GQ: A Graph Toolkit for Multicore Environments

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

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

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

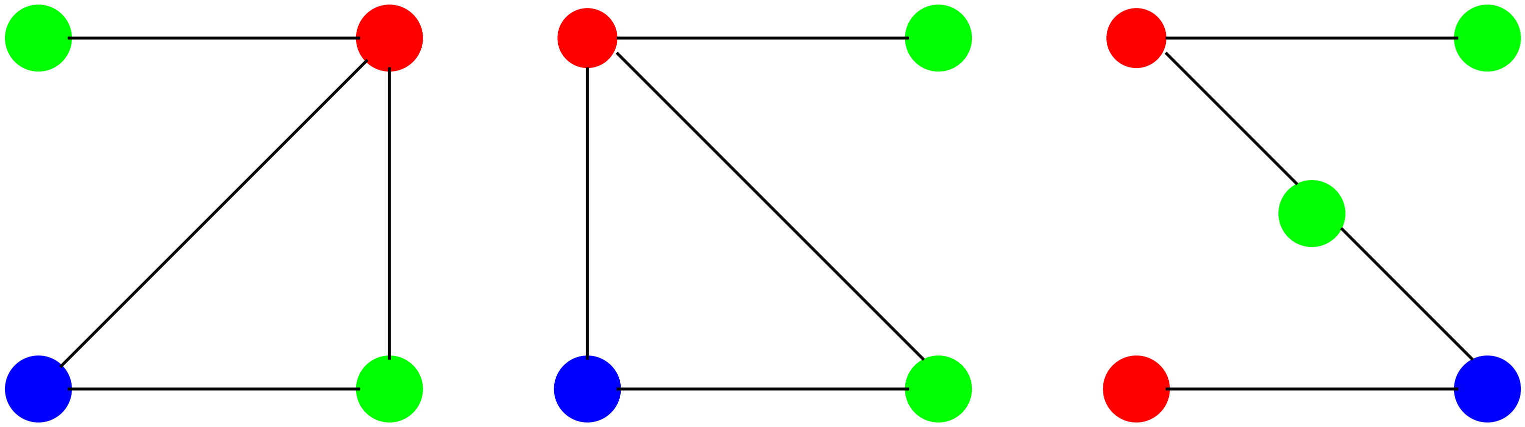
Graph Library Components

- Conservation in Networks
- Alignment of Networks
- Modularity in Networks
- Reputation/Rank
- Graph Grammars and Parsers
- Some Ongoing Work

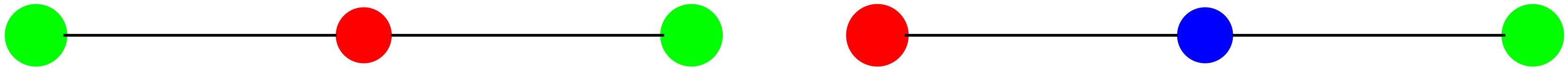
Conservation in Networks

- Given a collection of networks (say, protein 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 connected.
  - Frequency: The number of networks that contain a sub-network, is a coarse measure of statistical significance
- 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 and beyond)

Graph Analysis



Network database



Interaction patterns that are common to all networks

Problem Statement

- Given a set of nodes V, a set of edges 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

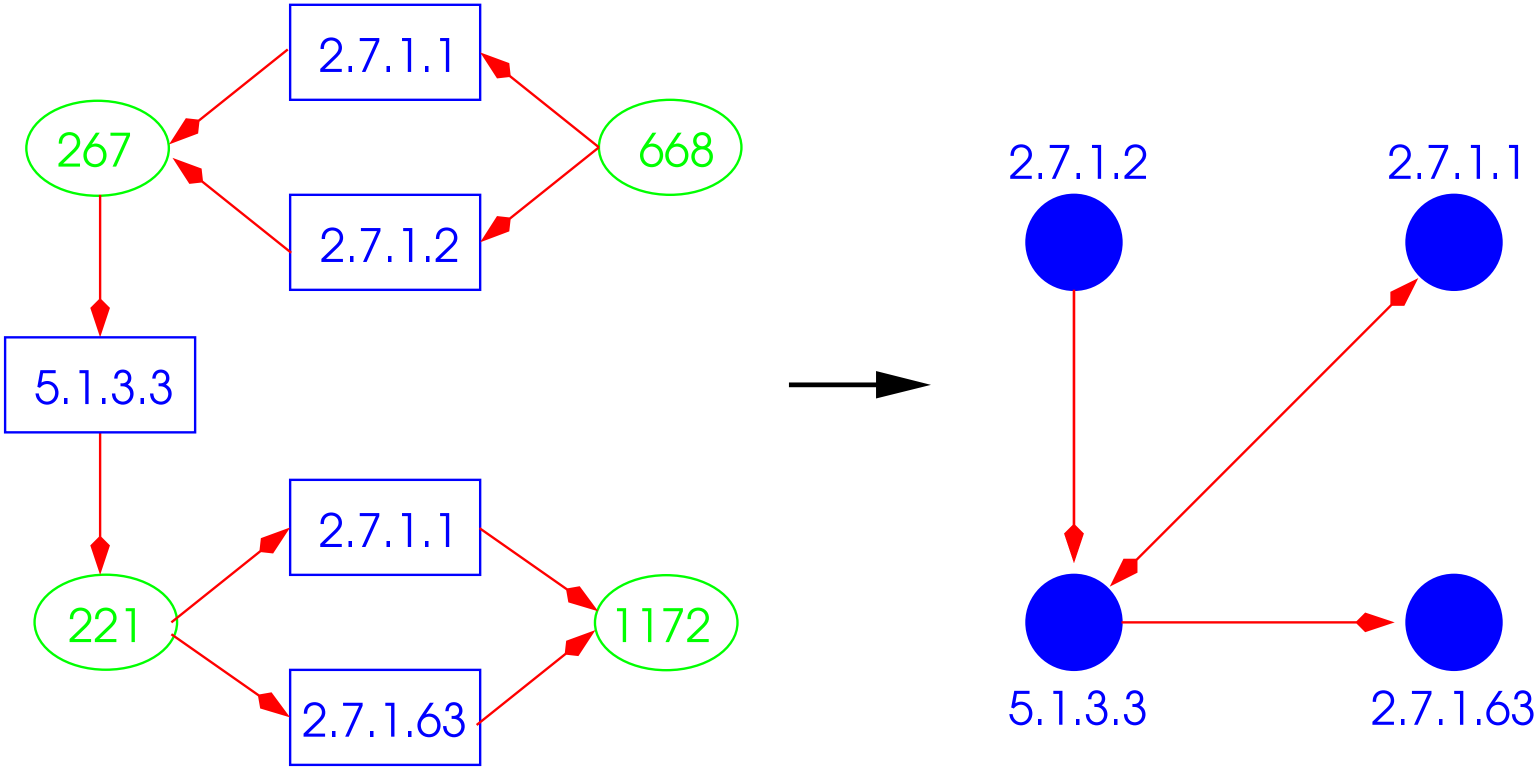
- Conserved sub-network discovery
  - Instance: A set of networks $\mathcal{G} = \{G_1 = (V_1, E_1, \mathcal{L}), G_2 = (V_2, E_2, \mathcal{L}), \ldots, G_m = (V_m, E_m, \mathcal{L})\}$, and a frequency threshold σ^*.
  - Problem: Let $H(S) = \{G_i : S \sqsubseteq G_i\}$ be the occurrence set of graph 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 ⇒ 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

