Data Integration

Lectures 16 & 17

Lectures Outline

- Goals for Data Integration
- Homogeneous data integration
  - time series data (Filkov et al. 2002)
- Heterogeneous data integration
  - microarray + sequence
  - microarray + protein
  - microarray + location
  - is integration always beneficial?
- Data Integration for Developmental Networks (Davidson et al., 2002)

Integrating data from various experiments should yield better understanding of the data compared to that of individual data sets.

Goals for Data Integration

We integrate data sets with specific goals in mind:

- better gene classification
- better gene clustering
- better regulatory networks

Methods used are the same (modeling):

- SVMs
- Bayesian inference
- Clustering/Classification
- Graph models and algorithms
- Statistical Significance

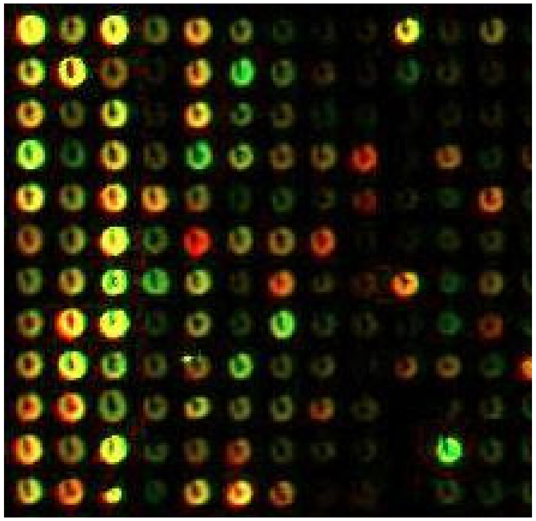
Homogeneous Data

- Expression Data (microarrays)
- Sequence Data
- Location Data (ChIP)
- Protein Expression Data

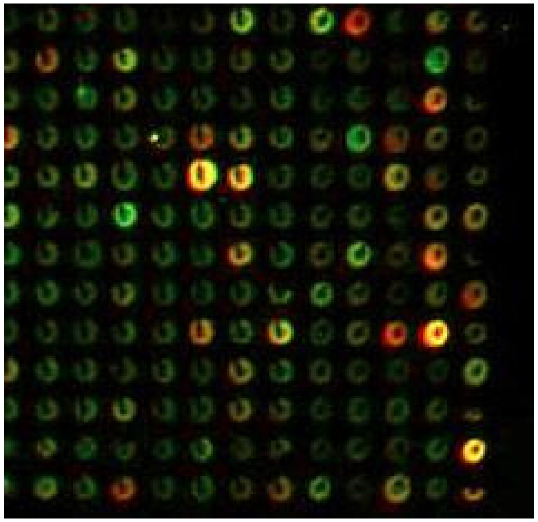
Common platforms for storage, retrieval and comparison across similar data type

Homogeneous Integration

Eg. Microarray expression data is compared across treatments to discover differential gene expression, i.e. genes that behave differently under treatment w.r.t control



treatment



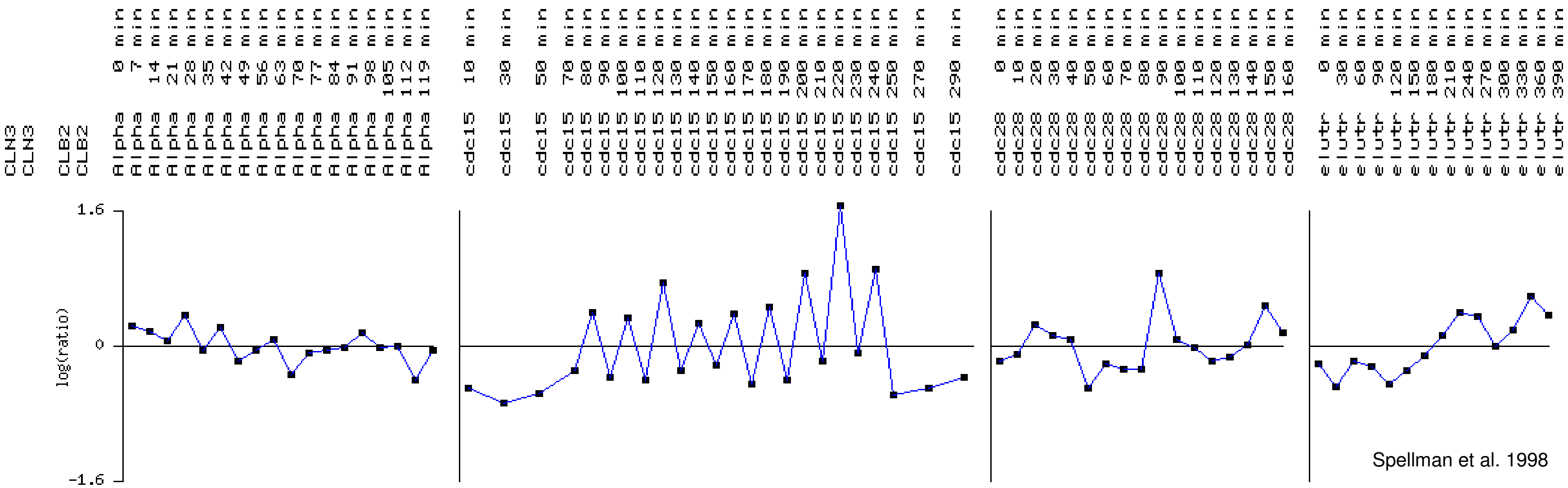
control

Homogeneous Integration: Time-series

(Filkov et al., 2002)

Yeast cells in different experiments are synchronized differently

Plot of ACT1 (YFL039C)



