Algorithms and Analyses of Molecular Interaction Networks

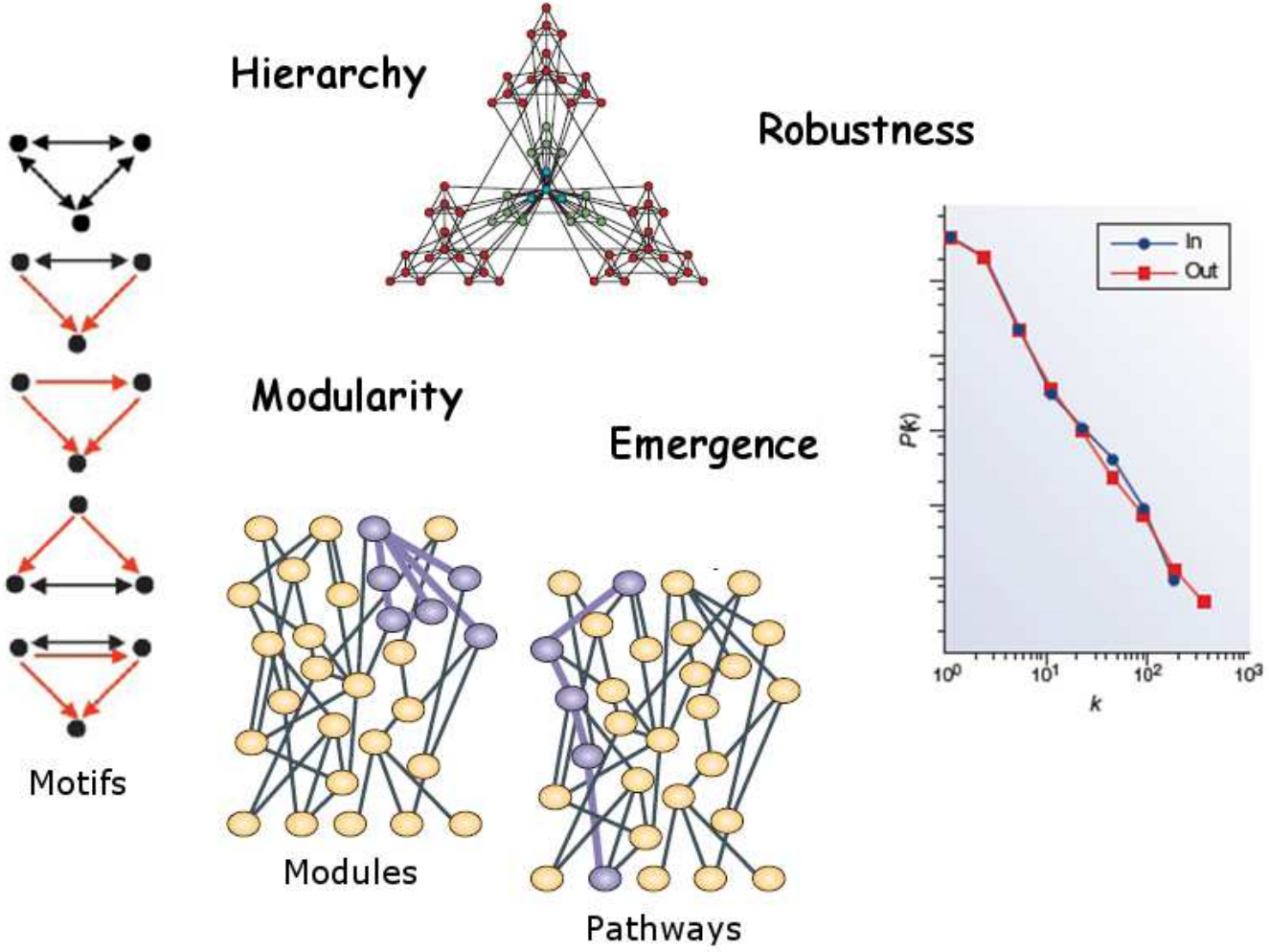
Ananth Grama
Center for Science of Information, Purdue University

With significant contributions from
Mehmet Koyuturk, Jayesh Pandey, Young Kim,
Shahin Mohammadi, and Shankar Subramaniam

Thanks to the Indian Institute of Science and
the US National Science Foundation.

Function & Topology in Molecular Networks

How does function relate to network topology?



Prior Work on Topology and Function

- Conservation (ISMB 04/Bioinf. 04)
- Alignment (RECOMB 05/JCB 06)
- Modularity (RECOMB 06/JCB 07)
- Inference (Bioinf. 06)
- Pathway Annotation (ISMB 07/Bioinf. 07, PSB 08)
- Network Abstractions/ Annotations (ECCB 08/ Bioinf. 08)
- Modularity and Domain Interactions (APBC 10/ BMC Bioinf. 10)
- Pathway Interaction Maps (PSB 12)
- Pathway Inference (ISMB 12)

Evolution of Molecular Interactions

- "Evolution thinks modular" (Vespignani, *Nature Gen.*, 2003)
- Cooperative tasks require all participating units
  - Selective pressure on preserving interactions & interacting proteins
- Proteins organized in cohesive patterns are highly conserved (Wuchty et al., *Nature Gen.*, 2003)
  - Functional modules are likely to be consistently conserved
- Orthologs of interacting proteins are likely to interact (Wagner, *Mol. Bio. Evol.*, 2001)
  - Conservation of interactions may provide clues on conservation of function
- Interacting proteins follow similar evolutionary trajectories (Pellegrini et al., *PNAS*, 1999)

Computational Analysis of Biological Networks

- Clustering
  - Interaction network: Proteins in functional modules densely interact with each other
  - Gene expression: Genes coding cooperating proteins are likely to be co-regulated
  - Phylogenetic profiles: Interacting proteins are likely to have co-evolved
- Graph Mining
  - Common topological motifs and frequent interaction patterns reveal conserved modularity
- Graph Alignment
  - Conservation/divergence of pathways, complexes, and functional modules

Frequent Interaction Patterns: Computational Problem

- 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 $uv \in E(S)$, $\phi(u)\phi(v) \in E(G)$.

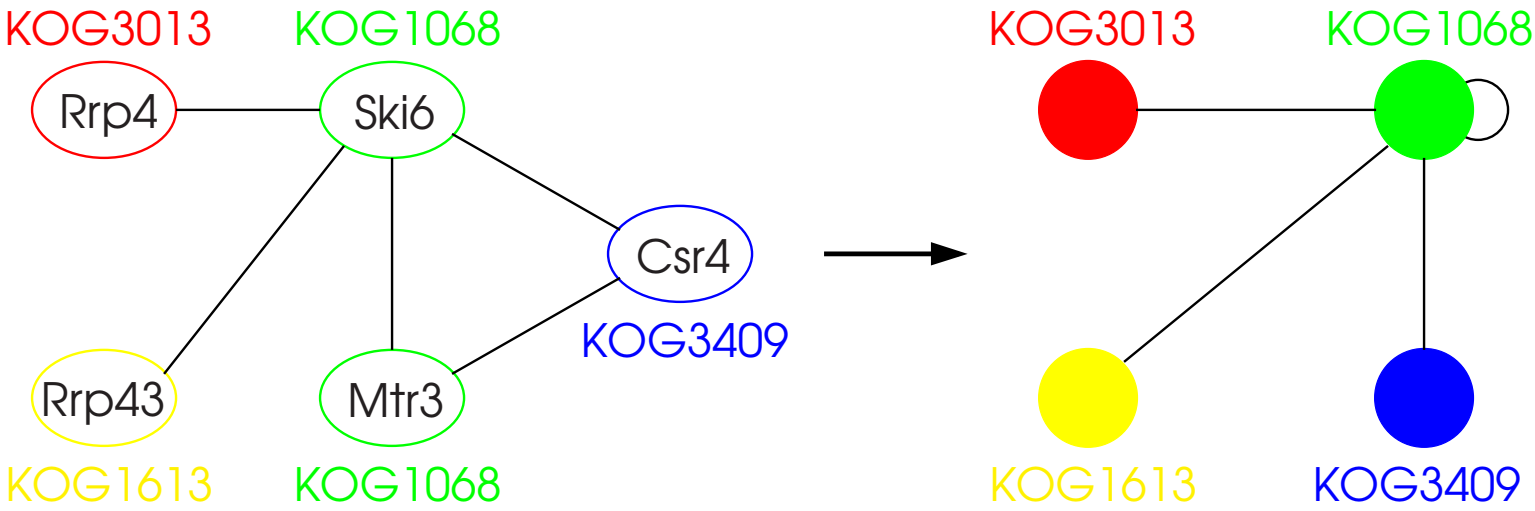
- Maximal frequent sub-network discovery
  - Instance: A set of interaction networks $\mathcal{G} = \{G_1 = (V_1, E_1, \mathcal{L}), G_2 = (V_2, E_2, \mathcal{L}), ..., G_m = (V_m, E_m, \mathcal{L})\}$, each belonging to a different organism, and a frequency threshold σ^*.
  - Problem: Let $H(S) = \{G_i : S \sqsubseteq G_i\}$ 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.

