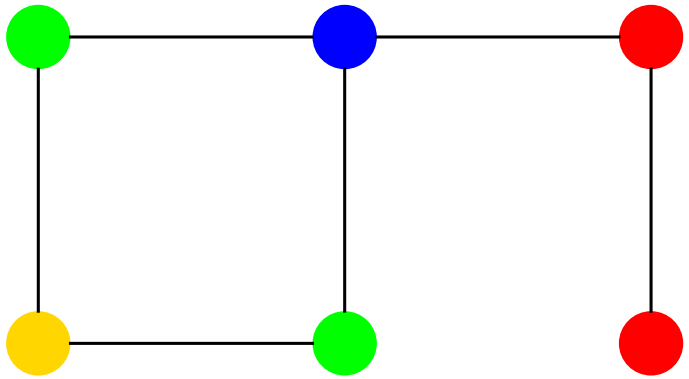
- Relating proteins in different organisms
  - Clustering: Orthologous proteins show sequence similarities
  - Phlyogenetic analysis: Allows multi-resolution analysis among distant species
  - Literature, ortholog databases
- Contraction
  - Interaction between proteins $\longrightarrow$ interaction between protein families
  - Must avoid distant paralogs



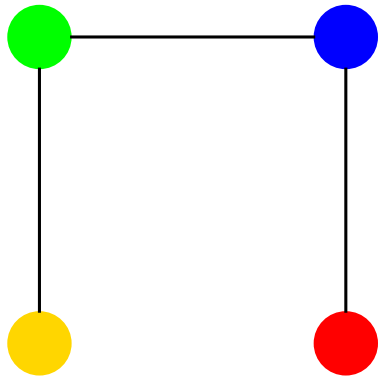
Preservation of Subgraphs

Theorem: Let $\tilde{G}$ be the ortholog-contracted graph obtained by contracting the orthologous nodes of graph G. Then, if S is a subgraph of G, $\tilde{S}$ is a subgraph of $\tilde{G}$.

Corollary: The ortholog-contracted representation of any frequent subgraph is frequent in the set of ortholog-contracted graphs.



G



$\tilde{G}$

Simplifying the Graph Analysis Problem

Labeling: Assign a unique label to each ortholog group. Observation: Since each label is unique in an ortholog-contracted graph, it is uniquely determined by the set of its edges.

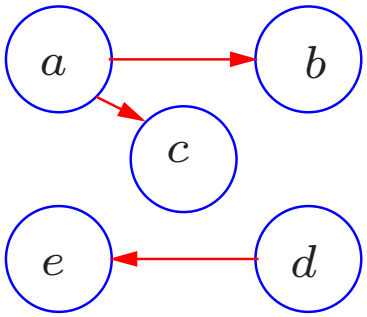
Maximal Frequent Subgraph Mining Problem

Given a set of labeled graphs $\{G_1, G_2, ..., G_m\}$, find all connected graphs S such that S is a subgraph of at least σm of the graphs (is frequent) and no supergraph of S is frequent (is maximal).

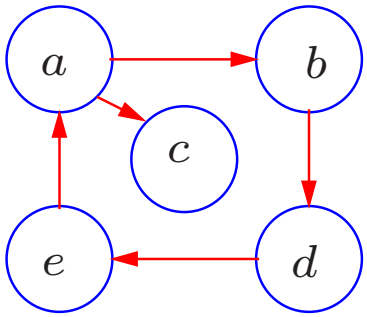
Maximal Frequent Edgeset Mining Problem

Given a set of edge (interaction) sets $\{E_1, E_2, ..., E_m\}$, find all connected edge sets F such that F is a subset of at least σm of the edge sets (is frequent) and no superset of F is frequent (is maximal).