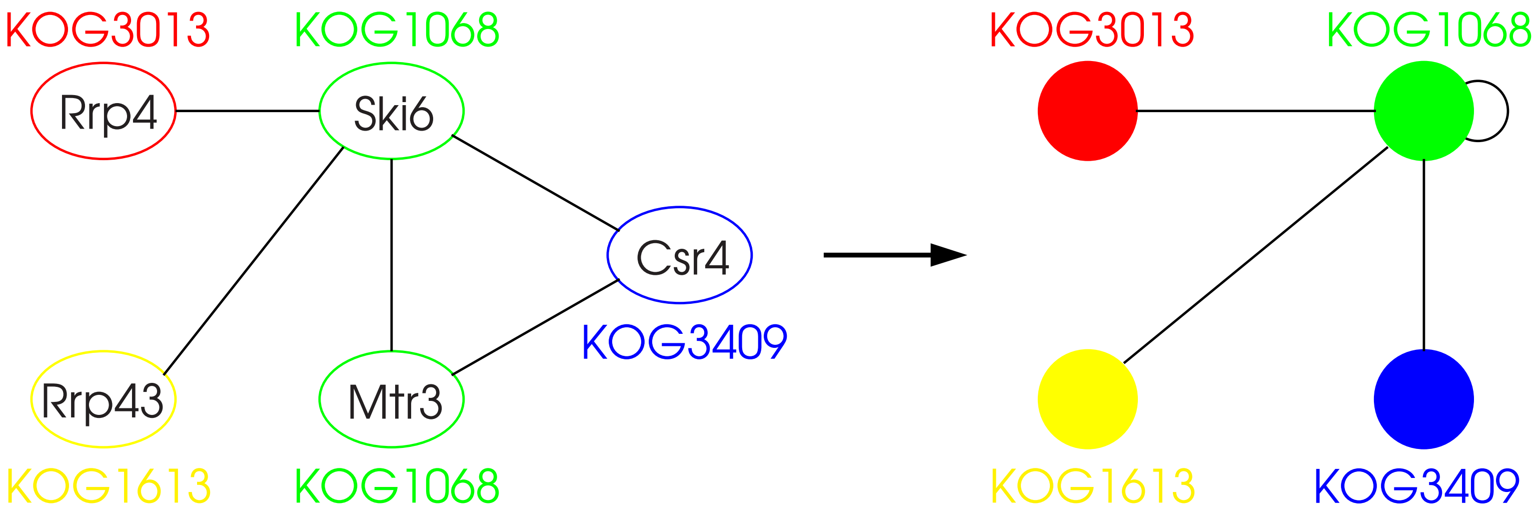
Ortholog Contraction in Real Applications (Metabolic Pathways)

- 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



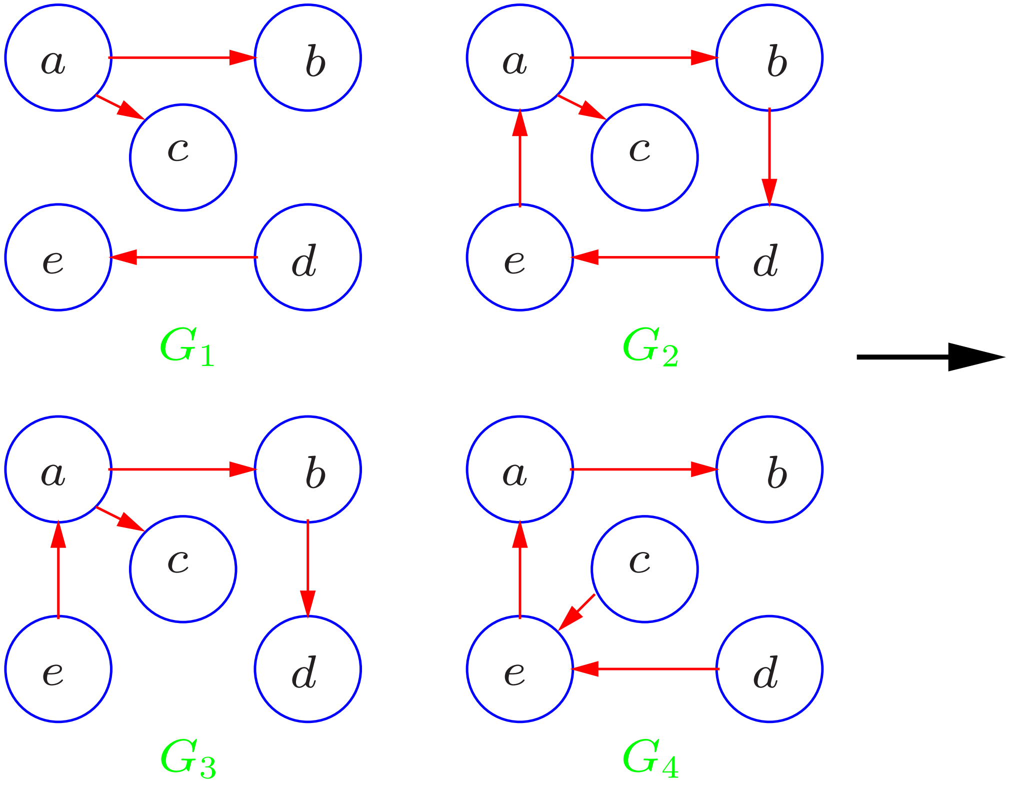
Ortholog Contraction in Real Applications (PPI Networks)

