Comparative Analysis of Molecular Interaction Networks

Ananth Grama
Coordinated Systems Lab and,
Computer Science Department,
Purdue University

http://www.cs.purdue.edu/people/ayg

Various parts of this work involve collaborations with Priya Vashishtha, Rajiv Kalia, Aiichiro Nakano (USC), William Goddard (CalTech), Yohan Kim and Shankar Subramaniam (UCSD), Vipin Kumar (UMN), Eric Jakobsson (UIUC), Larry Scott (IIT), Ahmed Sameh, Mete Sozen, Mehmet Koyuturk, Suresh Jagannathan (Purdue)

This work is supported by National Institutes of Health, National Science Foundation, Department of Energy, and Intel.

Outline

- Preamble: Who we are and what we do
- Molecular Interaction Networks
  - Modeling, evolution, problems, practical implications
- Algorithms for Analyzing Molecular Interaction Networks
  - *Analyzing* biological networks for *conserved molecular interaction patterns*
  - *Pairwise* Alignment *of protein-protein interaction networks*
  - Probabilistic *models/analyses for assessing* statistical significance
- Computational Synthesis of Interaction Networks
  - Inferring function from domain co-evolution
- Ongoing Work

Lab Overview

- Development of algorithmic and software substrates to solve fundamental problems in science and engineering.
- Research transcends software infrastructure (compilers, OS), algorithms (numerical and combinatorial), platforms (motes to petascale), and software (libraries to services).
- We focus on problems at the core of computing, but measure the value of our work in terms of its impact on science and engineering applications.
- All of our projects are in close collaboration with domain experts.

Lab Overview: Sample Projects

Model Reduction and Control of Large Structures

- Virtually all (new) large structures have some form passive or semi active control mechanisms.

Building	Control Mechanism	Damping Fr., Effective Damper Mass.
CN Tower, Toronto (533m).	Passive Tuned Mass Damper	
John Hancock Bldg, Boston (244m).	Two Passive Tuned Dampers	0.14 Hz, 2 x 300t, 4% damping ratio
Sydney Tower (305m)	Passive Tuned Pendulum	0.1, 0.5Hz, 220t
Rokko Island P&G, Kobe (117m)	Passive Tuned Pendulum	0.33 – 0.62Hz, 270t
Yokohama Landmark Tower (296m)	Active Tuned Mass Dampers (2)	0.185Hz, 340t
Shinjuku Park Tower, Tokyo (227m)	Active Tuned Mass Dampers (3)	330t
TYG Building, Atsugi (159m)	Tuned Liquid Dampers (720)	0.53 Hz, 18.2t

Model Reduction and Control of Large Structures

Pictured left is a passive fluid damper with bottom casing containing the bearings and oil used to absorb seismic energy. Pictured right is a semiactive damper with variable orifice damping (Picture credits Steven Williams).

Model Reduction and Control of Large Structures

The Dongting Lake Bridge is being retrofitted with MR dampers to control wind-induced vibration.

Model Reduction and Control of Large Structures

Objective: Develop the computational infrastructure to enable the next generation of civil infrastructure.

- Real-time sensing and actuation
- Model reduction and control
- In-network computation of control vectors

Model Reduction and Control of Large Structures

Pilot deployment at BOWEN labs

FM 433MHz

MICA2 motes with
ADXL 202

802.11b
Peer-to-Peer

Laser attached
via serial port to
Stargate computers

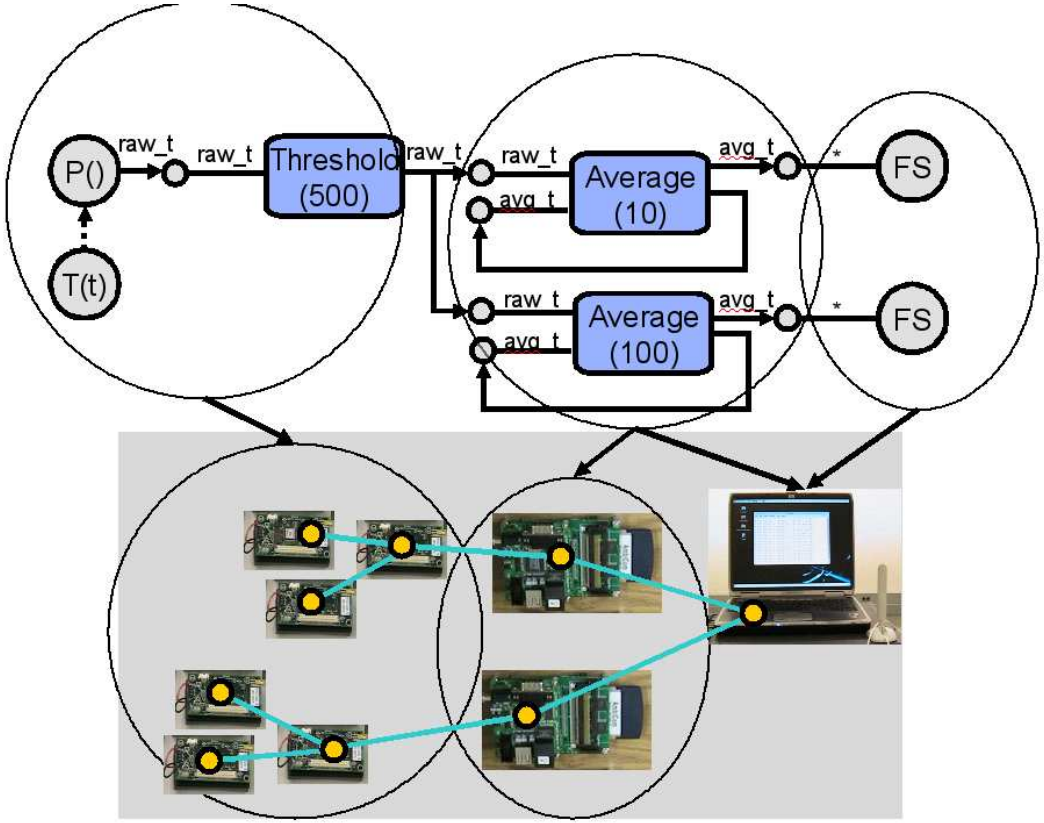
ECN Net

Internet

Currently laser readings
can be viewed for from
anywhere over the Internet
(conditioned on Firewall settings)

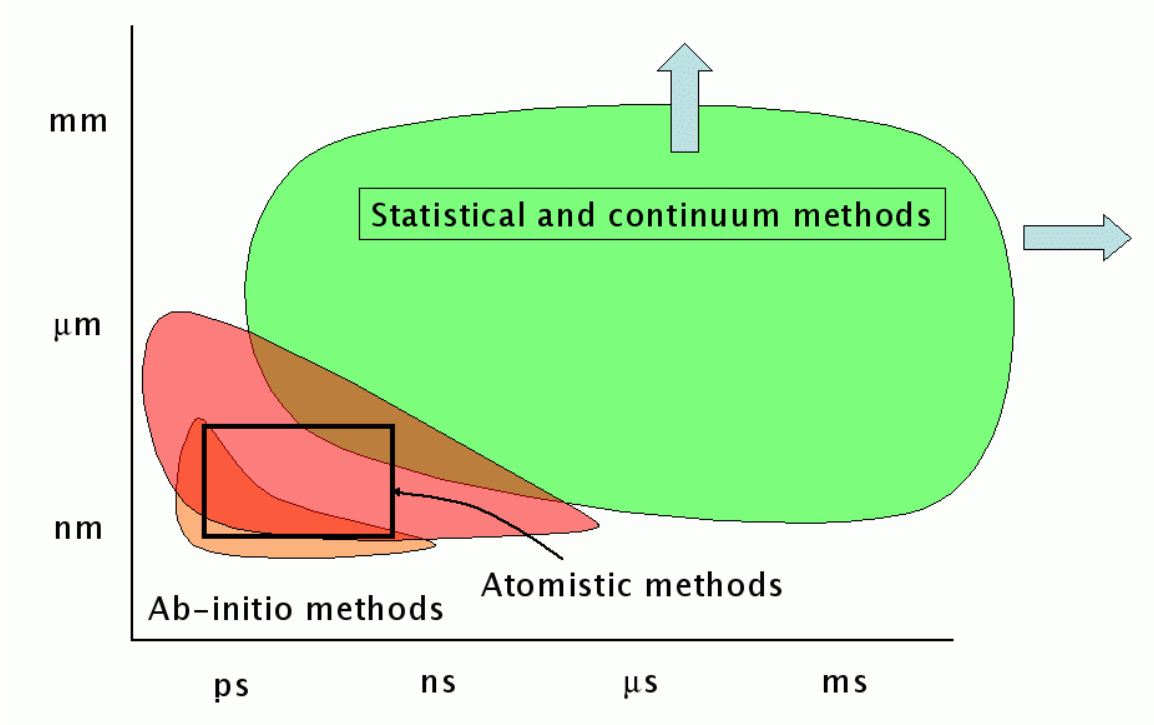
Model Reduction and Control of Large Structures

A critical component of embedded systems is the effort associated with application development. Our COSMOS environment fundamentally addresses this bottleneck.

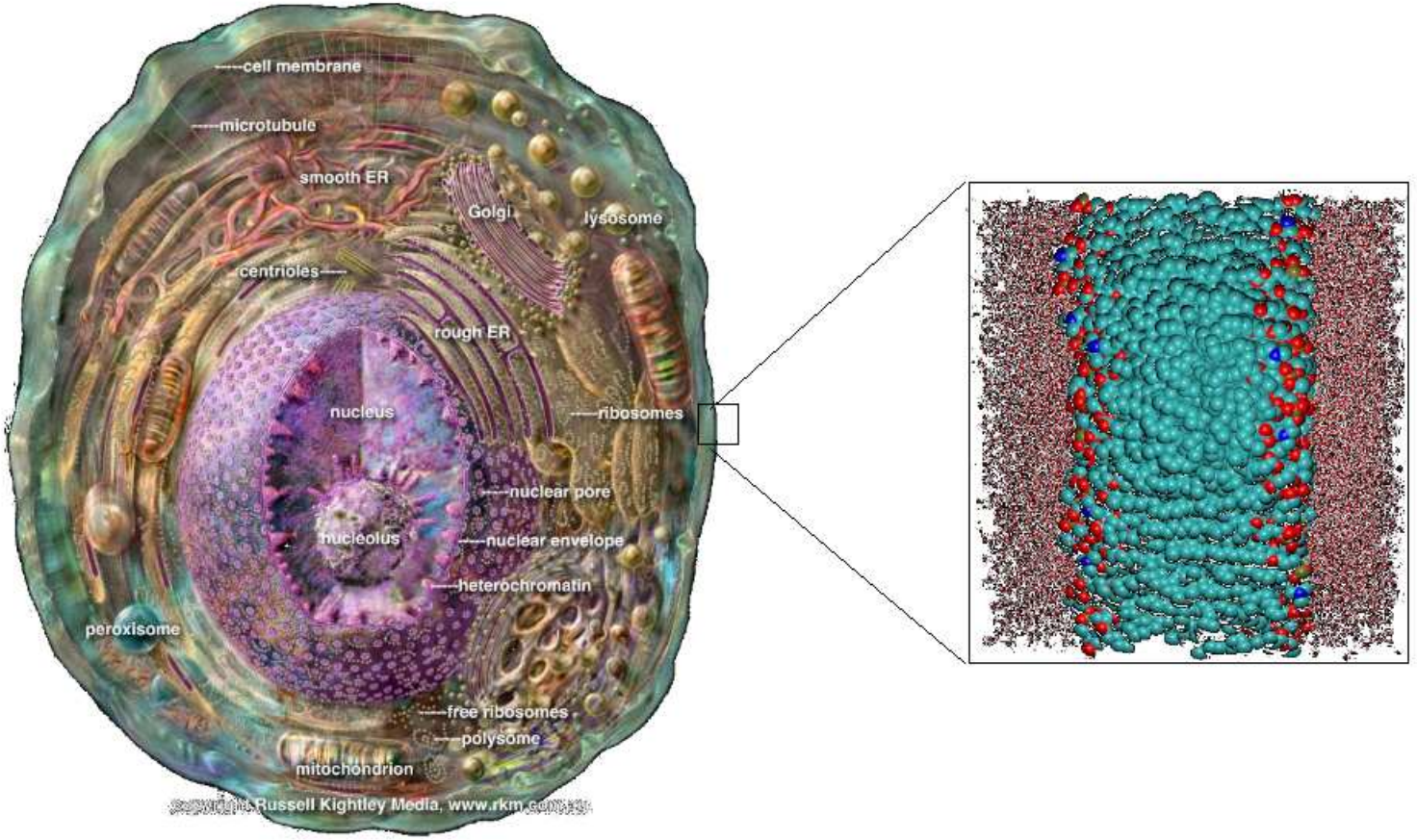


Simulation and Design of Materials

- Novel kernels for molecular simulations (fast electrostatics, periodic boundaries, solvers, integrators).
- New methods for multiscale modeling (ab-initio methods, reactive force fields, classical MD, and continuum methods).



Membrane Simulations



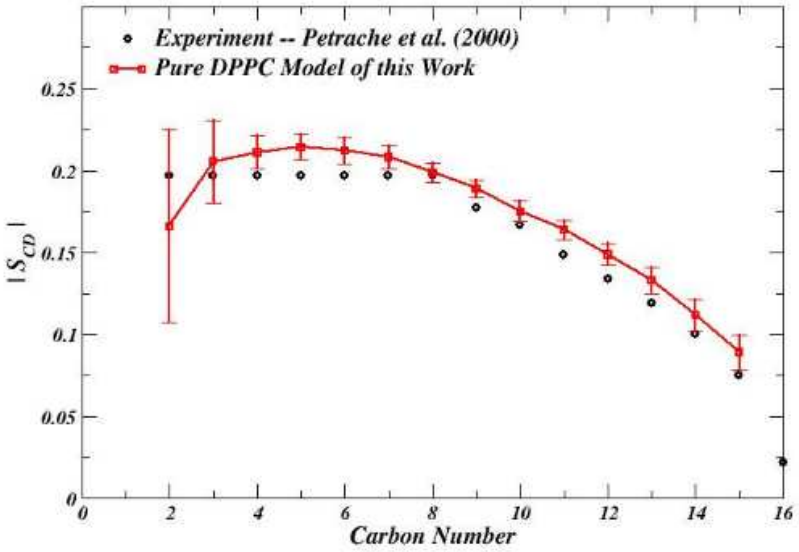
Membrane Simulations

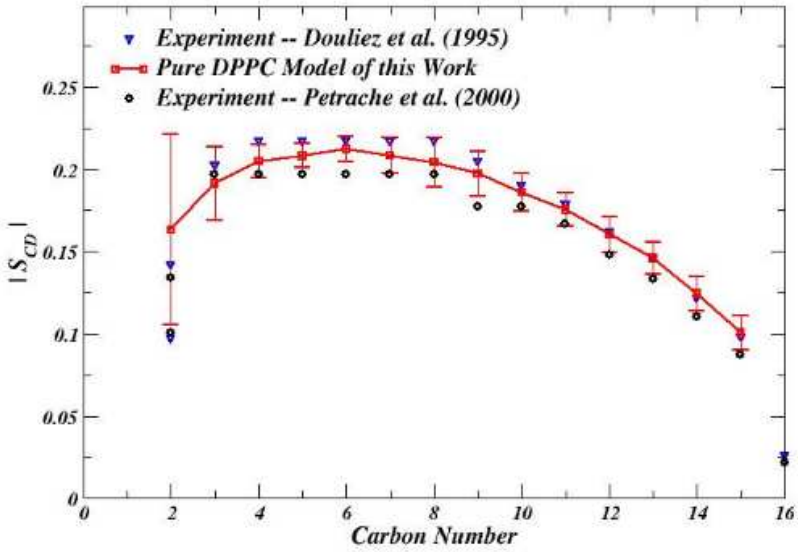
How real are they?