Ortholog Contraction

- Contract orthologous nodes into a single node
- No subgraph isomorphism
  - Graphs are uniquely identified by their edge sets
- 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
- Discovered frequent sub-networks are still biologically interpretable!
  - Interaction between proteins becomes interaction between ortholog groups
  - Ortholog-contraction may be thought of as going back in evolutionary history (to what point?)

Ortholog Contraction in PPI Networks

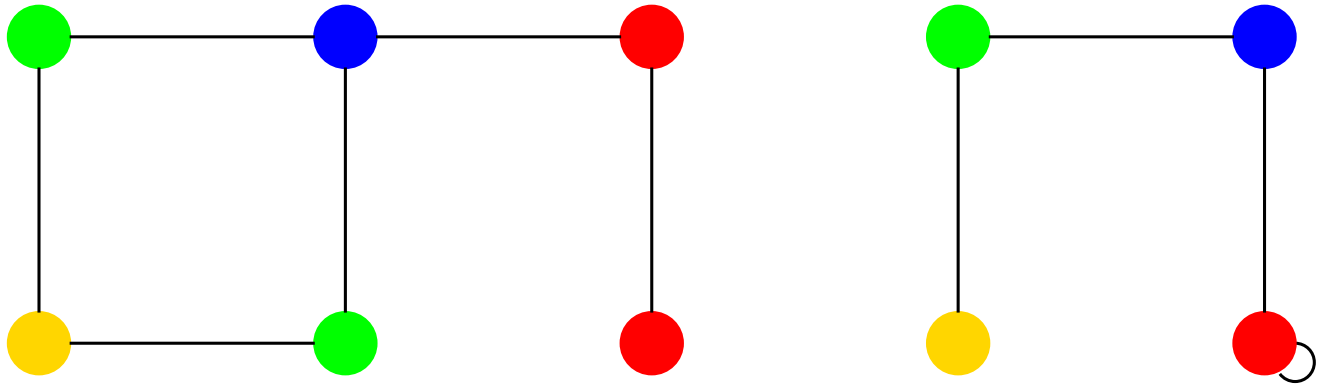
- Interaction between proteins → Interaction between ortholog groups or protein families



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.

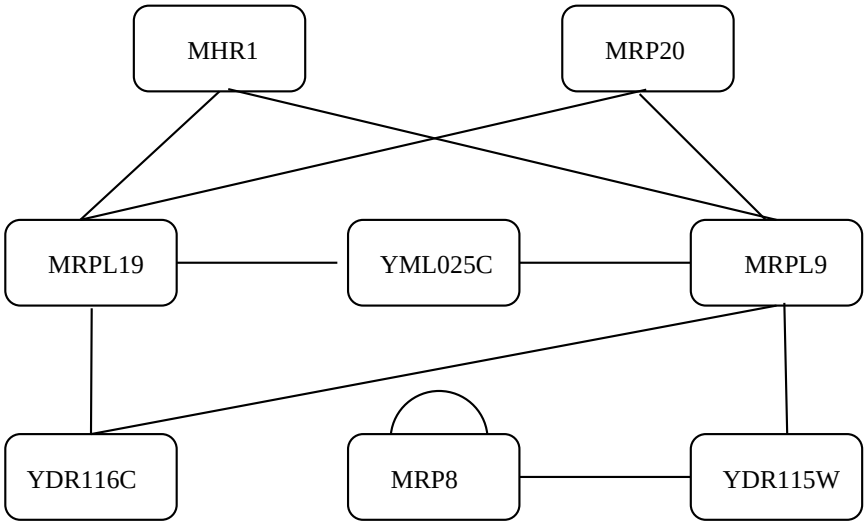


G $\tilde{G}$

Results: Mining PPI Networks

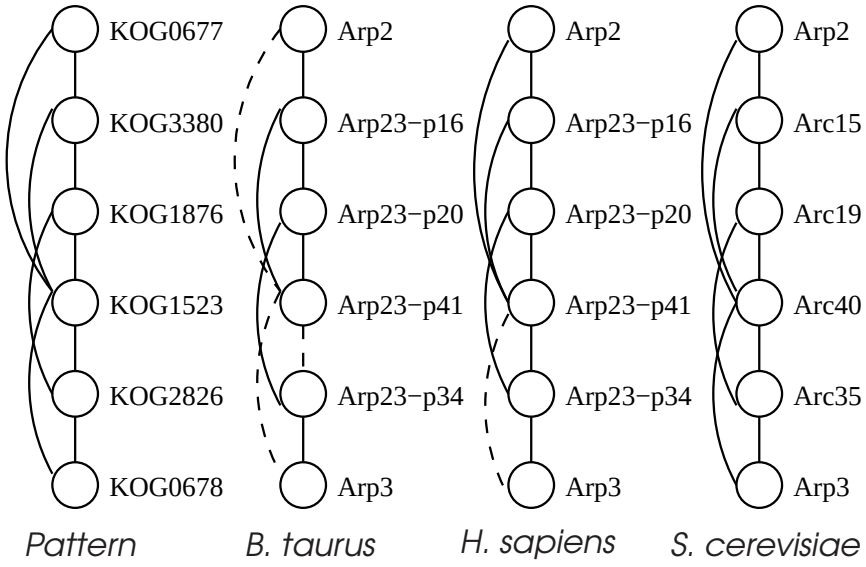
- PPI networks for 9 eukaryotic organisms derived from BIND and DIP
  - *A. thaliania, O. sativa, S. cerevisiae, C. elegans, D. melanogaster, H. sapiens, B. taurus, M. musculus, R. norvegicus*
  - # of proteins ranges from 288 (*Arabidopsis*) to 8577 (*fruit fly*)
  - # of interactions ranges from 340 (*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/homes/koyuturk/mule/`

Frequent Protein Interaction Patterns



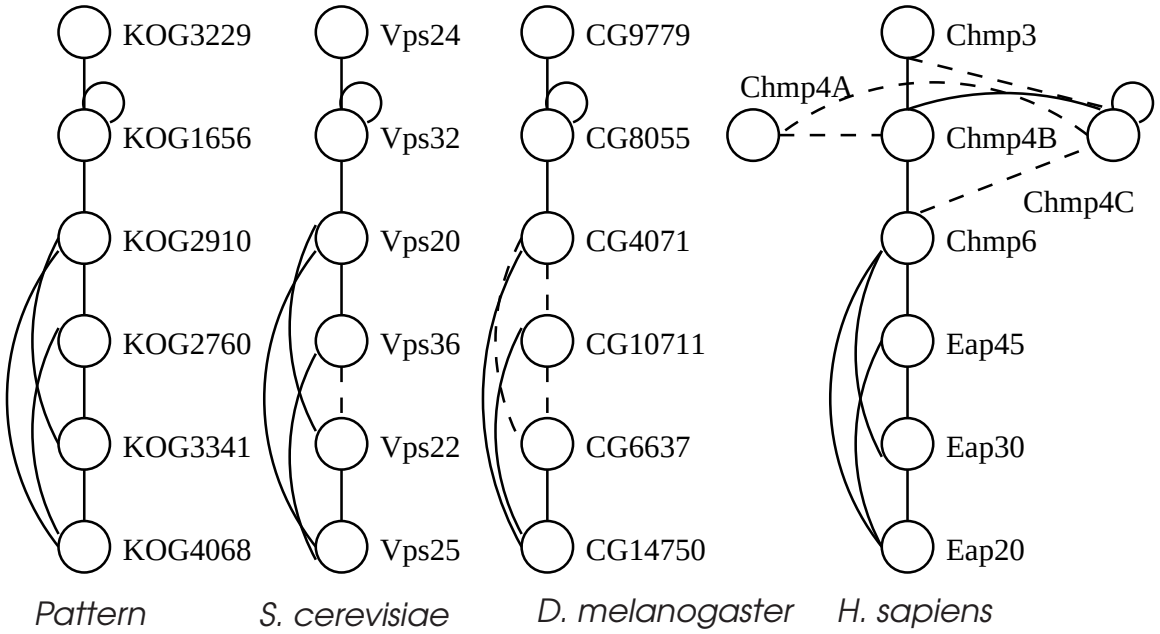
Small nuclear ribonucleoprotein complex ($p < 2e - 43$)

Frequent Protein Interaction Patterns



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

Frequent Protein Interaction Patterns



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

Modular Phylogenetics

- Top eight groups of three organisms that contain most frequent connected sub-networks and interactions

Organism set	# frequent sub-networks	# frequent interactions
C. elegans, D. melanogaster, H. sapiens	8	134
S. cerevisiae, D. melanogaster, H. sapiens	20	126
D. melanogaster, H. sapiens, M. musculus	17	86
S. cerevisiae, C. elegans, D. melanogaster	15	77
S. cerevisiae, C. elegans, H. sapiens	6	50
S. cerevisiae, H. sapiens, M. musculus	10	26
C. elegans, H. sapiens, M. musculus	5	23
H. sapiens, M. musculus, R. norvegicus	10	23

Runtime Characteristics

Comparison with isomorphism-based algorithms

FSG (Kuramochi & Karypis, *ICDM*, 2001), gSpan (Yan & Han, *KDD*, 2003)

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 patterns

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.			