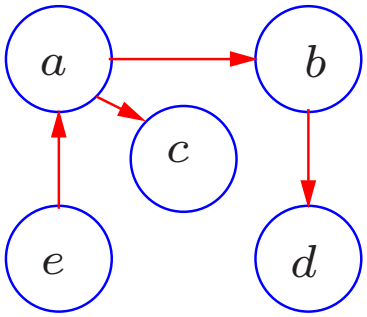
From Graphs to Edgesets



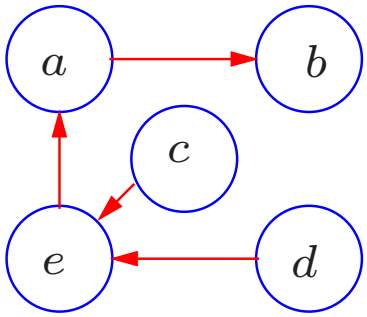
G_1



G_2



G_3



G_4

$F_1 = \{ab, ac, de\}$

$F_2 = \{ab, ac, bc, de, ea\}$

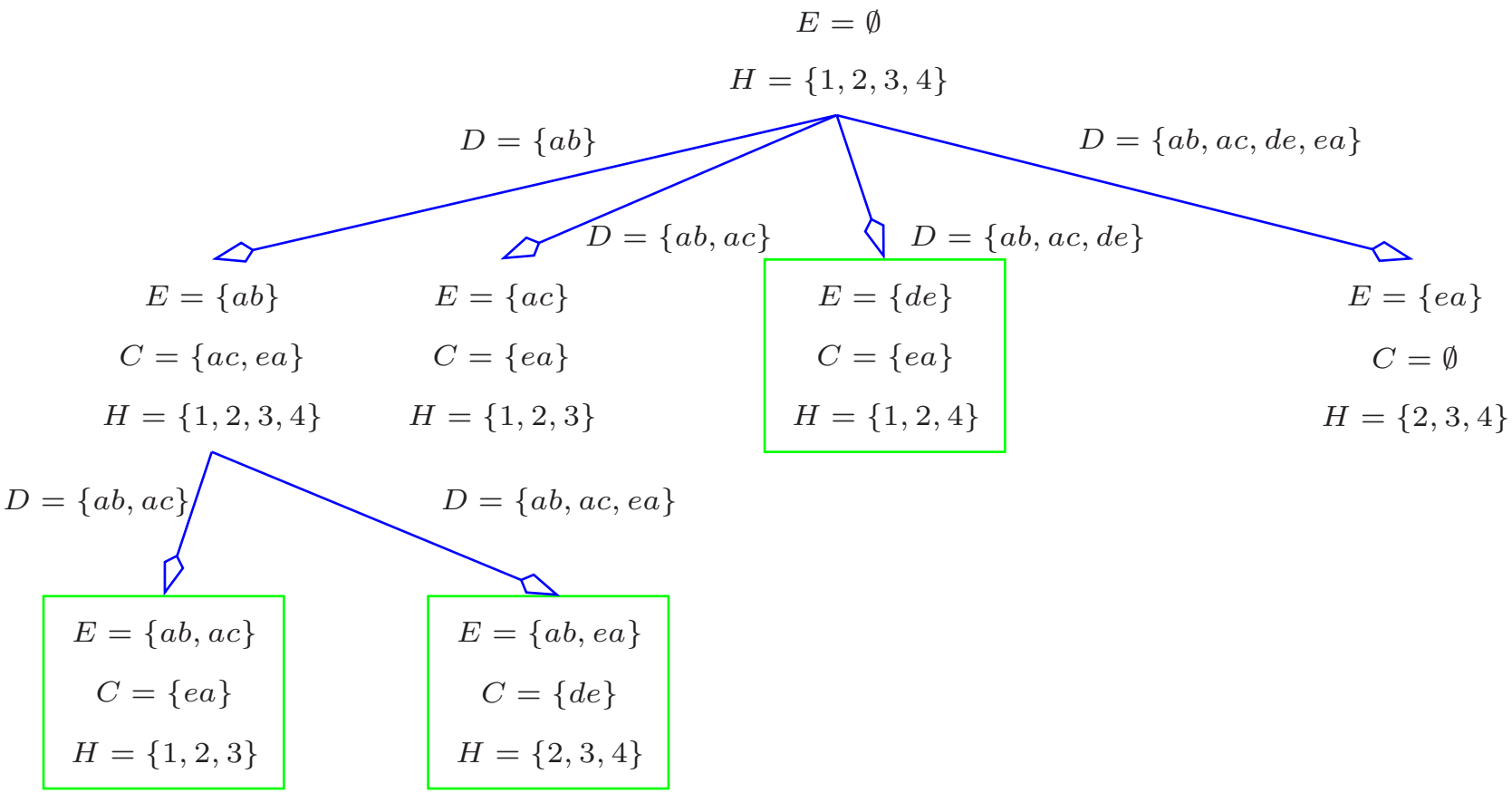
$F_3 = \{ab, ac, bc, ea\}$

$F_4 = \{ab, ce, de, ea\}$

We can construct the graphs on the left
if we know the sets on the right

MULE: Mining Uniquely Labeled Graphs

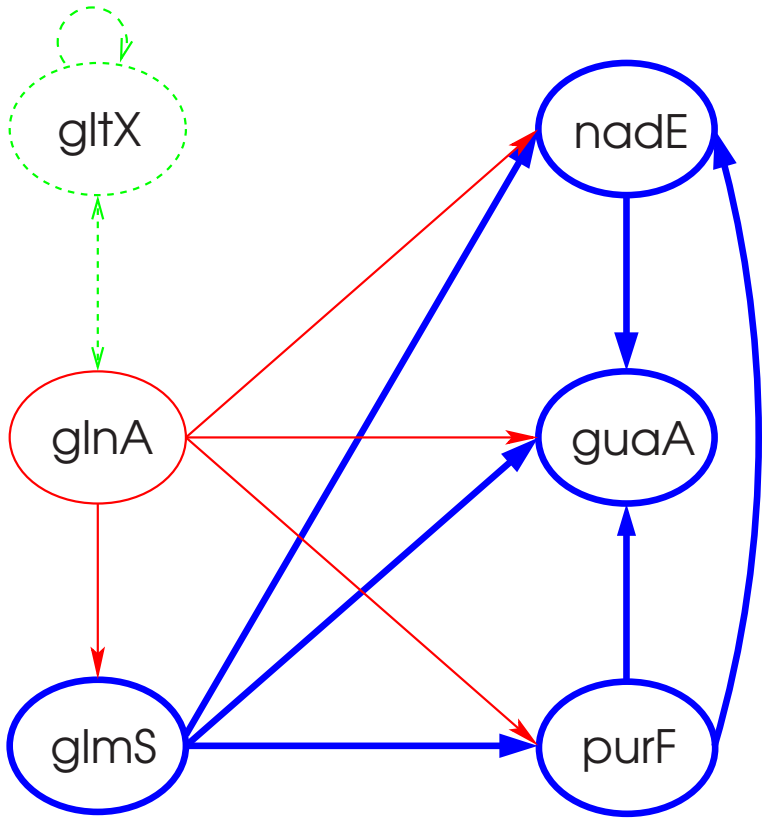
Depth-first enumeration of frequent subgraphs
using downward closure property



E: frequent subgraph, H: graphs that contain E,
D: already explored edges, C: edges to be added to E

Frequent Sub-Pathways in KEGG

Glutamate metabolism (155 organisms)



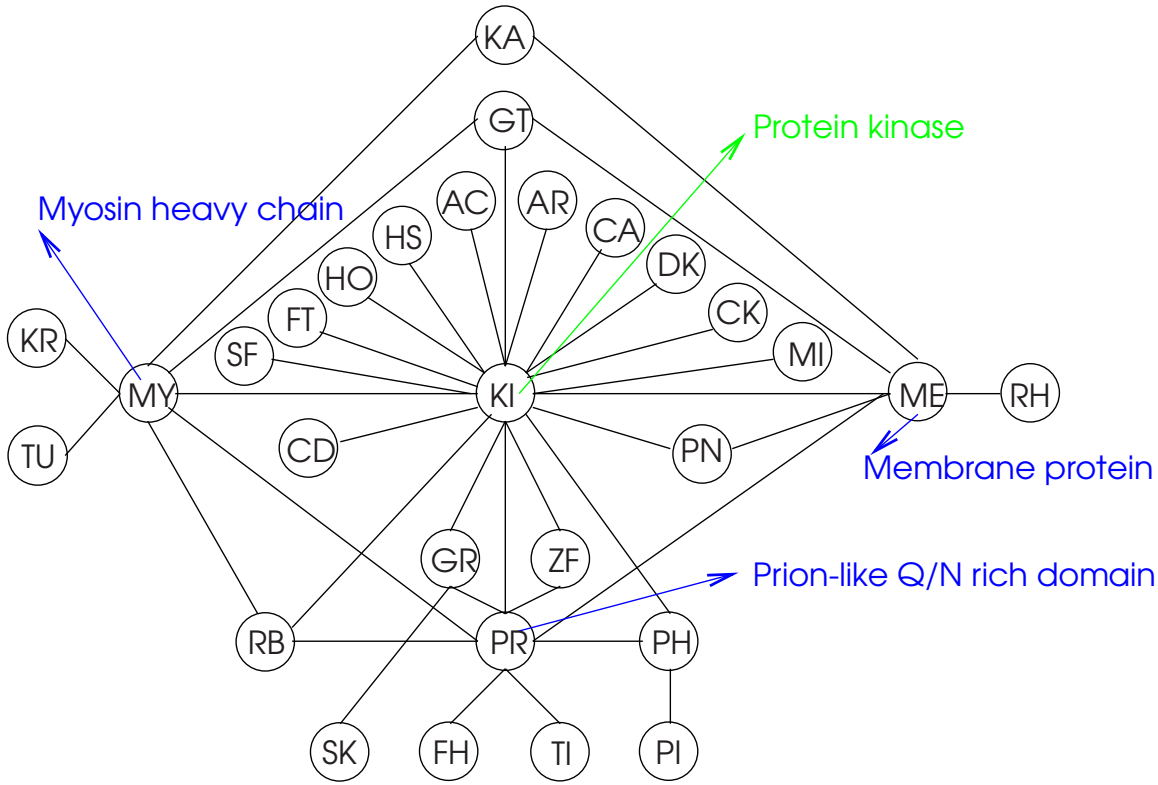
45 (29%) organisms

30 (19%) organisms

22 (14%) organisms

Frequent Interaction Patterns in DIP

- Protein interaction networks for 7 organisms
  - Ecoli, Hpylo, Scere, Celeg, Dmela, Mmusc, Hsapi
  - 44070 interactions between 16783 proteins
- Clustering with TribeMCL & node contraction
  - 30247 interactions between 6714 protein families