Lipid	Area per lipid (Å^2)	
	simulations	experiments
Dipalmitoylphosphatidylcholine (DPPC)	~ 62	~ 64
Dioleylphosphatidylcholine (DOPC)	~ 71	~ 72
Dipalmitoylphosphatidylserine (DPPS with Na^+ counter ions)	~ 53.6	~ 54
Sphingomyelin (18:0)	~ 53	~ 53
Cholesterol	~ 27†	~ 38*

Membrane Simulations

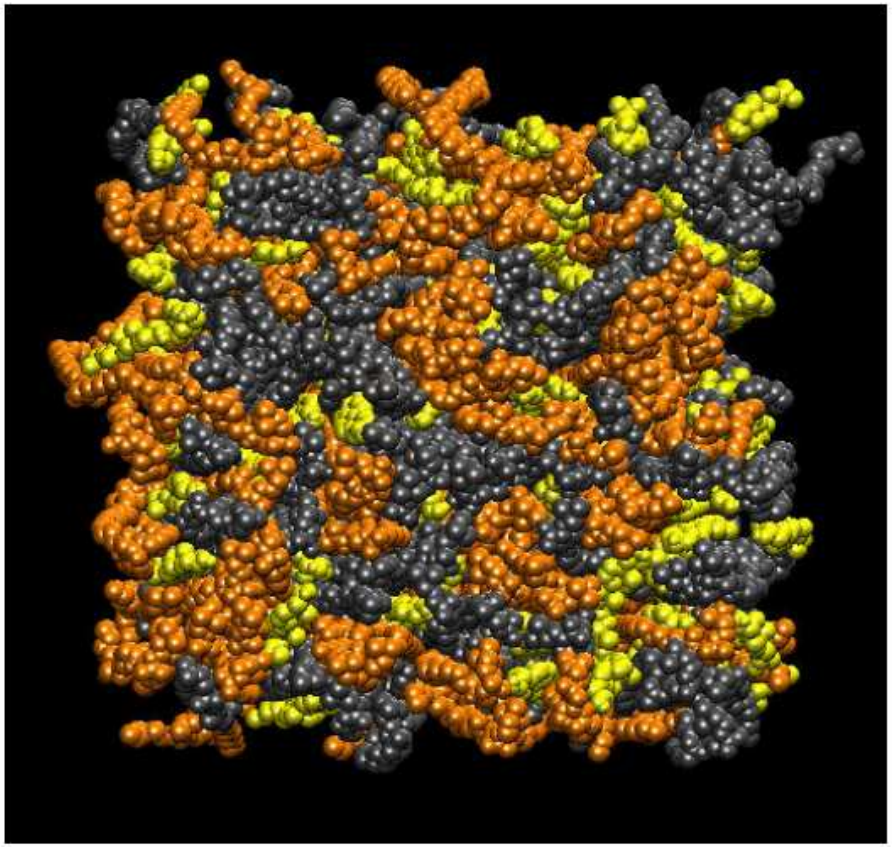
How real are they?

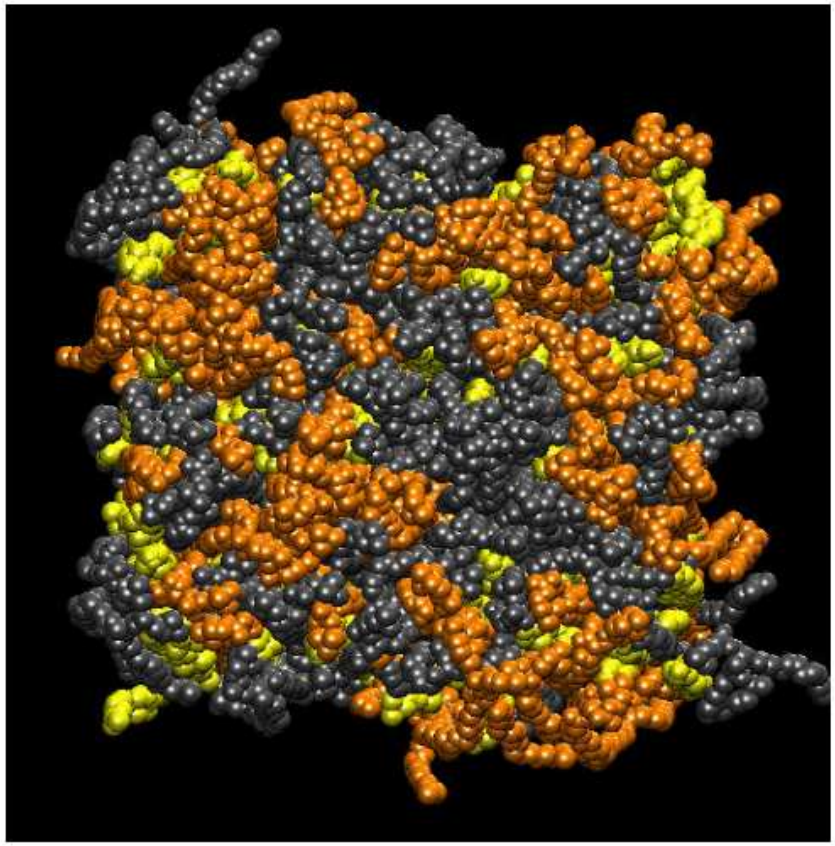




Simulation of 128 DPPC bilayer in SPC water

Membrane Simulations





Ternary mixture: 100 DOPC, 100 18:0 SM, 100 CHOL, 10000 water (43500 atoms), 250 ns.

Reactive Simulations

- Chemical reactions correspond to association and dissociation of chemical bonds.
  - Classical simulations cannot simulate reactions.
  - ab-initio methods calculate overlap of electron orbitals to investigate chemical reactions.
- ReaX force field postulates a classical bond order interaction to mimic the association and dissociation of chemical bonds.

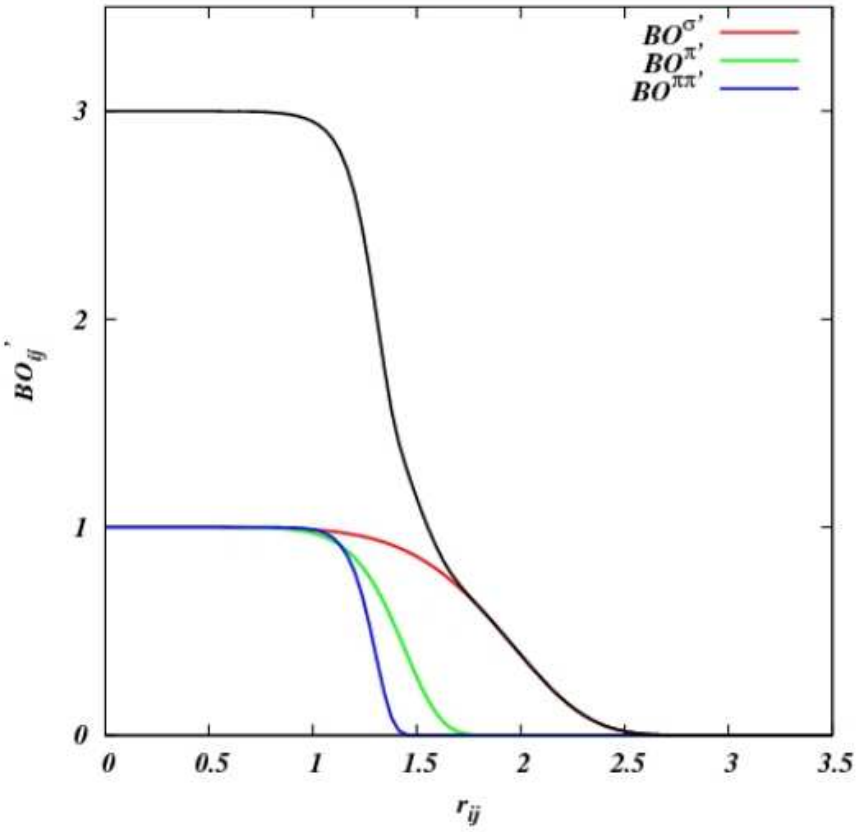
Reactive Simulations: Bond Order Interaction

- Uncorrected bond order:

$$BO_{ij}^{\alpha'}(r_{ij}) = \exp\left[a_\alpha \left(\frac{r_{ij}}{r_{0_\alpha}} \right)^{b_\alpha} \right]$$

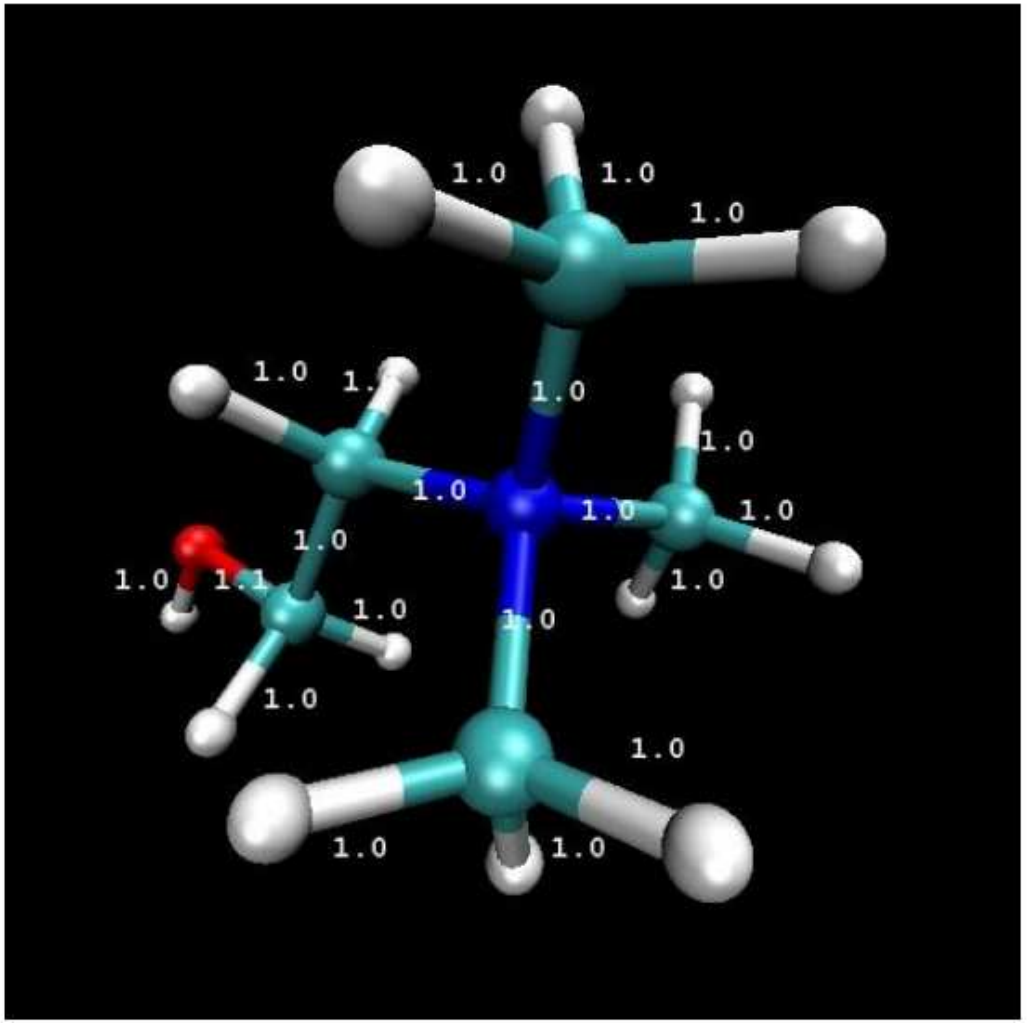
Where α is for
σ–σ, π–σ, and π–π bonds

- The total uncorrected bond order is sum of three types of bonds
- Bond order requires correction to account for the correct valency

