Using GOALIE to Analyze Time-course Expression Data and Reconstruct Kripke Structures

Marco Antoniotti
Department of Informatics, Systems and Communications
University of Milan Bicocca
ITALY

NYU CMACS NSF PI Meeting, New York, Oct 28-29 2010

Outline

- Interactions between experiments, data and interpretation
- Models of Biological Processes and Systems
  - Description (via controlled vocabularies and ontologies)
  - Reconstruction (via time-course analysis and statistical procedures)
  - Model Repositories
- Computational “Searches” for “models” (parameters, new interactions, etc)
  - Problems
    - Low sampling rate
    - Upsampling, optimization schemes
    - Models limitations

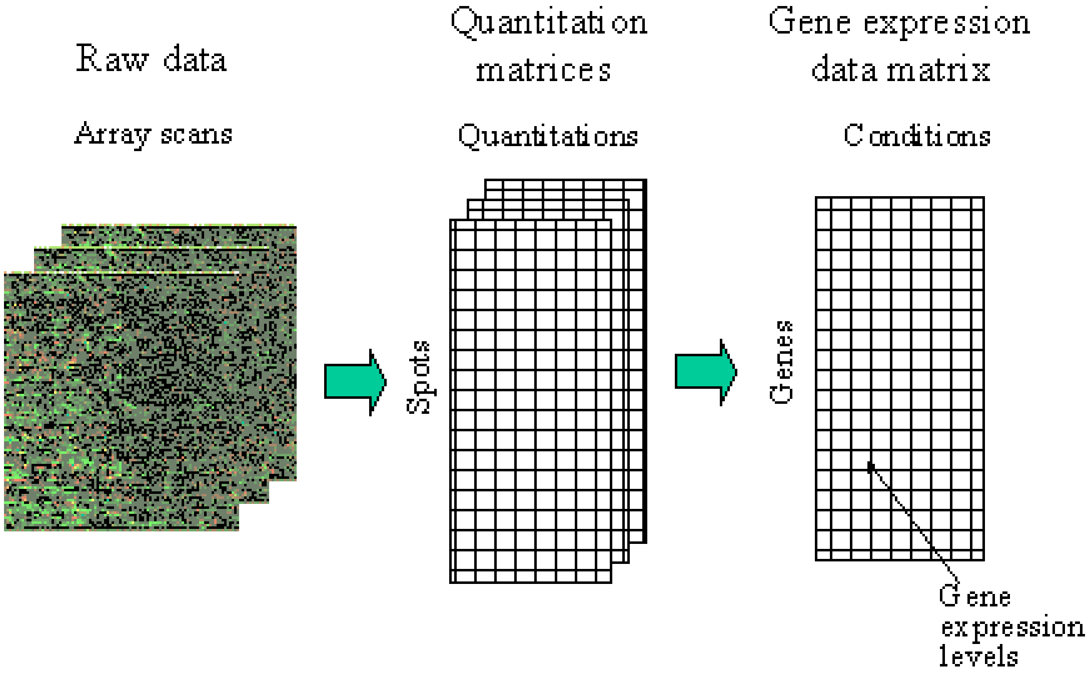
Analyzing Time-course Microarray Experiments

- Microrarray Experiments and Data
- “Enrichment” studies via Controlled Vocabularies and Ontologies (Gene Ontology and others)
- Model “reconstruction”
  - Similarity studies
  - Segmentation algorithms
  - Kernel methods
  - Results
- Future work

- Joint work with Bud Mishra, Courant NYU, Naren Ramakrishnan, Virginia Tech, Daniele Merico, University of Toronto, many others at NYU and UNIMIB

Microarray Experiments

- From laser scans readings, a numerical value corresponding to the relative expression of a “gene” is produced.
- When each raw data array scan corresponds to a given time-point under a specific condition, the final gene expression data matrix represents the temporal evolution of the gene expression.



Standard data-mining approaches to microarray data

- The results of microarray experiments have been studied by means of statistical techniques

- Aim:
  - To group together genes/probes that “behave similarly” under different experimental conditions (usually achieved by *clustering*)

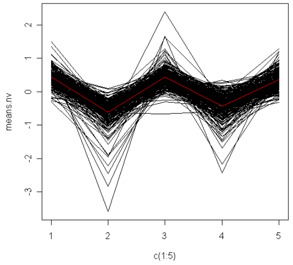
- Successful endeavor
  - Several tools and libraries are provided to perform this kind of studies
  - Several publications produced with results in this field
  - Many of the studies reported still contain a considerable amount of “hand curation”

Standard data-mining approaches to microarray data

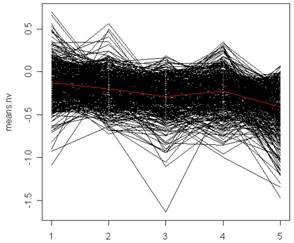
- The expression matrix is usually analyzed according to standard techniques:
  - Clustering

    enables to group together genes with a similar expression profile
  - **Gene Ontology** (GO) terms “Enrichment”

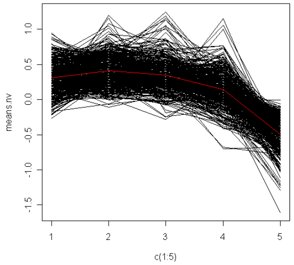
    enables to find statistically over-represented terms in given set of genes - i.e., clusters - thus providing some “functional” characterization
    - usually computed using some *statistical significance test*; e.g., Fisher’s exact test, Hypergeometric Test, Binomial Test, χ^2 Test, plus various corrections



- Ribosome
- Translation

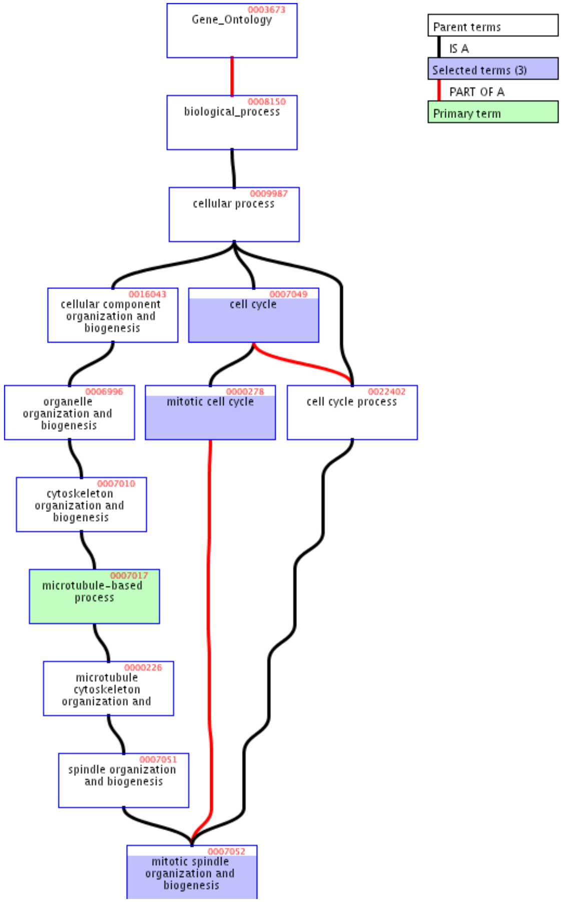


- Spindle
- Cell Wall
- Budding



- Glucose Transport

Gene Ontology (GO)