Runtime Characteristics

Comparison with isomorphism-based algorithms

Dataset	Minimum Support (%)	FSG Runtime (secs.)	FSG Largest pattern	FSG Number of patterns	MULE Runtime (secs.)	MULE Largest pattern	MULE Number of patterns
Glutamate	20	0.2	9	12	0.01	9	12
	16	0.7	10	14	0.01	10	14
	12	5.1	13	39	0.10	13	39
	10	22.7	16	34	0.29	15	34
	8	138.9	16	56	0.99	15	56
Alanine	24	0.1	8	11	0.01	8	11
	20	1.5	11	15	0.02	11	15
	16	4.0	12	21	0.06	12	21
	12	112.7	17	25	1.06	16	25
	10	215.1	17	34	1.72	16	34

Extraction of contracted patters

Glutamate metabolism, $\sigma = 8\%$				Alanine metabolism, $\sigma = 10\%$			
Size of contracted pattern	Extraction time (secs.) FSG	Extraction time (secs.) gSpan	Size of extracted pattern	Size of contracted pattern	Extraction time (secs.) FSG	Extraction time (secs.) gSpan	Size of extracted pattern
15	10.8	1.12	16	16	54.1	10.13	17
14	12.8	2.42	16	16	24.1	3.92	16
13	1.7	0.31	13	12	0.9	0.27	12
12	0.9	0.30	12	11	0.4	0.13	11
11	0.5	0.08	11	8	0.1	0.01	8
Total number of patterns: 56				Total number of patterns: 34			
Total runtime of FSG alone: 138.9 secs.				Total runtime of FSG alone :215.1 secs.			
Total runtime of MULE+FSG: 0.99+100.5 secs.				Total runtime of MULE+FSG: 1.72+160.6 secs.			
Total runtime of MULE+gSpan: 0.99+16.8 secs.				Total runtime of MULE+gSpan: 1.72+31.0 secs.			

Aligning Protein Interaction Networks

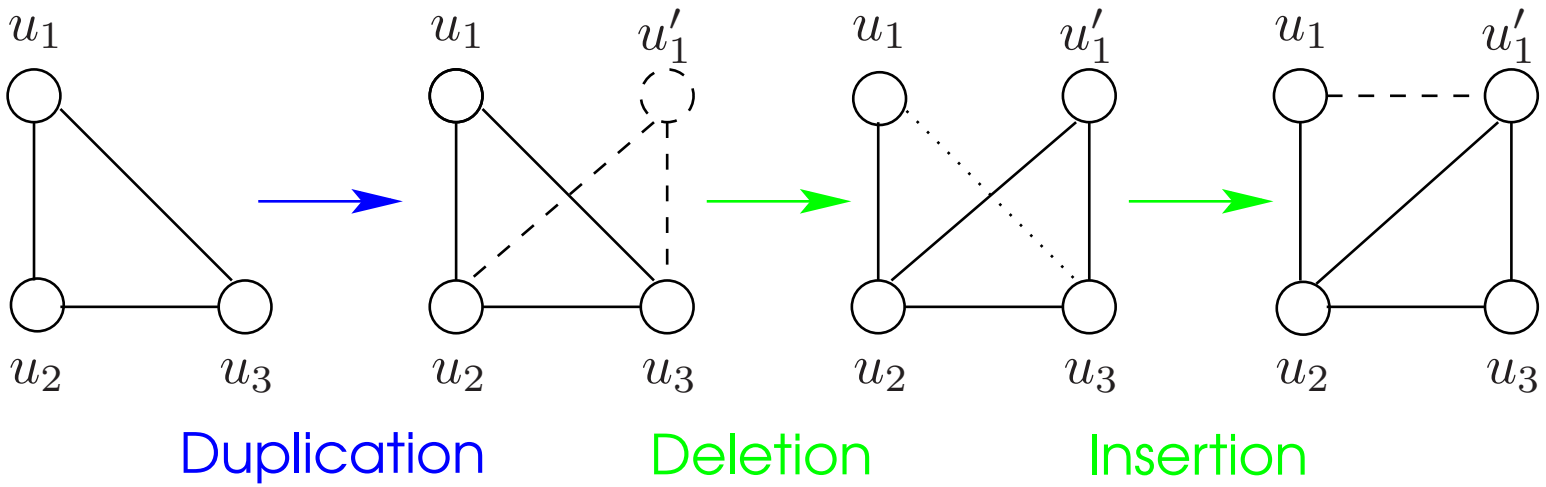
(Koyutürk, Grama, Szpankowski, RECOMB05)

- Defining graph alignment is difficult in general
  - Biological significance
  - Mathematical modeling
- Existing algorithms are based on simplified formulations
  - PathBLAST aligns pathways (linear chains) to render problem computationally tractable
  - Motif search algorithms look for small topological motifs, do not take into account conservation of proteins
- Our approach
  - Aligns subsets of proteins based on the observation that modules and complexes are conserved
  - Guided by models of evolution: Detailed understanding of conservation/ divergence

Evolution of Protein Interaction Networks

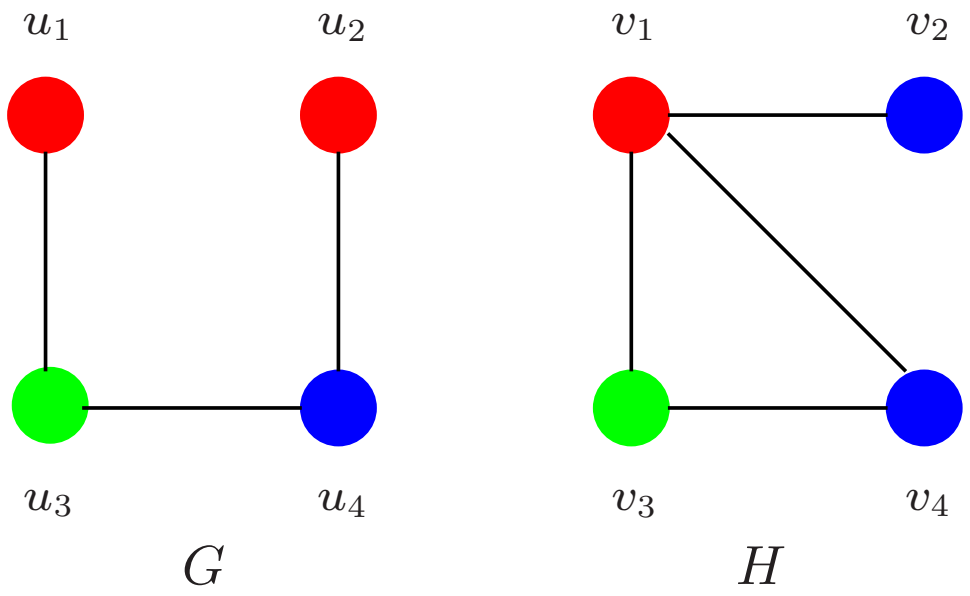
- Duplication/divergence models for the evolution of protein interaction networks
  - Interactions of duplicated proteins are also duplicated
  - Duplicated proteins rapidly lose interactions through mutations