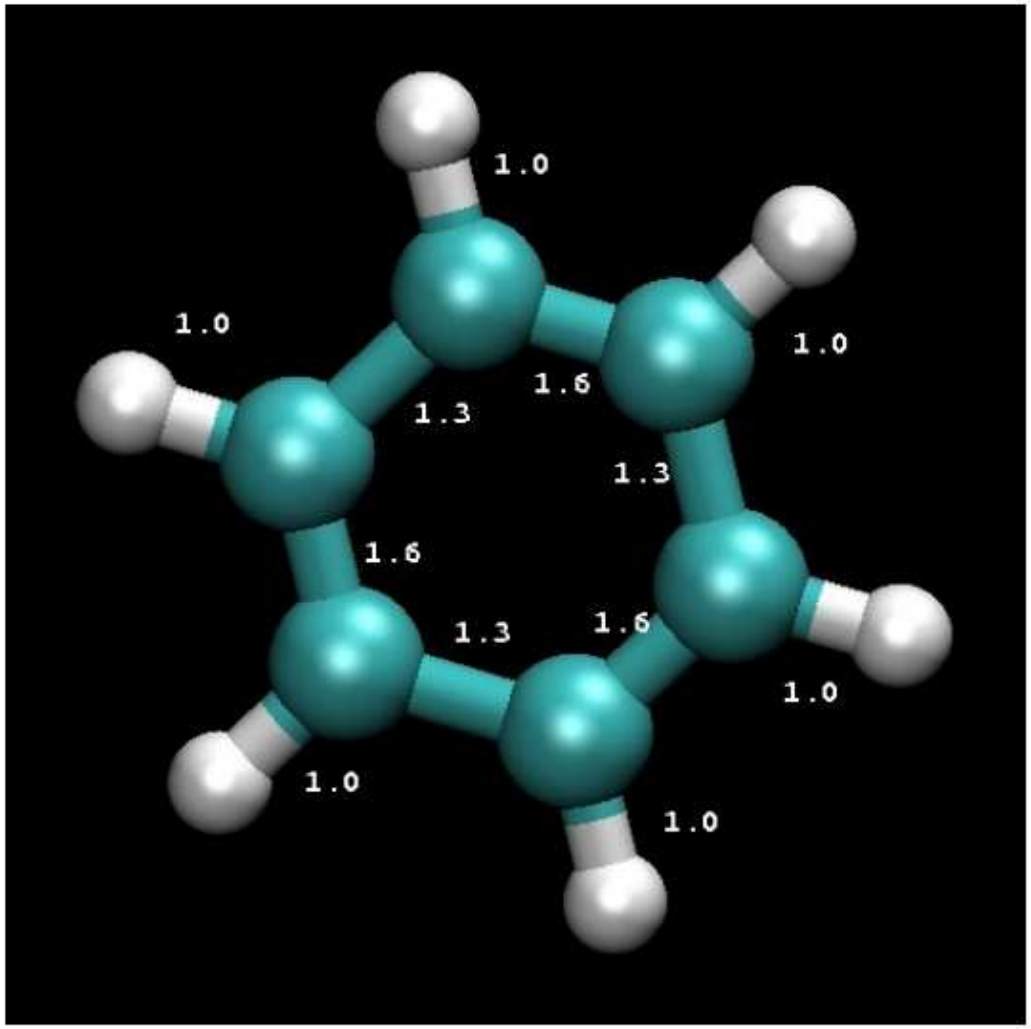


Bond order for C-C bond

Reactive Simulations: Bond Order of Choline



Reactive Simulations: Bond Order of Benzene



Reactive Simulations: Challenges

- Fast electrostatics and periodic boundaries.
- Charge equilibriation (qEq).
- Parallelism.

A comprehensive highly optimized code is currently being validated on a variety of systems for release.

Other Efforts

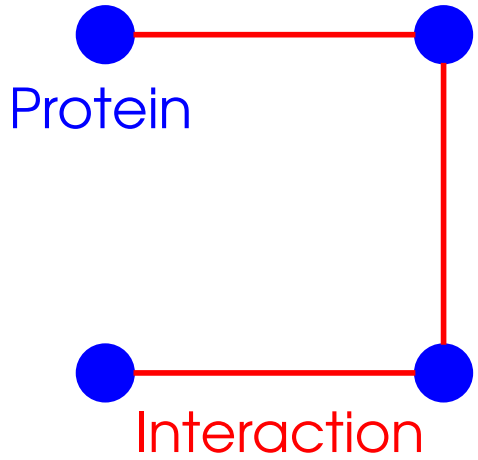
- Scaling solvers to petaFLOPS (DARPA/HPCS) and Next-Generation Solvers (NSF/CISE).
- Algorithmic Asynchrony and Scaling (NSF/CISE).
- Large-scale Storage Systems (NSF/STI).
- Speculation and Multicore Architectures (Intel).
- Modeling Corrossion and Cracking (DoE/SciDAC).

Outline

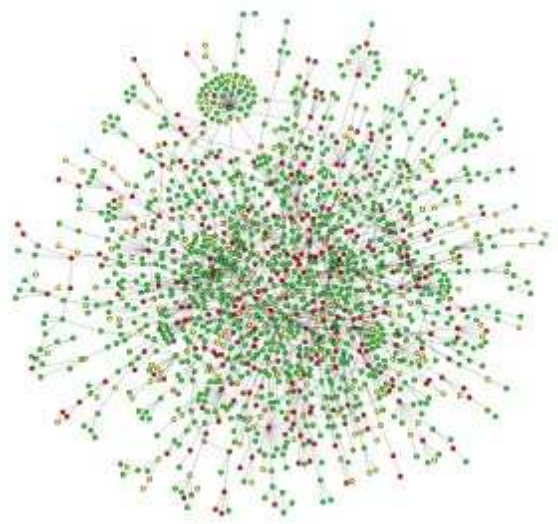
- Preamble: Who we are and what we do
- Molecular Interaction Networks
  - Modeling, evolution, problems, practical implications
- Algorithms for Analyzing Molecular Interaction Networks
  - *Analyzing* biological networks for *conserved molecular interaction patterns*
  - *Pairwise* Alignment *of protein-protein interaction networks*
  - Probabilistic *models/analyses for assessing* statistical significance
- Computational Synthesis of Interaction Networks
  - Inferring function from domain co-evolution
- Ongoing Work

Protein-Protein Interaction (PPI) Networks

- Interacting proteins can be identified via high-throughput screening
  - Two-hybrid
  - Mass spectrometry
  - Tandem affinity purification (TAP)



Undirected Graph Model

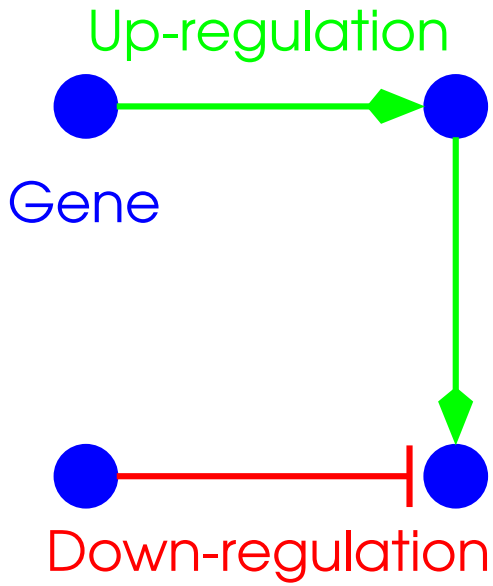


S. Cerevisiae PPI network

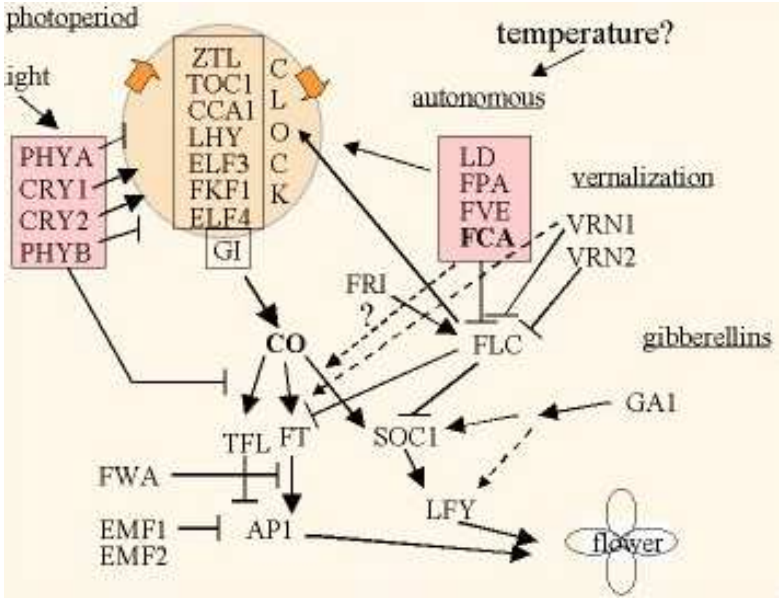
(Jeong et al., *Nature*, 2001)

Gene Regulatory Networks

- Expression of genes is dynamically orchestrated through genes controlling each other's transcription
  - Computationally induced from gene expression data and/or sequence level analysis



Boolean Network Model

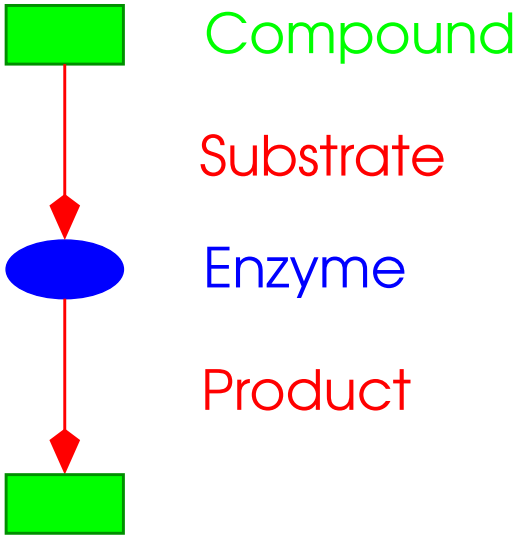


Genetic network that controls flowering time in *A. Thaliania*

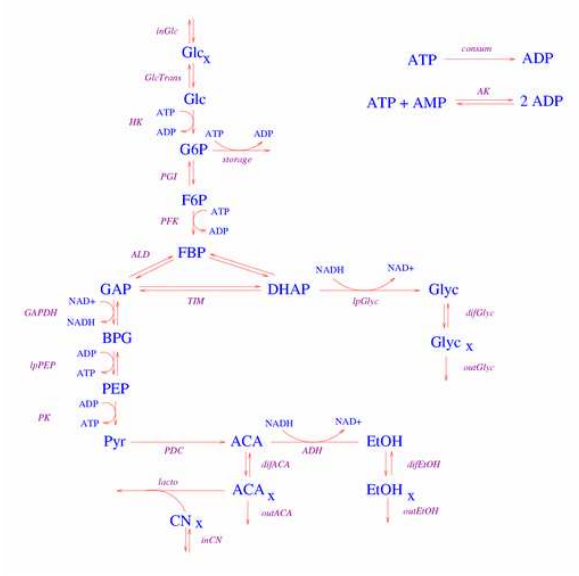
(Blazquez et al, *EMBO Reports*, 2001)

Metabolic Pathways

- Chains of reactions that perform a particular metabolic function
  - Reactions are linked to each other through substrate-product relationships
  - Experimentally derived & computationally extended



Directed Hypergraph Model



Glycolysis pathway in *S. Cerevisiae*
(Hynne et al., *Biophys. Chem.*, 2001)

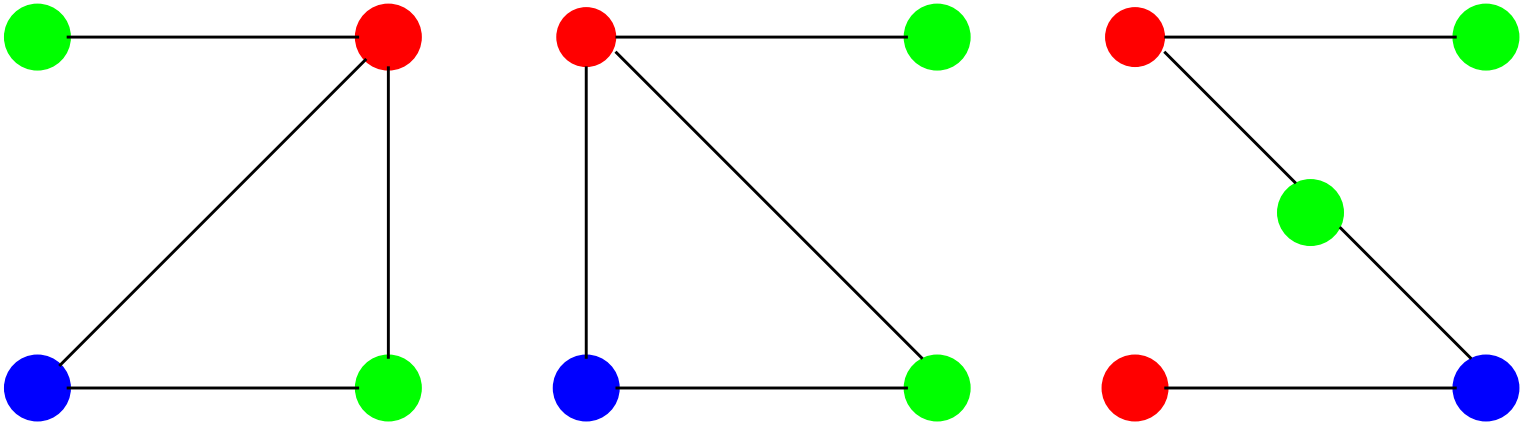
Evolution of Molecular Interactions

- "Evolution thinks modular" (Vespignani, *Nature Gen.*, 2003)
- Cooperative tasks require all participating units
  - Selective pressure on preserving interactions & interacting proteins
  - Interacting proteins follow similar evolutionary trajectories (Pellegrini et al., *PNAS*, 1999)
- Orthologs of interacting proteins are likely to interact (Wagner, *Mol. Bio. Evol.*, 2001)
  - Conservation of interactions may provide clues relating to conservation of function
- Modular conservation and alignment hold the key to critical structural, functional, and evolutionary concepts in systems biology

Conserved Interaction Patterns

- Given a collection of interaction networks (belonging to different species), find sub-networks that are common to an interesting subset of these networks (Koyutürk, Grama, & Szpankowski, *ISMB*, 2004)
  - A sub-network is a group of interactions that are tied to each other (connected)
  - Frequency: The number of networks that contain a sub-network, is a coarse measure of statistical significance
- Computational challenges
  - How to relate molecules (proteins) in different organisms?
  - Requires solution of the intractable subgraph isomorphism problem
  - Must be scalable to potentially large number of networks
  - Networks are large (in the range of $10K$ edges)

Graph Analysis



Network database

Interaction patterns that are common to all networks

Relating Proteins in Different Species

- Ortholog Databases
  - PPI networks: COG, Homologene, Pfam, ADDA
  - Metabolic pathways: Enzyme nomenclature
  - Reliable, but conservative
  - Domain families rely on domain information, but the underlying domains for most interactions are unknown $\Rightarrow$ Multiple node labels
- Sequence Clustering
  - Cluster protein sequences and label proteins according to this clustering
  - Flexible, but expensive and noisy
- Labels may span a large range of functional relationships, from protein families to ortholog groups
  - Without loss of generality, we call identically labeled proteins as orthologs

Problem Statement

- Given a set of proteins V, a set of interactions E, and a many-to-many mapping from V to a set of ortholog groups $\mathcal{L} = \{l_1, l_2, ..., l_n\}$, the corresponding interaction network is a labeled graph $G = (V, E, \mathcal{L})$.
  - $v \in V(G)$ is associated with a set of ortholog groups $L(v) \subseteq \mathcal{L}$.
  - $uv \in E(G)$ represents an interaction between u and v.
- S is a sub-network of G, *i.e.*, $S \sqsubseteq G$ if there is an injective mapping $\phi : V(S) \rightarrow V(G)$ such that for all $v \in V(S)$, $L(v) \subseteq L(\phi(v))$ and for all $uv \in E(S)$, $\phi(u)\phi(v) \in E(G)$.