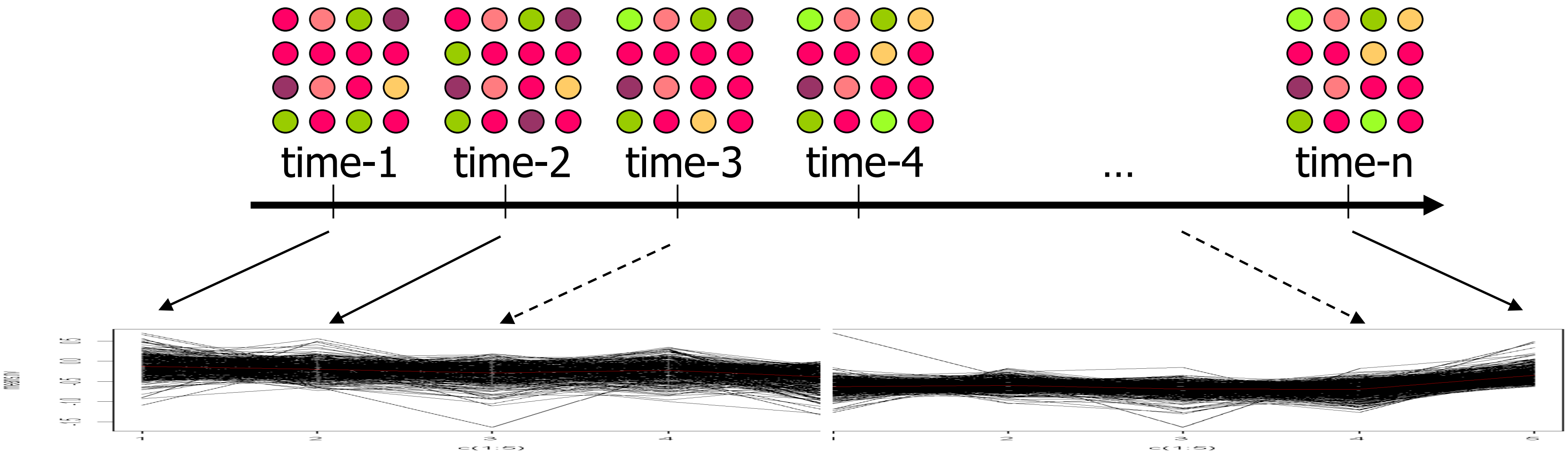

0003673
Gene_Ontology
0008150
biological_process
0009987
cellular process
0016043
cellular component organization and biogenesis
0007049
cell cycle
0006996
organelle organization and biogenesis
0000278
mitotic cell cycle
0022402
cell cycle process
0007010
cytoskeleton organization and biogenesis
0007017
microtubule-based process
0000226
microtubule cytoskeleton organization and
0007051
spindle organization and biogenesis
0007052
mitotic spindle organization and biogenesis
Parent terms
IS A
Selected terms (3)
PART OF A
Primary term

- GO is a controlled vocabulary for the functional annotation of genes
- GO is composed by three independent classifications, each of them having a hierarchical DAG structure
  - **MF**: Molecular Function (biochemical activity and molecule type)
  - **BP**: Biological Process
  - **CC**: Cellular Component

www.geneontology.org

Time-course microarray data

- Clustering is performed with all time-points together spanning the whole time-course



time-1

time-2

time-3

time-4

...

time-n

- This amounts to assume that if genes are co-regulated across some time-points, they will also be co-regulated throughout the whole time-course
- However, co-regulation may be interrupted at a certain point
  - Different short-time and long-time response, e.g., *DNA damage*
  - Multiple-stages transcriptional program, e.g., *development*

GOALIE: a twist on “enrichment” studies

- GOALIE introduces a twist on enrichment studies by taking into account possible temporal variations of biological processes in time-course measurements

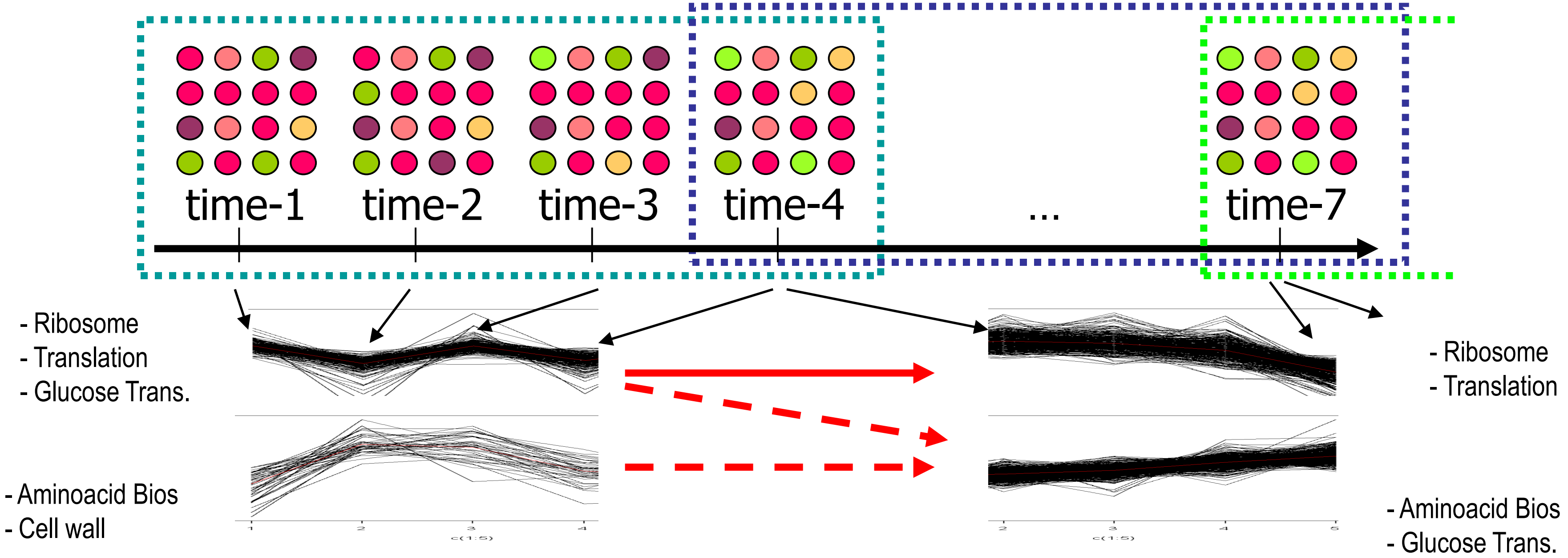
- The key observation is that an “enrichment” of a set of genes/probes may vary depending on the length of the (time) vector of measurements

- GOALIE assumes that the a time-course experiment has been broken down into windows and that each window has been clustered separately

- Afterward the enrichment of each cluster in a window is compared with the enrichment of clusters in neighboring windows and all the possible relations are built in a DAG
  - GOALIE provides several interfaces to explore, summarize and compare the DAGs pertaining to different experiments

Piece-wise approach to time-course microarray data

- We split the time-course into discrete windows,
- Then compute clusters for each window separately,
- Finally reconnect clusters from adjacent windows exploiting similarity of Gene Ontology cluster enrichments



- Ribosome
- Translation
- Glucose Trans.

- Aminoacid Bios
- Cell wall

- Ribosome
- Translation

- Aminoacid Bios
- Glucose Trans.

Computational Modules

- In order to enhance the GOALIE software we concentrated on the components computational modules
- Computational modules are required for:
  1. Clustering (*Clique* [Shamir et al.], K-means, SVM, SOMs etc.; tool *Genesis* from TU-Graz and many other ones)
  2. Segmentation (PNAS 2010 [Ramakrishnan et al.]
  3. Gene Ontology (GO) enrichment (Fisher's exact test etc.)
  4. Computing similarity among clusters from adjacent time-windows, based on GO enrichment (*ex-novo* – Kernel function)
  5. Select only relevant connections among clusters (*ex-novo*)
- In the rest of this presentation, the focus will be on the Kernel approach developed for module #4; #5 has been published in (CaOR 2010 [Antoniotti et al.])