- This provides us with a simplified basis for solving a very hard problem



Aligning Protein Interaction Networks: Input

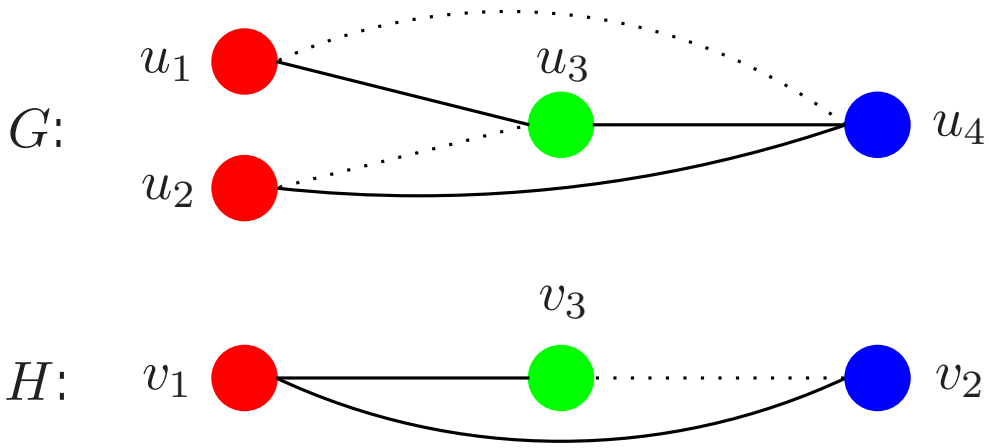
- PPI networks $G(U, E)$ and $H(V, F)$
- Sparse similarity function $S(u, v)$ for all $u, v \in U \cup V$
  - If $S(u, v) > 0$, u and v are potentially orthologous



Identical color $\Rightarrow S > 0$

Local Alignment Induced by Subsets of Proteins

- Alignment induced by protein subset pair $P = \{\tilde{U} \in U, \tilde{V} \in V\}$: $\mathcal{A}(\mathcal{P}) = \{\mathcal{M}, \mathcal{N}, \mathcal{D}\}$
  - A match $\in \mathcal{M}$ corresponds to two pairs of homolog proteins from each protein subset such that both pairs interact in both PPI networks. A match is associated with score μ.
  - A mismatch $\in \mathcal{N}$ corresponds to two pairs of homolog proteins from each PPI network such that only one pair is interacting. A mismatch is associated with penalty ν.
  - A duplication $\in D$ corresponds to a pair of homolog proteins that are in the same protein subset. A duplication is associated with penalty δ.



Alignment induced by protein subset pair

$\{\{u_1, u_2, u_3, u_4\}, \{v_1, v_2, v_3\}\}$

Pairwise Local Alignment of PPI networks

- Alignment score:
  $\sigma(\mathcal{A}(P)) = \sum_{M \in \mathcal{M}} \mu(M) - \sum_{N \in \mathcal{N}} \nu(N) - \sum_{D \in \mathcal{D}} \delta(D)$
  - Matches are rewarded for conservation of interactions
  - Duplications are penalized for differentiation after split
  - Mismatches are penalized for divergence and experimental error

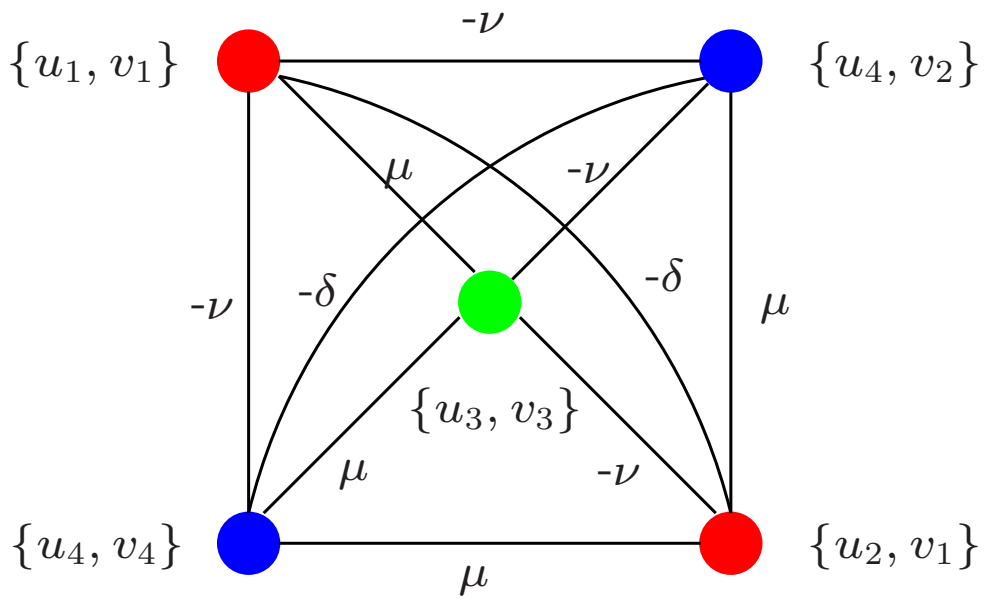
- All scores and penalties are functions of similarity between associated proteins

- Problem: Find all protein subset pairs with statistically significant alignment score.
  - High scoring protein subsets are likely to correspond to conserved modules or complexes

- A graph equivalent to BLAST

Weighted Alignment Graph $\mathbf{G(V, E)}$

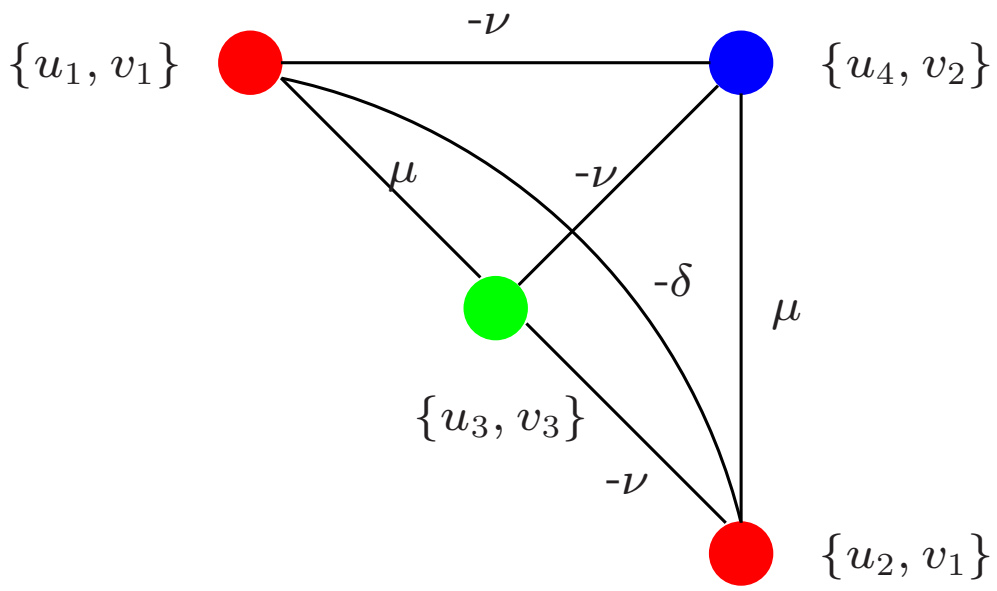
- $\mathbf{V}$ consists all pairs of ortholog proteins $\mathbf{v} = \{u \in U, v \in V\}$
- An edge $\mathbf{vv'} = \{uv\}\{u'v'\}$ in $\mathbf{E}$ is a
  - match edge if $uu' \in E$ and $vv' \in V$, with weight $w(\mathbf{vv'}) = \mu(uv, u'v')$
  - mismatch edge if $uu' \in E$ and $vv' \notin V$ or vice versa, with weight $w(\mathbf{vv'}) = -\nu(uv, u'v')$
  - duplication edge if $S(u, u') > 0$ or $S(v, v') > 0$, with weight $w(\mathbf{vv'}) = -\delta(u, u')$ or $w(\mathbf{vv'}) = -\delta(v, v')$



Maximum Weight Induced Subgraph Problem

- Definition: (MAWISH)
  - Given graph $\mathbf{G}(\mathbf{V}, \mathbf{E})$ and a constant ϵ, find $\tilde{\mathbf{V}} \in \mathbf{V}$ such that $\sum_{\mathbf{v},\mathbf{u} \in \tilde{\mathbf{V}}} w(\mathbf{vu}) \geq \epsilon$.
  - NP-complete

- Theorem: (MAWISH $\equiv$ Pairwise alignment)
  - If $\tilde{\mathbf{V}}$ is a solution for the MAWISH problem on $\mathbf{G}(\mathbf{V}, \mathbf{E})$, then $P = \{\tilde{U}, \tilde{V}\}$ induces an alignment $\mathcal{A}(P)$ with $\sigma(\mathcal{A}) \geq \epsilon$, where $\tilde{\mathbf{V}} = \tilde{U} \times \tilde{V}$.