Computational Problem

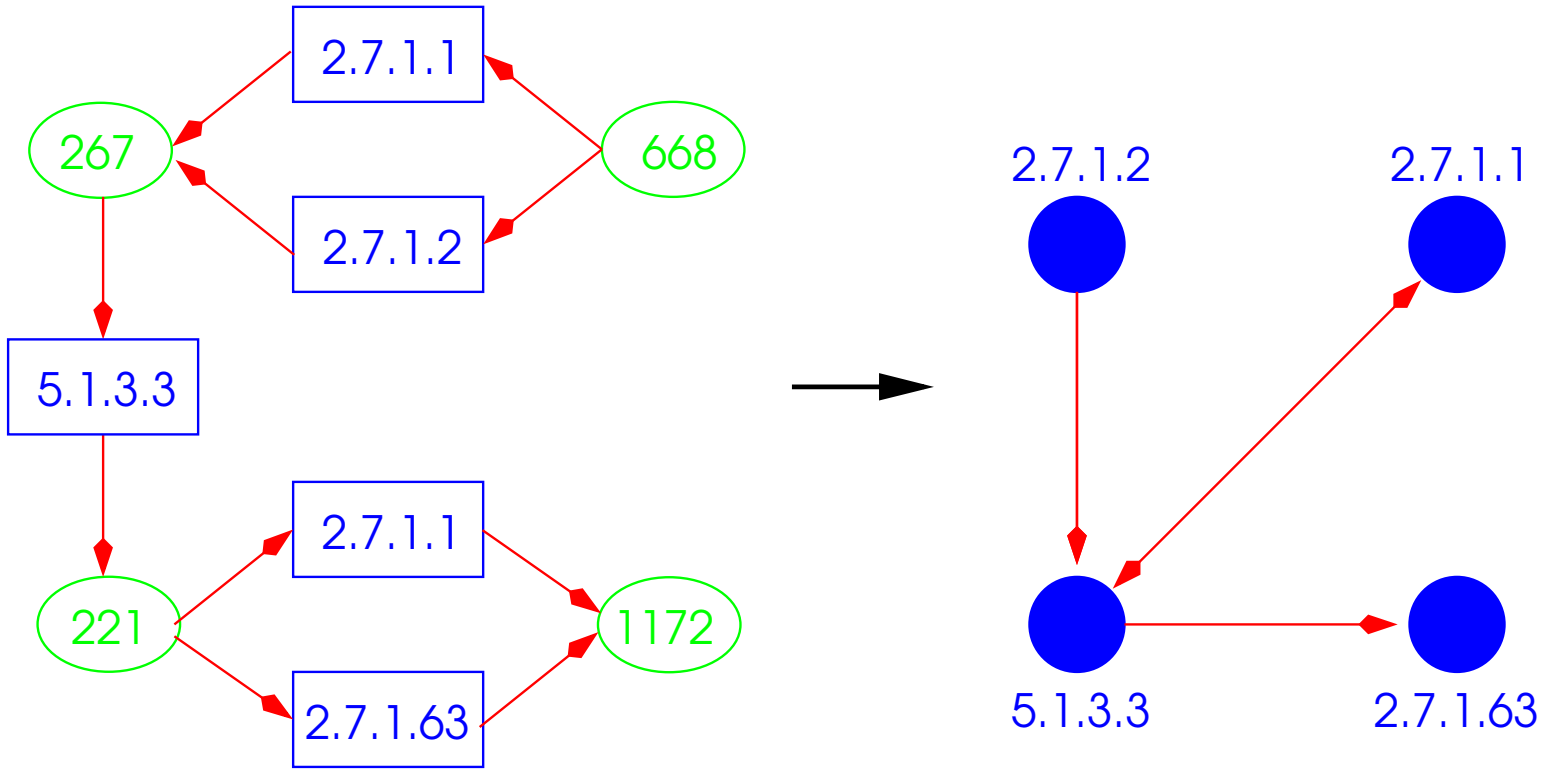
- Conserved sub-network discovery
  - Instance: A set of interaction networks $\mathcal{G} = \{G_1 = (V_1, E_1, \mathcal{L}), G_2 = (V_2, E_2, \mathcal{L}), ..., G_m = (V_m, E_m, \mathcal{L})\}$, each belonging to a different organism, and a frequency threshold σ^*.
  - Problem: Let $H(S) = \{G_i : S \sqsubseteq G_i\}$ be the occurrence set of graph S. Find all connected subgraphs S such that $|H(S)| \geq \sigma^*$, *i.e.*, S is a frequent subgraph in $\mathcal{G}$ and for all $S' \sqsupset S$, $H(S) \neq H(S')$, *i.e.*, S is maximal.

Algorithmic Insight: Ortholog Contraction

- Contract orthologous nodes into a single node
- No subgraph isomorphism
  - Graphs are uniquely identified by their edge sets
- Key observation: Frequent sub-networks are preserved $\Rightarrow$ No information loss
  - Sub-networks that are frequent in general graphs are also frequent in their ortholog-contracted representation
  - Ortholog contraction is a powerful pruning heuristic
- Discovered frequent sub-networks are still biologically interpretable!
  - Interaction between proteins becomes interaction between ortholog groups

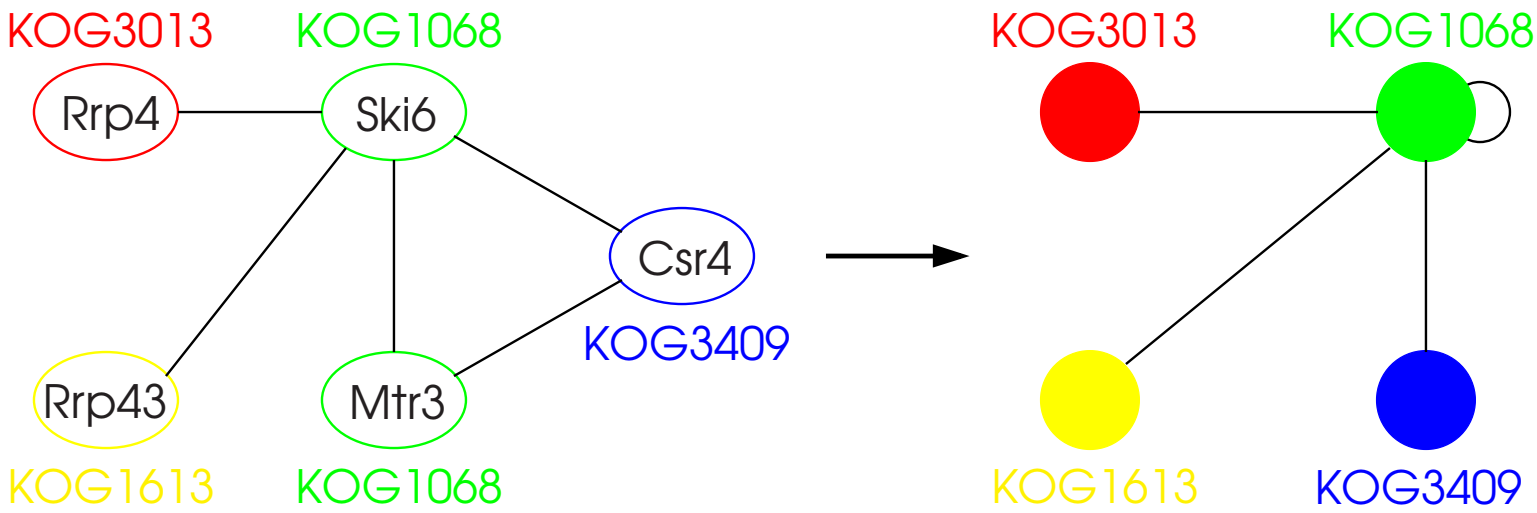
Ortholog Contraction in Metabolic Pathways

- Directed hypergraph → uniquely-labeled directed graph
  - 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



Ortholog Contraction in PPI Networks

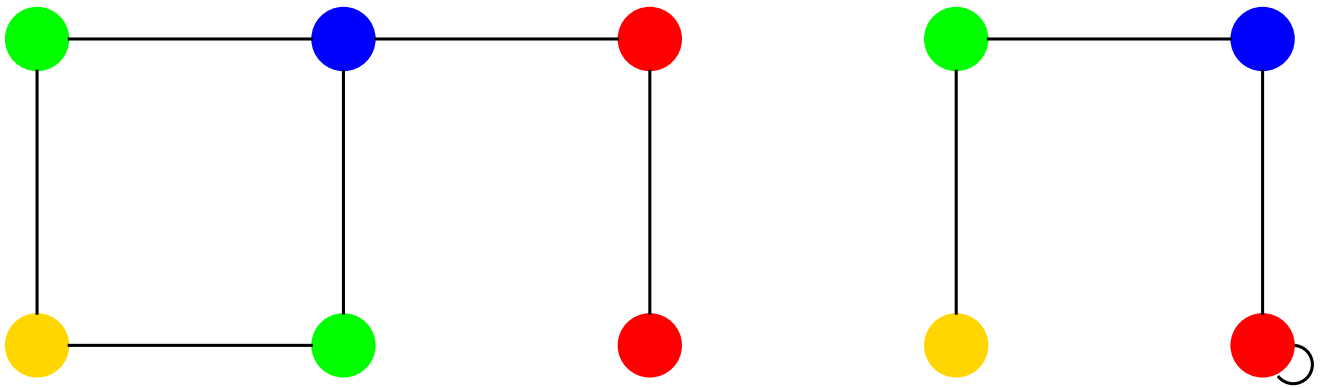
- Interaction between proteins → Interaction between ortholog groups or protein families



Preservation of Sub-networks

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

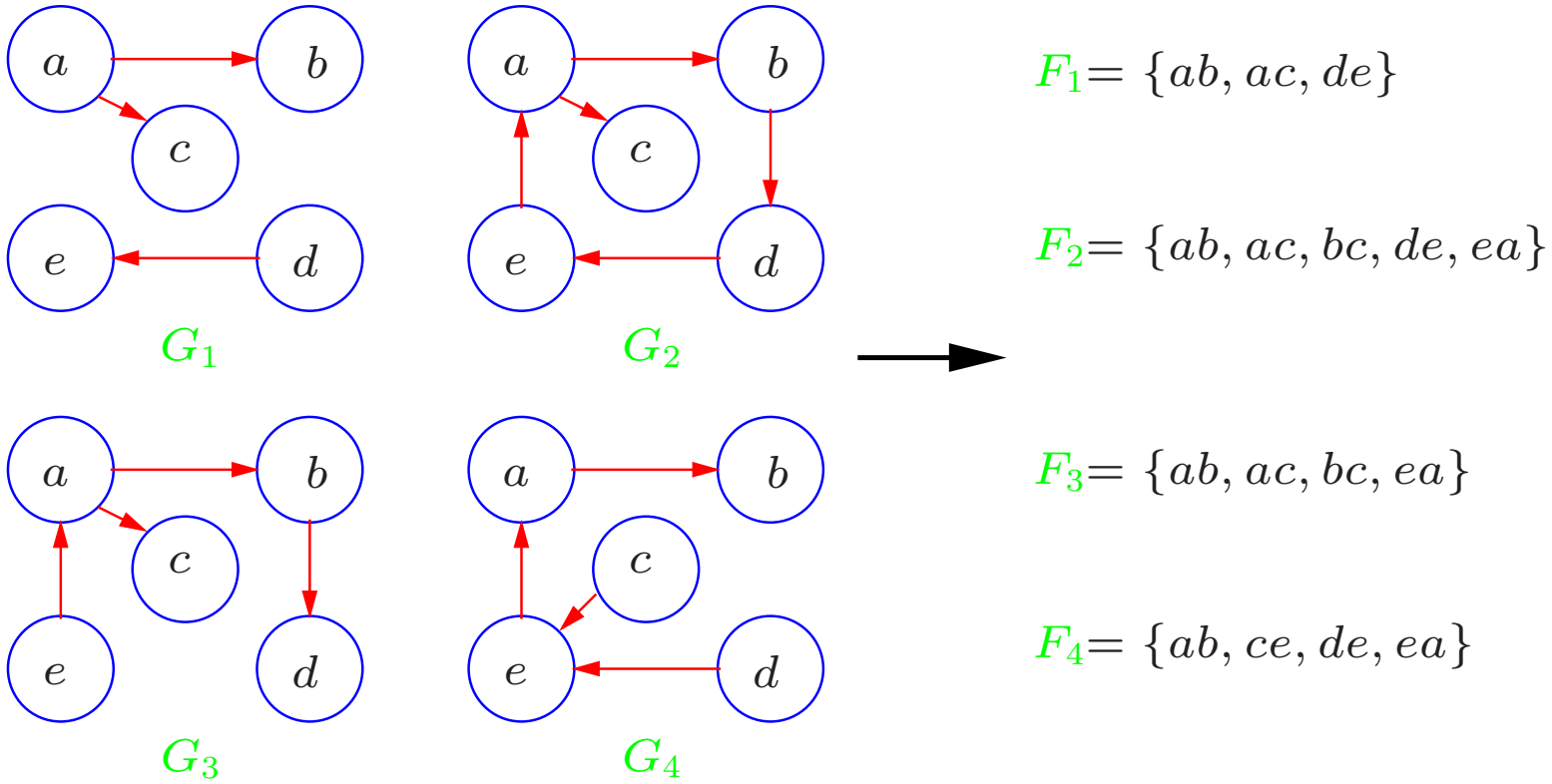
Corollary: The ortholog-contracted representation of any frequent sub-network is also frequent in the set of ortholog-contracted graphs.



