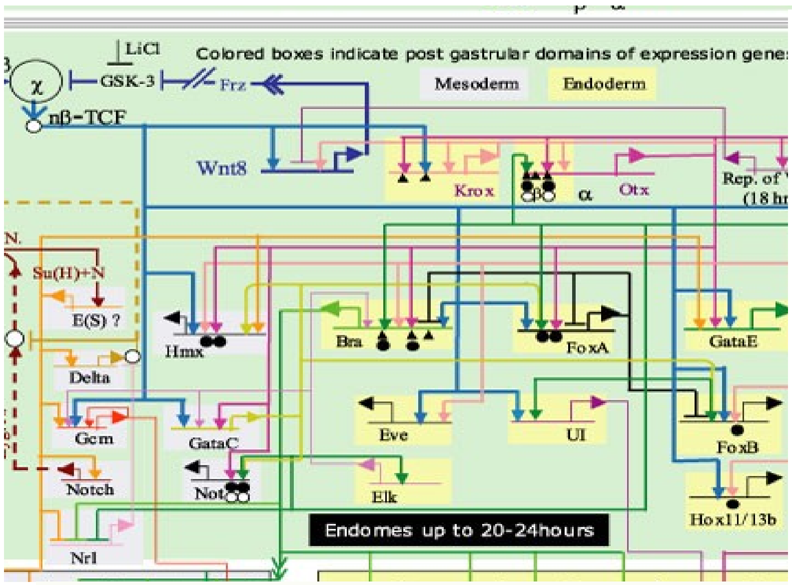
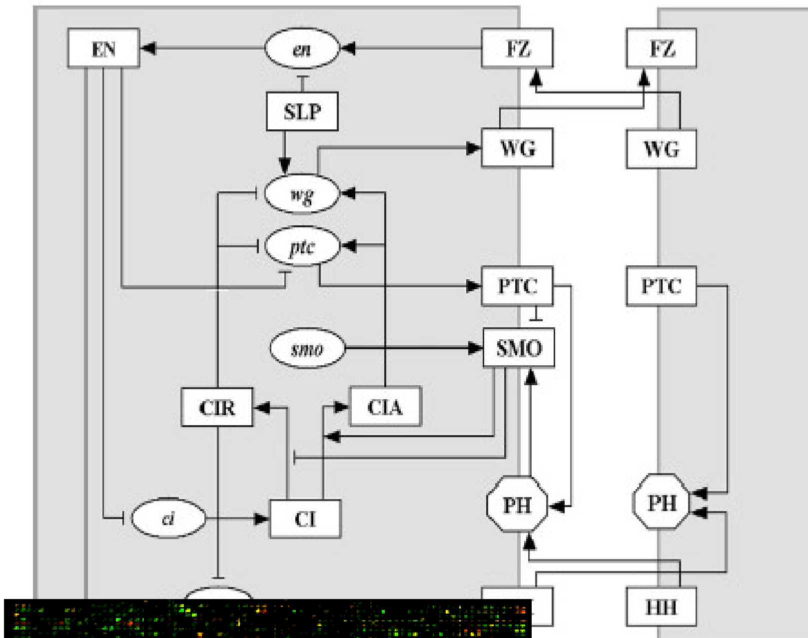
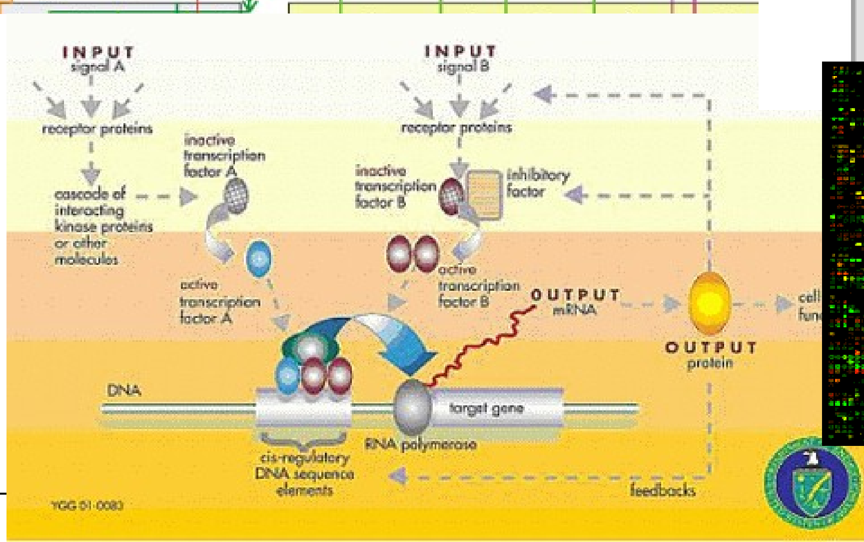
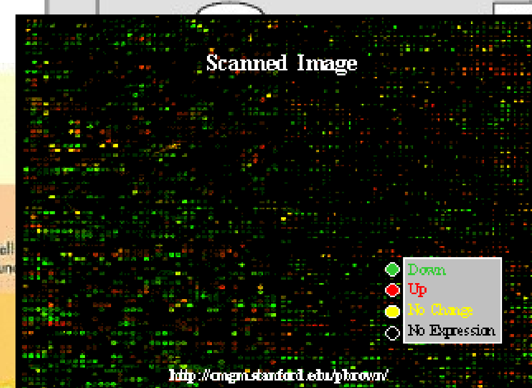