A Greedy Algorithm for MAWISH

- Greedy graph growing
  - Start with a heavily connected node, put it in $\tilde{\mathbf{V}}$
  - Choose $\mathbf{v}$ that is most heavily connected to $\tilde{\mathbf{V}}$ and put it in $\tilde{\mathbf{V}}$ until no $\mathbf{v}$ is positively connected to $\tilde{\mathbf{V}}$.
  - If total weight of the subgraph induced by $\tilde{\mathbf{V}}$ is statistically significant, return $\tilde{\mathbf{V}}$
  - Works in linear time.

- As modules and complexes are densely connected within the module and loosely connected to the rest of the network, this algorithm is expected to be effective.

- For all local alignments, remove discovered subgraph and run the greedy algorithm again.

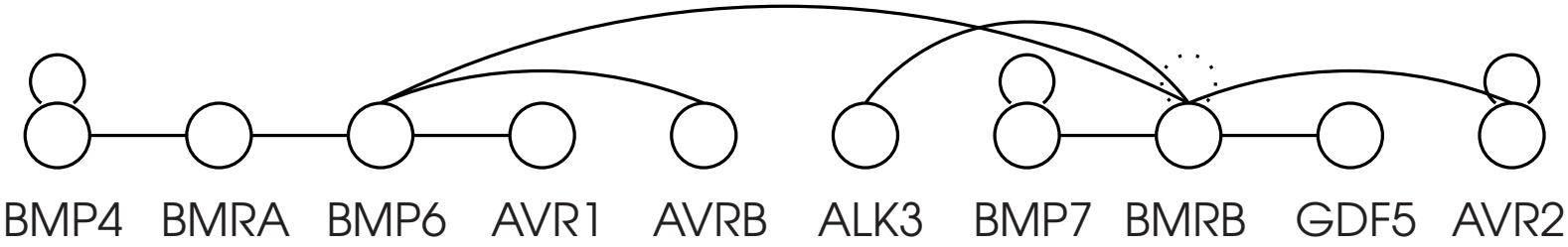
- If the number of homologs for each protein is constant, construction of alignment graph and solution of the MAWISH takes $O(|E| + |F|)$ time.

Scoring Matches, Mismatches and Duplications

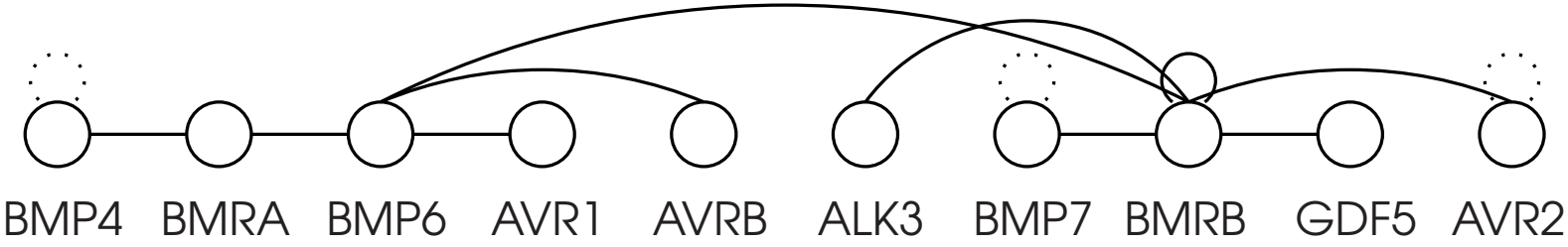
- Quantizing similarity between two proteins
  - Confidence in two proteins being orthologous (paralogous)
  - BLAST E-value: $S(u,v) = log_{10}\frac{p(u,v)}{p_{random}}$, where $p(u,v)$ is the probability of true homology between u and v, given BLAST E-value
  - Ortholog clustering: $S(u,v) = c(u)c(v)$, where $0 \leq c(u) \leq 1$ is the confidence of the INPARANOID algorithm in assigning u to its corresponding cluster
- Match score
  - Two interactions are orthologous only if both interacting partners are orthologous
  - $\mu(uu', vv') = \bar{\mu} \min\{S(u,v), S(u',v')\}$
- Mismatch penalty
  - $\nu(uu', vv') = \bar{\nu} \min\{S(u,v), S(u',v')\}$
- Duplication penalty
  - $\delta(u,u') = \bar{\delta}(d - S(u,u'))$

Alignment of Human and Mouse PPI Networks

Homo Sapiens



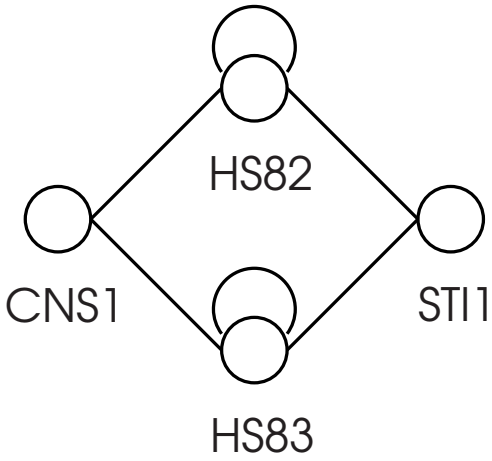
Mus Musculus



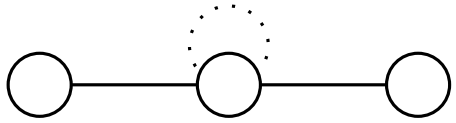
A conserved subnet that is part of
transforming growth factor beta receptor signaling pathway

Alignment of Yeast and Fly PPI Networks

Saccharomyces Cerevisiae



Drosophila Melanogaster



A conserved subnet that is part of response to stress

Penalties need to be relaxed while analyzing distant species

Ongoing Work on PPI Network Alignment

- Assessing statistical significance
  - Constructing a refence model based on models of evolution
- BLAST-like search queries for network alignment
  - Given a query graph, find all high-scoring local alignments in a database of PPI networks
- Multiple Graph Alignment (CLUSTAL, BLASTCLUST)
  - How to combine graph mining and pairwise alignment

Inferring Functional Modules from Phylogenetic Information

(Kim, Koyutürk, Topkara, Grama, Subramaniam, ECCB05 (submitted))

- Functionally related proteins are likely to have co-evolved
  - Construct phylogenetic profile for each genome: Vector of E-values signifying existence of an orthologous protein in each organism
  - Identify pairwise functional associations based on mutual information between phylogenetic profiles (Pellegrini et al. (1999))
  - Mutual information:
    $I(X,Y) = H(X) - H(X|Y) = \sum_x \sum_y p(x,y) \log(p(x,y)/p(x)p(y))$
  - Shown to identify functionally associated protein pairs at a coarser level than high-throughput methods

- However, domains, not proteins, co-evolve
  - How can we incorporate domain information to enhance performance of phylogeny-based interaction prediction?