G $\tilde{G}$

Simplifying the Graph Analysis Problem

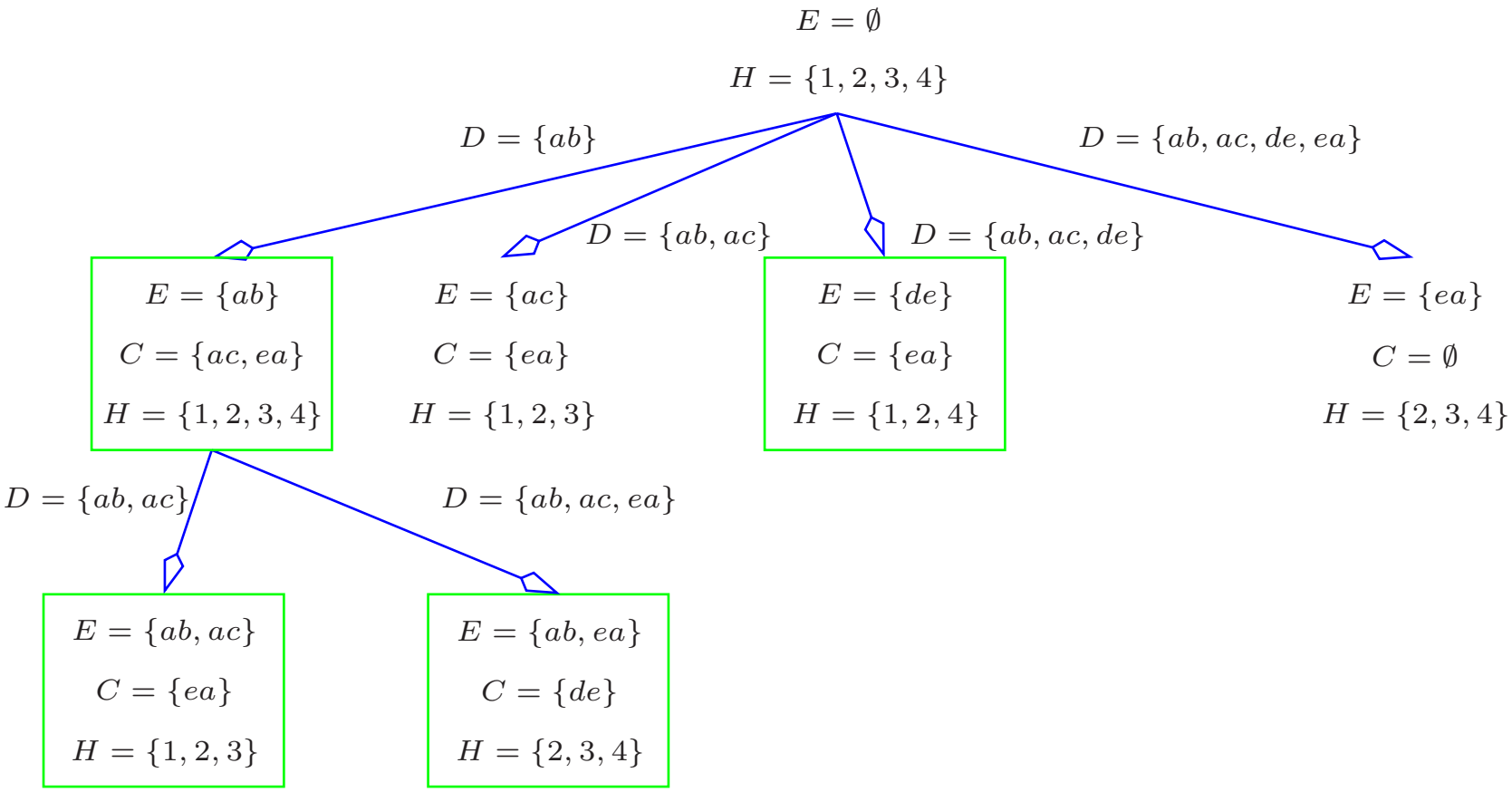
- Observation: An ortholog-contracted graph is uniquely determined by the set of its edges.
  - Conserved Sub-network Discovery Problem → Frequent Edge set Discovery Problem



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
  - An edge set is frequent only if all of its subsets are frequent
  - Generate edge sets (sub-networks) from small to large, pruning supersets of infrequent sets
- No redundancy
- No subgraph enumeration

MULE: Analyzing Ortholog-Contracted Networks

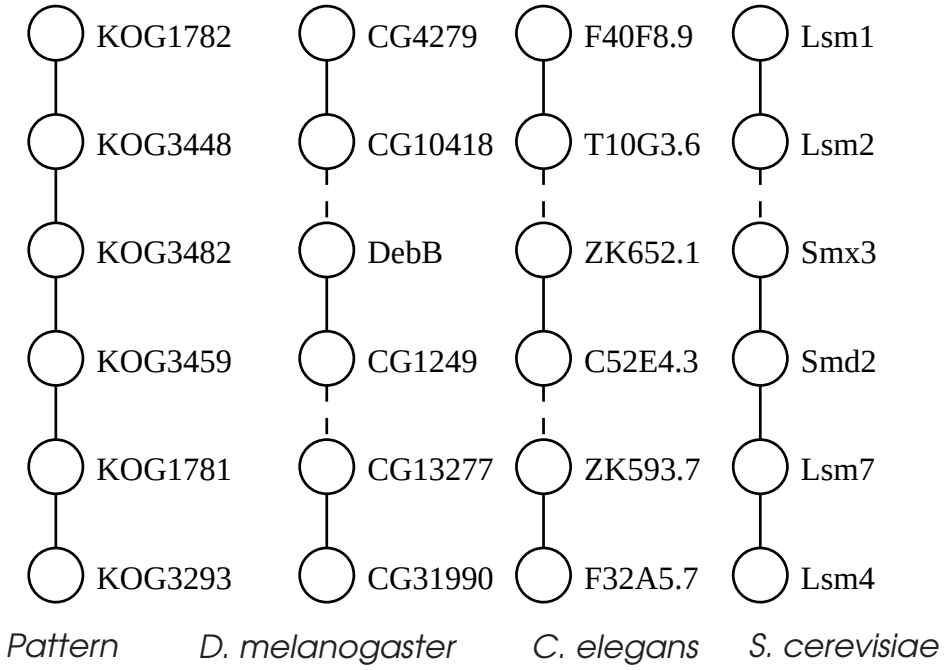


Sample run of MULE for identifying maximal sub-networks that are common to at least 3 organisms

Results: Analyzing PPI Networks

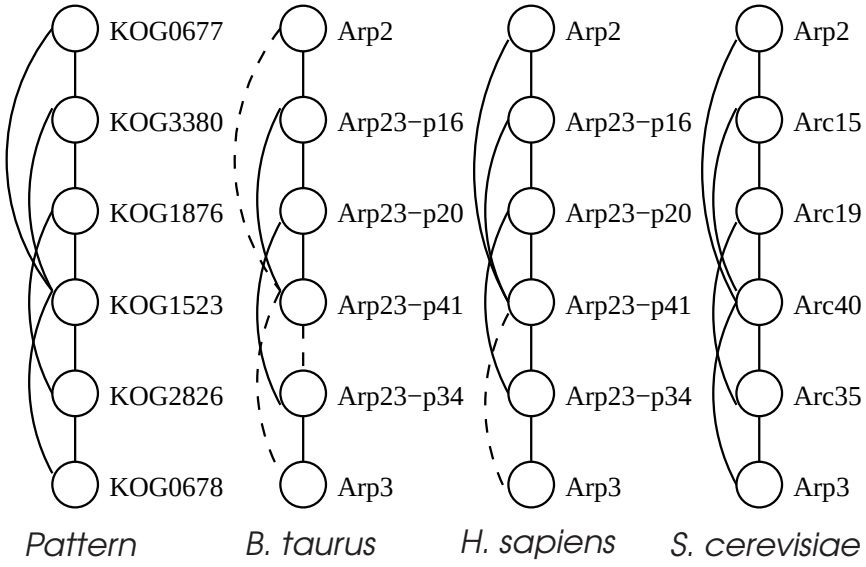
- PPI networks for 9 eukaryotic organisms derived from BIND and DIP
  - *A. thaliania, O. sativa, S. cerevisiae, C. elegans, D. melanogaster, H. sapiens, B. taurus, M. musculus, R. norvegicus*
  - # of proteins ranges from 288 (*Arabidopsis*) to 8577 (*fruit fly*)
  - # of interactions ranges from 340 (*rice*) to 28829 (*fruit fly*)
- Ortholog contraction
  - Group proteins according to existing COG ortholog clusters
  - Merge Homologene groups into COG clusters
  - Cluster remaining proteins via BLASTCLUST
  - Ortholog-contracted *fruit fly* network contains 11088 interactions between 2849 ortholog groups
- MULE is available at `http://www.cs.purdue.edu/pdsl/`

Conserved Protein Interaction Patterns



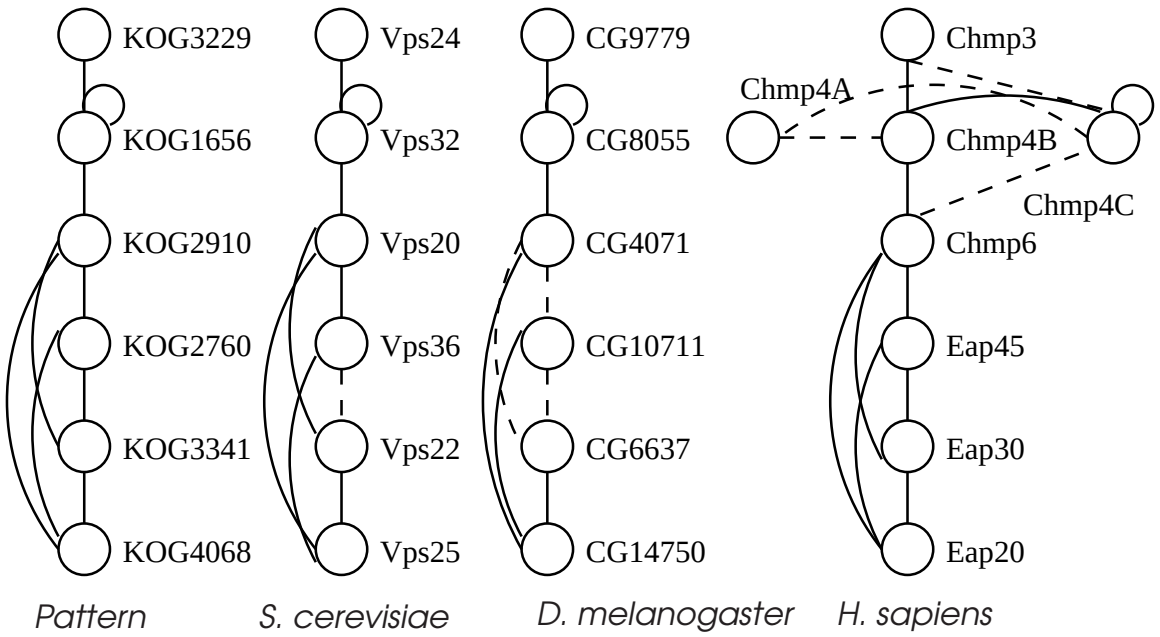
Small nuclear ribonucleoprotein complex ($p < 2e - 43$)

Conserved Protein Interaction Patterns



Actin-related protein Arp2/3 complex ($p < 9e - 11$)

Conserved Protein Interaction Patterns



Endosomal sorting ($p < 1e - 78$)

Runtime Characteristics

Comparison with isomorphism-based algorithms

FSG (Kuramochi & Karypis, *IEEE TKDE*, 2004), gSpan (Yan & Han, *KDD*, 2003)

		FSG			MULE		
Dataset	Minimum Support (%)	Runtime (secs.)	Largest pattern	Number of patterns	Runtime (secs.)	Largest pattern	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 patterns

Glutamate metabolism, $\sigma = 8\%$				Alanine metabolism, $\sigma = 10\%$			
Size of contracted pattern	Extraction time (secs.) FSG	gSpan	Size of extracted pattern	Size of contracted pattern	Extraction time (secs.) FSG	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.			

Discussion

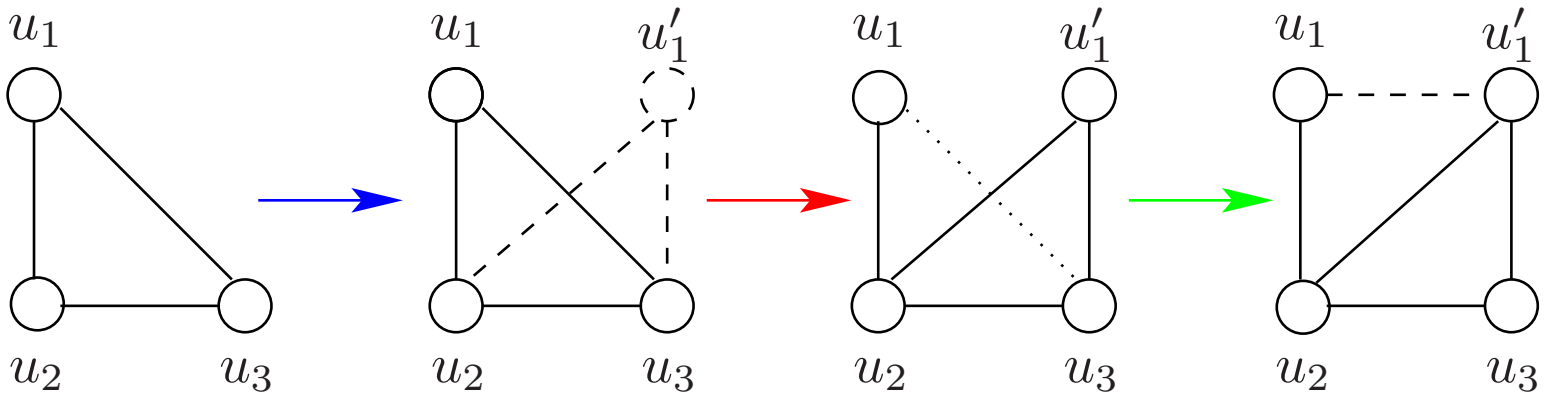
- Ortholog contraction is fast & scalable
  - Graph cartesian product based methods (Sharan et al., *PNAS*, 2004), (Koyutürk et al., *RECOMB*, 2005) create m^n product nodes for an ortholog group that has m proteins in each of n organisms
  - Ortholog contraction represents the same group with only n contracted nodes
  - Isomorphism-based graph analysis algorithms do not scale to large networks
- Ortholog contraction implicitly accounts for noise by eliminating false positives by thresholding frequency, and false negatives by contraction
- Frequency-based approach is not easily extensible to weighted graphs (Zhou et al., *ISMB*, 2005)