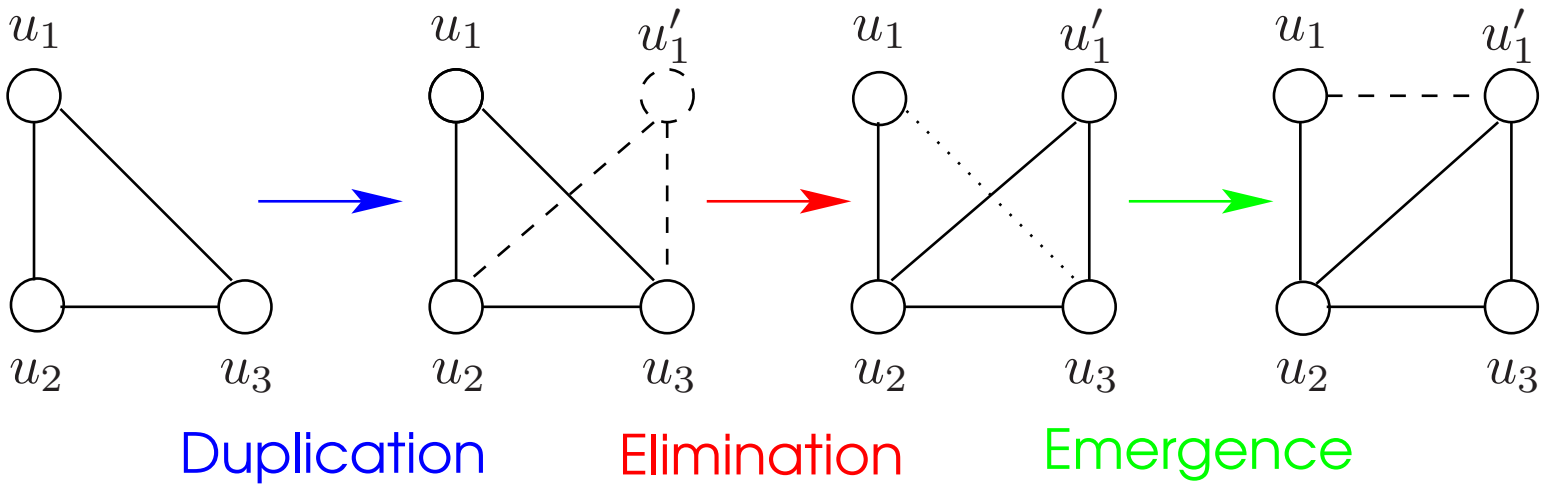
Pairwise Alignment of PPI Networks

- Given two PPI networks that belong to two different organisms, identify sub-networks that are similar to each other
  - Biological meaning
  - 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, Kim, Topkara, Subramaniam, Szpankowski, & Grama, *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

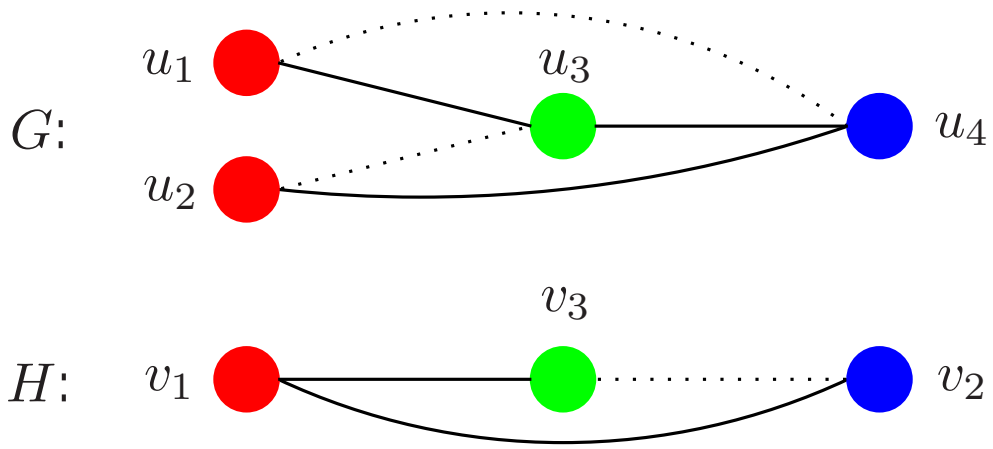
- Duplication/divergence models for the evolution of protein interaction networks
  - Interactions of duplicated proteins are also duplicated
  - Duplicated proteins rapidly lose interactions through mutations

- 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.



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?)

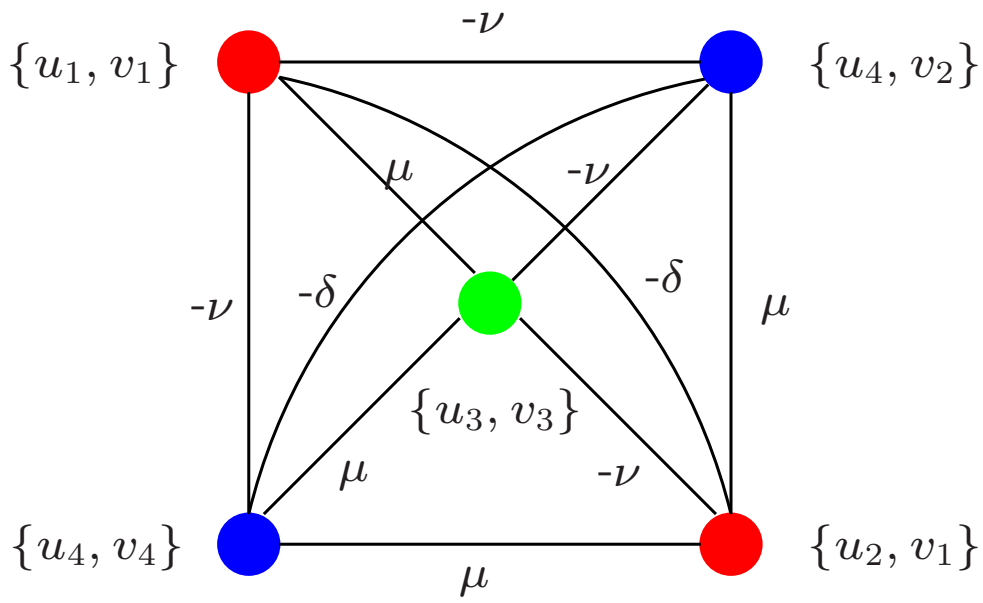
- Scores are functions of similarity between associated proteins