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



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



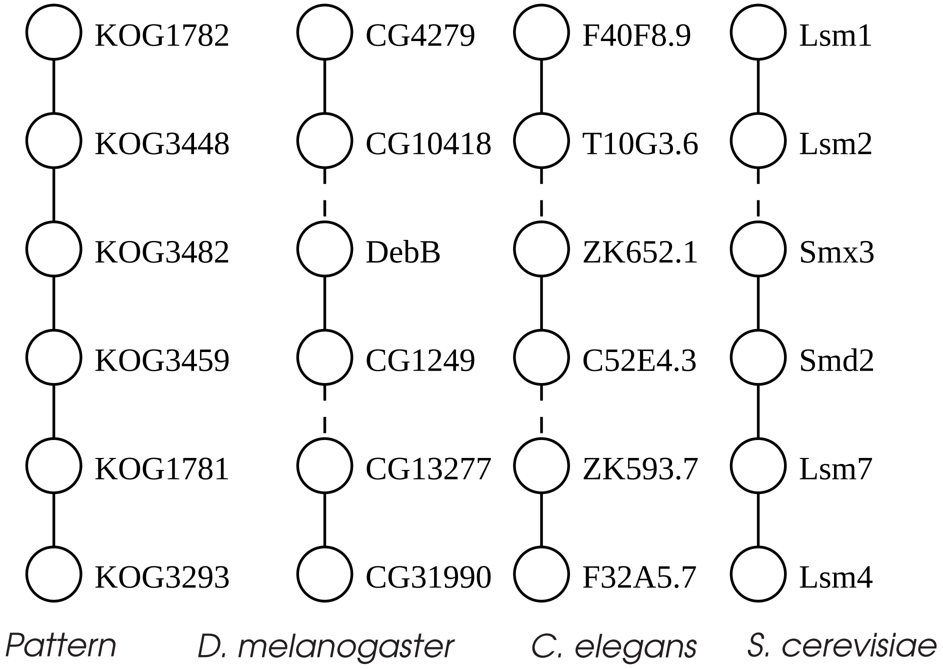
$F_1 = \{ab, ac, de\}$

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

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

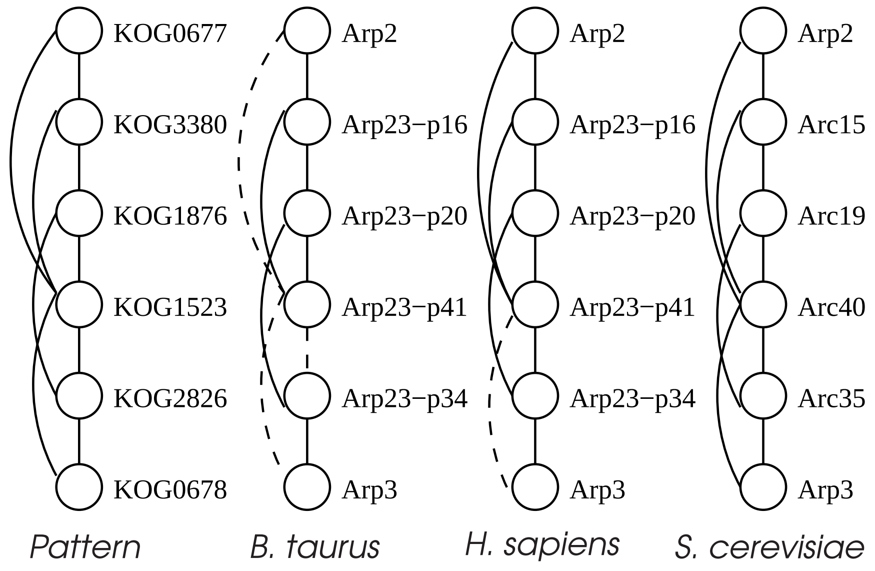
$F_4 = \{ab, ce, de, ea\}$

Conserved Protein Interaction Patterns



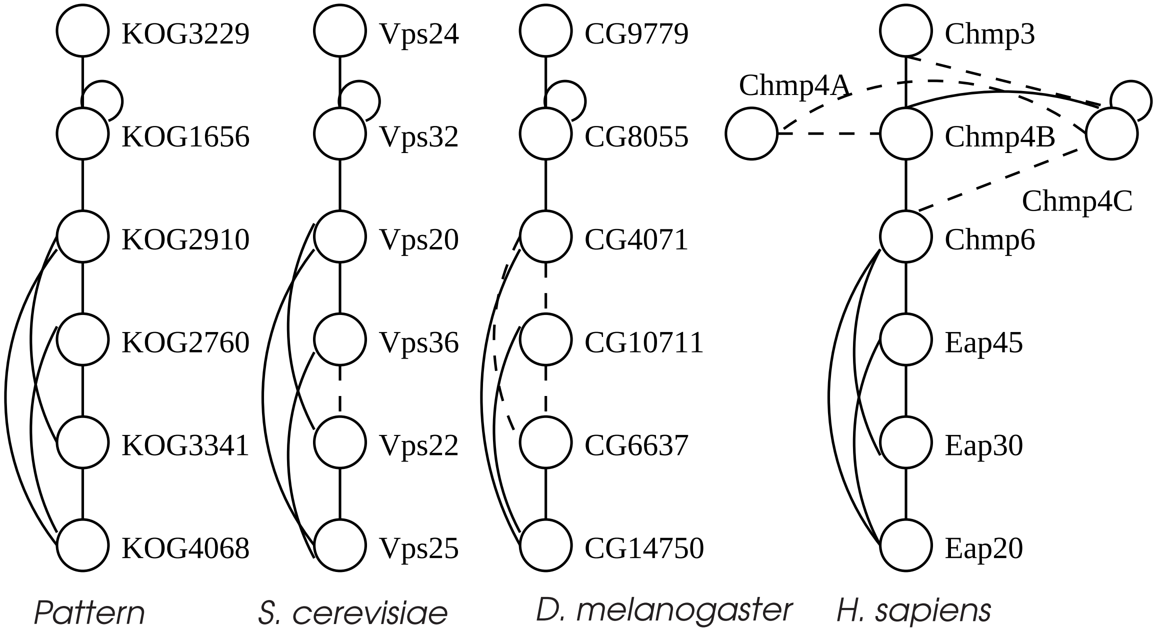
Small nuclear ribonucleoprotein complex ($p < 2e - 43$)

Conserved Protein Interaction Patterns



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

Conserved Protein Interaction Patterns



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

Runtime Characteristics

Dataset	Minimum Support (%)	Runtime (secs.)	Largest pattern	Number of patterns	Runtime 2 Cores	Runtime 4 Cores
Glutamate	12	0.10	13	39	0.08	0.07
	10	0.29	15	34	0.16	0.10
	8	0.99	15	56	0.58	0.37
Alanine	16	0.06	12	21	0.05	0.07
	12	1.06	16	25	0.57	0.33
	10	1.72	16	34	0.90	0.52

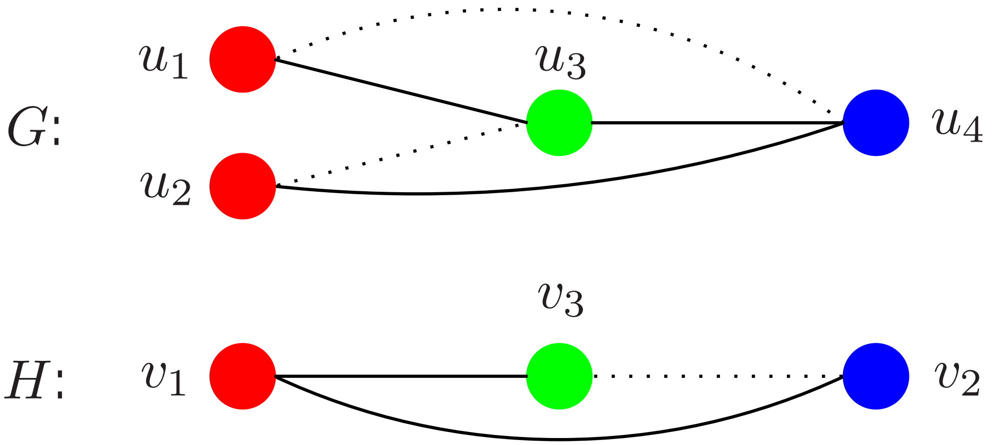
All times on a 2.66 MHz i7 Processor.

Alignment of Networks

- Given two networks, identify sub-networks that are similar to each other.
  - Optimization function
  - 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)

Match, Mismatch, and Duplication

- Establishing a Cost Measure
  - A match $\in \mathcal{M}$ corresponds to two pairs of homologous nodes such that both pairs interact in both networks. A match is associated with score μ.
  - A mismatch $\in \mathcal{N}$ corresponds to two pairs of homologous nodes such that only one pair is interacting. A mismatch is associated with penalty ν.
  - A duplication $\in D$ corresponds to a pair of homologous nodes in the same network. A duplication is associated with score δ.