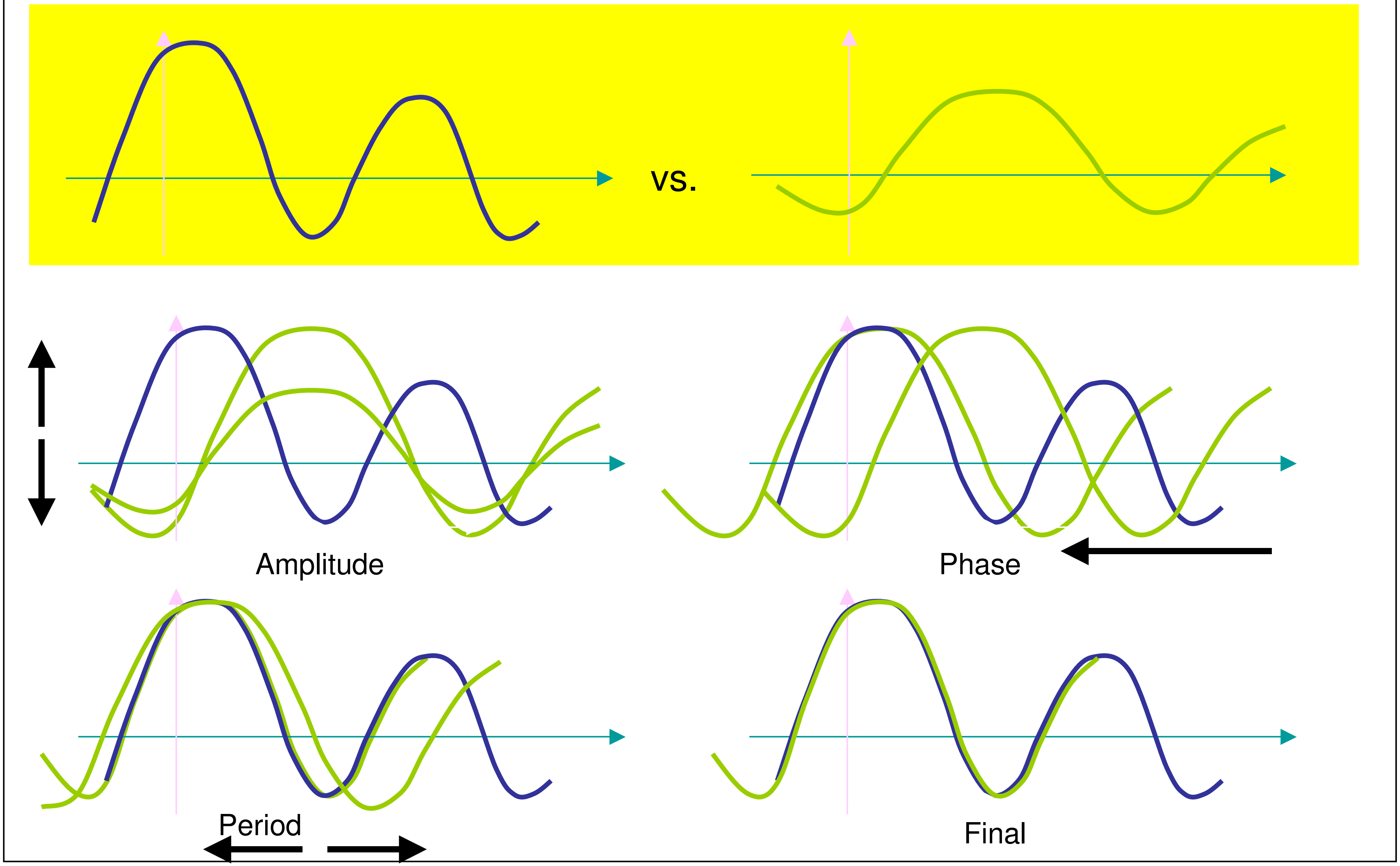
Spellman et al. 1998

How to integrate time-series data?

Warp the curves so that they all have the same

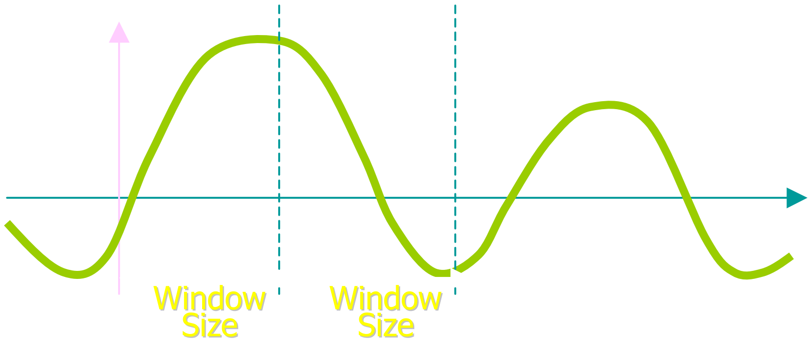
- amplitude
- phase and
- period

Warping Time-series



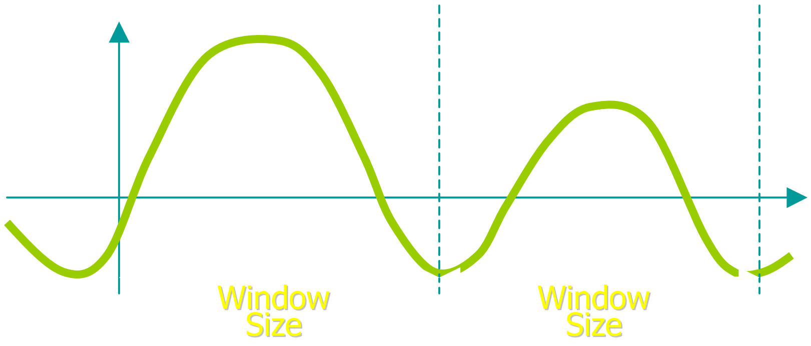
Period and Phase Warp

1. Assume Most Genes are Periodic
2. Perform auto-correlation studies to find period and phase shift
3. Correct for correlation significance in short sequences

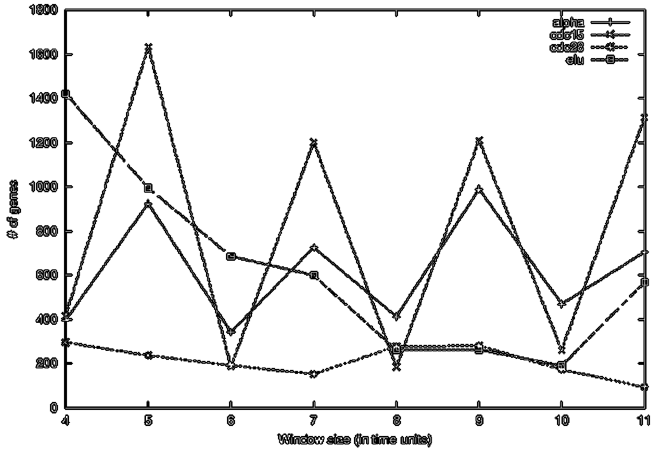


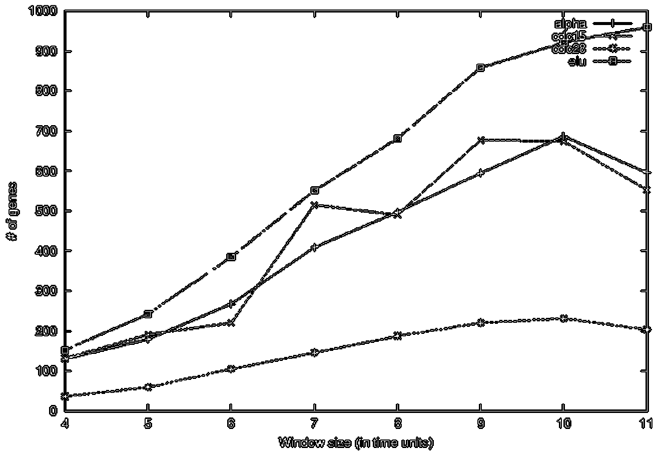
Window length different from cell cycle length => small correlation