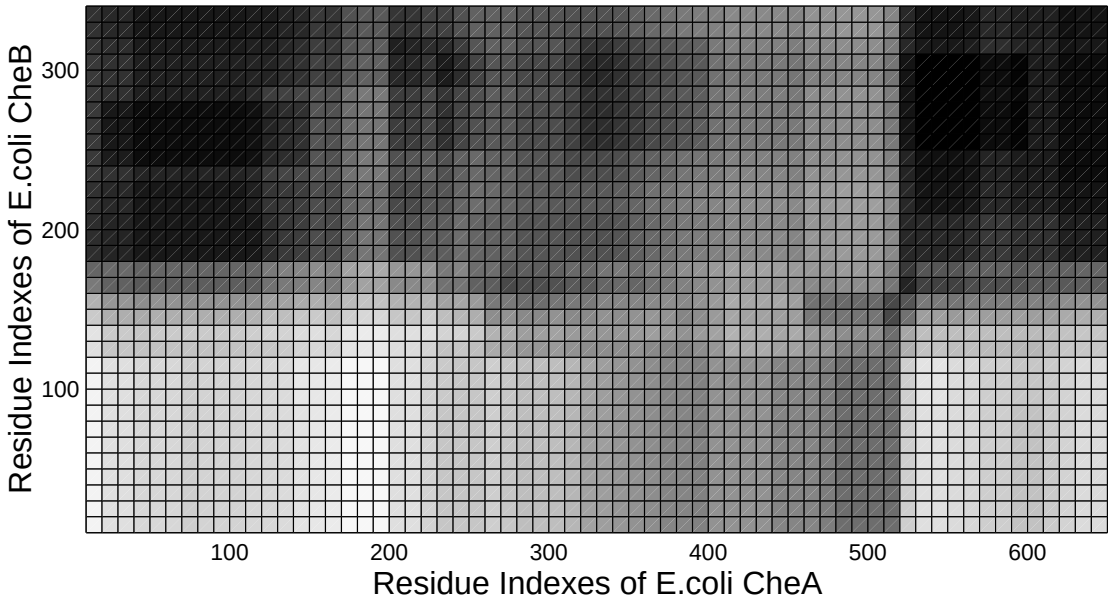
Identification of Co-evolved Domains

- While sequence information is widely available, domain information is not generally comprehensive
- Approximating domains between fixed-size segments (Kim & Subramaniam (2004))
  - Chop proteins into overlapping (*e.g.*, 30 residues) fixed-size (*e.g.*, 120 residues) segments
  - Construct phylogenetic profile for each segment, find maximum-mutual-information segment pair for each protein pair
  - Improves single-profile based approach
  - However, there is no fixed domain size
- Can we identify domains from phylogenetic information as well?
  - Residue phylogenetic profiles!

Residue-Level Phylogenetic Analysis

- Residue phylogenetic profile
  - For each residue r_{ij} on protein P_i, the existence of r_{ij} in genome G_k is signified by the minimum e-value of alignments between P_i and G_k that contain r_{ij}
- Mutual information matrix
  - Matrix of mutual information between any pair of residues each from one protein
  - $M(P_i, P_j) = [m_{kl}]$,
    where $m_{kl} = I(profile(r_{ik}), profile(r_{jl}))$
- A sufficiently large contiguous submatrix of the mutual information matrix that contains consistently high entries may correspond to a pair of co-evolved domains.

Mutual Information Matrix



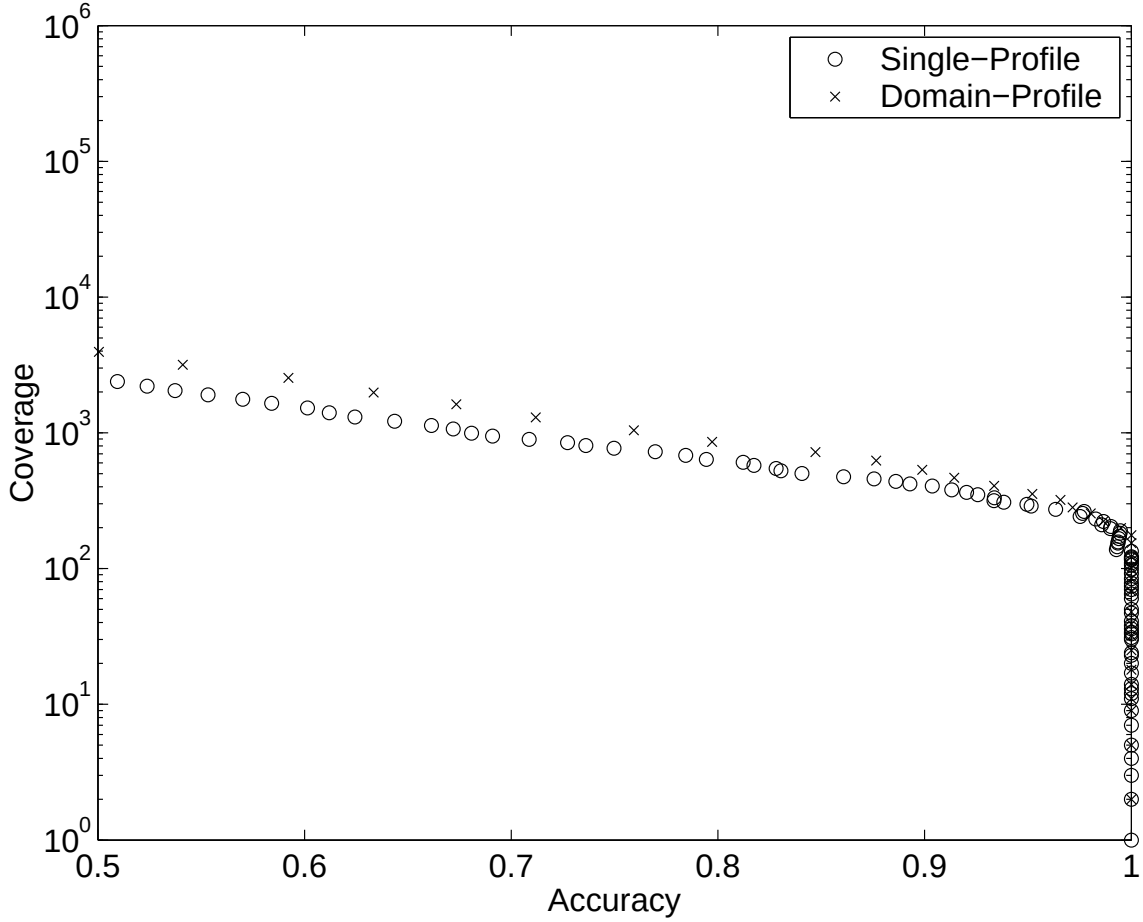
Mutual information matrix for proteins CheA and CheB in E-coli.
Darker pixels indicate higher mutual information.

Co-evolved domains identified by dark submatrices!

Clustering Residue Phylogenetic Profiles

- Cluster residues to identify co-evolved domains
- For each protein pair
  - Downsample residues of each protein (for computational efficiency)
  - Construct residue phylogenetic profiles
  - Compute mutual information matrix
  - Identify sufficiently large contiguous submatrices of mutual information matrix with consistently high mutual-information scores
  - Set phylogenetic association score of the two proteins to the maximum of mutual information of such matrices
- Can be used for domain identification as well!