Computing “Similarity” Using Graph Kernels

- The results of the first three steps of the algorithm consist in the “enrichment” of each cluster by a set of representative labels (GO terms)
- Next we want to see how similar two clusters are based on this labeling
- **Note**
  - This check may be useful to a biologist trying to track biological processes over time; e.g., trying to see which genes are involved in a certain process as time evolves
  - From a more abstract point of view this is a procedure that measures how two objects are similar
    - The similarity between the two objects is done in a **re-described** space (possibly with lower dimensionality)
    - In our case there is some more structure we want to exploit

Computing "Similarity" Using Graph Kernels

- Peculiarities of our method
  - Our objects are clusters ordered in a time-course
  - The labeling by GO terms does have a structure imposed by their hierarchical arrangement in a DAG
- Previous work
  - Similarity between objects of this kind is computed using various measures
  - In the specific case of labeling of gene sets, flat lists of symbols were used
    - Similarity computed Jaccard index

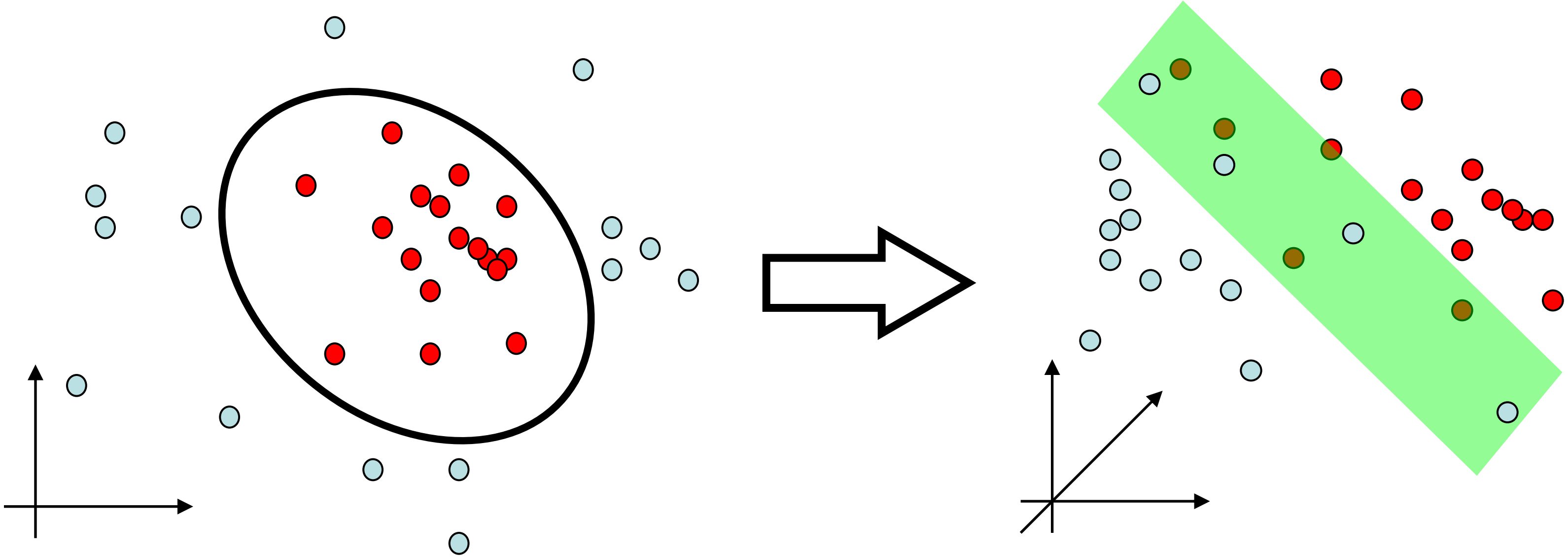
$$J(X,Y) = 1 - \frac{|X \cap Y|}{|X \cup Y|}$$

- Graph kernels can instead be used to take into account the DAG nature of the GO labels
  - Question: what is the performance of our Graph Kernel method w.r.t. a simple Jaccard index calculation?

Kernel Methods

When the existence of a non-linear pattern prevents from using a linear classification algorithm, the problem can be solved introducing a mapping function Φ which projects the problem in a higher dimension space, where the pattern is linear

$$\Phi : R^N \rightarrow R^M \ (M > N)$$



Kernel methods

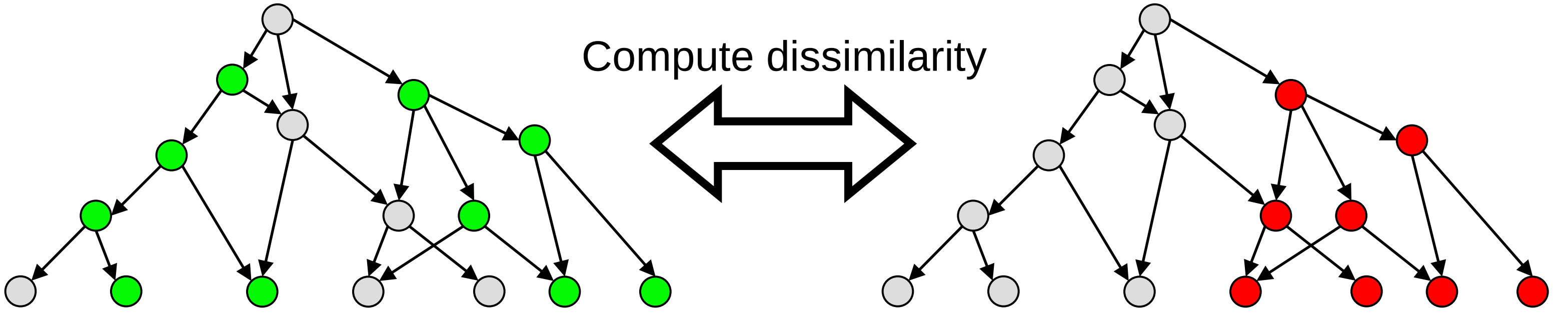
- How to perform the mapping?
  - We don't really have to know the mapping Φ if we introduce a **Kernel function *k***

$$k(x, y) = \langle \phi(x), \phi(y) \rangle_F$$

  - The internal product between the remapped points is compute by *k* thus avoiding the explicit computation of Φ (the so called **Kernel Trick**)
- In order to be a proper Kernel, a function must be positive semi-definite and symmetric (Mercer's Theorem)
- A Kernel function can also be used to induce a dissimilarity function (that's exactly what we do)

A Kernel Function for Gene Ontology Graph Comparison

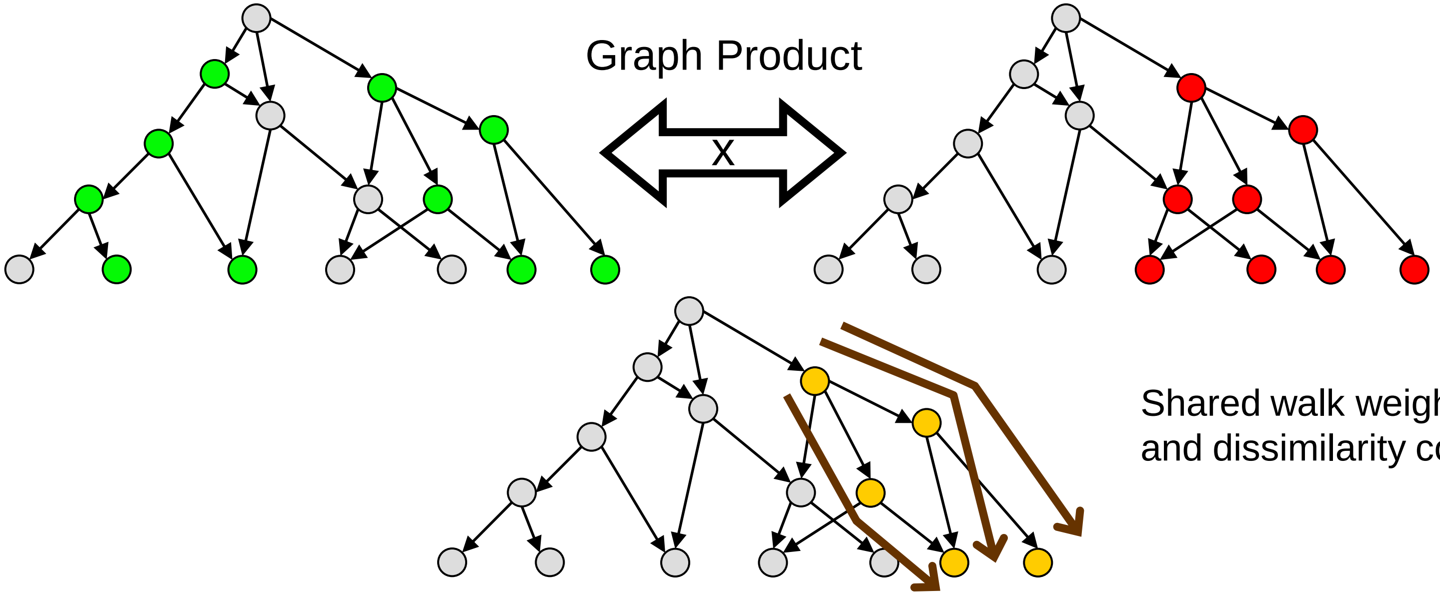
- Input: GO enrichment graph; i.e., sub-graphs of the overall GO taxonomy for each cluster
  - Each vertex is identified by a label - the GO term name - which is then used for walk matching
  - Each vertex has also an associated *p*-value label, from Fisher's exact test, which is then used to compute a dissimilarity score between the walks
    - We work on GO sub-graphs (forests), obtained by filtering in only the terms with *p*-value < *significance threshold*