Window length equal to cell cycle length => large correlation



After correcting for chance the data sets periods are predicted correctly





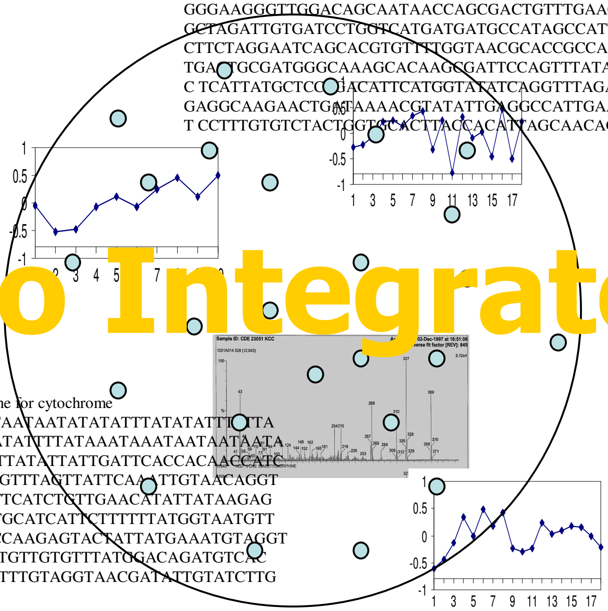
Data Set	Period Observed	Period Detected	Dt	# samples	# full orfs
Alpha	66 ± 11min	70 ± 7min	7	18	3361
Cdc28 (Cho)	90 ± 10min	100 ± 10min	10	17	1188
Cdc15	70 ± 10min	90 ± 10min	10/20	24	3453
elu	---	---	30	14	4753

(phase shift determined similarely...)

Heterogeneous Data Integration

- DNA Sequence
- Microarray
- Proteomics

Important to Integrate!



>gi|12004594|gb|AF217406.1| Saccharomyces cerevisiae uridine nucleosidase (URH1) gene, complete cds
ATGGAATCTGCTGATTTTTTTACCTCACGAAACTTATTAAAACAGATAATTTCCCTCATCTGCAAGGTTG
GGGAAGGGTTGGACAGCAATAACCAGCGACTGTTTGAAGAAAAGATGACTGTTAGTAAAATACCCATATG
GCTAGATTGTGATCCTGGTCATGATGATGCCATAGCCATTTTATTAGGCTGTTTCCATCCAGCTTTCAAT
CTTCTAGGAATCAGCACGTGTTTTGGTAACGCACCGCCAGAGAATACTGACTACAACGCCCGTTCTCTTT
TGA TGCGATGGGCAAAGCACAAGCGATTCCAGTTTATAAAGGCGCACAGAGACCTTGGAAAAGGGAAC
C TCATTATGCTCC GACATTCATGGTATATCAGGTTTAGACGGCACTTCTTTGCTACCTAAGCCAACATTT
GAGGCAAGAACTGAT AAAACGTATATTGAGGCCATTGAAGAGGCGATTCTAGCTAACAATGGAGAGATA
T CCTTTGTGTCTACTGGTGCACTTACCACATTAGCAACAGTTTTTAGGTGTAAACCATACCTAAAAAAATC

>gi|13534|emb|V00696.1|MISC16 Yeast (S. cerevisiae) mitochondrial gene for cytochrome
ATATATATAATTATAAATATATATATATATAATAAGTATTAATTAATAATATATATTTATATATTT TTA
TTAATTAATATATATAAAATATTAGTAATAAATAATATTATTAATATTTTATAAATAAATAATAATAATA
TGGCATTTAGAAAATCAAATGTGTATTTAAGTTTAGTGAATAGTTATATTATTGATTCACCACAACCATC
ATCAATTAATTATTGATGAAATATGGGTTCATTATTAGGTTTATGTTTAGTTATTCAA TTGTAACAGGT
ATTTTTATGGCTATGCATTATTCATCTAATATTGAATTAGCTTTTTCATCTGTTGAACATATTATAAGAG
ATGTGCATAATGGTTATATTTTAAGATATTTACATGCAAATGGTGCATCATTCTTTTTTATGGTAATGTT
TATGCATATGGCTAAAGGTTTATATTATGGTTCATATAGATCACCAAGAGTACTATTATGAAATGTAGGT
GTTATTATTTTCATTTTAACTATTGCTACAGCTTTTTTAGGTTATTGTTGTGTTTATGGACAGATGTCAC
ATTGAGGTGCACTAGTTATTACTAATTTATTCTCAGCAATTCCATTTGTAGGTAACGATATTGTATCTTG

Yeast Genes

Why Does It Pay to Integrate?

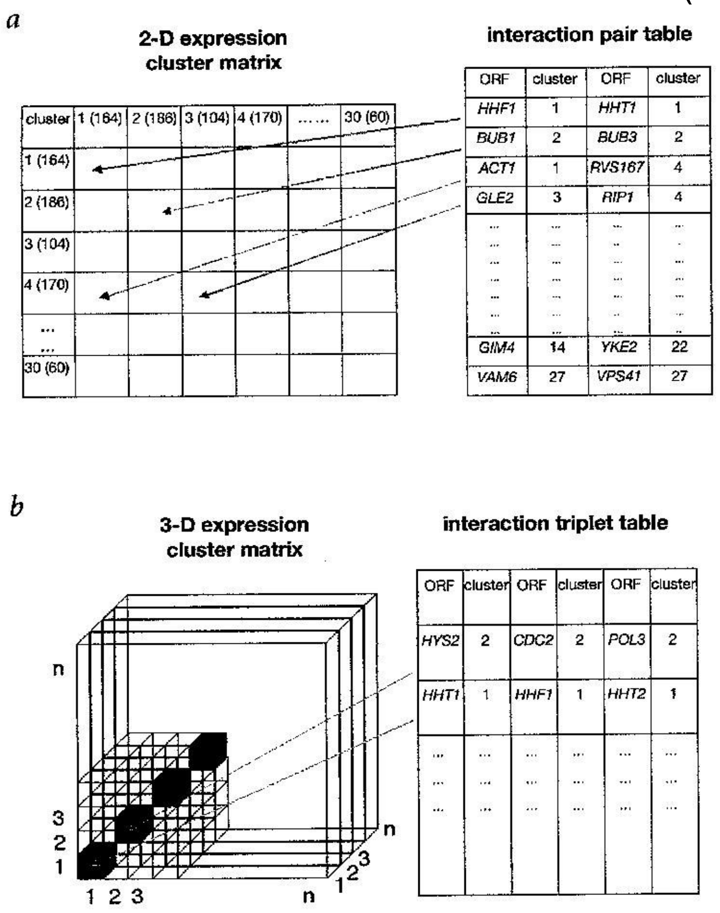
- Gifford, Computational Functional Genomics, Lecture 18
- “Multiple independent constraints can dramatically increase the significance of otherwise elusive effects”
- Dependent vs. Independent