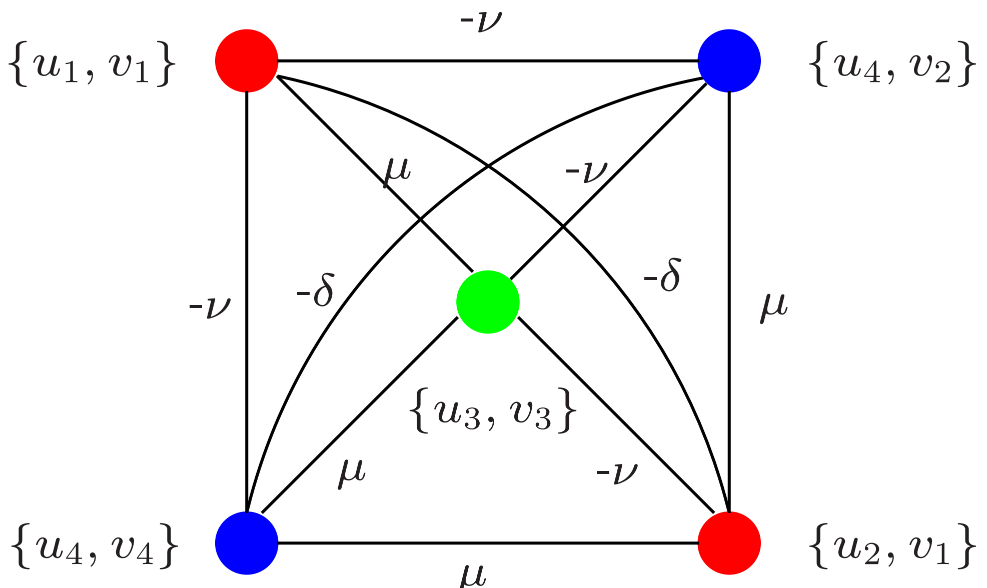


Pairwise Alignment of Networks as an Optimization Problem

- 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 divergence.
- Problem: Find all subnet pairs with significant alignment score
- A graph equivalent to BLAST

Weighted Alignment Graph

- $\mathbf{G(V,E)}$: $\mathbf{V}$ consists of all pairs of homologous nodes $\mathbf{v} = \{u \in U, v \in V\}$
- 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

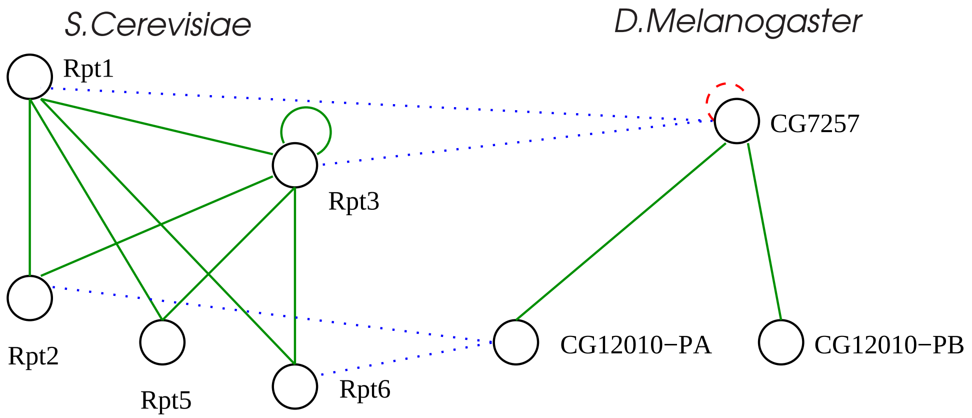
- Definition: (MAWISH)
  - Given graph $\mathcal{G}(\mathcal{V},\mathcal{E})$ and a constant ϵ, find $\tilde{\mathcal{V}} \in \mathcal{V}$ such that $\sum_{\mathbf{v},\mathbf{u}\in\tilde{\mathcal{V}}} w(\mathbf{vu}) \geq \epsilon$.
  - NP-complete by reduction from Maximum-Clique
- Theorem: (MAWISH $\equiv$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

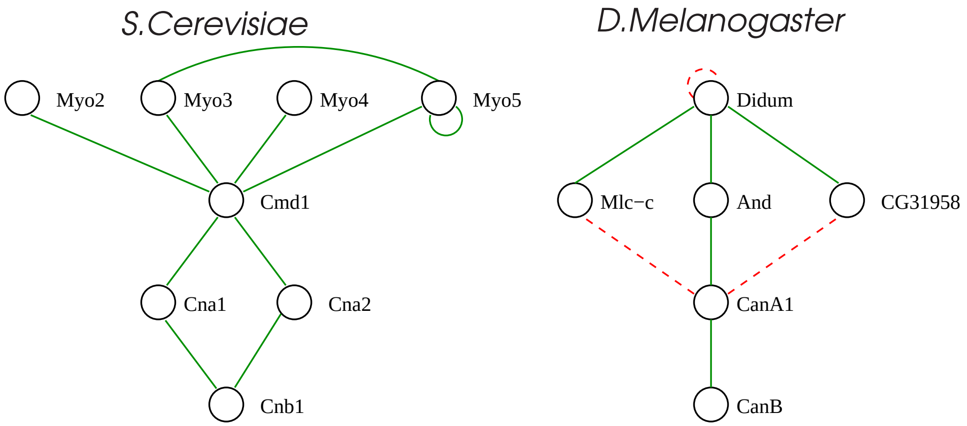
Rank	Score	z-score	# Proteins	# Matches	# Mismatches	# Dups.
1	15.97	6.6	18 (16, 5)	28	6	(4, 0)
	protein amino acid phosphorylation (69%) JAK-STAT cascade (40%)					
2	13.93	3.7	13 (8, 7)	25	7	(3, 1)
	endocytosis (50%) / calcium-mediated signaling (50%)					
5	8.22	13.5	9 (5, 3)	19	11	(1, 0)
	invasive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



Runtime Characteristics

Dataset	Number of Patterns	Runtime (secs.)	2 Cores	4 Cores
Yeast/Fruit Fly	8	0.16	0.12	0.10
	13	1.80	1.02	0.68
	20	2.93	1.61	0.94

All times on a 2.66 MHz i7 Processor.