Alignment of PPI Networks

- Given two PPI networks that belong to two different organisms, identify sub-networks that are similar to each other
  - Biological implications
  - Mathematical modeling
- Existing algorithms
  - PathBLAST aligns pathways (linear chains) to simplify the problem while maintaining biological meaning (Kelley et al., *PNAS*, 2004)
  - NetworkBLAST compares conserved complex model with null model to identify significantly conserved subnets (Sharan et al., *J. Comp. Biol.*, 2005)
- Our approach (Koyutürk et al., *RECOMB*, 2005) (Koyutürk et al., *J. Comp. Biol.*, 2006)
  - Guided by models of evolution
  - Scores evolutionary events
  - Identifies sets of proteins that induce high-scoring sub-network pairs

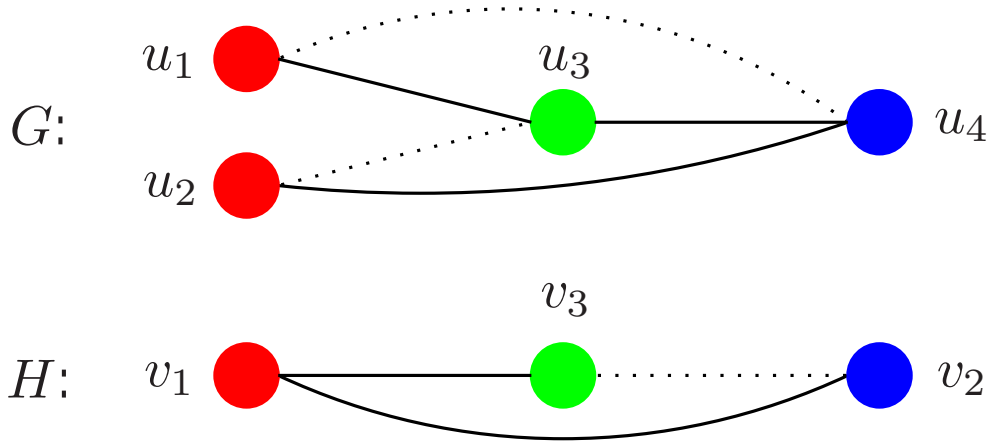
Evolution of PPI Networks

- Duplication/divergence models for the evolution of protein interaction networks
  - Interactions of duplicated proteins are also duplicated
  - Duplicated proteins rapidly lose interactions through mutations
- Allows defining and scoring evolutionary events as graph-theoretical concepts



Match, Mismatch, and Duplication

- Evolutionary events as graph-theoretic concepts
  - A match $\in \mathcal{M}$ corresponds to two pairs of homolog proteins from each organism such that both pairs interact in both PPIs. A match is associated with score μ.
  - A mismatch $\in \mathcal{N}$ corresponds to two pairs of homolog proteins from each organism 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 organism. A duplication is associated with score δ.



Scoring Matches, Mismatches and Duplications

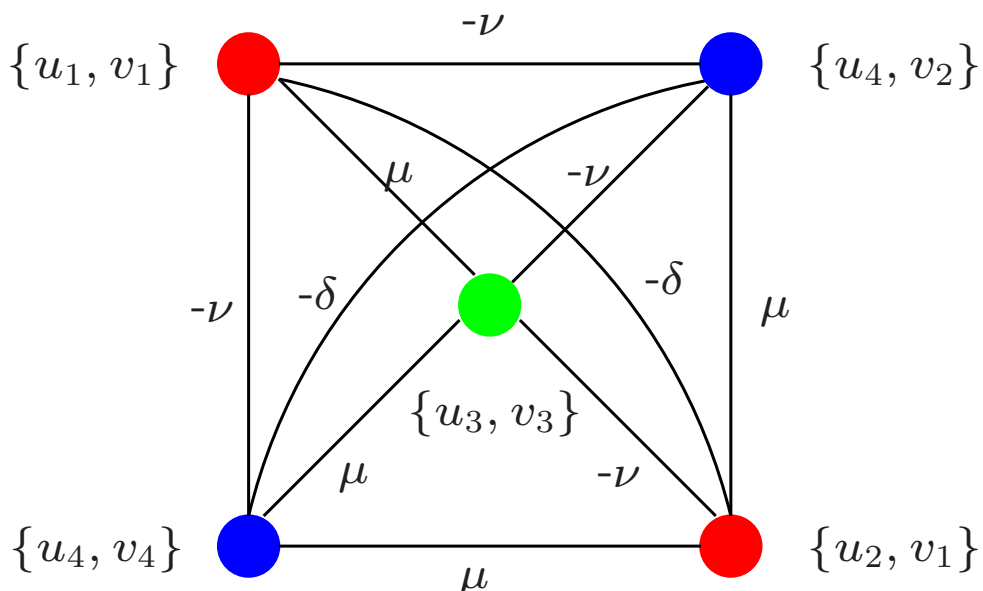
- Quantizing similarity between two proteins
  - Confidence in two proteins being orthologous
  - BLAST E-value: $S(u,v) = log_{10}\frac{p(u,v)}{p_{random}}$
  - Ortholog clustering: $S(u,v) = c(u)c(v)$
- Match score
  - $\mu(uu', vv') = \bar{\mu}\min\{S(u,v), S(u',v')\}$
- Mismatch penalty
  - $\nu(uu', vv') = \bar{\nu}\min\{S(u,v), S(u',v')\}$
- Duplication score
  - $\delta(u,u') = \bar{\delta}(\hat{\delta} - S(u,u'))$
  - $\hat{\delta}$ specifies threshold for sequence similarity to be considered functionally conserved

Pairwise Alignment of PPIs as an Optimization Problem

- 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 rewarded/penalized for functional conservation/differentiation after split
  - Mismatches are penalized for functional divergence (what about experimental error?)
- Scores are functions of similarity between associated proteins
- Problem: Find all protein subset pairs with significant alignment score
  - High scoring protein subsets are likely to correspond to conserved modules
- A graph equivalent to BLAST

Weighted Alignment Graph

- $\mathbf{G(V, E)}$: $\mathbf{V}$ consists of all pairs of homolog 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

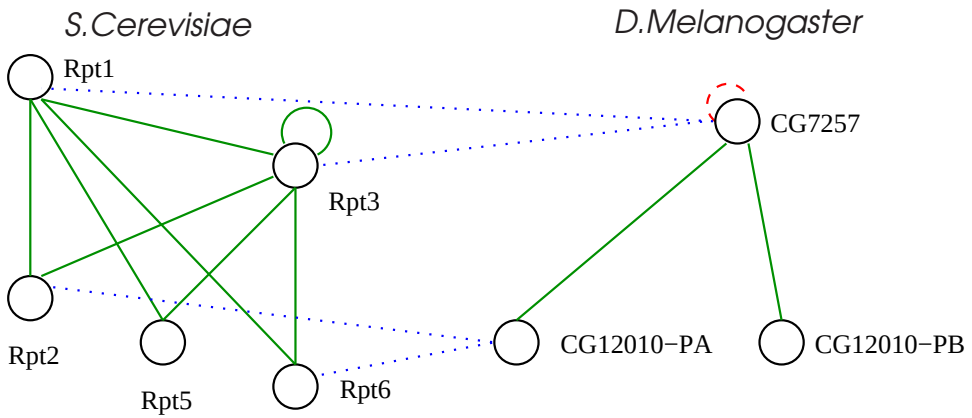
- Definition: (MAWISH)
  - Given graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and a constant ϵ, find $\tilde{\mathcal{V}} \in \mathcal{V}$ such that $\sum_{\mathbf{v},\mathbf{u}\in\tilde{\mathcal{V}}} w(\mathbf{vu}) \geq \epsilon$.
  - NP-complete by reduction from Maximum-Clique
- Theorem: (MAWISH $\equiv$ Pairwise alignment)
  - If $\tilde{\mathcal{V}}$ is a solution for the MAWISH problem on $\mathcal{G}(\mathcal{V}, \mathcal{E})$, then $P = \{\tilde{U}, \tilde{V}\}$ induces an alignment $\mathcal{A}(P)$ with $\sigma(\mathcal{A}) \geq \epsilon$, where $\tilde{\mathcal{V}} = \tilde{U} \times \tilde{V}$.
- Solution: Local graph expansion
  - Greedy graph growing + iterative refinement
  - Linear-time heuristic
- Source code available at `http://www.cs.purdue.edu/pdsl/`

Alignment of Yeast and Fruit Fly PPI Networks

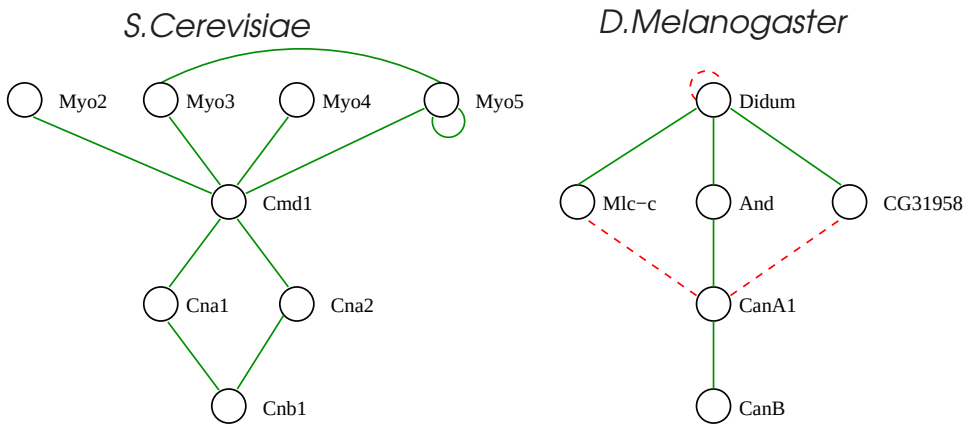
Rank	Score	z-score	# Proteins	# Matches	# Mismatches	# Dups.
1	15.97	6.6	18 (16, 5)	28	6	(4, 0)
	protein amino acid phosphorylation (69%)					
	JAK-STAT cascade (40%)					
2	13.93	3.7	13 (8, 7)	25	7	(3, 1)
	endocytosis (50%) / calcium-mediated signaling (50%)					
5	8.22	13.5	9 (5, 3)	19	11	(1, 0)
	invasive growth (sensu Saccharomyces) (100%)					
	oxygen and reactive oxygen species metabolism (33%)					
6	8.05	7.6	8 (5, 3)	12	2	(0, 1)
	ubiquitin-dependent protein catabolism (100%)					
	mitosis (67%)					
21	4.36	6.2	9 (5, 4)	18	13	(0, 5)
	cytokinesis (100%, 50%)					
30	3.76	39.6	6 (3, 5)	5	1	(0, 6)
	DNA replication initiation (100%, 80%)					

Subnets Conserved in Yeast and Fruit Fly

Proteosome regulatory particle subnet



Calcium-dependent stress-activated signaling pathway



Discussion

- Comparison to other approaches: NetworkBlast (Sharan et al., *PNAS*, 2005), NUKE (Novak et al., *Genome Informatics*, 2005)
  - Much faster than NetworkBLAST, but provides less coverage
  - Comparable to NUKE depending on speed vs coverage trade-off
- Scores evolutionary events
  - Flexible, allows incorporation of different evolutionary models, experimental bases, target structures
  - Somewhat ad-hoc, what is a good weighting of scores?

Analytical Assessment of Statistical Significance

- What is the significance of a dense component in a network?
- What is the significance of a conserved component in multiple networks?
- Existing techniques
  - Mostly computational (*e.g.*, Monte-Carlo simulations)
  - Compute probability that the pattern exists rather than a pattern with the property (*e.g.*, size, density) exists
  - Overestimation of significance

Random Graph Models

- Interaction networks generally exhibit power-law property (or exponential, geometric, etc.)
- Analysis simplified through independence assumption (Itzkovitz et al., *Physical Review*, 2003)
- Independence assumption may cause problems for networks with arbitrary degree distribution
- $P(uv \in E) = d_u d_v / |E|$, where d_u is expected degree of u, but generally $d_{\max}^2 > |E|$ for PPI networks
- Analytical techniques based on simplified models (Koyutürk, Grama, Szpankowski, *RECOMB*, 2006)
  - Rigorous analysis on $G(n, p)$ model
  - Extension to piecewise $G(n, p)$ to capture network characteristics more accurately

Significance of Dense Subgraphs

- A subnet of r proteins is said to be ρ-dense if $F(r) \geq \rho r^2$, where $F(r)$ is the number of interactions between these r proteins
- What is the expected size of the largest ρ-dense subgraph in a random graph?
  - Any ρ-dense subgraph with larger size is statistically significant!
- $G(n, p)$ model
  - n proteins, each interaction occurs with probability p
  - Simple enough to facilitate rigorous analysis
  - If we let $p = d_{\max}/n$, largest ρ-dense subgraph in $G(n, p)$ stochastically dominates that in a graph with arbitrary degree distribution
- Piecewise $G(n, p)$ model
  - Few proteins with many interacting partners, many proteins with few interacting partners
  - Captures the basic characteristics of PPI networks
  - Analysis of $G(n, p)$ model immediately generalized to this model

Largest Dense Subgraph

- Theorem: If G is a random graph with n nodes, where every edge exists with probability p, then

$$\lim_{n\to\infty} \frac{R_\rho}{\log n} = \frac{1}{\kappa(p,\rho)} \qquad (pr.), \tag{1}$$

where

$$\kappa(p,\rho) = \rho \log \frac{\rho}{p} + (1-\rho)\log \frac{1-\rho}{1-p}. \tag{2}$$

More precisely,

$$P(R_\rho \geq r_0) \leq O\left(\frac{\log n}{n^{1/\kappa(p,\rho)}}\right), \tag{3}$$

where

$$r_0 = \frac{\log n - \log\log n + \log \kappa(p,\rho)}{\kappa(p,\rho)} \tag{4}$$

for large n.