1. Classification

- Simple, intuition based classifications
  - compare to the leukemia classification of Golub et al.,
- Machine Learning classifiers (SVMs)
  - compare to Cristianini et al.

Eg. Gene Expression + Protein Interaction Data

(Ge et al., 2001)



2-D expression cluster matrix

cluster	1 (164)	2 (186)	3 (104)	4 (170)	……	30 (60)
1 (164)						
2 (186)						
3 (104)						
4 (170)						
... ...						
30 (60)						

interaction pair table

ORF	cluster	ORF	cluster
HHF1	1	*HHT1*	1
BUB1	2	*BUB3*	2
ACT1	1	*RVS167*	4
GLE2	3	*RIP1*	4
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
...	...	...	...
GIM4	14	*YKE2*	22
VAM6	27	*VPS41*	27

3-D expression cluster matrix

interaction triplet table

ORF	cluster	ORF	cluster	ORF	cluster
HYS2	2	*CDC2*	2	*POL3*	2
HHT1	1	*HHF1*	1	*HHT2*	1
...	...	...	...	...	...
...	...	...	...	...	...
...	...	...	...	...	...

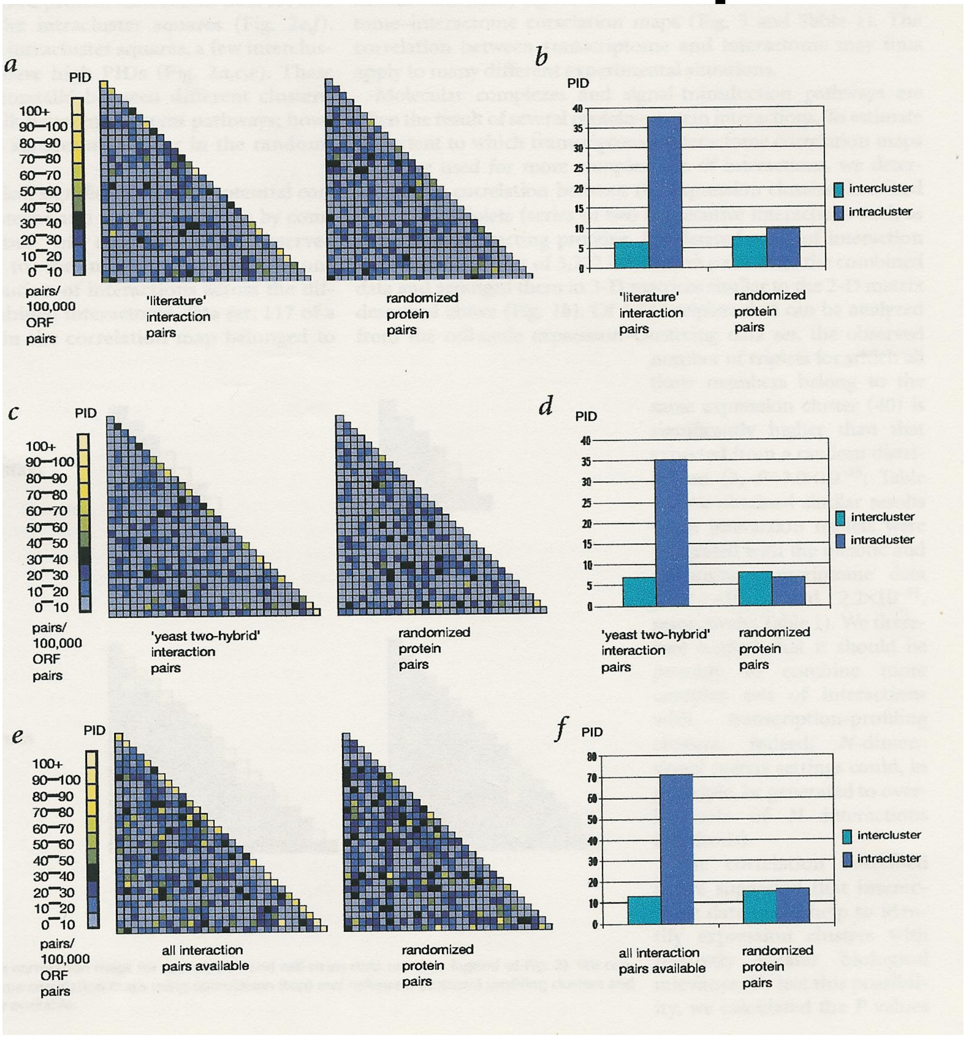
Goals

1. To compare the levels of interaction between proteins encoded by co-expressed genes vs. proteins not encoded by co-expressed genes
2. Improved modeling of protein-protein interactions

Methods

Calculate protein interaction density, and corresponding significance within and between co-expressed clusters of genes