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 $d^2_{\max} > |E|$ 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

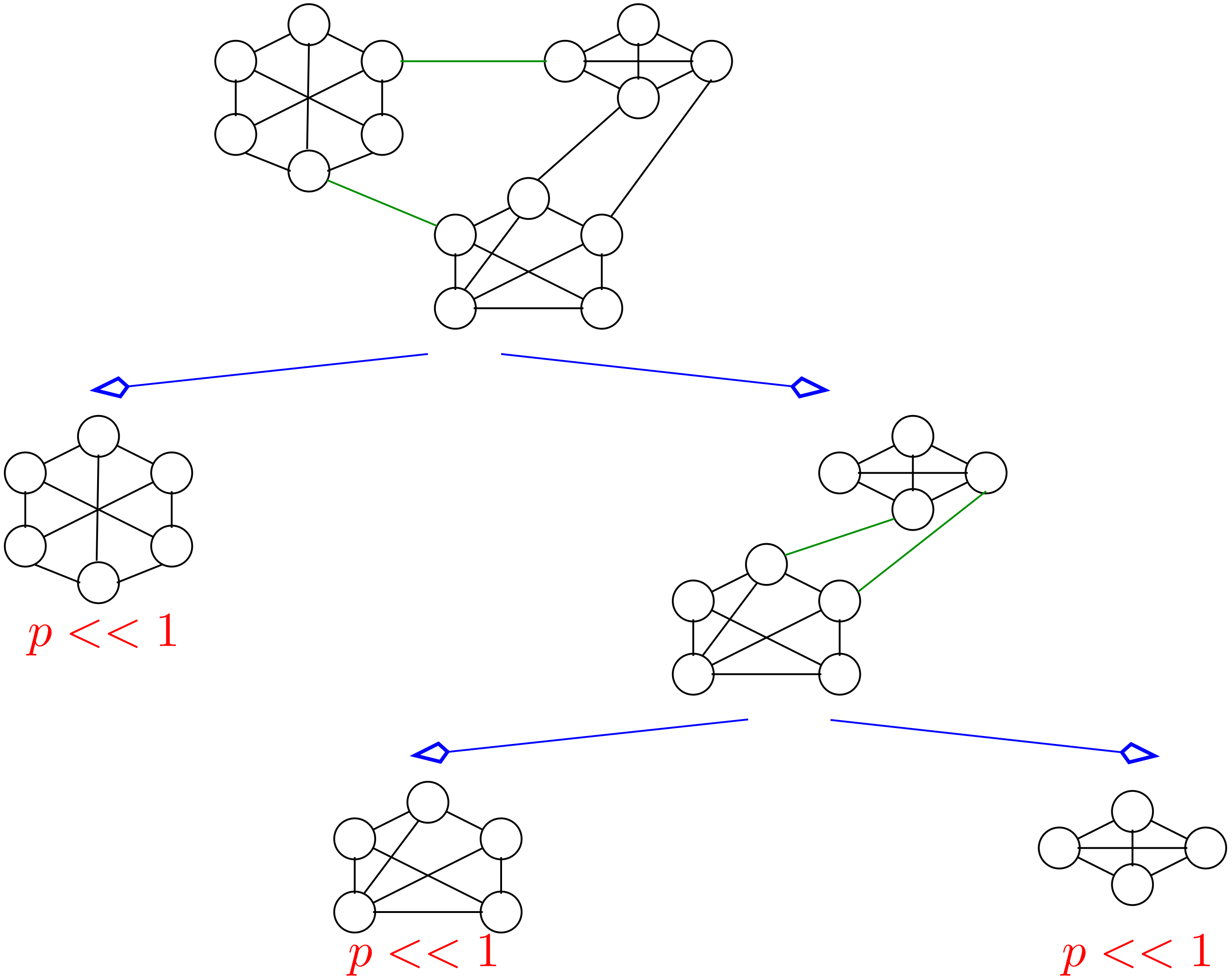
$$r_1 = \frac{\log n - \log\log n + 2n_h \log B + \log \kappa(p_l,\rho) - \log e + 1}{\kappa(p_l,\rho)}$$

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)

Runtime Characteristics

Dataset	Number of Clusters	Runtime (secs.)	2 Cores	4 Cores
Yeast PPI	11	4.80	2.64	1.60
	18	7.32	3.70	1.99
	26	10.19	5.61	2.90

All times on a 2.66 MHz i7 Processor.

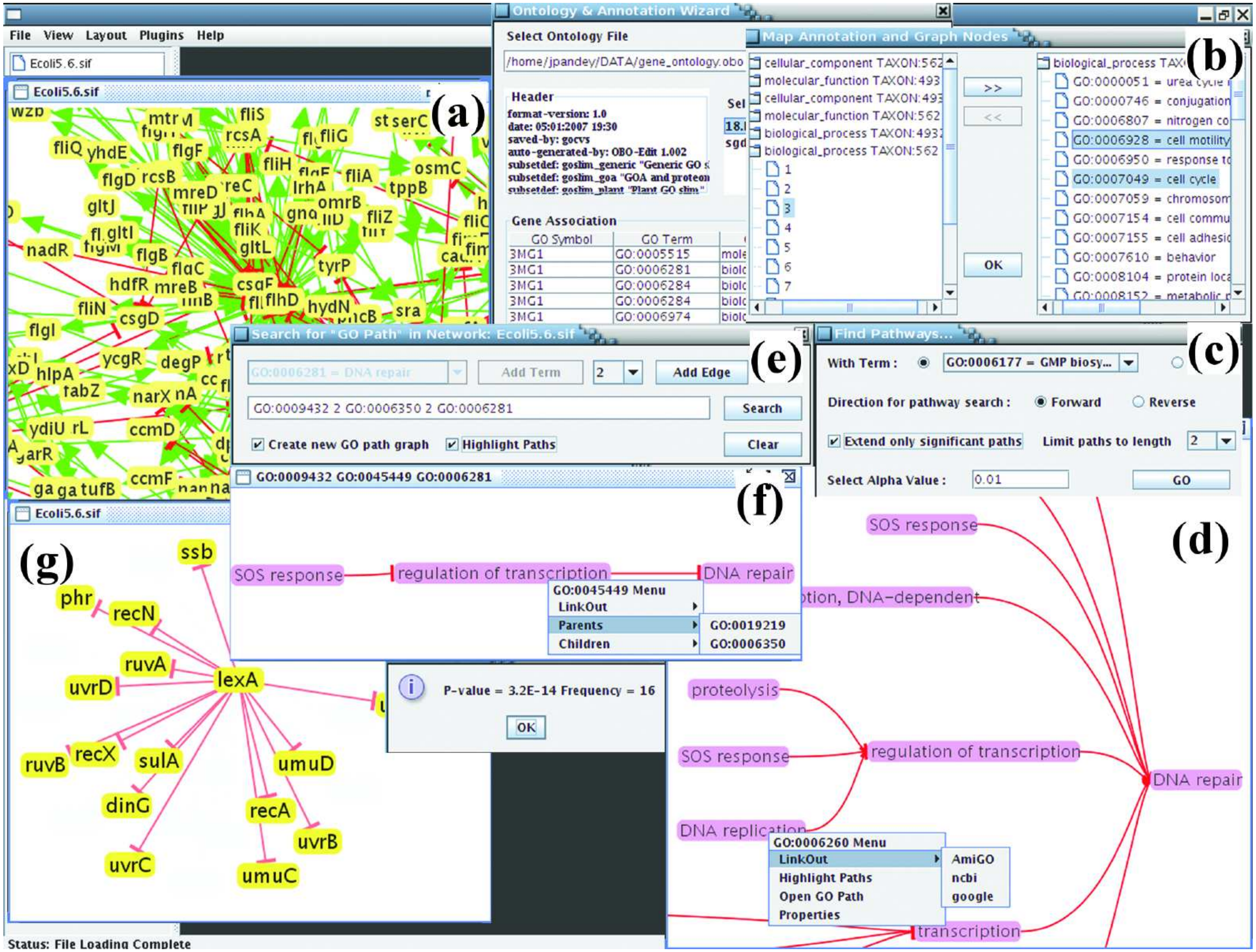
Functional Annotation of Pathways

- Identifying Significant Pathways
- Annotations and Metrics
- Application to Protein/Domain Interaction Networks
- Implementation and Results

Node Annotation

- Node annotation is in the form of an ontology.
- For instance, Gene Ontology provides a library of molecular annotations (we refer to each annotation class as a functional attribute).