- 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

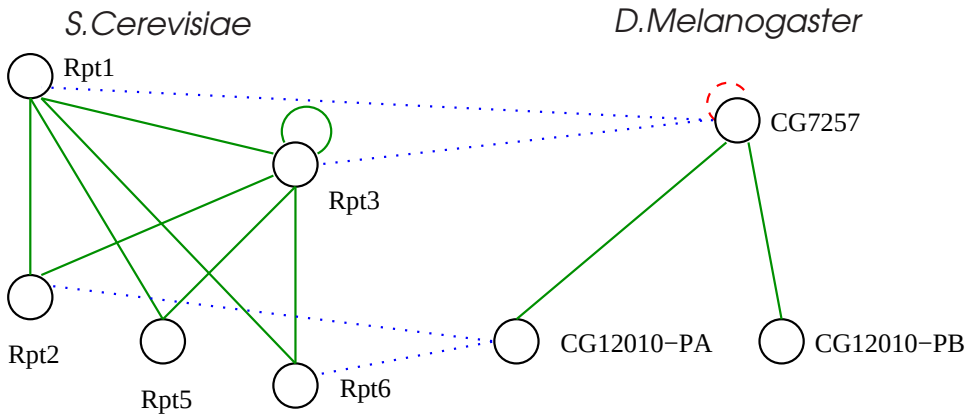
- Definition: (MAWISH)
  - Given graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ and a constant ϵ, find $\tilde{\mathcal{V}} \in \mathcal{V}$ such that $\sum_{\mathbf{v},\mathbf{u} \in \tilde{\mathcal{V}}} w(\mathbf{vu}) \geq \epsilon$.
  - NP-complete
- Theorem: (MAWISH $\equiv$ Pairwise alignment)
  - 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 partitioning
  - Greedy graph growing + iterative refinement
  - Linear-time heuristic
- Source code available at `http://www.cs.purdue.edu/homes/koyuturk/mawish/`

Alignment of Yeast and Fruit Fly PPI Networks

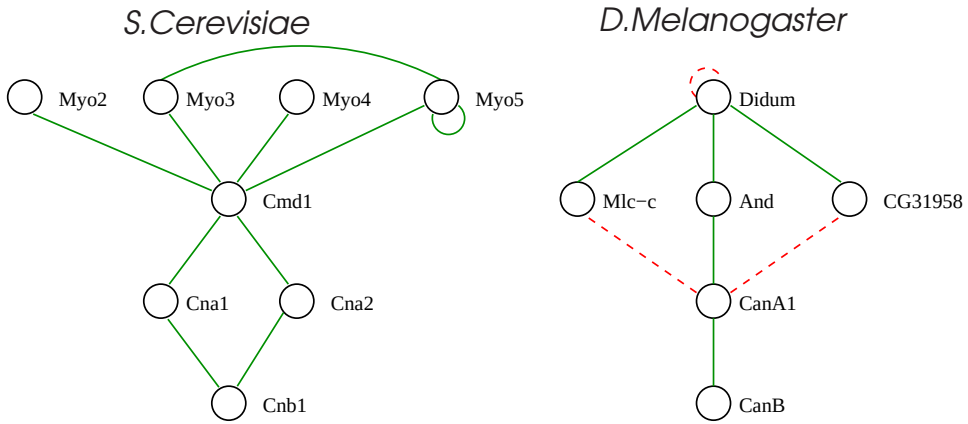
Rank	Score	z-score	# Proteins	# Matches	# Mismatches	# Dups.
1	15.97	6.6	18 (16, 5)	28	6	(4, 0)
	protein amino acid phosphorylation (69%)					
	JAK-STAT cascade (40%)					
2	13.93	3.7	13 (8, 7)	25	7	(3, 1)
	endocytosis (50%) / calcium-mediated signaling (50%)					
5	8.22	13.5	9 (5, 3)	19	11	(1, 0)
	invasive 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



Statistical Significance of Modularity

- 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
  - PPI networks generally exhibit power-law property (or exponential, geometric, etc.)
  - Analysis simplified through independence assumption
  - 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)

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

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.

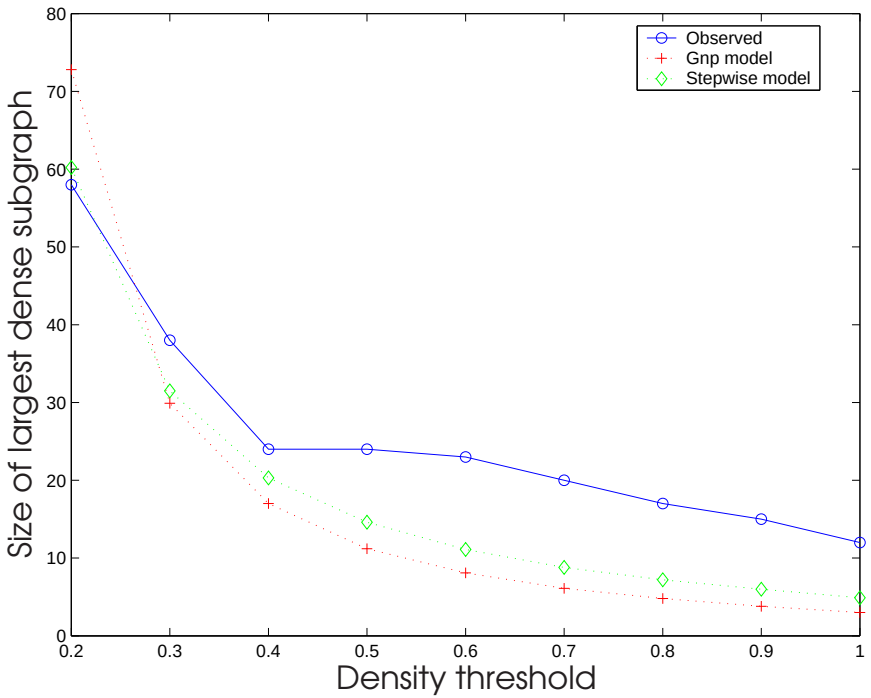
Generalizing Results to Complex Models

- Piecewise $G(n, p)$ model
  - Few proteins with many interacting partners, many proteins with few interacting partners
  - Captures the basic characteristics of PPI networks
  - The size of largest dense subgraph is still proportional to $\log n$
- More general models
  - Increasing the number of pieces, we approach models with characteristic degree distributions
  - Analysis of power-law graphs in progress
- Multiple networks: Conservation
  - Superpose graphs based on sequence homology

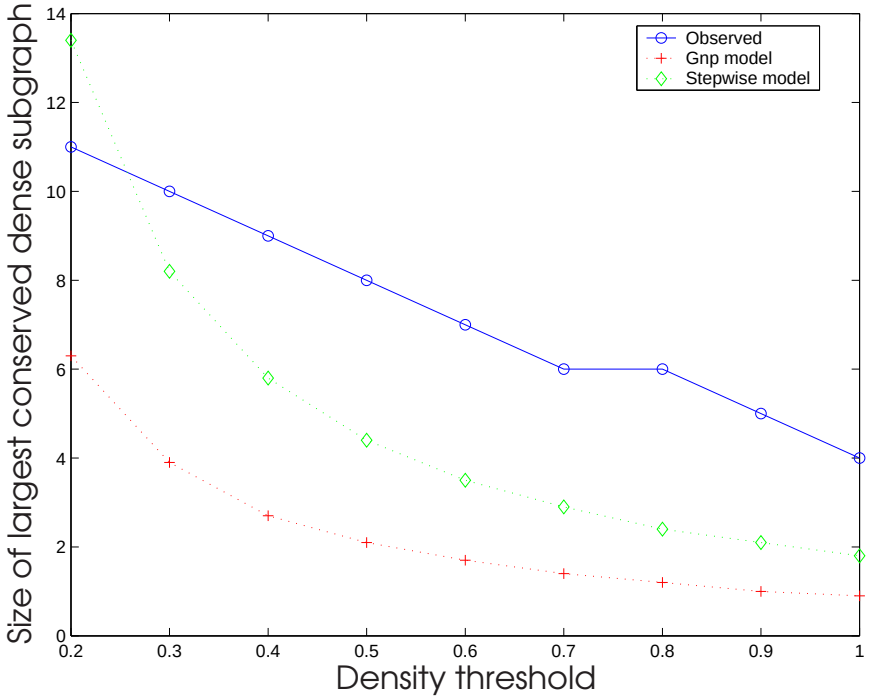
Algorithms Based on Statistical Significance

- 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
- 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

Largest Dense Subgraph for Varying Density



Yeast PPI network



Yeast & *Fruit Fly* PPI networks

Pathway Organization: Genetic Interactome

Double mutants exhibit unexpected phenotypes, as compared to joint single mutations.