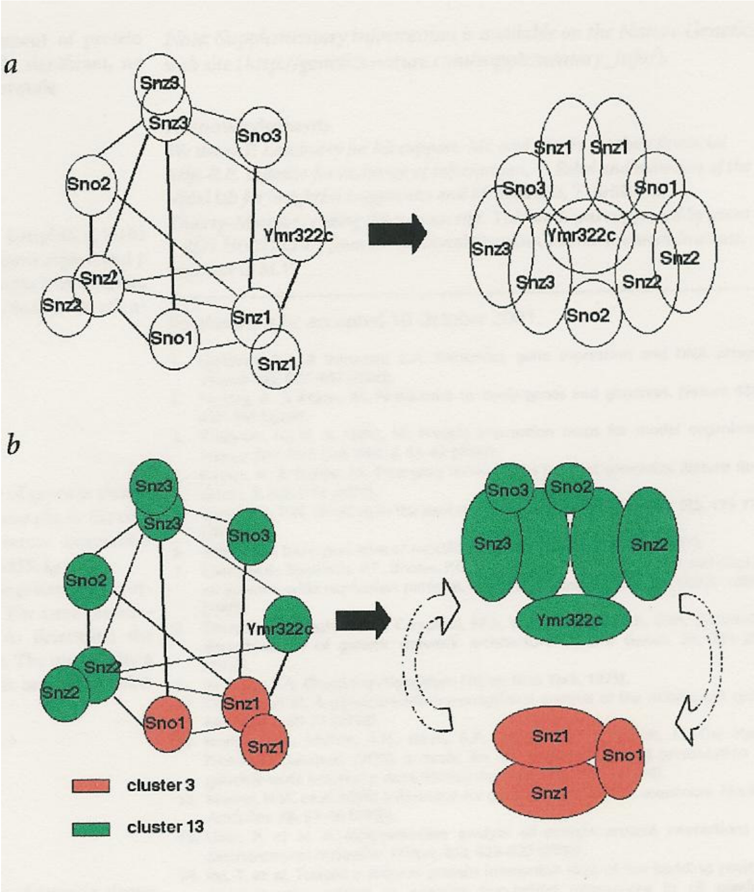
Transcriptome – Interactome Correlation Maps



More Knowledge Yields Better Models

a) Protein-protein interaction data

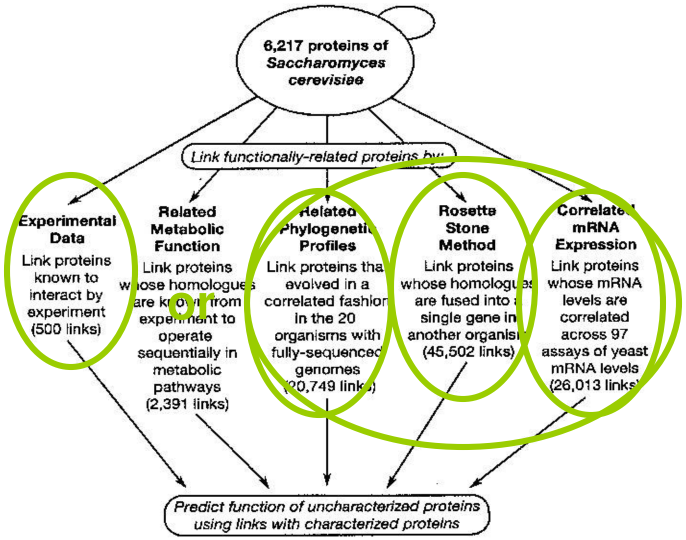
b) Protein Interaction + Gene Expression Data



Stress response proteins

Eg. Protein Function Prediction

(Marcotte et al., 1999)


6,217 proteins of *Saccharomyces cerevisiae*
Link functionally-related proteins by:
Experimental Data
Link proteins known to interact by experiment (500 links)
Related Metabolic Function
Link proteins whose homologues are known from experiment to operate sequentially in metabolic pathways (2,391 links)
or
Related Phylogenetic Profiles
Link proteins that evolved in a correlated fashion in the 20 organisms with fully-sequenced genomes (20,749 links)
Rosetta Stone Method
Link proteins whose homologues are fused into a single gene in another organism (45,502 links)
Correlated mRNA Expression
Link proteins whose mRNA levels are correlated across 97 assays of yeast mRNA levels (26,013 links)
Predict function of uncharacterized proteins using links with characterized proteins

Combining various strategies to link functionally related proteins. Total: 93750 links

Link confidence:

- highest confidence (4130 links)
- high confidence (19521 links)
- rest

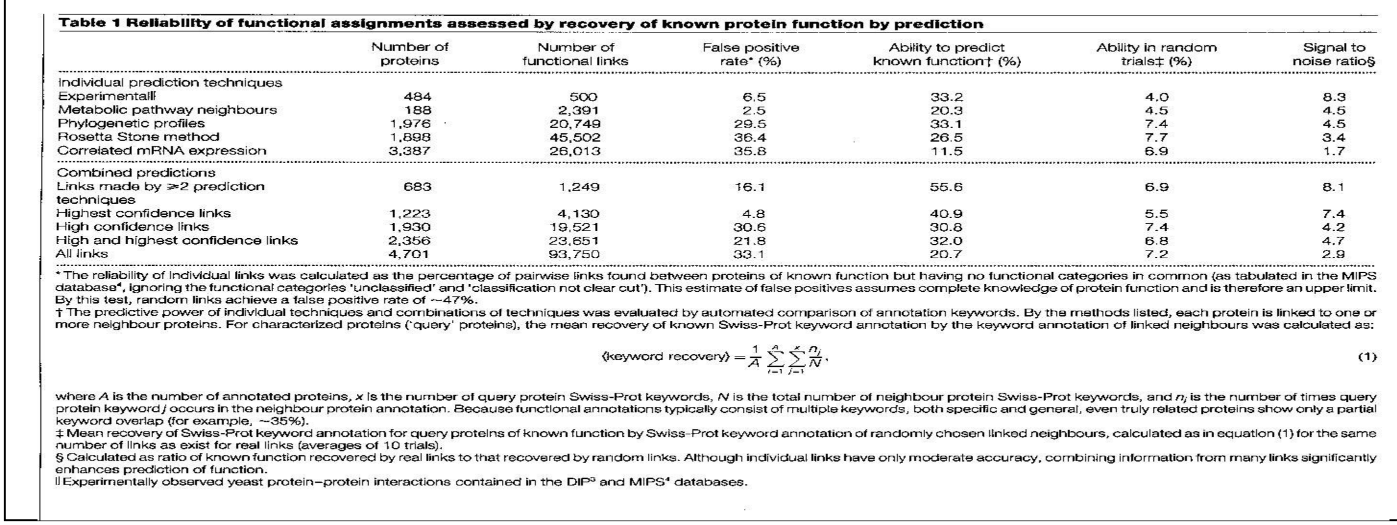


Table 1 Reliability of functional assignments assessed by recovery of known protein function by prediction

	Number of proteins	Number of functional links	False positive rate* (%)	Ability to predict known function† (%)	Ability in random trials‡ (%)	Signal to noise ratio§
Individual prediction techniques						
Experimental‖	484	500	6.5	33.2	4.0	8.3
Metabolic pathway neighbours	188	2,391	2.5	20.3	4.5	4.5
Phylogenetic profiles	1,976	20,749	29.5	33.1	7.4	4.5
Rosetta Stone method	1,898	45,502	36.4	26.5	7.7	3.4
Correlated mRNA expression	3,387	26,013	35.8	11.5	6.9	1.7
Combined predictions						
Links made by ⩾2 prediction techniques	683	1,249	16.1	55.6	6.9	8.1
Highest confidence links	1,223	4,130	4.8	40.9	5.5	7.4
High confidence links	1,930	19,521	30.6	30.8	7.4	4.2
High and highest confidence links	2,356	23,651	21.8	32.0	6.8	4.7
All links	4,701	93,750	33.1	20.7	7.2	2.9

* The reliability of individual links was calculated as the percentage of pairwise links found between proteins of known function but having no functional categories in common (as tabulated in the MIPS database[4], ignoring the functional categories 'unclassified' and 'classification not clear cut'). This estimate of false positives assumes complete knowledge of protein function and is therefore an upper limit. By this test, random links achieve a false positive rate of ~47%.