Colored dots represent GO terms with p-value < significance threshold

A Kernel Function for Gene Ontology Graph Comparison

- The computation (informally) proceeds in the following way
  1. We compute the (direct) graph product between the two GO sub-graphs
  2. We identify common walks in the product GO sub-graph
  3. We compute a weighted dissimilarity score for each walk
  4. We sum all the walk dissimilarities to get the total dissimilarity



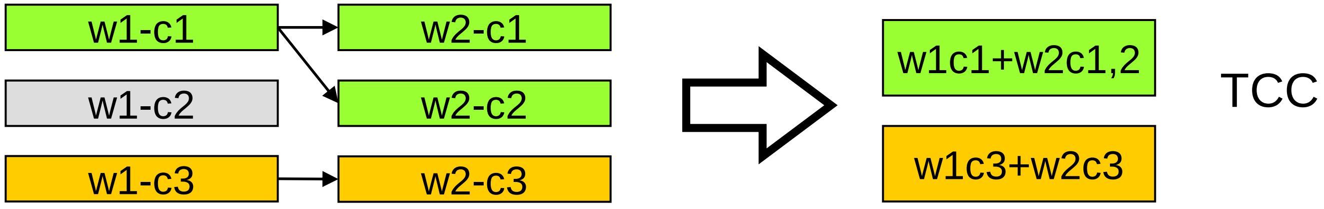
Graph Product

Shared walk weighting and dissimilarity comp.

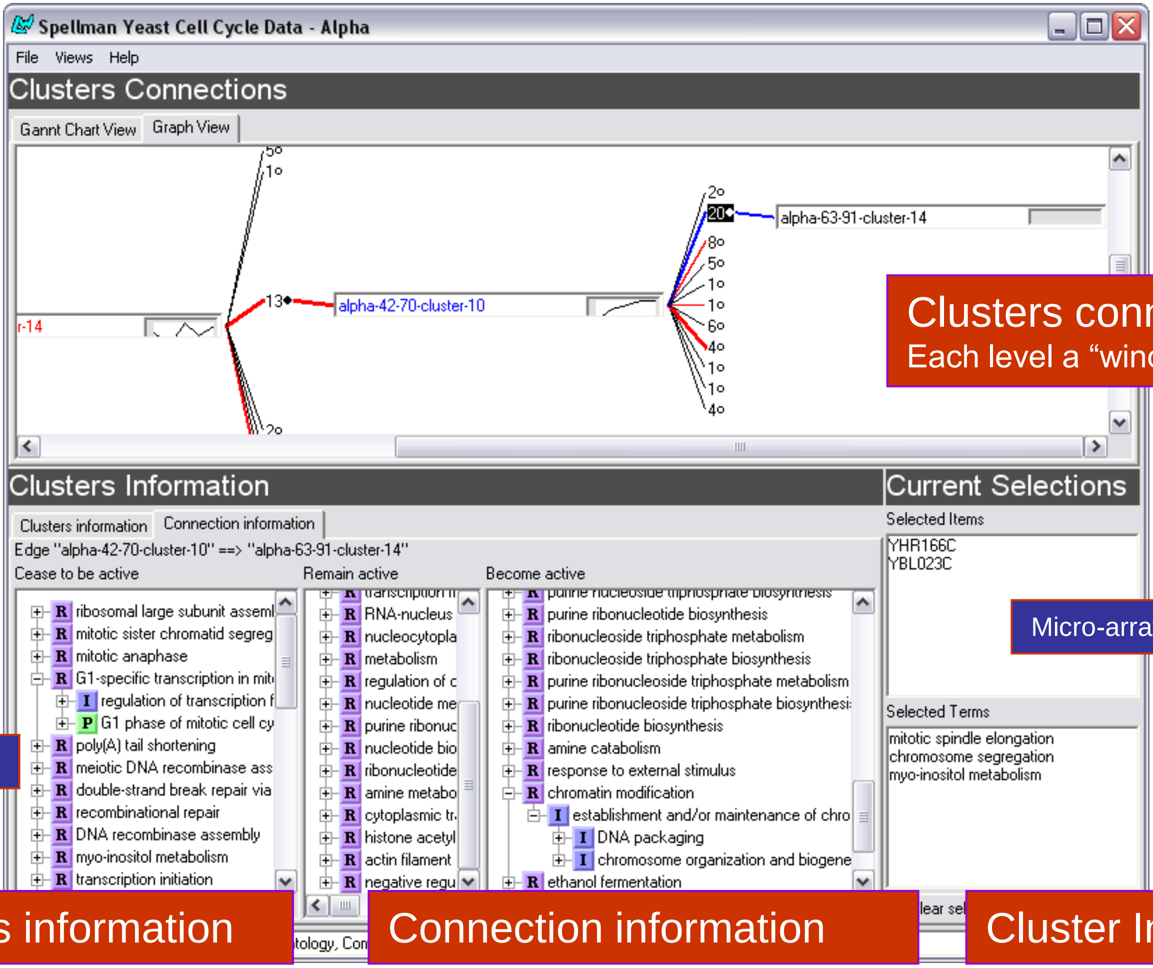
A Kernel function for Gene Ontology graph comparison

- What are the advantages of our approach?
  - We explicitly take into account the hierarchical structure of GO cluster enrichments (Zoppis et al. 07 ISBRA)

- Next we concentrated on evaluating our approach
  - For a benchmark for our Kernel function we set up a comparison with a Jaccard Coefficient-based dissimilarity, working on GO enrichments as flat lists of terms
    - Once the dissimilarities are computed with both methods, we select only significant similarity patterns among clusters from adjacent windows (*)
  - We also consider a model manually curated by an expert
  - To quantitatively assess performance, we adopt the Loganantharaj et al (BMC Bioinformatics, 2006) **Total Cluster Cohesiveness** (TCC) score, which enables to assess the homogeneity of a cluster in terms of its GO terms; we compute TCC for groups of connected clusters (Merico et al. 07 KES-WIRN)



GOALIE Interface



Spellman Yeast Cell Cycle Data - Alpha

Clusters Connections

Clusters Information

Edge "alpha-42-70-cluster-10" ==> "alpha-63-91-cluster-14"