Comparison of Domain-Profile and Single-Profile Methods



Accuracy vs Coverage

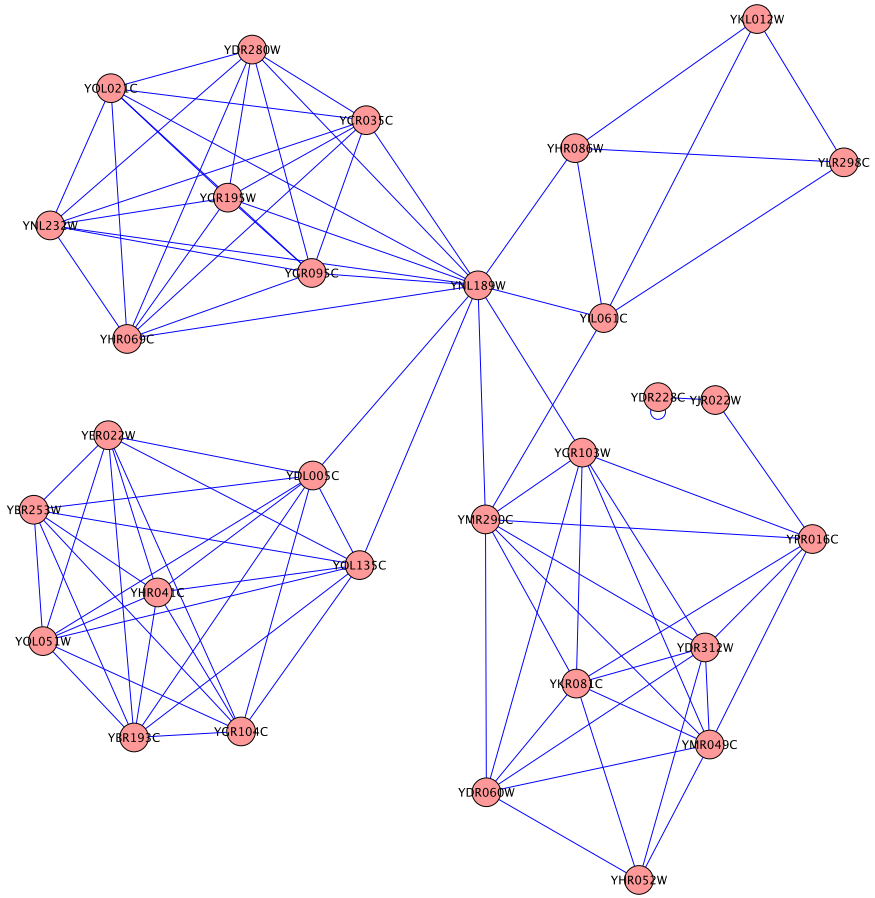
Accuracy: Fraction of true-positives among all predicted functional associations

Coverage: Number of functionally associated protein pairs that are identified by the algorithm

Phylogenetic Analysis of Computationally Identified Functional Modules

- "Comprehensive" PPI network for Saccharomyces Cerevisiae is available
- We can identify functional modules in this network based on density of interactions
  - Proteins in a functional module are expected to densely interact with each other
- Whole genome sequences for many other species are available
  - 12 yeast species: *S. Bayanus, S. Kluyveri, S. Kudriavzevii, S. Paradoxus, S. Mikatae, S. Castellii, K. Lactis, D. Hansenii, A. Gossypii, C. Glabrata, Y. Lipolytica, S. Pombe*
- Can we analyze the conservation of S. Cerevisiae modules by projecting them on other yeast species based on sequence comparison?

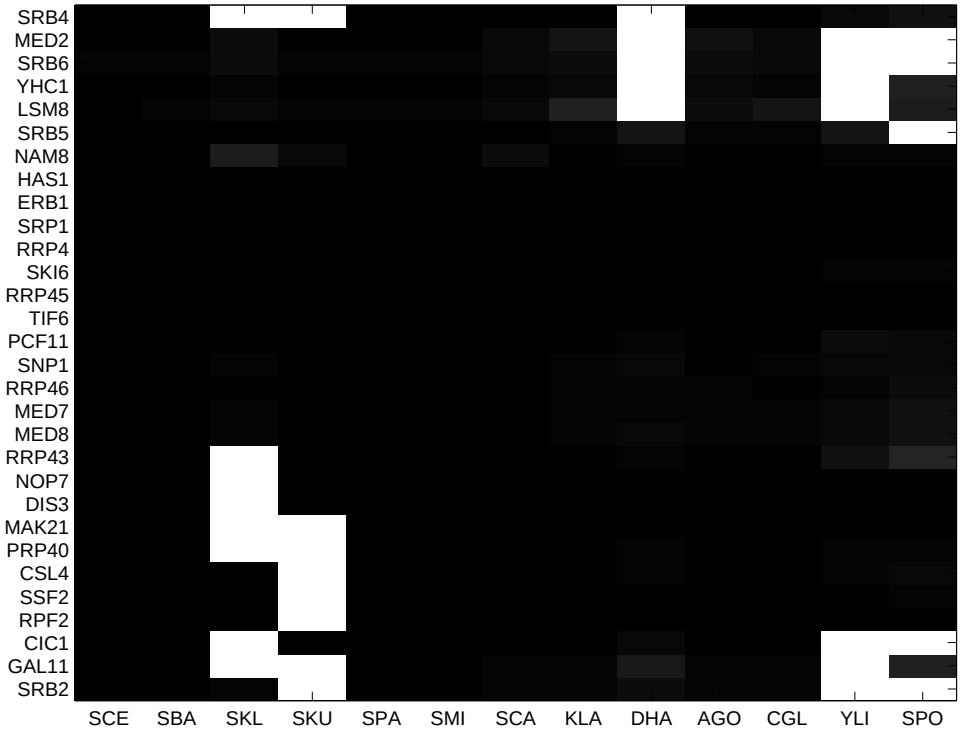
Computational Identification of Functional Modules on *S. Cerevisiae* PPI Network



A group of densely interacting proteins involved in RNA processing as identified by the MCODE algorithm

Module Phylogenetic Matrix

- For each protein in module, find orthologs in other yeast proteomes
  - If BLAST E-value for the best match of protein P_i in organism G_j is E_{ij}, set the $(i, j)^{th}$ entry of module phylogenetic matrix to $1 - 1/\log(E_{ij})$.