Narada Functionality


File View Layout Plugins Help
Ecoli5.6.sif
(a)
wzb mtr fliS fliQ yhdE rcsA fliG st serC flgF fliH osmC flgD rcsB fliA mreD reC lrhA tppB gltJ flhA gnq omrB fliZ nadR fliK fliC flgB gltL fim flaC tyrP hdfR mreB csgF fliN flhD hydN sra csgD flgI ycgR degP hlpA tabZ narX nA ydiU rL ccmD yarR ga tufB ccmF
Ontology & Annotation Wizard
Map Annotation and Graph Nodes
(b)
>>
<<
OK
Find Pathways...
(c)
GO
(d)
Search for "GO Path" in Network: Ecoli5.6.sif
(e)
Add Term
Add Edge
Search
Clear
(f)
SOS response
regulation of transcription
DNA repair
GO:0045449 Menu
LinkOut
Parents
Children
GO:0019219
GO:0006350
Ecoli5.6.sif
(g)
ssb phr recN ruvA uvrD lexA ruvB recX sulA umuD dinG recA uvrB uvrC umuC
OK
SOS response
tion, DNA-dependent
proteolysis
SOS response
regulation of transcription
DNA repair
DNA replication
GO:0006260 Menu
LinkOut
Highlight Paths
Open GO Path
Properties
AmiGO
ncbi
google
transcription

Select Ontology File

/home/jpandey/DATA/gene_ontology.obo

Header

format-version: 1.0
date: 05:01:2007 19:30
saved-by: gocvs
auto-generated-by: OBO-Edit 1.002
subsetdef: goslim_generic "Generic GO s
subsetdef: goslim_goa "GOA and proteon
subsetdef: goslim_plant "Plant GO slim"

Gene Association

GO Symbol	GO Term	
3MG1	GO:0005515	mole
3MG1	GO:0006281	biolo
3MG1	GO:0006284	biolo
3MG1	GO:0006284	biolo
3MG1	GO:0006974	biolo

cellular_component TAXON:562
molecular_function TAXON:493
cellular_component TAXON:493
molecular_function TAXON:562
biological_process TAXON:493
biological_process TAXON:562
1
2
3
4
5
6
7

biological_process TAX
GO:0000051 = urea cycle
GO:0000746 = conjugation
GO:0006807 = nitrogen co
GO:0006928 = cell motility
GO:0006950 = response to
GO:0007049 = cell cycle
GO:0007059 = chromosom
GO:0007154 = cell commu
GO:0007155 = cell adhesic
GO:0007610 = behavior
GO:0008104 = protein loca
GO:0008152 = metabolic p

With Term : ● GO:0006177 = GMP biosy... ○

Direction for pathway search : ● Forward ○ Reverse

☑ Extend only significant paths Limit paths to length 2

Select Alpha Value : 0.01

GO:0006281 = DNA repair 2

GO:0009432 2 GO:0006350 2 GO:0006281

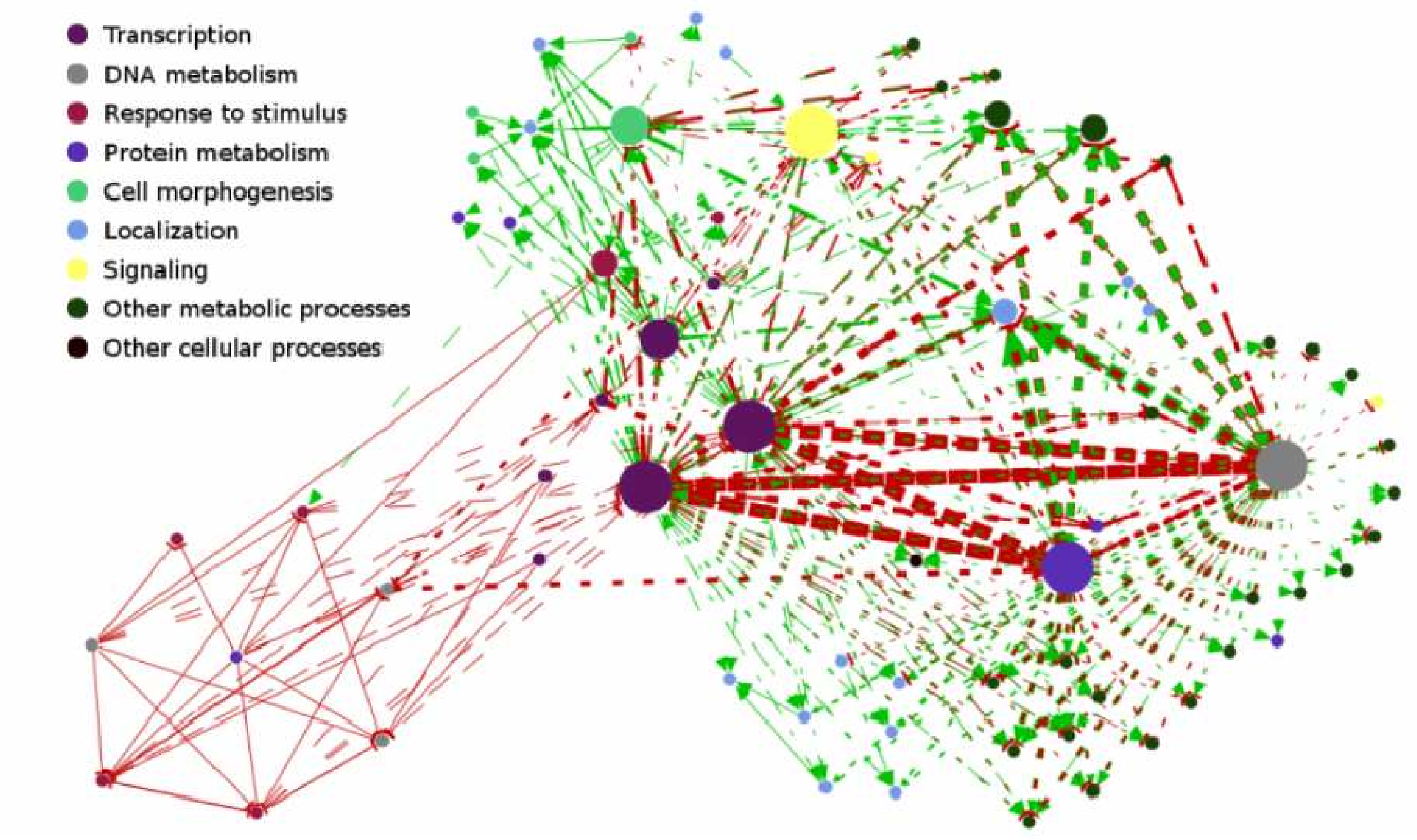
☑ Create new GO path graph ☑ Highlight Paths