Cease to be active
- R ribosomal large subunit asseml
- R mitotic sister chromatid segreg
- R mitotic anaphase
- R G1-specific transcription in mit
  - I regulation of transcription f
  - P G1 phase of mitotic cell cy
- R poly(A) tail shortening
- R meiotic DNA recombinase ass
- R double-strand break repair via
- R recombinational repair
- R DNA recombinase assembly
- R myo-inositol metabolism
- R transcription initiation

Remain active
- R transcription in
- R RNA-nucleus
- R nucleocytopla
- R metabolism
- R regulation of c
- R nucleotide me
- R purine ribonuc
- R nucleotide bio
- R ribonucleotide
- R amine metabo
- R cytoplasmic tr
- R histone acetyl
- R actin filament
- R negative regu

Become active
- R purine nucleoside triphosphate biosynthesis
- R purine ribonucleotide biosynthesis
- R ribonucleoside triphosphate metabolism
- R ribonucleoside triphosphate biosynthesis
- R purine ribonucleoside triphosphate metabolism
- R purine ribonucleoside triphosphate biosynthesi
- R ribonucleotide biosynthesis
- R amine catabolism
- R response to external stimulus
- R chromatin modification
  - I establishment and/or maintenance of chro
    - I DNA packaging
    - I chromosome organization and biogene
- R ethanol fermentation

tology, Con

Current Selections

Selected Items

YHR166C
YBL023C

Selected Terms

mitotic spindle elongation
chromosome segregation
myo-inositol metabolism

lear sel

Clusters connection tree

Each level a “window”

Micro-array accessions

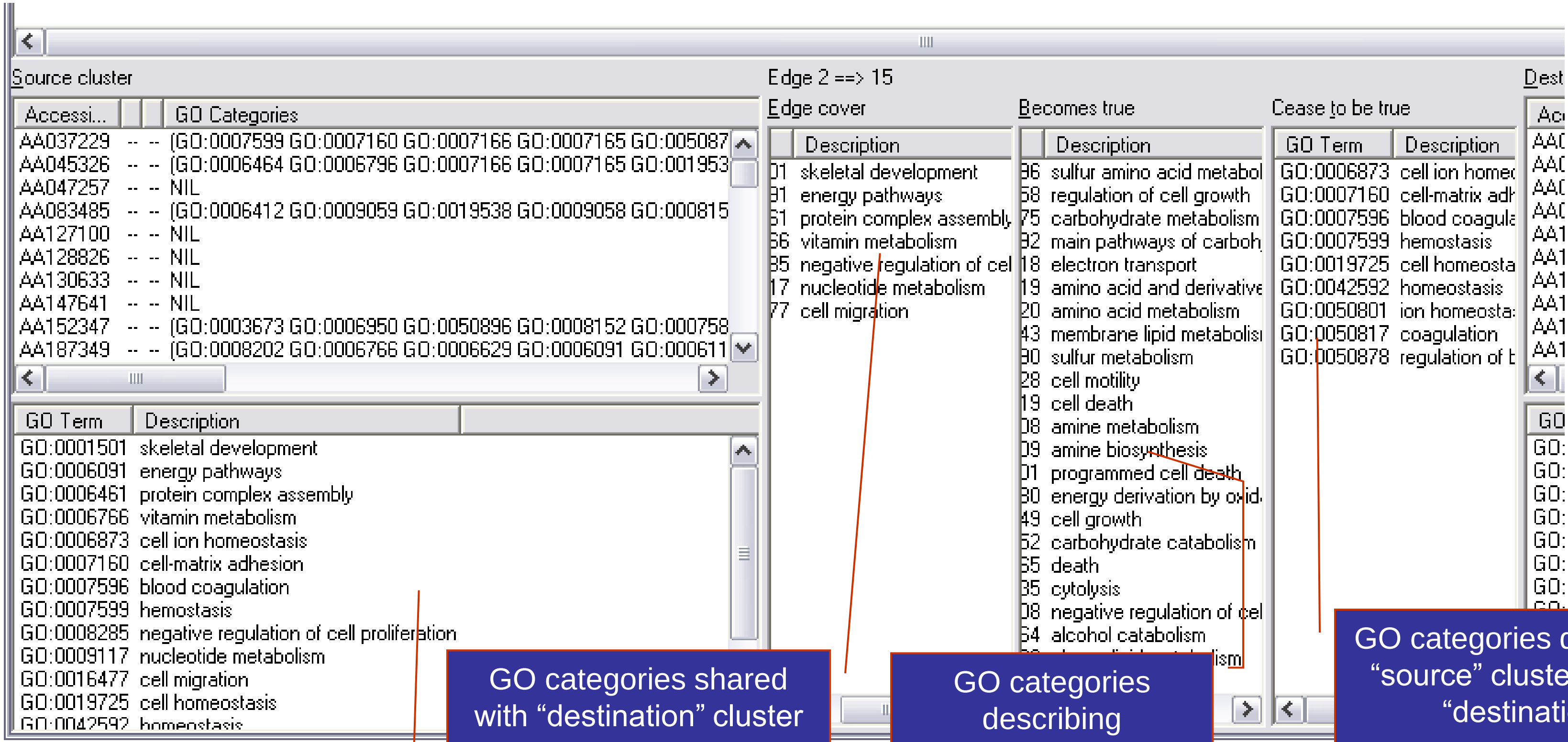
GO categories

Clusters information

Connection information

Cluster Information

GOALIE Interface



Source cluster

Accessi...		GO Categories
AA037229	-- --	(GO:0007599 GO:0007160 GO:0007166 GO:0007165 GO:005087
AA045326	-- --	(GO:0006464 GO:0006796 GO:0007166 GO:0007165 GO:001953
AA047257	-- --	NIL
AA083485	-- --	(GO:0006412 GO:0009059 GO:0019538 GO:0009058 GO:000815
AA127100	-- --	NIL
AA128826	-- --	NIL
AA130633	-- --	NIL
AA147641	-- --	NIL
AA152347	-- --	(GO:0003673 GO:0006950 GO:0050896 GO:0008152 GO:000758
AA187349	-- --	(GO:0008202 GO:0006766 GO:0006629 GO:0006091 GO:000611

GO Term	Description
GO:0001501	skeletal development
GO:0006091	energy pathways
GO:0006461	protein complex assembly
GO:0006766	vitamin metabolism
GO:0006873	cell ion homeostasis
GO:0007160	cell-matrix adhesion
GO:0007596	blood coagulation
GO:0007599	hemostasis
GO:0008285	negative regulation of cell proliferation
GO:0009117	nucleotide metabolism
GO:0016477	cell migration
GO:0019725	cell homeostasis
GO:0042592	homeostasis

Edge 2 ==> 15

Edge cover

	Description
01	skeletal development
91	energy pathways
61	protein complex assembly
66	vitamin metabolism
85	negative regulation of cel
17	nucleotide metabolism
77	cell migration

Becomes true

	Description
96	sulfur amino acid metabol
58	regulation of cell growth
75	carbohydrate metabolism
92	main pathways of carboh
18	electron transport
19	amino acid and derivative
20	amino acid metabolism
43	membrane lipid metabolis
90	sulfur metabolism
28	cell motility
19	cell death
08	amine metabolism
09	amine biosynthesis
01	programmed cell death
80	energy derivation by oxid
49	cell growth
52	carbohydrate catabolism
65	death
35	cytolysis
08	negative regulation of cel
64	alcohol catabolism
	lism

Cease to be true

GO Term	Description
GO:0006873	cell ion homeo
GO:0007160	cell-matrix adh
GO:0007596	blood coagula
GO:0007599	hemostasis
GO:0019725	cell homeosta
GO:0042592	homeostasis
GO:0050801	ion homeosta
GO:0050817	coagulation
GO:0050878	regulation of b

Dest

Ac
AA0
AA0
AA0
AA0
AA1
AA1
AA1
AA1
AA1
AA1

GO
GO:
GO:
GO:
GO:
GO:
GO:
GO:

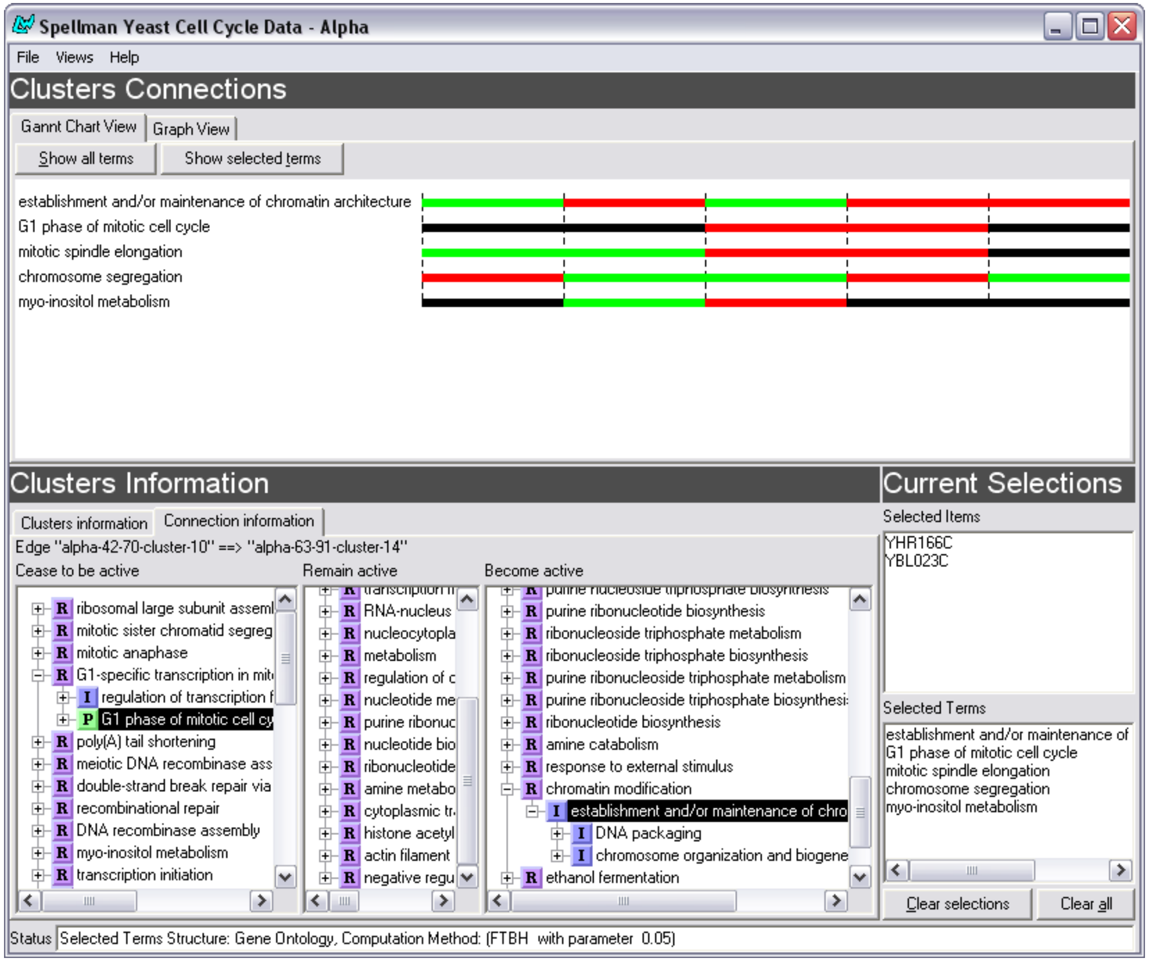
GO categories shared with “destination” cluster

GO categories describing “destination” cluster but not “source”

GO categories describing “source” cluster but not “destination”

GO categories describing genes in “source” cluster

GOALIE Interface



Clusters Connections

establishment and/or maintenance of chromatin architecture
G1 phase of mitotic cell cycle
mitotic spindle elongation
chromosome segregation
myo-inositol metabolism

Clusters Information

Edge "alpha-42-70-cluster-10" ==> "alpha-63-91-cluster-14"

Cease to be active
- R ribosomal large subunit asseml
- R mitotic sister chromatid segreg
- R mitotic anaphase
- R G1-specific transcription in mit
  - I regulation of transcription f
  - P G1 phase of mitotic cell cy
- R poly(A) tail shortening
- R meiotic DNA recombinase ass
- R double-strand break repair via
- R recombinational repair
- R DNA recombinase assembly
- R myo-inositol metabolism
- R transcription initiation

Remain active
- R transcription in
- R RNA-nucleus
- R nucleocytopla
- R metabolism
- R regulation of c
- R nucleotide me
- R purine ribonuc
- R nucleotide bio
- R ribonucleotide
- R amine metabo
- R cytoplasmic tr
- R histone acetyl
- R actin filament
- R negative regu

Become active
- R purine nucleoside triphosphate biosynthesis
- R purine ribonucleotide biosynthesis
- R ribonucleoside triphosphate metabolism
- R ribonucleoside triphosphate biosynthesis
- R purine ribonucleoside triphosphate metabolism
- R purine ribonucleoside triphosphate biosynthesi:
- R ribonucleotide biosynthesis
- R amine catabolism
- R response to external stimulus
- R chromatin modification
  - I establishment and/or maintenance of chro
    - I DNA packaging
    - I chromosome organization and biogene
- R ethanol fermentation

Current Selections

Selected Items

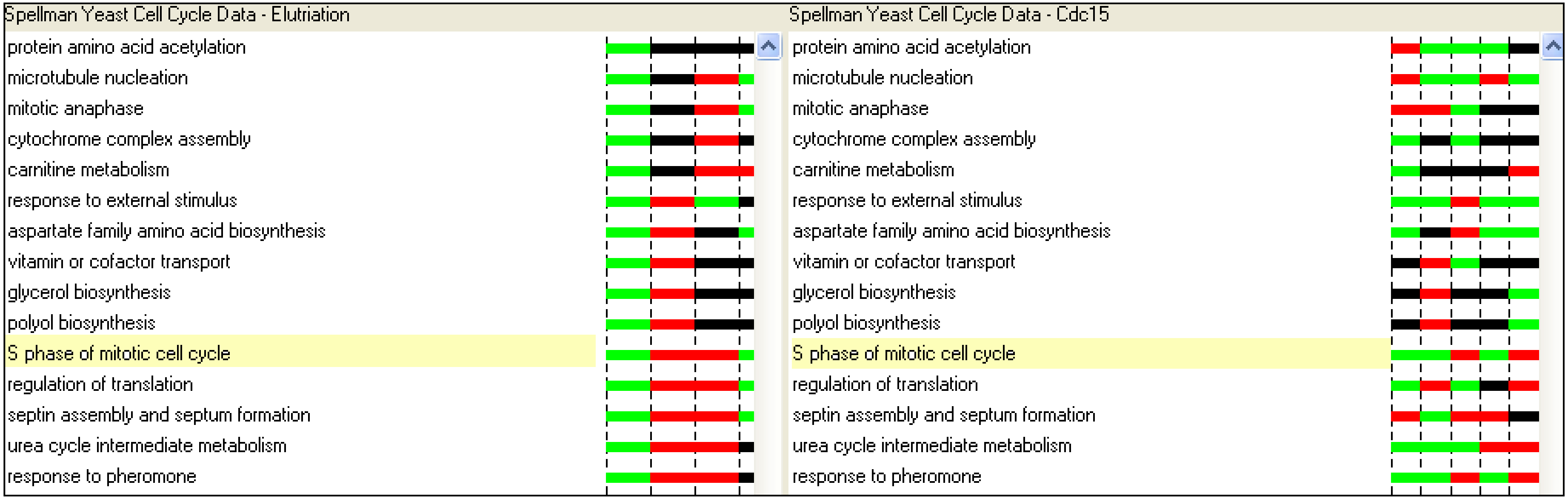
YHR166C
YBL023C

Selected Terms

establishment and/or maintenance of
G1 phase of mitotic cell cycle
mitotic spindle elongation
chromosome segregation
myo-inositol metabolism