† The predictive power of individual techniques and combinations of techniques was evaluated by automated comparison of annotation keywords. By the methods listed, each protein is linked to one or more neighbour proteins. For characterized proteins ('query' proteins), the mean recovery of known Swiss-Prot keyword annotation by the keyword annotation of linked neighbours was calculated as:

$$\langle \text{keyword recovery} \rangle = \frac{1}{A} \sum_{i=1}^{A} \sum_{j=1}^{x} \frac{n_j}{N}, \qquad (1)$$

where A is the number of annotated proteins, x is the number of query protein Swiss-Prot keywords, N is the total number of neighbour protein Swiss-Prot keywords, and n_j is the number of times query protein keyword j occurs in the neighbour protein annotation. Because functional annotations typically consist of multiple keywords, both specific and general, even truly related proteins show only a partial keyword overlap (for example, ~35%).

‡ Mean recovery of Swiss-Prot keyword annotation for query proteins of known function by Swiss-Prot keyword annotation of randomly chosen linked neighbours, calculated as in equation (1) for the same number of links as exist for real links (averages of 10 trials).

§ Calculated as ratio of known function recovered by real links to that recovered by random links. Although individual links have only moderate accuracy, combining information from many links significantly enhances prediction of function.

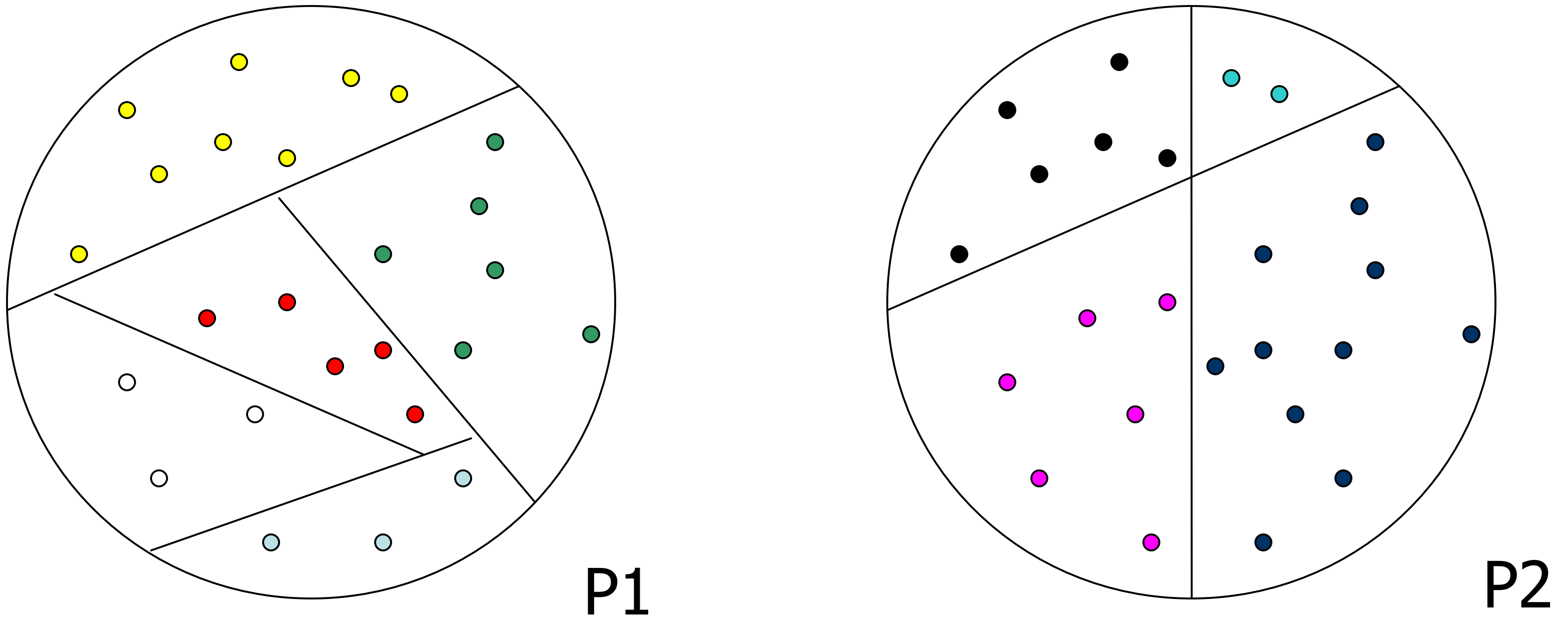
‖ Experimentally observed yeast protein–protein interactions contained in the DIP[3] and MIPS[4] databases.

2. Integrating Clusterings

(Filkov and Skiena, 2003)

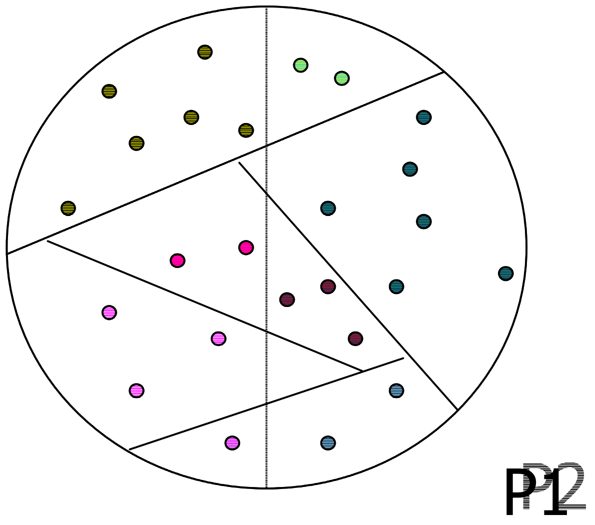
Data sets are usefully summarized as clusterings