GO:0009432 GO:0045449 GO:0006281

P-value = 3.2E-14 Frequency = 16

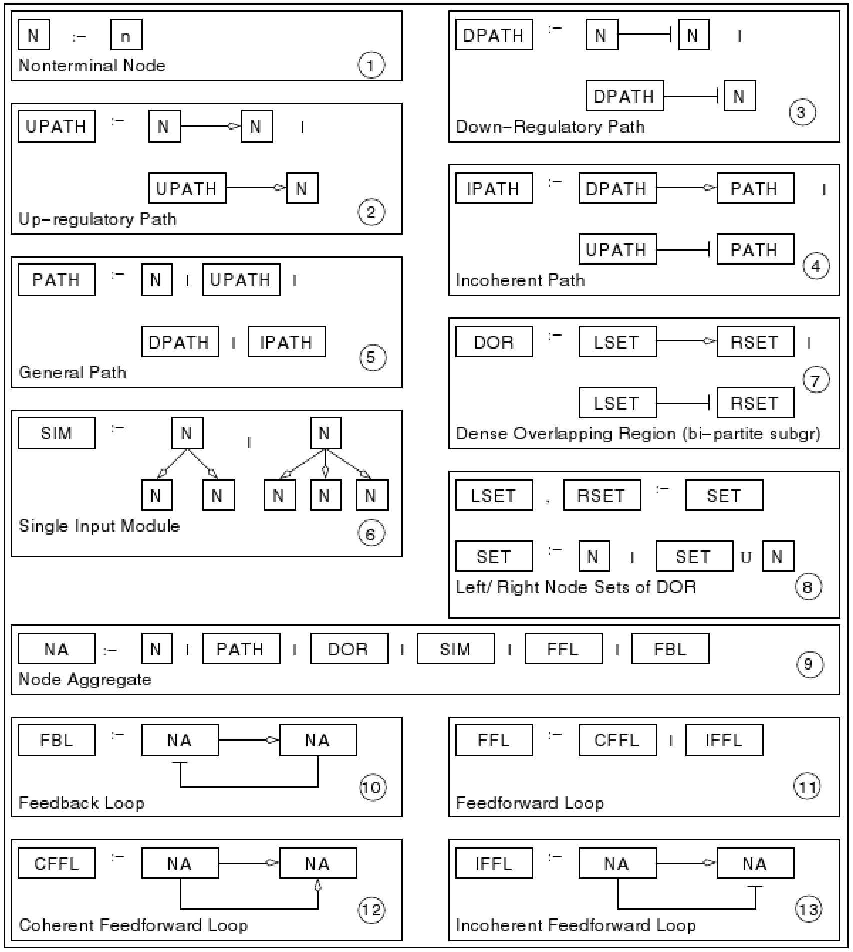
Status: File Loading Complete

Narada Network Annotation



* Parallel implementation ongoing.