Status: Selected Terms Structure: Gene Ontology, Computation Method: (FTBH with parameter 0.05)

GOALIE Interface

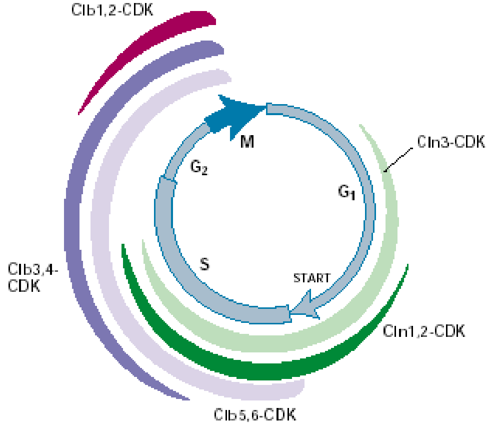


Spellman Yeast Cell Cycle Data - Elutriation				
protein amino acid acetylation	🟩	⬛	⬛	⬛
microtubule nucleation	🟩	⬛	🟥	🟩
mitotic anaphase	🟩	⬛	🟥	🟩
cytochrome complex assembly	🟩	⬛	🟥	⬛
carnitine metabolism	🟩	⬛	🟥	🟥
response to external stimulus	🟩	🟥	🟩	⬛
aspartate family amino acid biosynthesis	🟩	🟥	⬛	🟩
vitamin or cofactor transport	🟩	🟥	⬛	⬛
glycerol biosynthesis	🟩	🟥	⬛	⬛
polyol biosynthesis	🟩	🟥	⬛	⬛
S phase of mitotic cell cycle	🟩	🟥	🟥	🟩
regulation of translation	🟩	🟥	🟥	🟩
septin assembly and septum formation	🟩	🟥	🟥	🟩
urea cycle intermediate metabolism	🟩	🟥	🟥	⬛
response to pheromone	🟩	🟥	🟥	⬛

Spellman Yeast Cell Cycle Data - Cdc15					
protein amino acid acetylation	🟥	🟩	🟩	🟩	⬛
microtubule nucleation	🟥	🟩	🟩	🟥	🟩
mitotic anaphase	🟥	🟥	🟩	⬛	⬛
cytochrome complex assembly	🟩	⬛	🟩	⬛	⬛
carnitine metabolism	🟩	⬛	⬛	⬛	🟥
response to external stimulus	🟩	🟩	🟥	🟩	🟩
aspartate family amino acid biosynthesis	🟩	⬛	🟥	🟩	🟩
vitamin or cofactor transport	⬛	🟥	🟩	⬛	⬛
glycerol biosynthesis	⬛	🟥	⬛	⬛	🟩
polyol biosynthesis	⬛	🟥	⬛	⬛	🟩
S phase of mitotic cell cycle	🟩	🟩	🟥	🟩	🟥
regulation of translation	🟩	🟥	🟩	⬛	🟥
septin assembly and septum formation	🟥	🟩	🟥	🟥	⬛
urea cycle intermediate metabolism	🟩	🟩	🟩	🟥	🟥
response to pheromone	🟩	🟩	🟥	🟩	🟥

GOALIE summary comparison view of two cell cycle experiments

Yeast Cell Cycle benchmark



- Cell Cycle is a multi-stage phenomenon (phases), therefore co-regulation patterns may change across time
  - In [Ramakrishnan et al. 2010] we consider different datasets regarding YCC and Yeast Metabolic Cycle
  - In particular, we consider two windows: G1>S and G2>M>G1
- We use Spellman microarray yeast cell cycle data (1998; a well known benchmark for testing novel analysis tools and methods)
  - CDC15-mutant synchronization
  - ALPHA factor synchronization

Comparison results using KL segmentation

Yeast "Metabolic" Cycle Segmentation Comparison: 8 segments inferred

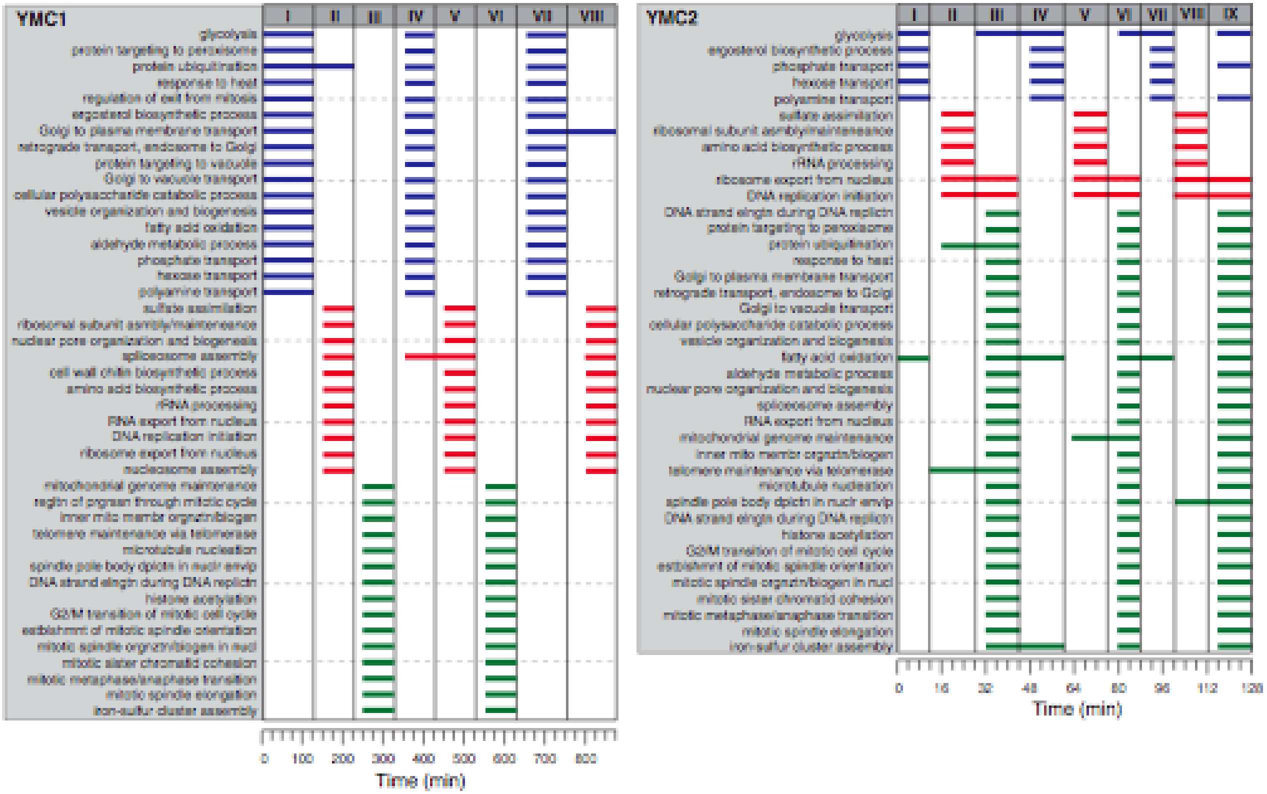
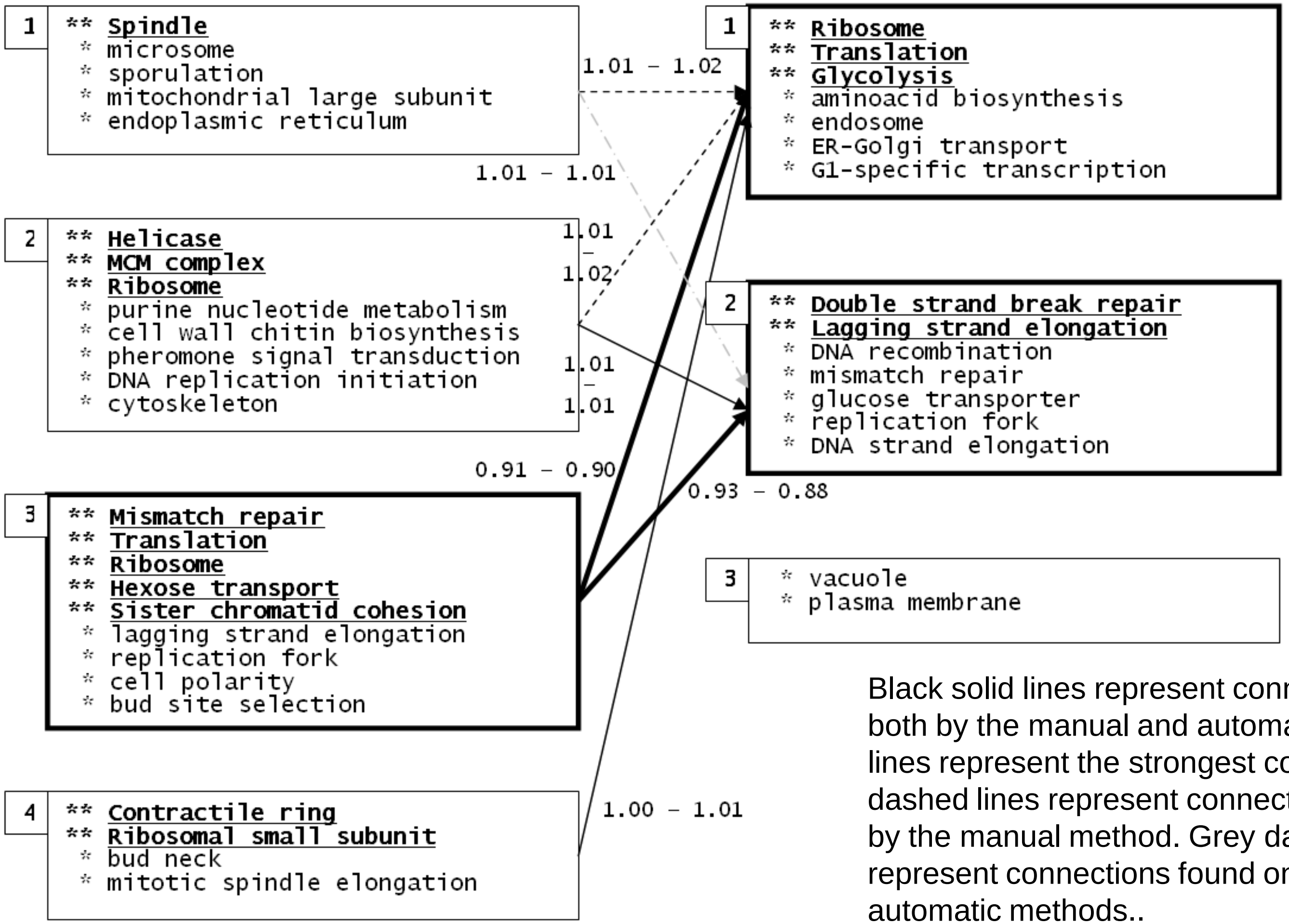