Definition 1. • ***Negative Interactions***: *more severe phenotype than expected*

- *Also known as aggravating or synergistic*

• ***Positive Interactions***: *Less severe phenotype than expected*

- *Also known as alleviating or epistatic*

Most commonly used:

Phenotype : Growth rate

Model : Multiplicative null model

Organization of Genetic Interactions

Definition 2. • *Between-Pathway Model*

- *Among genes participating in redundant functions*

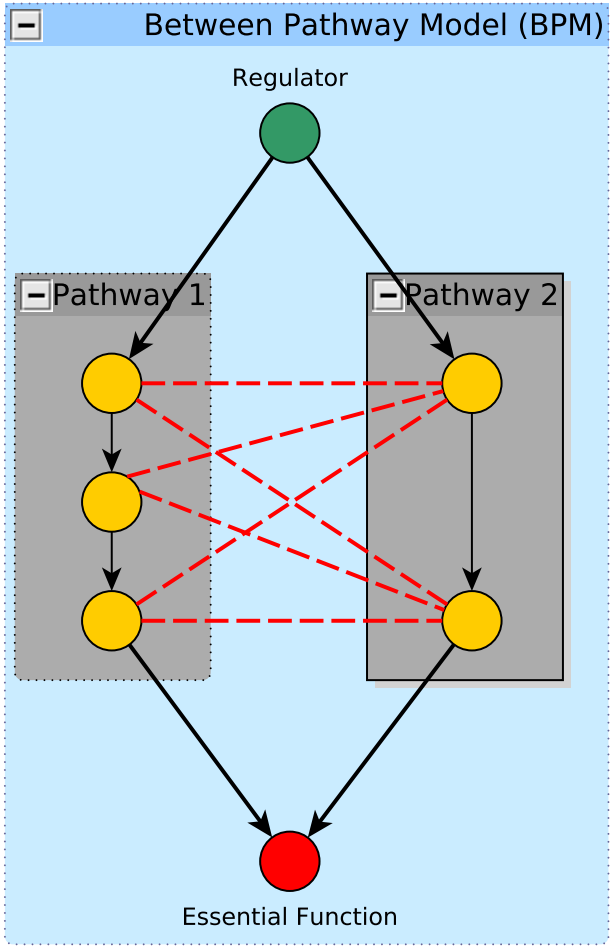
• *Within-Pathway Model*

- *Among genes with additive effect*

• *Indirect Effect*

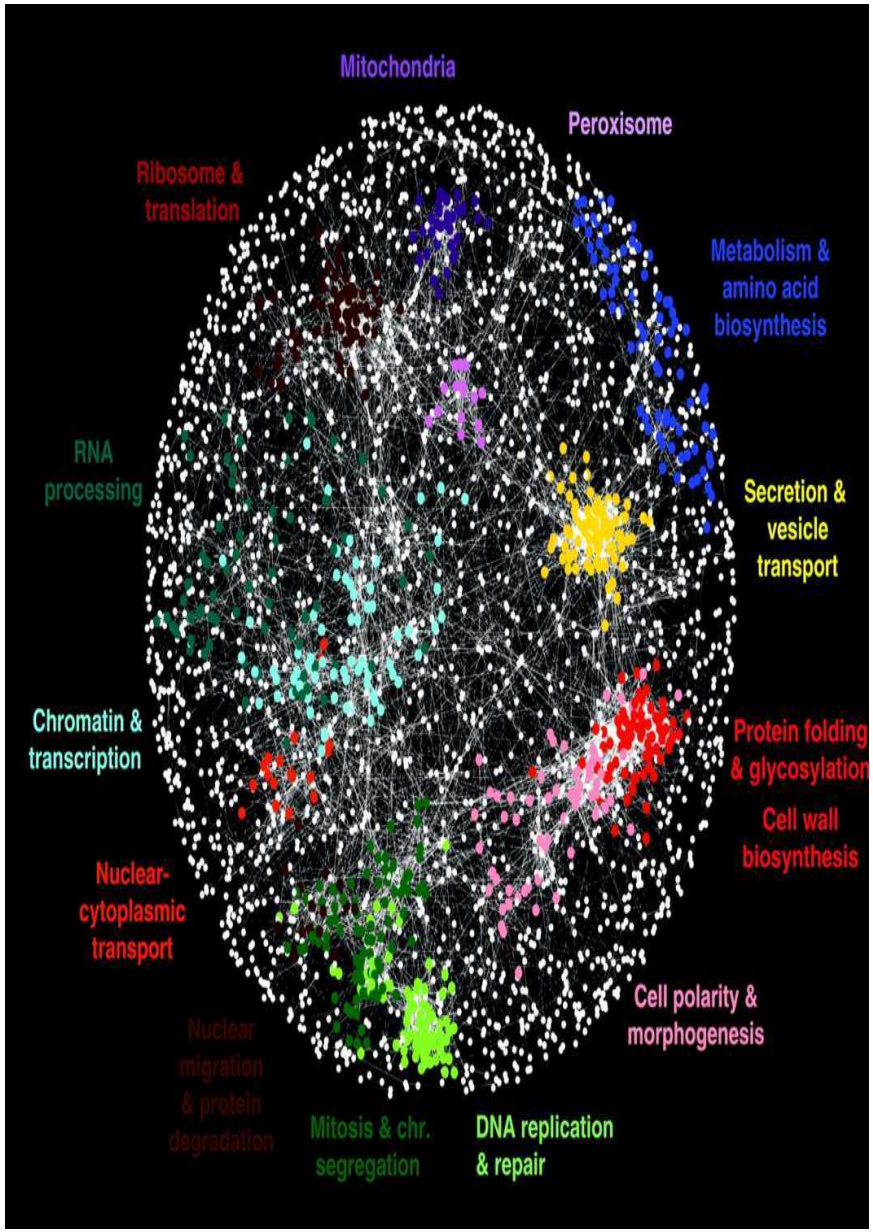
- *Among genes with distant functions that are not directly related*

Between-Pathway Model (BPM)



- Bi-cliquish structure
- Have been used to:
  1. Predict co-pathway membership of gene pairs
  2. Extract redundant pathways

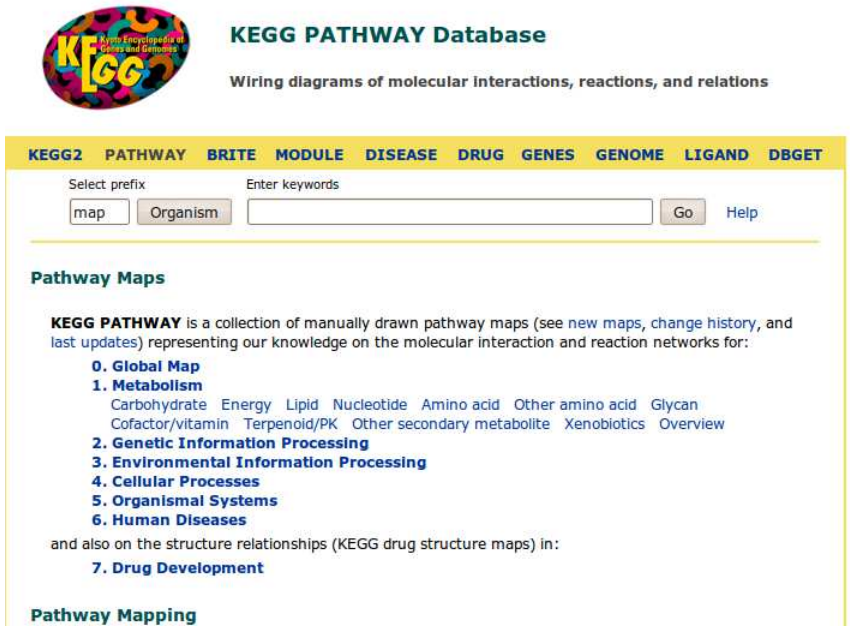
The Genetic Landscape of a Cell



Adopted from Costanzo et al., 2010

- Baker's yeast, *Saccharomyces cerevisiae*
- Synthetic Genetic Array (SGA)
- 1712 query genes
  1. 1378 null alleles of non-essential genes
  2. 334 hypomorphic or conditional alleles of essential genes
- 3885 array strains

Functional Annotations

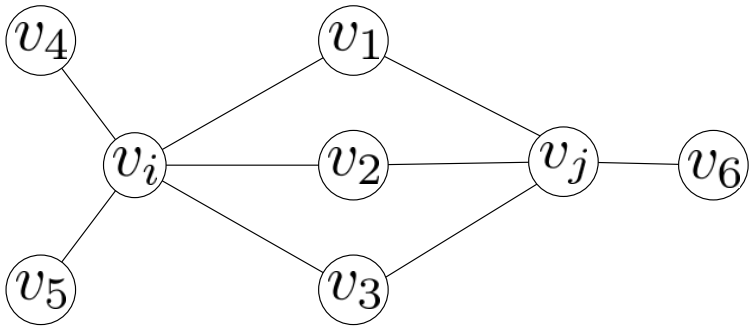


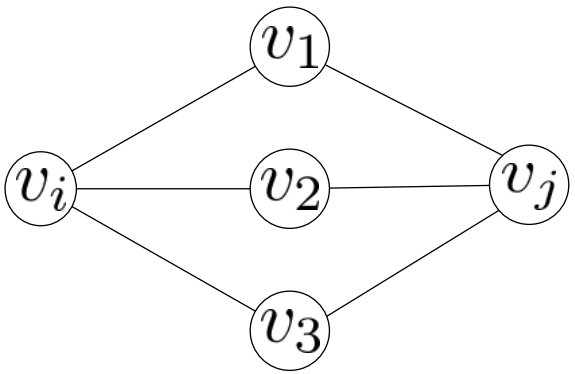
- KEGG Pathway Database
- Annotations for 1026 genes in the experiment
- 96 Pathways
  - 80 pathways after filtering pathways with less than 10 genes.