- Functional
- Structural
- Data Driven

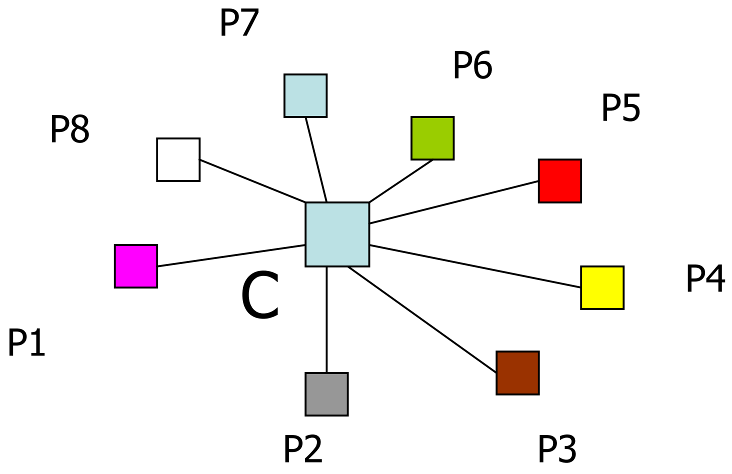


By using multiple clusterings we can learn more, but how?

Overlapping two clusterings is useful, but can we generalize it?



Problem: ***Find a Consensus Clustering that describes the given clusterings well***

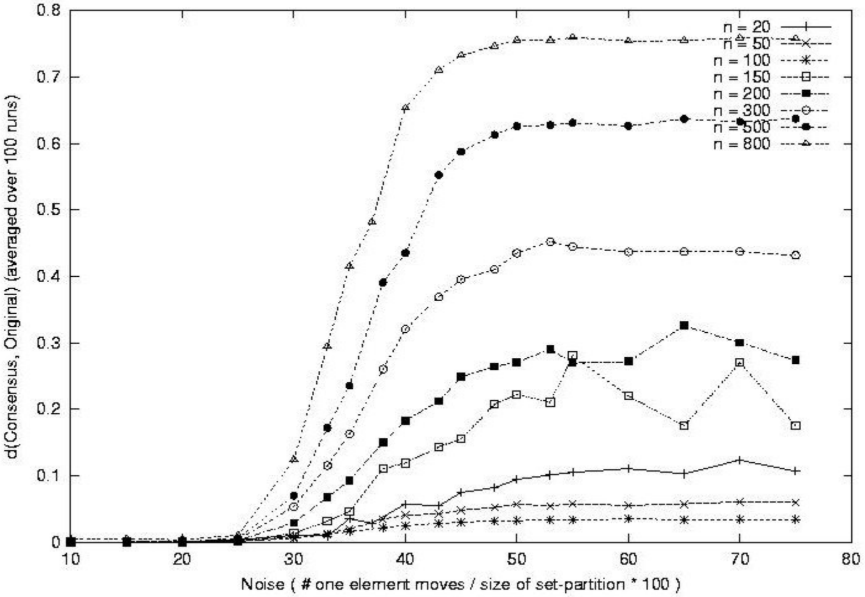


Approach: Integrate the clusterings of data by minimizing the sum of distances between them and a consensus

$$\boxed{\min S = \sum_i d(P_i, C)}$$

Solving the Consensus Problem

- min S consensus is NP-complete even for a very simple distance function (Rand Index)
- Simple Heuristics based on random element move between clusters work well on large data sets
- a measure of benefit of integration



(artificial data)

- Integrating Spellman's Data
  - Alpha, Avg. SoD = 0.1121
  - cdc15, Avg. SoD = 0.1042
  - elu, Avg. SoD = 0.1073
  - Overall, Avg. SoD = 0.107, benefit
- Spellman + Phylogeny = No benefit
- Spellman + Yeast Stress = Benefit

(real data)

3. Gene Network Inference

- Data Integration for Link (Graph) Modeling in General
- Probabilistic Setting
- Each data source is an “expert” proposing a model
- Independent experts: easy (Gifford 2002)
  - independent significances, p_1, p_2
  - combined significance, $p=f(p_1,p_2)$

Graph Models

- Dependent experts (Hartemink et al., 2002)
  - – Joint probability distributions
  - – Bayesian Networks
  - – Model scoring
    - Maximizing a Bayesian scoring function
    - simulated annealing optimizer
    - averaging over high-scoring models
  - – Location+expression data used as <u>priors</u>