Piecewise $G(n,p)$ model

- The size of largest dense subgraph is still proportional to $\log n/\kappa$ with a constant factor depending on number of hubs

- Model:

$$P(uv \in E(G)) = \begin{cases} p_h & \text{if } u,v \in V_h \\ p_l & \text{if } u,v \in V_l \\ p_b & \text{if } u \in V_h, v \in V_l \text{ or } u \in V_l, v \in V_h \end{cases}$$

- Result:
  Let $n_h = |V_h|$. If $n_h = O(1)$, then $P(R_n(\rho) \geq r_1) \leq O\left(\frac{\log n}{n^{1/\kappa(p_l,\rho)}}\right)$, where

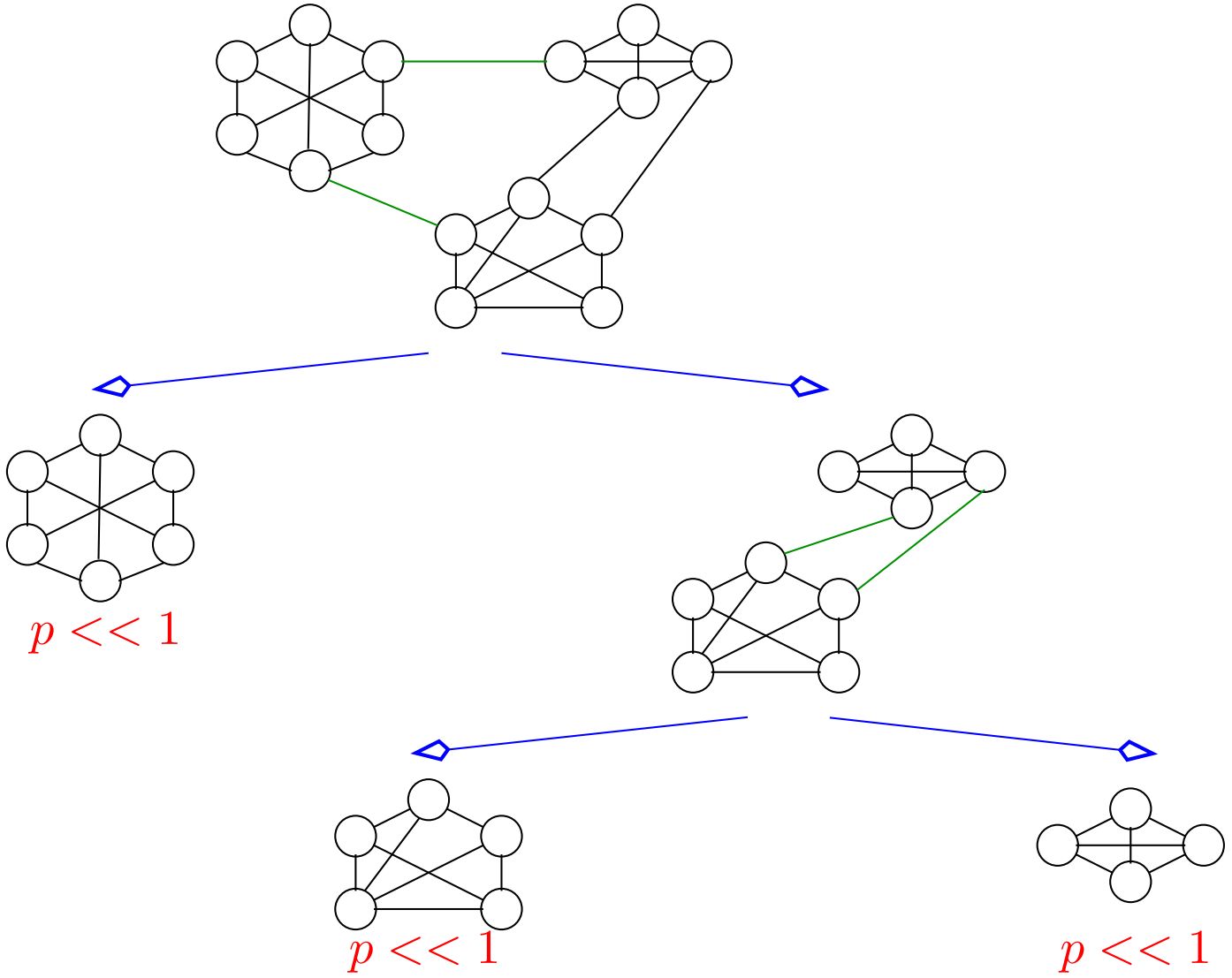
$$r_1 = \frac{\log n - \log\log n + 2n_h \log B + \log \kappa(p_l,\rho) - \log e + 1}{\kappa(p_l,\rho)}$$

  and $B = \frac{p_b q_l}{p_l} + q_b$, where $q_b = 1 - p_b$ and $q_l = 1 - p_l$.

Algorithms Based on Statistical Significance

- Identification of topological modules
- Use statistical significance as a stopping criterion for graph clustering heuristics
- HCS Algorithm (Hartuv & Shamir, *Inf. Proc. Let.*, 2000)
  - Find a minimum-cut bipartitioning of the network
  - If any of the parts is dense enough, record it as a dense cluster of proteins
  - Else, further partition them recursively
- SIDES: Use statistical significance to determine whether a subgraph is sufficiently dense
  - For given number of proteins and interactions between them, we can determine whether those proteins induce a significantly dense subnet

SIDES **Algorithm**



SIDES is available at `http://www.cs.purdue.edu/pdsl`

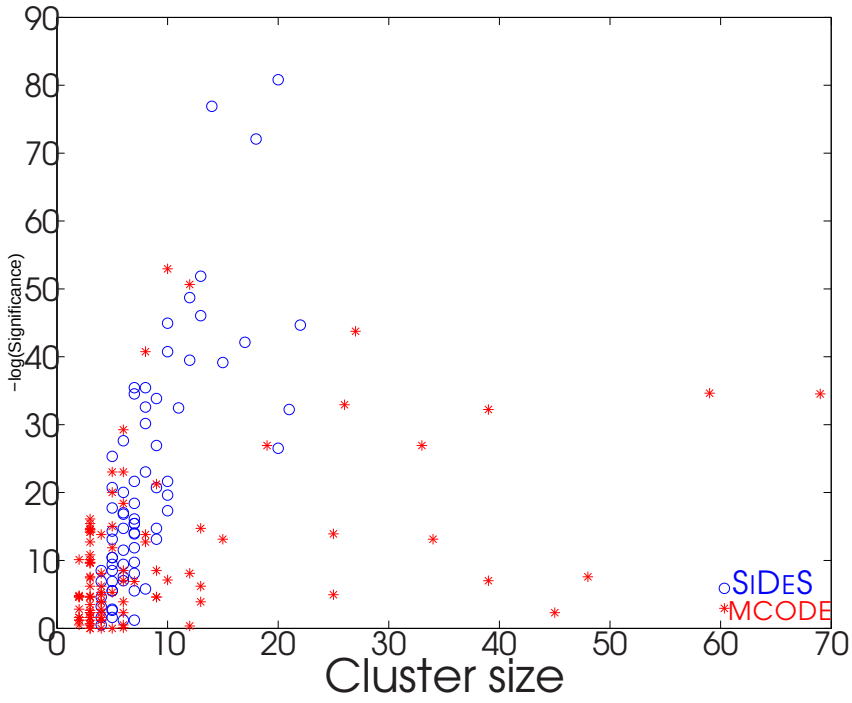
Performance of SIDES

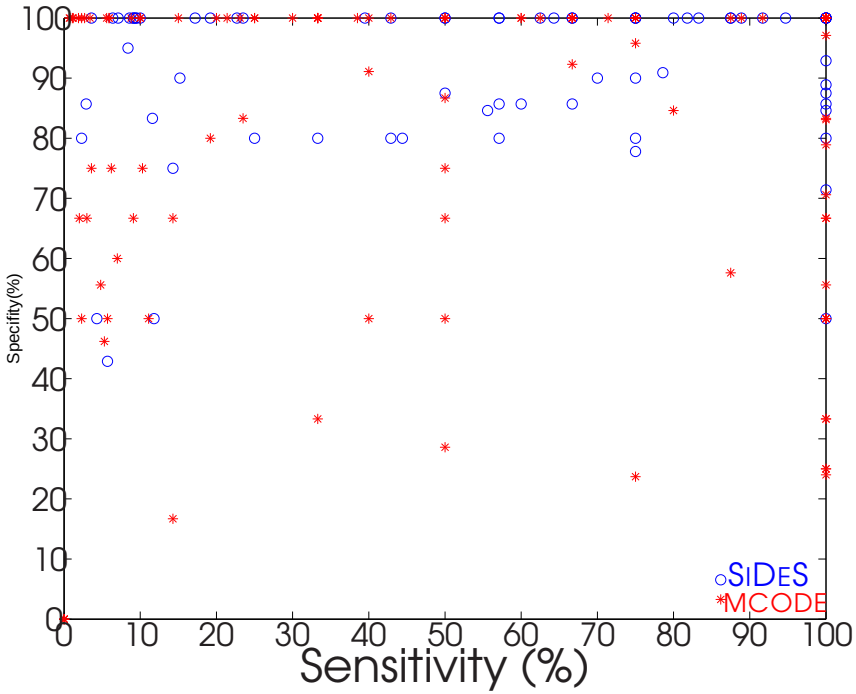
- Biological relevance of identified clusters is assessed with respect to Gene Ontology (GO)
  - Estimate the statistical significance of the enrichment of each GO term in the cluster
- Quality of the clusters with respect to GO annotations
  - Assume cluster C containing n_C genes is associated with term T that is attached to n_T genes and n_{CT} of genes in C are attached to T
  - specificity = $100 \times n_{CT}/n_C$
  - sensitivity = $100 \times n_{CT}/n_T$

	SIDES			MCODE		
	Min.	Max.	Avg.	Min.	Max.	Avg.
Specificity (%)	43.0	100.0	91.2	0.0	100.0	77.8
Sensitivity (%)	2.0	100.0	55.8	0.0	100.0	47.6

Comparison of SIDES with MCODE (Bader & Hogue, *BMC Bioinformatics*, 2003)

Performance of SIDES





Size vs Significance

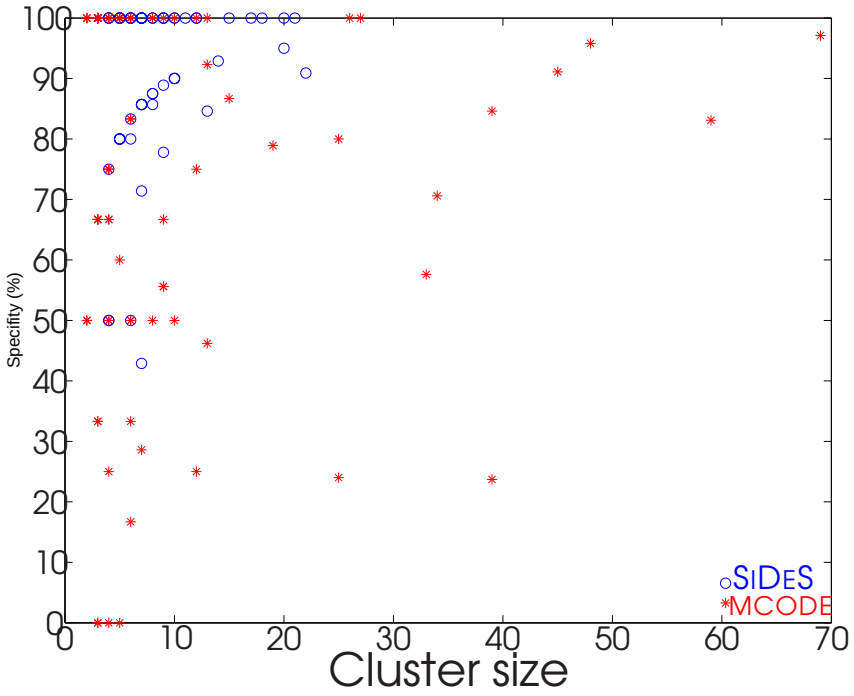
Sensitivity vs Specifity

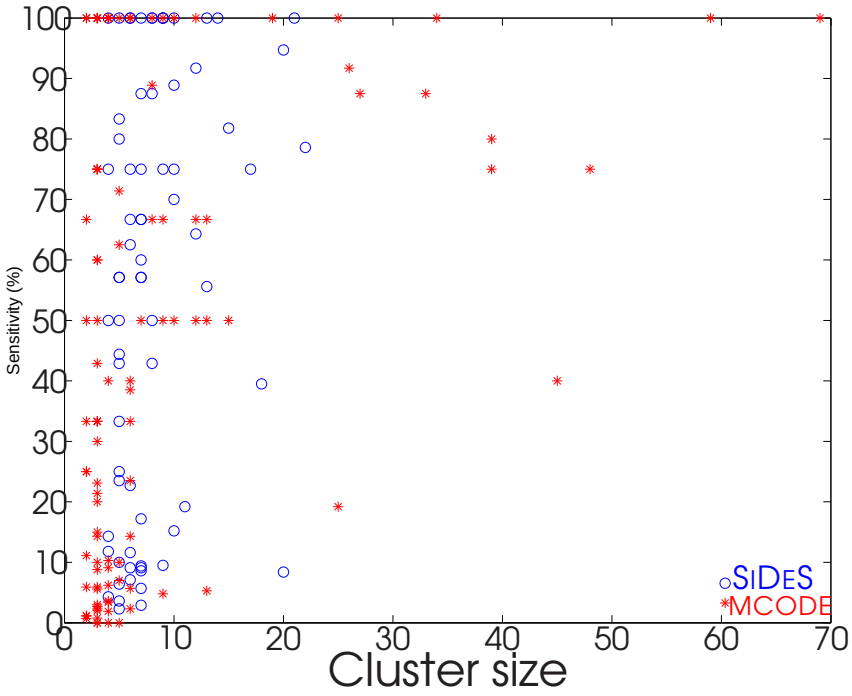
Correlation

SIDES: 0.76

MCODE: 0.43

Performance of SIDES





Size vs Specificity

Size vs Sensitivity

Correlation

SIDES: 0.22
MCODE: -0.02

SIDES: 0.27
MCODE: 0.36

Ongoing Work

- Domain identification through interactions.
- Statistically overepresented subnets (network motifs).
- Phylogenies of networks.
- Network assembly.

References

- Mehmet Koyutürk, Wojciech Szpankowski, and Ananth Grama, Assessing Significance of Connectivity and Conservation in Protein Interaction Networks, *Journal of Computational Biology* (in press).

- Mehmet Koyuturk, Yohan Kim, Shankar Subramaniam, Wojciech Szpankowski, and Ananth Grama, "Detecting conserved interaction patterns in biological networks", *Journal of Computational Biology*, 13(7), 1299-1322, 2006.

- Mehmet Koyuturk, Yohan Kim, Umut Topkara, Shankar Subramaniam, Wojciech Szpankowski, and Ananth Grama, Pairwise Alignment of Protein Interaction Networks, *Journal of Computational Biology*, 13(2), 182-199, 2006.

- Yohan Kim, Mehmet Koyuturk, Umut Topkara, Ananth Grama, and Shankar Subramaniam, Inferring Functional Information from Domain Co-evolution, *Bioinformatics*, 22(1), pp. 40-49, 2006.