Local Neighborhood Similarity: A Predictor of Co-Pathway Membership

Similarity prediction methods

- Number of Shared Neighbors
- Congruence Score
- Pearson Correlation of Interaction Profiles





Both v_i and v_j have three shared neighbors. However, in the first case their congruence score is almost 0.6, while in the second case it is approximately 2 (assuming a graph of size 10).

Evaluating Ranking Methods

Given a pathway P_A and a cut size (target set) l.

Definition 3.

$$\begin{aligned} P-value(X=k) &= Prob(k \leq X) \\ &= HGT(k|N, N_A, l) \\ &= \sum_{x=k}^{min(N_A,l)} \frac{C(l,x)C(N-l, N_A-x)}{C(N, N_A)} \end{aligned}$$

X: Random variable denoting the number of true positives in a random sample, N: Total number of gene pairs, N_A: Number of gene pairs in pathway A, l: Size of target set

Minimum HyperGeometric (mHG) Score

Target size unknown:

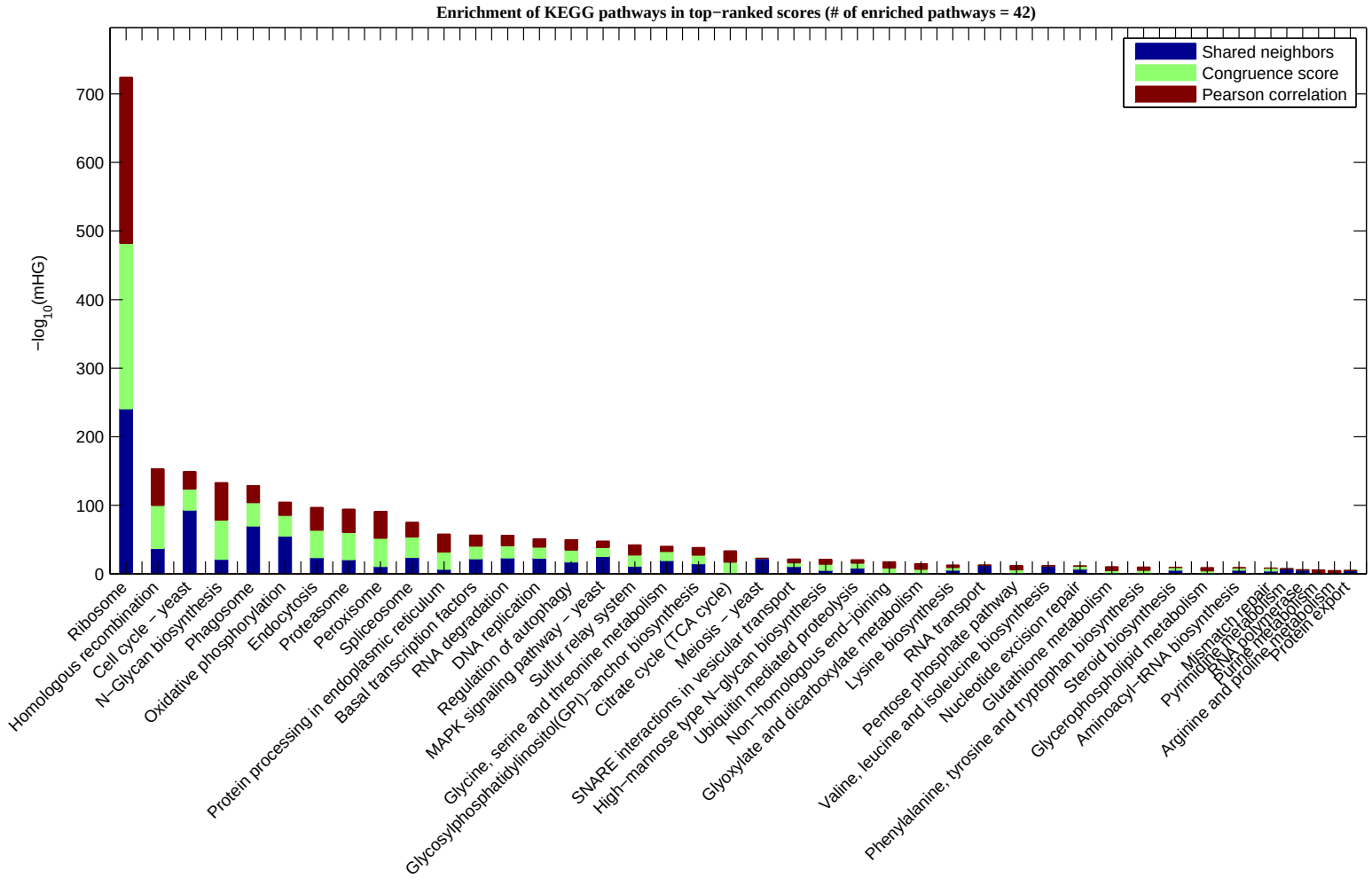
Definition 4. *The Minimum HyperGeometric (mHG) score*

$$mHG(\lambda) = min_{1 \leq l \leq N} HGT(b_l(\lambda); N, N_A, l),$$

where $b_l(\lambda) = \sum_{i=1}^{l} \lambda_i$

λ_i is 1 if both of the genes in the i^{th} ranked gene pair are members of P_A, and 0 otherwise.

- mHG Adjusted for Multiple Comparison


Enrichment of KEGG pathways in top−ranked scores (# of enriched pathways = 42)
Shared neighbors
Congruence score
Pearson correlation
−log10(mHG)
0
100
200
300
400
500
600
700
Ribosome
Homologous recombination
Cell cycle − yeast
N−Glycan biosynthesis
Phagosome
Oxidative phosphorylation
Endocytosis
Proteasome
Peroxisome
Spliceosome
Protein processing in endoplasmic reticulum
Basal transcription factors
RNA degradation
DNA replication
Regulation of autophagy
MAPK signaling pathway − yeast
Sulfur relay system
Glycine, serine and threonine metabolism
Glycosylphosphatidylinositol(GPI)−anchor biosynthesis
Citrate cycle (TCA cycle)
Meiosis − yeast
SNARE interactions in vesicular transport
High−mannose type N−glycan biosynthesis
Ubiquitin mediated proteolysis
Non−homologous end−joining
Glyoxylate and dicarboxylate metabolism
Lysine biosynthesis
RNA transport
Pentose phosphate pathway
Valine, leucine and isoleucine biosynthesis
Nucleotide excision repair
Phenylalanine, tyrosine and tryptophan biosynthesis
Glutathione metabolism
Glycerophospholipid metabolism
Steroid biosynthesis
Aminoacyl−tRNA biosynthesis
Pyrimidine metabolism
Mismatch repair
RNA polymerase
Purine metabolism
Arginine and proline metabolism
Protein export

Highlights

Basic Idea

Heterogeneous performance of co-pathway membership predictions $\Longleftrightarrow$ Existence of specific structure around enriched pathways

- Decomposing neighborhood of each pathway
- Inferring lethal crosstalk among pathways

Modified Congruence Score (MCS)