Graph Grammars and Parsing



N :- n
Nonterminal Node 1

UPATH :- N → N | UPATH → N
Up-regulatory Path 2

PATH :- N | UPATH | DPATH | IPATH
General Path 5

SIM :- N (N N) | N (N N N)
Single Input Module 6

DPATH :- N ⊣ N | DPATH ⊣ N
Down-Regulatory Path 3

IPATH :- DPATH → PATH | UPATH ⊣ PATH
Incoherent Path 4

DOR :- LSET → RSET | LSET ⊣ RSET
Dense Overlapping Region (bi-partite subgr) 7

LSET , RSET :- SET
SET :- N | SET U N
Left/ Right Node Sets of DOR 8

NA :- N | PATH | DOR | SIM | FFL | FBL
Node Aggregate 9

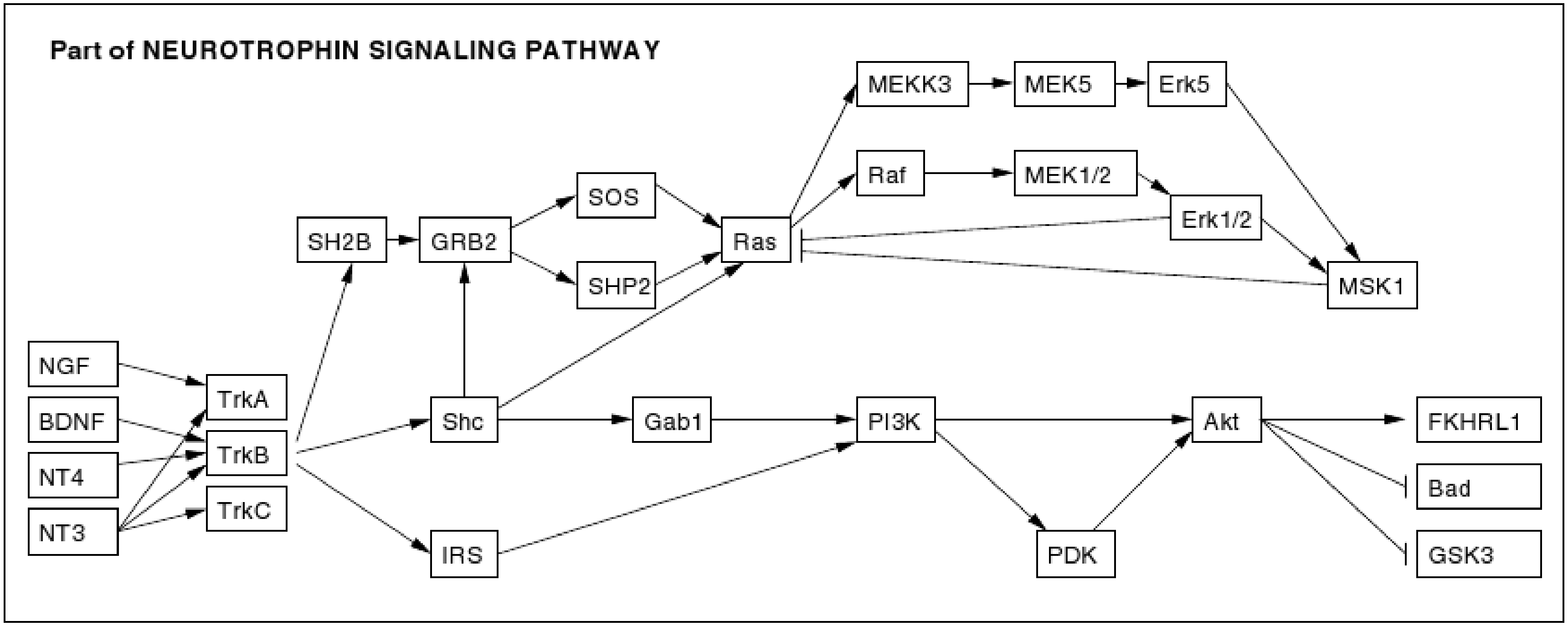
FBL :- NA → NA
Feedback Loop 10

FFL :- CFFL | IFFL
Feedforward Loop 11

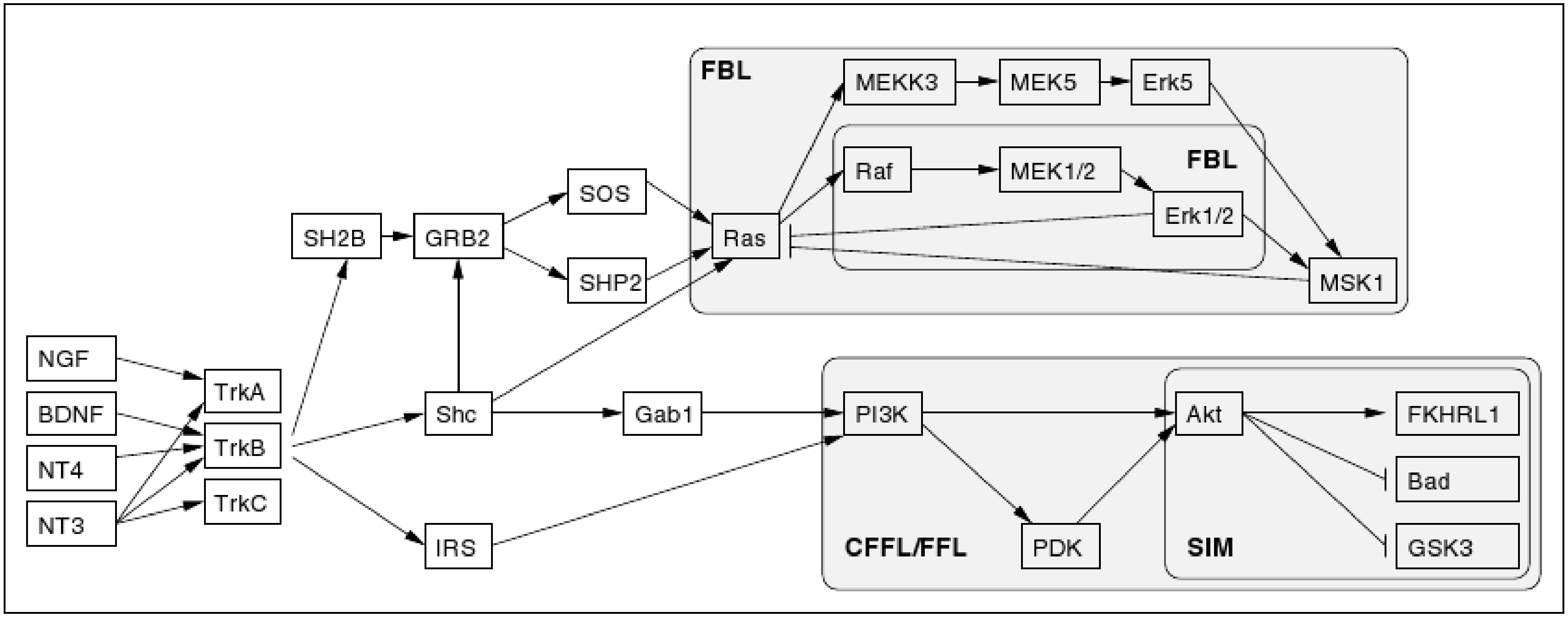
CFFL :- NA → NA
Coherent Feedforward Loop 12

IFFL :- NA → NA
Incoherent Feedforward Loop 13

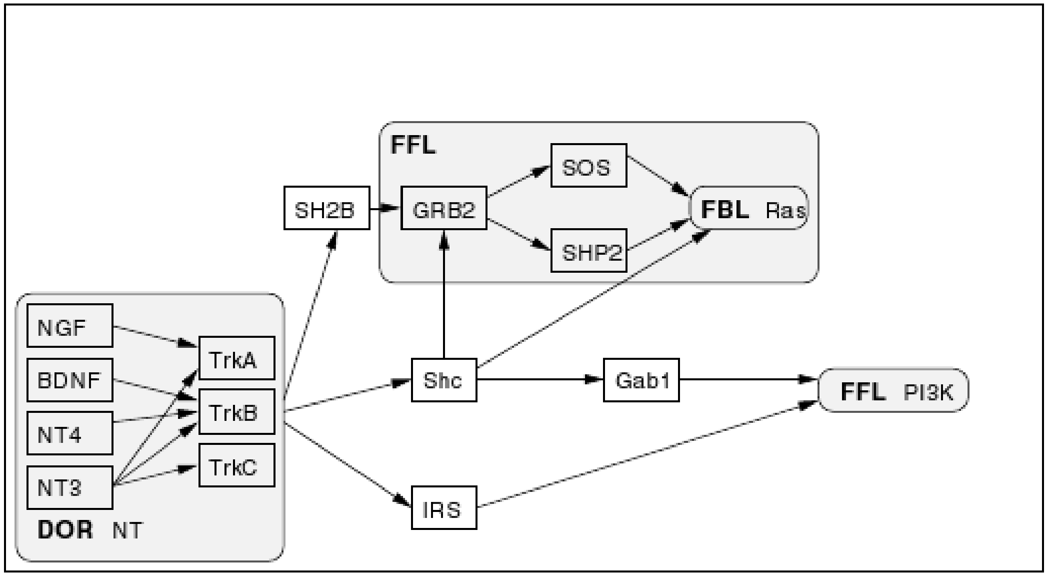
Graph Grammars and Parsing

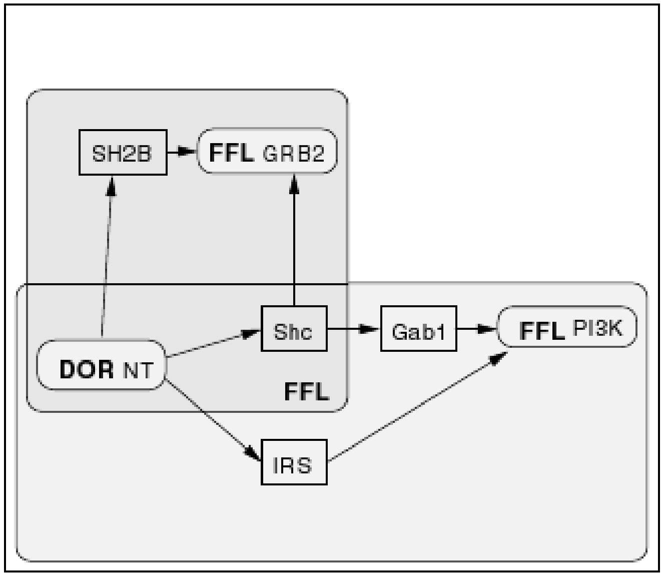


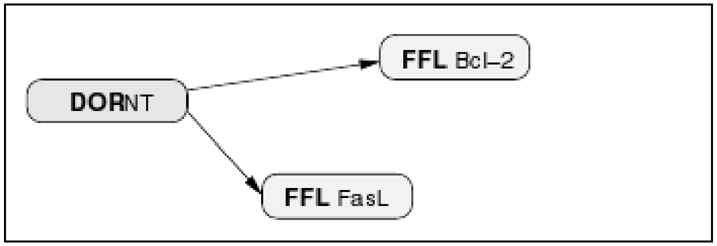
Graph Grammars and Parsing



Graph Grammars and Parsing







Interpretation of Parse:
The input Neurotrophin (sub)pathway primarily consists of two feed–forward loops, feeding from a dense overlapping region. An examination of the literature verifies this to be the case. The first feed–forward loop feeds the Bcl–2 apoptosis regulator and the second feed–forward loop regulates the FASL gene, which is a tumor necrosis factor.

Graph Grammars and Parsing: Status

- Serial parser complete.
- Parallel parser currently in implementation.
- Grammar inference currently under implementation.

Science of Information

“An NSF STC focused on post-Shannon Information theory.“