References

- Mehmet Koyutürk, Ananth Grama, and Wojciech Szpankowski, An Efficient Algorithm for Detecting Frequent Subgraphs in Biological Networks, *Bioinformatics*, Vol. 20, Suppl. 1, pp i200-i207, 2004.

- Mehmet Koyuturk, Ananth Grama, and Wojciech Szpankowski, Assessing Significance of Connectivity and Conservation in Protein Interaction Networks, *10th International Conference on Research in Computational Molecular Biology (RECOMB)*, LNBI 3909, pp. 45-59, 2006.

- Mehmet Koyutürk, Ananth Grama and Wojciech Szpankowski, Pairwise Local Alignment of Protein Interaction Networks Guided by Models of Evolution, RECOMB 2005.

Available from `http://www.cs.purdue.edu/pdsl`

Thank you!

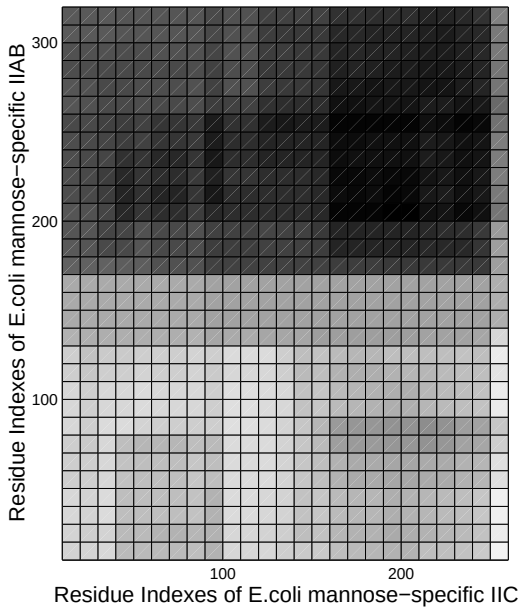
Phylogenetic Analysis for Predicting Interactions

- 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., *PNAS*, 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

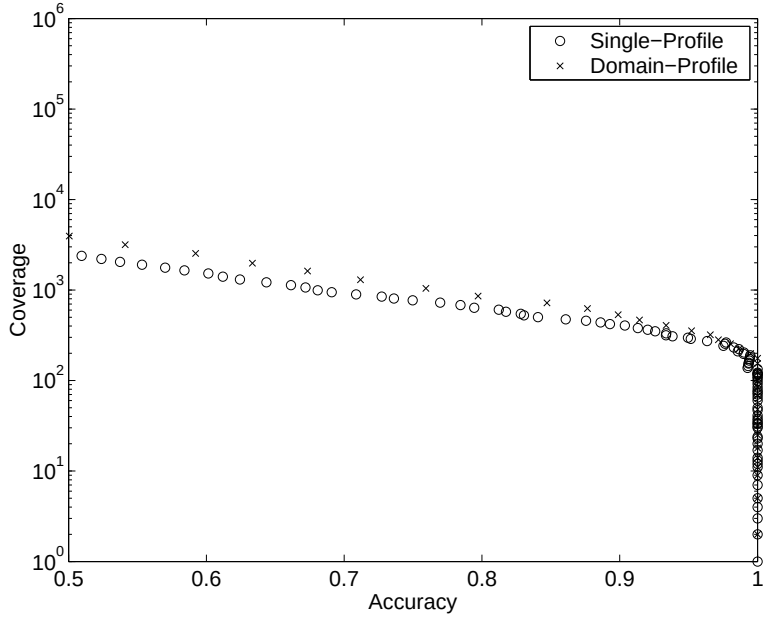
- However, domains, not proteins, co-evolve
  - How can we incorporate domain information to enhance performance of phylogeny-based interaction prediction?

Inferring Function from Domain Co-evolution

- Residue-level phylogenetic analysis (Kim, Koyutürk, Topkara, Grama, & Subramaniam, *Bioinformatics*, 2006)
  - No a-priori knowledge about domains
  - Construct residue phylogenetic profiles from local alignment results
  - Construct mutual information matrix
  - High-information contiguous submatrices that are sufficiently large correspond to putative co-evolved domains



Mutual Information Matrix



Single vs. Domain Profile

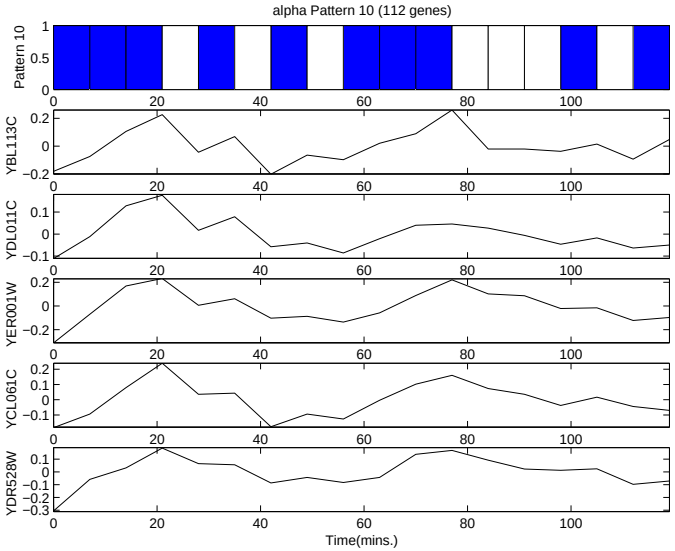
Conclusion

- A lot to learn from comparative network analysis
- We have fast algorithms
- We need
  - Enhanced & more detailed network models
  - High quality & comprehensive data
  - Detailed statistical models

Other Work on Pattern Identification

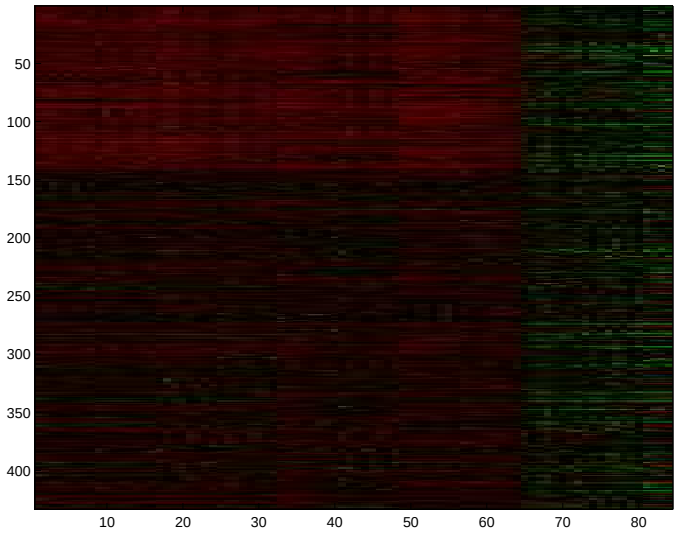
- PROXIMUS: Non-orthogonal decomposition of binary matrices
  (Koyutürk, Grama, Ramakrishnan, *IEEE TKDE*, 2005)
  - Find a compact set of vectors that represent the entire matrix
  - Recursive decomposition through rank-one approximations
  - Fast (linear-time) iterative heuristics for computing approximations
  - Source code available at
    `http://www.cs.purdue.edu/homes/pdsl/`

Patterns of regulation



"Algorithms for bounded-error correlation of high dimensional data in microarray experiments"
(Koyutürk, Grama, Szpankowski: *CSB'03*.)

Biclustering

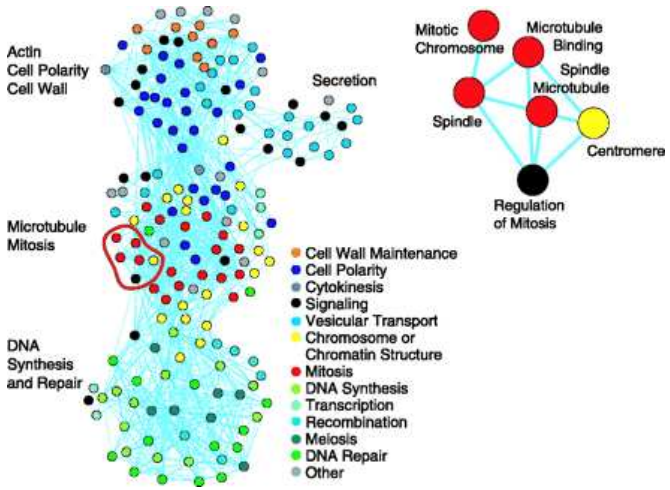


"Biclustering gene-feature matrices for statistically significant dense patterns"
(Koyutürk, Grama, Szpankowski: *CSB'04*.)

Identifying "Canonical" Regulatory Pathways

- Can we derive rules in terms of GO terms, *e.g.*, $P_i \rightarrow P_j \dashv P_k$?
  - Statistical challenge: Such patterns have to be significantly abundant
  - Computational challenge: When statistical significance is the basis (as opposed to frequency), monotonicity properties (*e.g.*, downward closure) no longer hold!
  - Our approach: conditional significance, *i.e.*, evaluate significance of a pattern based on the background constructed by its substructures

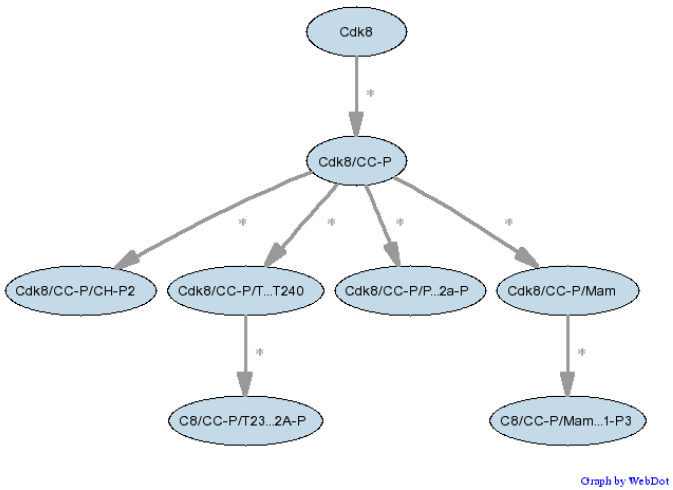
- Final goal: Database of (computationally derived) canonical modules and pathways



A network of GO terms (Tong et al., *Science*, 2004)

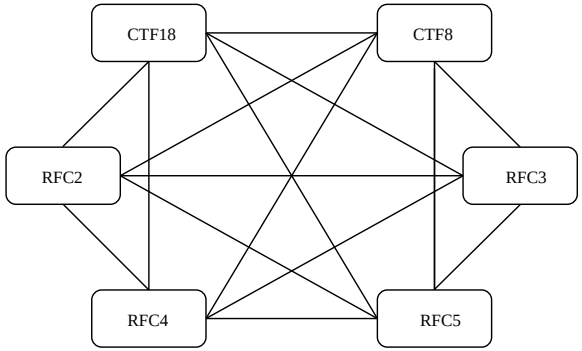
Cell as a State Machine

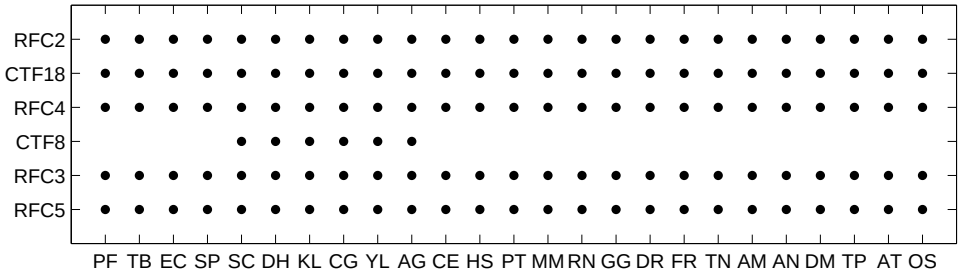
- Signaling pathways can be modeled as a series of transitions between states of protein or peptide molecules, non-protein molecules, (non-)protein complexes, and modules
  - Signaling Gateway provides a database of network states for proteins, a mirror is available to our group via our collaboration with S. Subramaniam
- Constructing signaling pathways from state information for individual molecules
  - Smallest common supergraph problem
  - Identification of specified pathways



State diagram for Cdk8 protein (from Molecule Pages)

Modular Phylogenetics

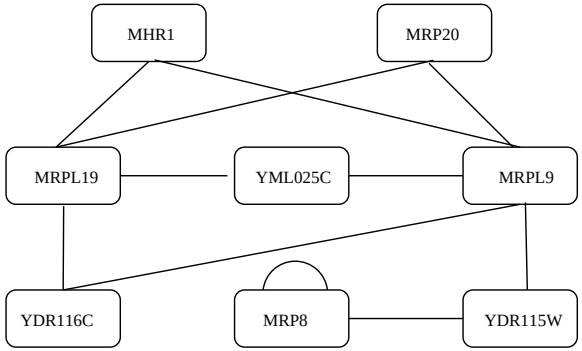


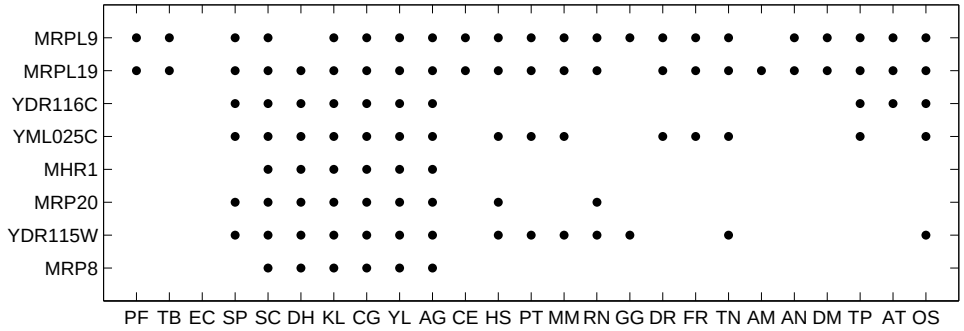


Replication Factor C complex identified on yeast PPI network by MCODE (Bader & Hogue, *BMC Bioinformatics*, 2003) algorithm and the phylogenetic profiles of its proteins on 25 eukaryotic genomes

Conserved in all eukaryotic species!

Modular Phylogenetics





A component of mitochondrial ribosome identified on yeast PPI network by MCODE algorithm and the phylogenetic profiles of its proteins on 25 eukaryotic genomes

Conserved in only yeast species!

- Models and algorithms for quantifying, analyzing, and evaluating modular conservation and divergence across species