Evaluating Neighborhood Overlap of Gene Pairs With Respect to a Given Pathway

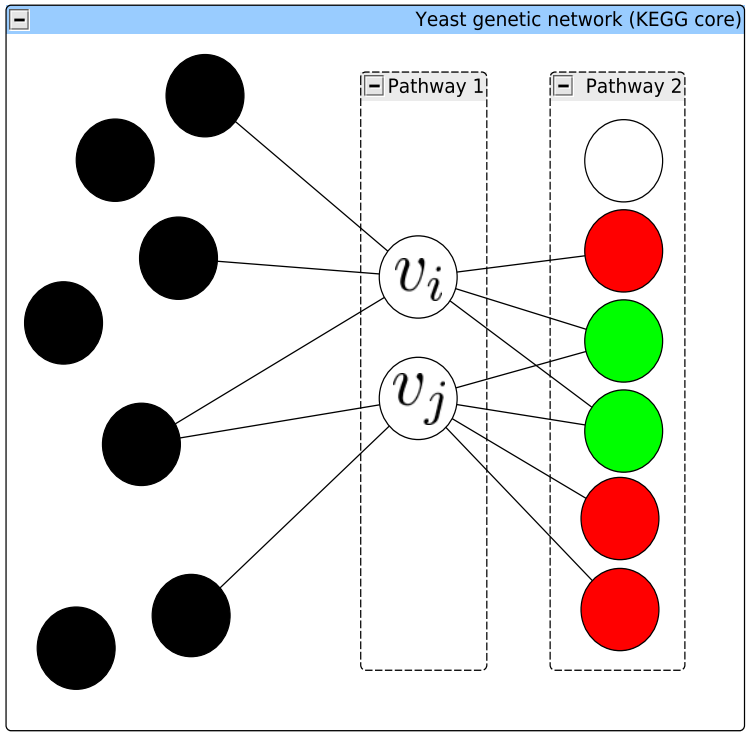
Definition 5.

$$
\begin{aligned}
P-value(X=k_{ij}^B) &= Prob(k_{ij}^B \leq X) \\
&= HGT(k_{ij}^B | n_B, d_i^B, d_j^B) \\
&= \sum_{x=k_{ij}^B}^{\min(d_i^B, d_j^B)} \frac{C(d_j^B, x) C(n_B - d_j^B, d_i^B - x)}{C(n_B, d_i^B)}
\end{aligned}
$$

MCS is defined as $-log_{10}$ of the P-value.

Modified Congruence Score (MCS): continued

Example 1

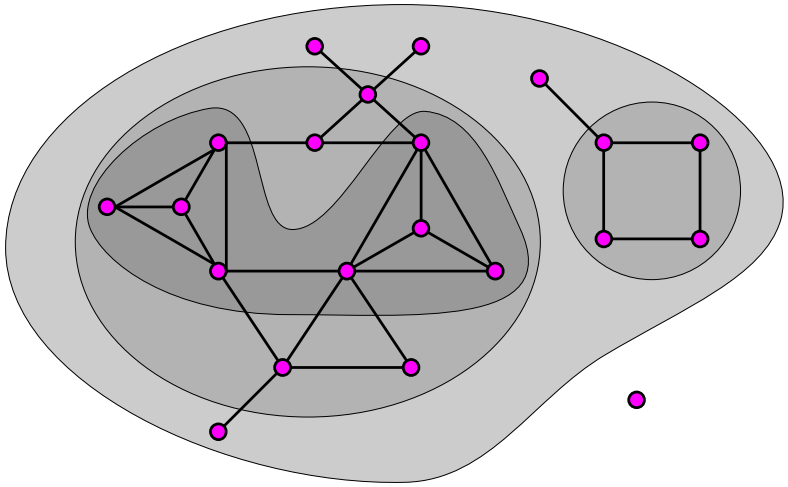


A sample neighborhood configuration for v_i and v_j. Here $n = 15, D_i = 6, D_j = 5, n_B = 6, d_i = 3, d_j = 4$ and $k = 2$.

Constructing Neighborhood Overlap Graph For a Given Pathway Pair

Definition 6. *The neighborhood overlap graph (NOG) of a given pathway P_A with respect to pathway P_B, denoted by $H_{A \to B} = (V_H, E_H)$, is an unweighted, undirected graph defined over same vertices as P_A. In this graph, there is a link between vertices v_i and v_j if the network structure around them with respect to P_B is statistically significant .*

Pruning neighborhood overlap graph, finding cohesive subgraphs, and identifying interaction ports



Adopted from Batagelj and Zaversnik, 2002

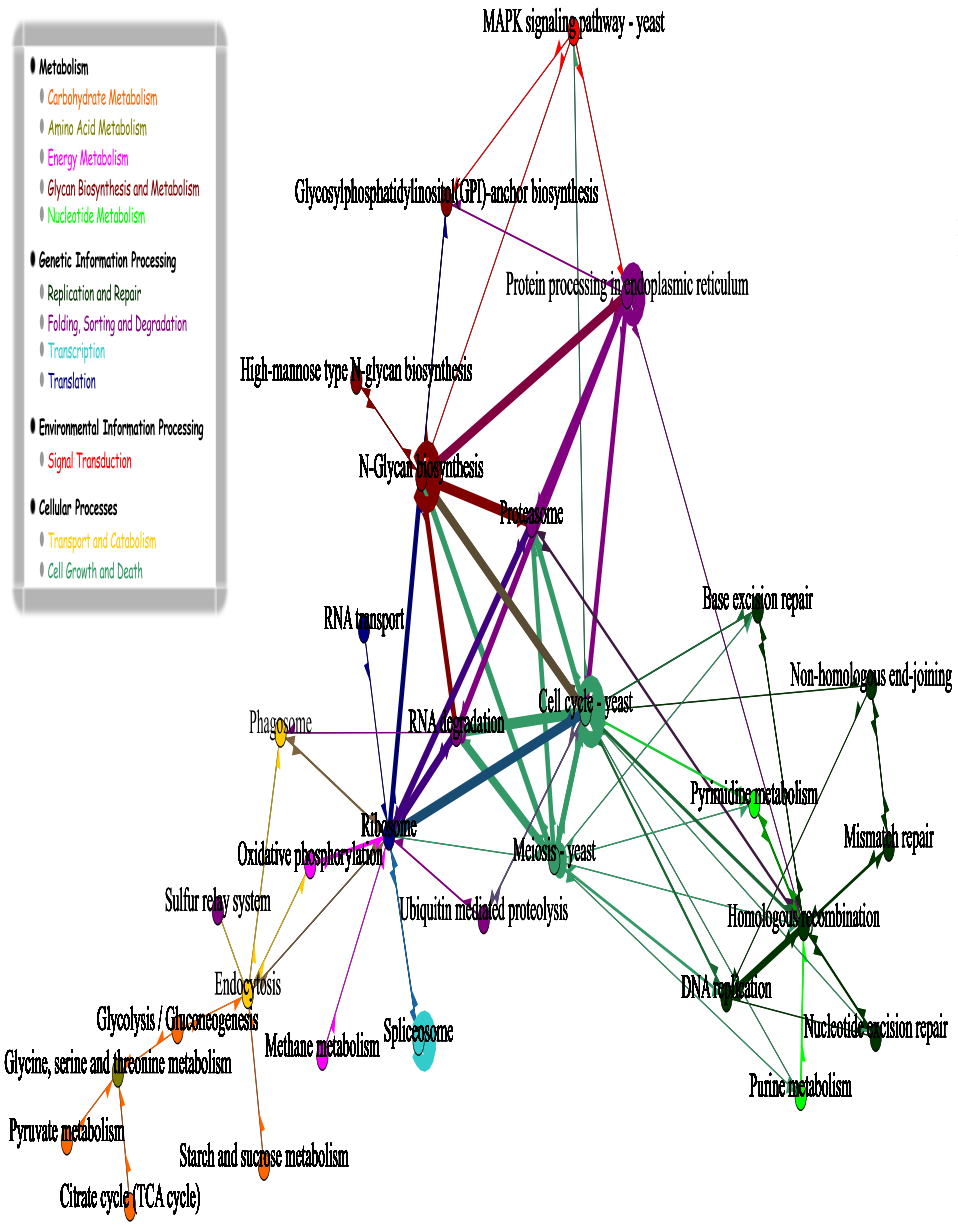
1. Iterative peeling of K-shells
   Pruning hairy components