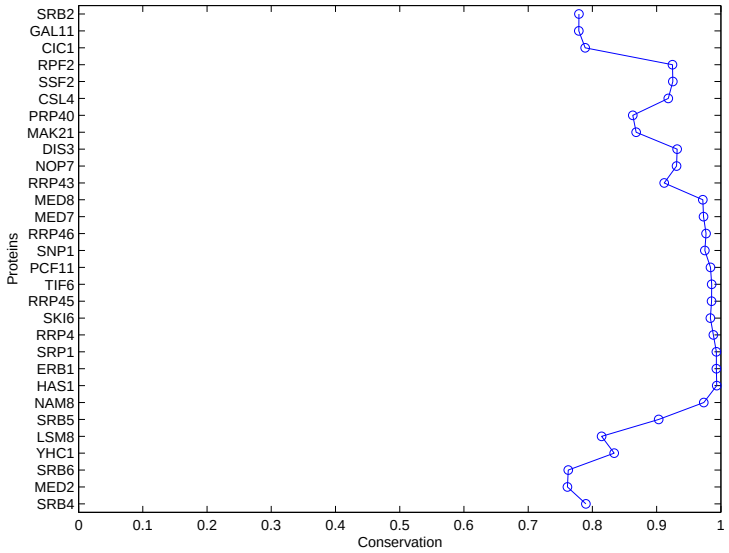


Module phylogenetic matrix for the module in previous slide

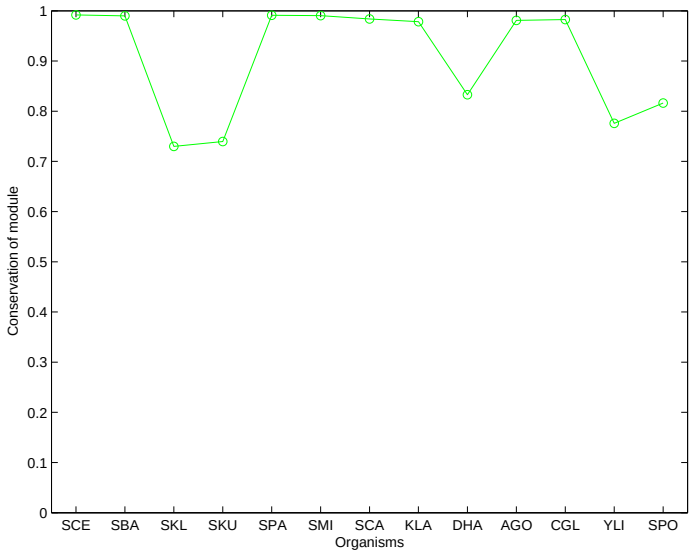
Rows: Proteins, Columns: Organisms

Darker box indicates higher significance

Module Conservation



First principal component in protein space



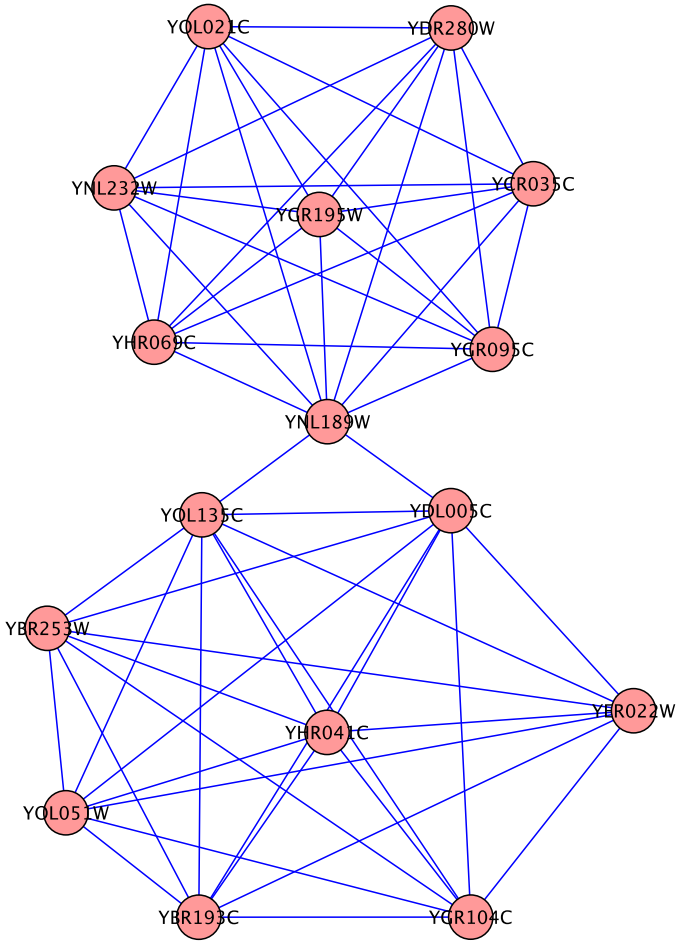
First principal component in organism space

- Principal component analysis provides a broad idea about conservation
  - MED8, MED7, RRP46, SNP1, PCF11, TIF6, RRP25, SKI6, RRP4, SRP1, ERB1, and HAS1 are almost perfectly conserved in all genomes, while others are partially conserved
  - S. Bayanus, S. Paradoxus, S. Mikatae, S. Castellii, K. Lactis, A. Gossypii, and Y. Lipolytica contain orthologs of almost all proteins in the module, while others contain only some of them

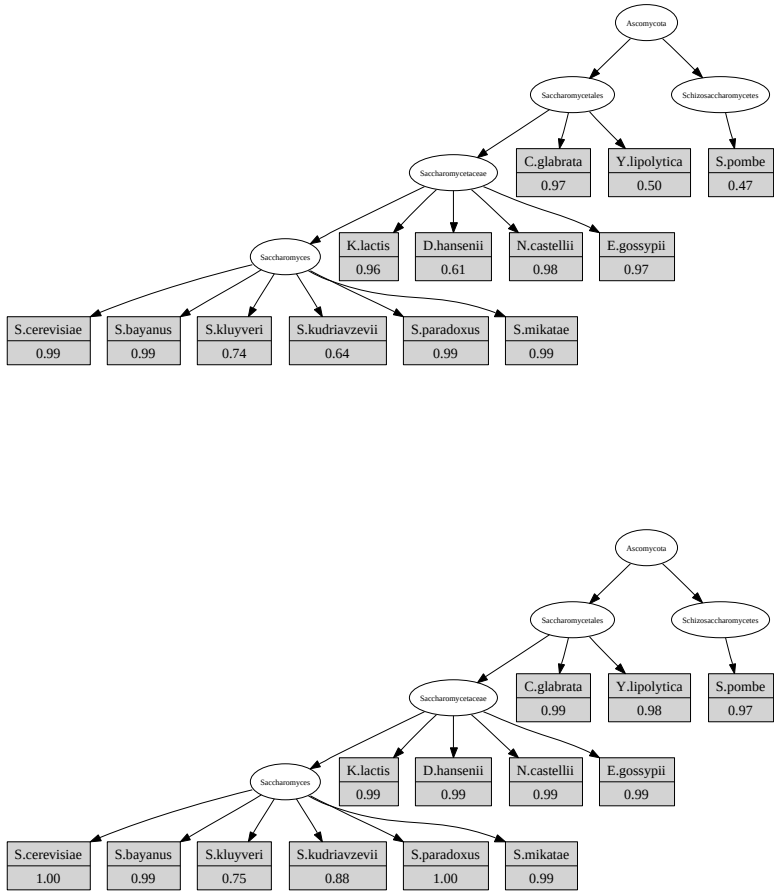
- Which proteins are conserved in which organisms? Why?

An Example for Module Specification

Module



Conservation on NCBI Taxonomy



Lower sub-module (RNA polymerase II transcription mediator activity) is completely conserved in Y. Lipolytica, S. Pombe, and D. Hansenii, while the upper sub-module (3′-5′-exoribonuclease activity) almost disappears in these organisms

Ongoing Work on Module Phylogenetics

- Which proteins are conserved in which organisms? Why?
  - Does partial loss of proteins in a module imply loss of function?
  - Are there modules that are divided due to functional divergence?
  - Are there modules that are completely lost due to functional divergence?
  - If a module is completely conserved, what does this imply in terms of functional conservation and evolutionary pressure?
  - Are module-specific phylogenetic trees consistent with the whole-proteome phylogenotic tree?
  - How is topology in PPI network related to conservation?

- How do we quantify the conservation of a module in a given organism?

So many questions, so little time!