Fig. 4. Segmentation resulting from the GOALIE analysis of transcriptional profiling datasets evaluating the rhythmical growth of *S. cerevisiae* (YMC1: diploid CEN.PK122, nutrient-limited conditions; YMC2: diploid IFO0233, not nutrient limited). The time line of each experiment is shown with each hash mark indicating a sampling point. GOALIE accurately determined the G1, S, and G2/M phases of the cell cycle, respectively. Note that the genes associated with each segment were culture and strain-dependent.

Results



1
** Spindle
* microsome
* sporulation
* mitochondrial large subunit
* endoplasmic reticulum

2
** Helicase
** MCM complex
** Ribosome
* purine nucleotide metabolism
* cell wall chitin biosynthesis
* pheromone signal transduction
* DNA replication initiation
* cytoskeleton

3
** Mismatch repair
** Translation
** Ribosome
** Hexose transport
** Sister chromatid cohesion
* lagging strand elongation
* replication fork
* cell polarity
* bud site selection

4
** Contractile ring
** Ribosomal small subunit
* bud neck
* mitotic spindle elongation

1
** Ribosome
** Translation
** Glycolysis
* aminoacid biosynthesis
* endosome
* ER-Golgi transport
* G1-specific transcription

2
** Double strand break repair
** Lagging strand elongation
* DNA recombination
* mismatch repair
* glucose transporter
* replication fork
* DNA strand elongation

3
* vacuole
* plasma membrane

Inferred cluster connections

Black solid lines represent connections found both by the manual and automatic methods; Bold lines represent the strongest connections. Black dashed lines represent connections found only by the manual method. Grey dash-dotted lines represent connections found only by the automatic methods..

Results

Results overview

- Main results were generated for Alpha subset (2 windows), displaying a substantial convergence between the three methods
  - Numerical results are comparable with Jaccard method
  - Kernel method is more “correct” from the information point of view
  - Kernel method is more computationally intensive
- Preliminary results were also generated for CDC15 subset, displaying a better performance of Kernel over Jaccard

Results (Alpha subset)

Distance	**TCC**	**threshold**
Jaccard	94.28	0.05
Jaccard	92.95	0.01
Jaccard	92.95	0.005
Kernel	92.95	0.01
Kernel	94.63	0.05
Manual	92.27	N/A

Problems

- Low sampling rate: biological experiments usually have a way too low sampling rate
  - Ok for long term observations at equilibrium
  - Not ok for transients and discontinuities detection
    - Assumption: transients and discontinuities are interesting

- Solutions
  - Upsampling after fitting the data to a set of interpolating functions (rational functions or polynomials)
  - Merging of different data sources
    - Several institutions and databanks (e.g., GEO) contain several experiments
    - “Related” experiments can be combined to yield a Virtual Time-Course Experiment that organized the extant corpus of knowledge

Current and future research

- Connection ordering between clusters
  - Method based on optimization of (average) entropy orders connections according to a decrease in the uncertainty of the result graph Kernel similarity between the labeling of two clusters (Antoniotti et al. CaOR 2010)
    - “Complementary” with work on segmentan based on KL divergence published in Ramakrishnan et al. PNAS 2010

- Sample classification (i.e. VTE reconstruction) can be performed if there is an appropriate model of the underlying biological system
  - Ontology research
    - Signs Symptoms Findings Workshop in Milan, 3-4 September 2009

Current and future research

- Temporal Series Reconstruction is a hard problem (deterministically akin to the Traveling Salesman Problem)
  - Bar-Joseph models based on EM optimization procedure
  - Magwene and Kim procedure based on heuristic MST built on top of PQ-trees
  - Lack of data points is a problem

- Prediction Models
  - What happens if we “extend” a time course in the future?

Acknowledgements

- BiMiB Lab, Dipartimento Informatica Sistemistica Comunicazione Milano-Bicocca `bimib.disco.unimib.it`
  - I. Zoppis, M. Carreras, G. Genta, G. Mauri, A. Farinaccio, L. Vanneschi
- Courant Bioinformatics Group New York University
  - S. Kleinberg, A. Sundstrom, A. Witzel, S. Paxia, B. Mishra
- Virginia Tech
  - S. Tadepalli, N. Ramakrishnan
- IFOM, Milan
  - M. Gariboldi, J. Reid, M.Pierotti
- Bader Lab, Donnelly Centre for Cellular and Biomolecular Research, University of Toronto
  - G. Bader, D. Merico
- Virtual Physiological Human Network of Excellence, European Commission FP7
- Regione Lombardia
- National Science Foundation EMT Program
- European Commission Marie Curie Program FP6

Thank you!