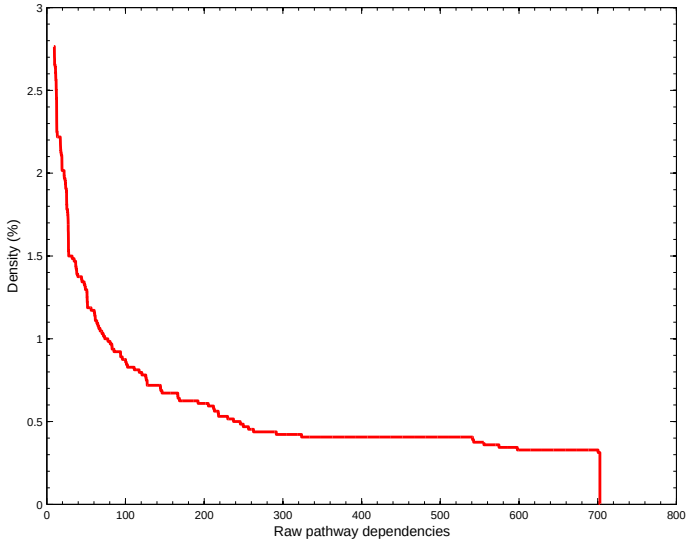
2. Connected components in each core

3. Evaluating the significance of components

   - Evaluating significance using ER random graph model

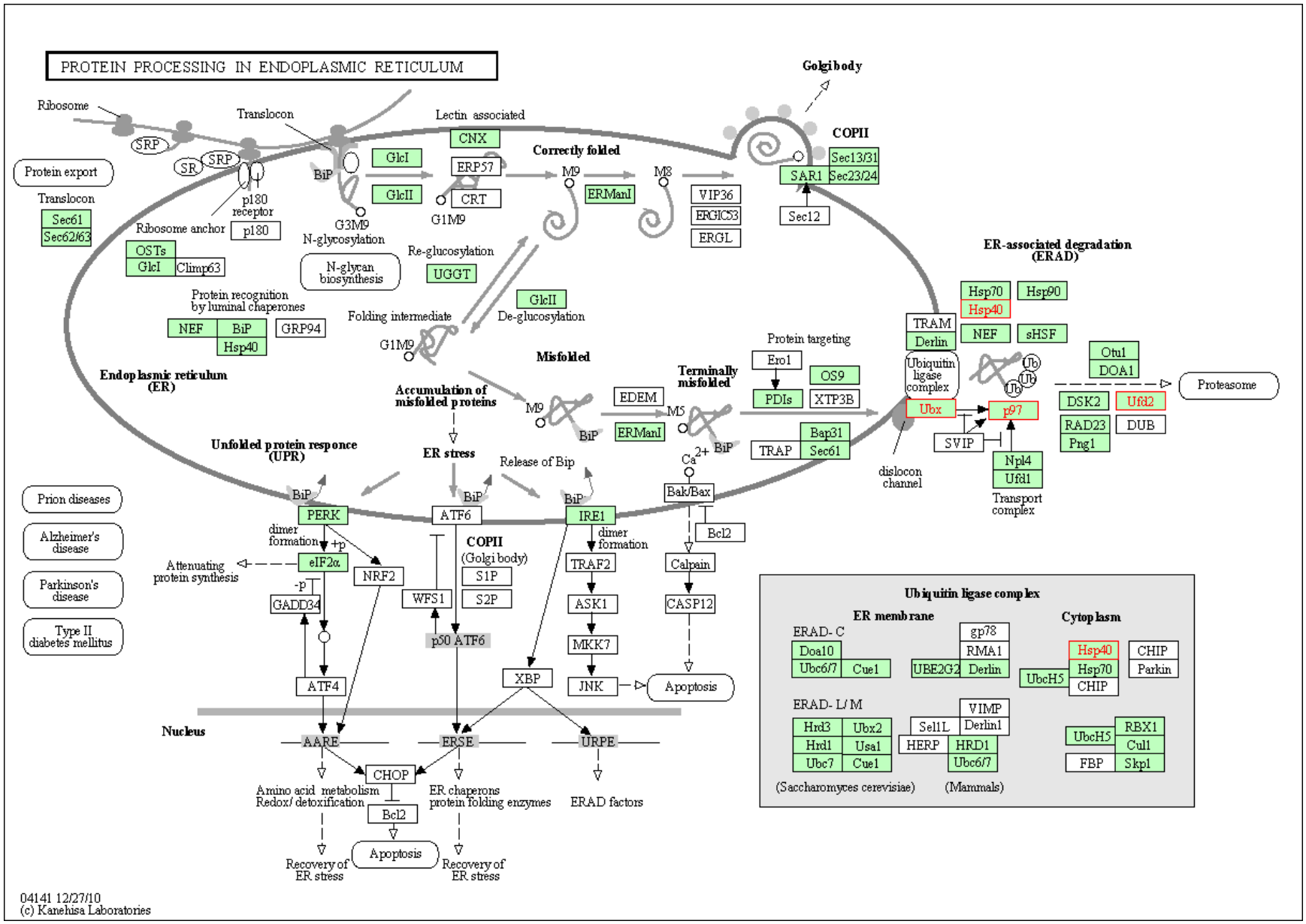
KEGG Crosstalk Map



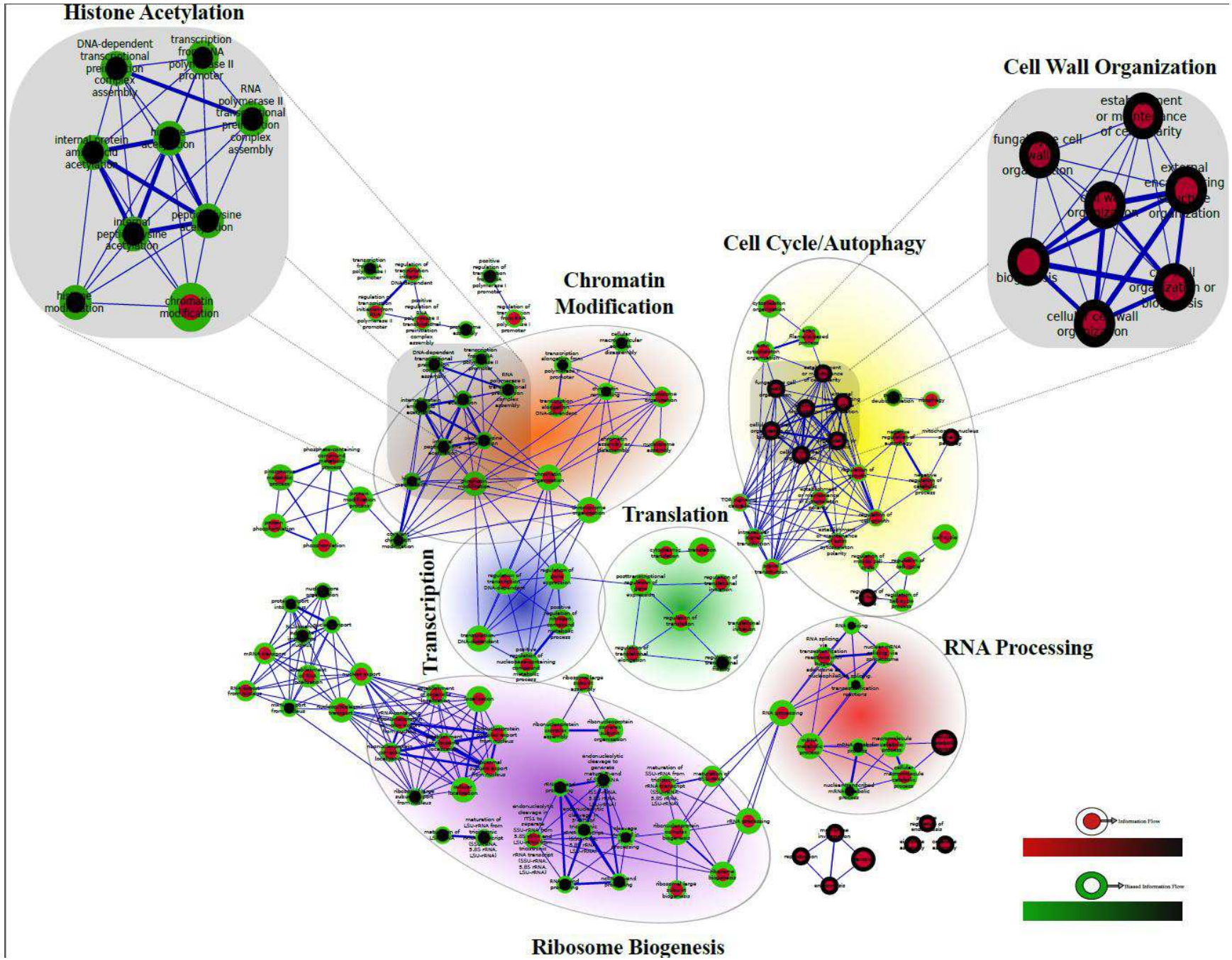


Interaction Port Case Study

Crosstalk Between Protein Processing in ER and Proteasome



Ongoing Work: The Interaction Map of Aging



Ongoing Work

- The surprisingness of choice in networks.
- Tissue-specific alignments.

Thanks

- To the US National Science Foundation for their funding.
- To our hosts at IISc for all of their support and funding.
- To all of my collaborators.
- To you for your attention!