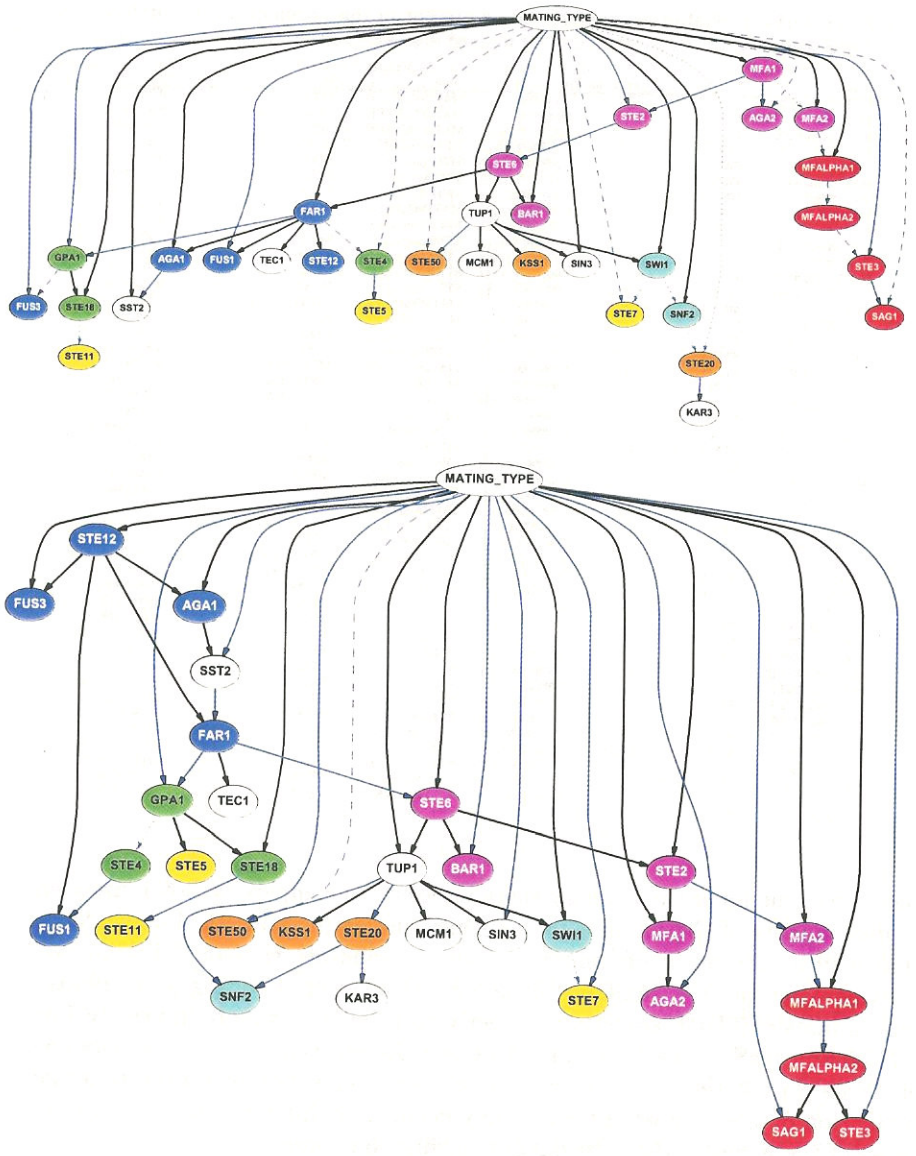
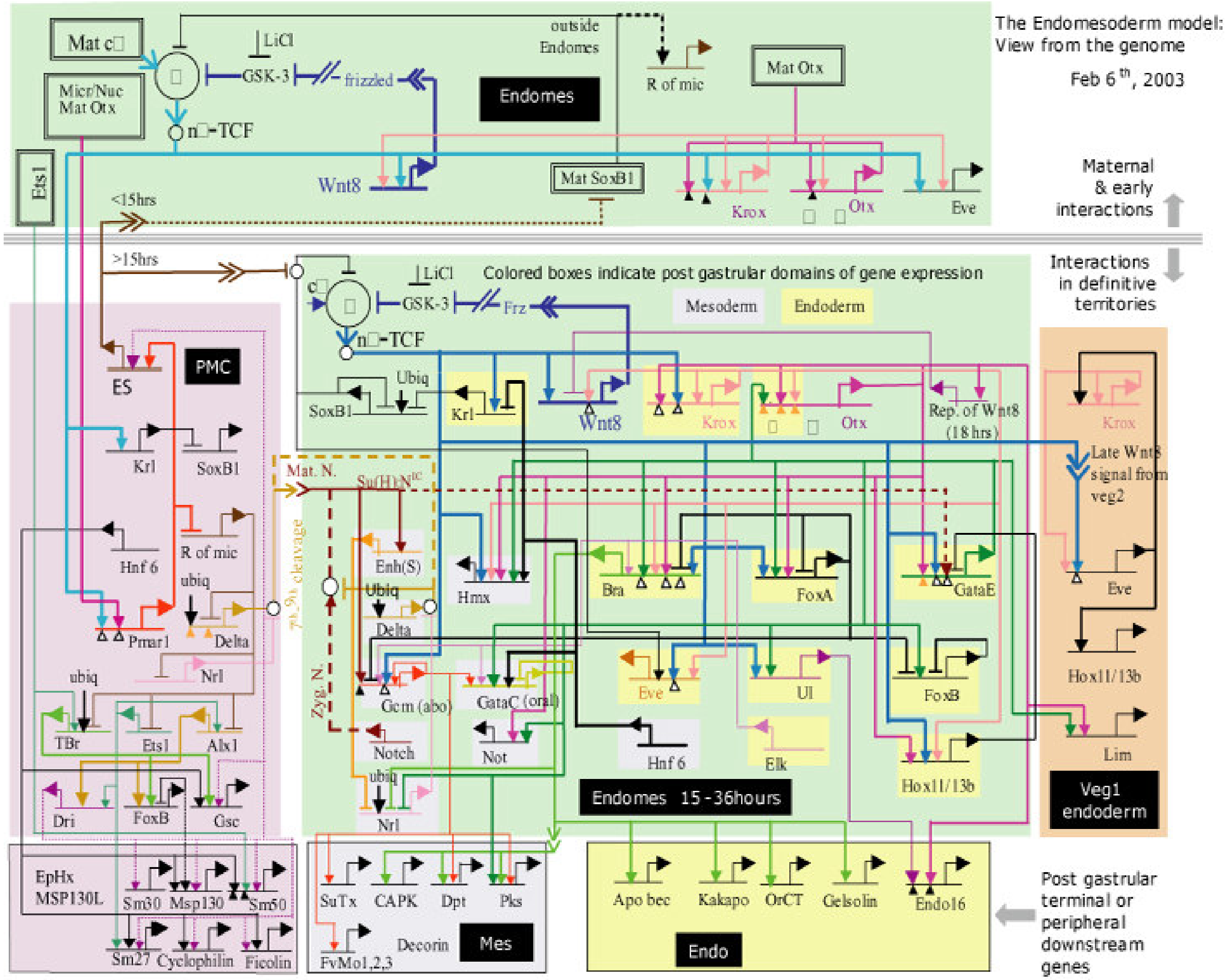


Figure 2. Bayesian network models learned by model averaging over the 500 highest scoring models visited during the unconstrained and constrained simulated annealing search runs, respectively. Edges are included in the figure if and only if their posterior probability exceeds 0.5. Node and edge color descriptions are included in the text.

4. Putting It All Together

Davidson et al., 2002



The Endomesoderm model: View from the genome

Feb 6 th, 2003

Bibliography

- Chiang et al., *Visualizing Associations Between Genome Sequences and Gene Expression Data Using Genome-Mean Expression Profiles*, Bioinformatics, v. 17, 2001, S49-S55.
- Davidson et al., A Genomic Regulatory Network for Development. Science 295 (5560): 1669-2002
- Filkov et al., *Analysis Techniques for Microarray Time-Series Data*, Journal of Computational Biology 9(2): 317-330 (2002).
- Filkov and Skiena, Integrating Heterogeneous Data Sets via Consensus Clustering, 2003 (in progress)
- Ge et al., *Correlation Between Transcriptome and Interactome Mapping Data from Saccharomyces Cerevisiae,* Nature Genetics, v. 29, 2001, 482-486.
- Hartemnik et al., *Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models*, Pacific Symposium on Biocomputing 2002.
- Marcotte et al., *A Combined Algorithm for Genome-wide Prediction of Protein Function*, Nature, v. 402, 1999, 83-86.
- Pavlidis et al., *Learning Gene Functional Classification from Multiple Data Types*, Journal of Computational Biology, v. 9, 2002, 401-411.
- Giffrod's course at MIT: Comp. Funct. Genomics, 2002 http://www.psrg.lcs.mit.edu/6892/handouts/lecture-18-1.pdf