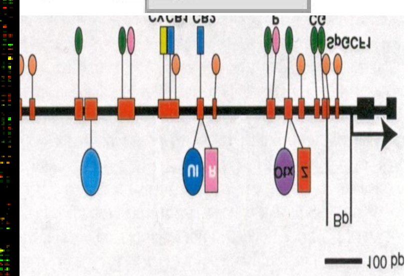
# ECS 234: Genomic Data Integration



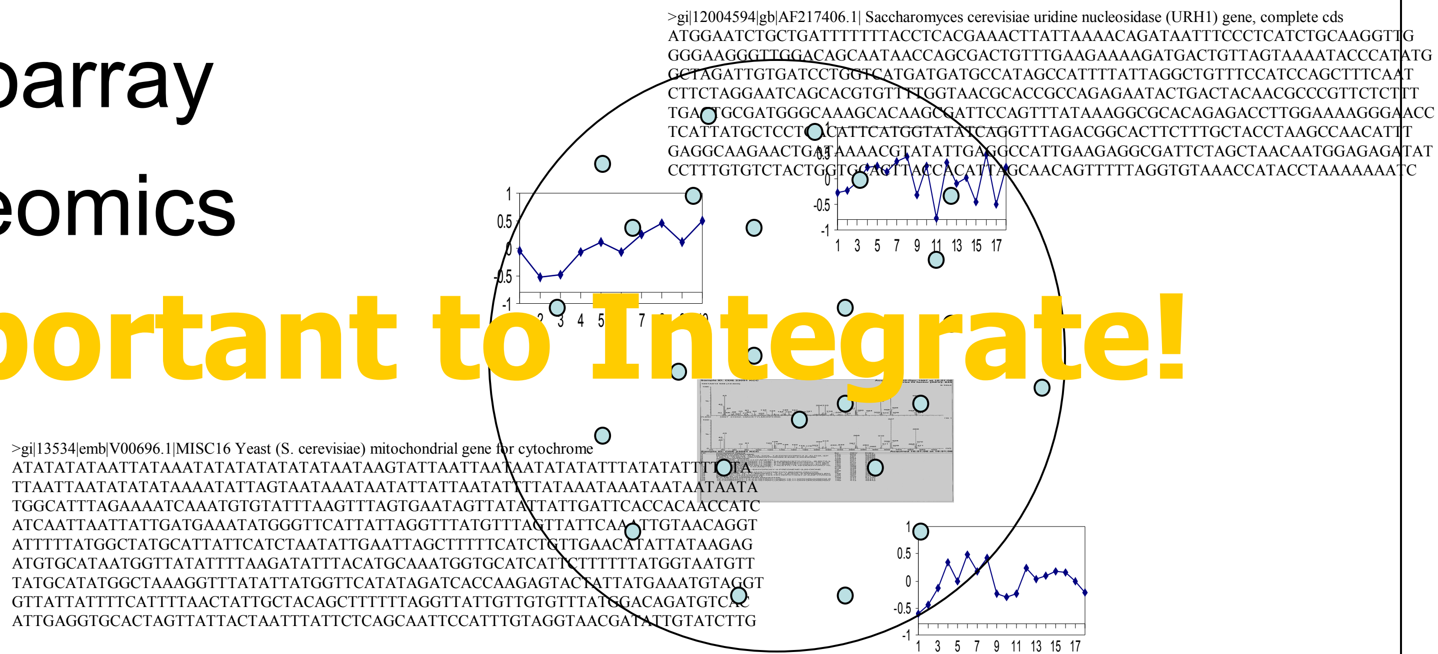



# Heterogeneous Data Integration

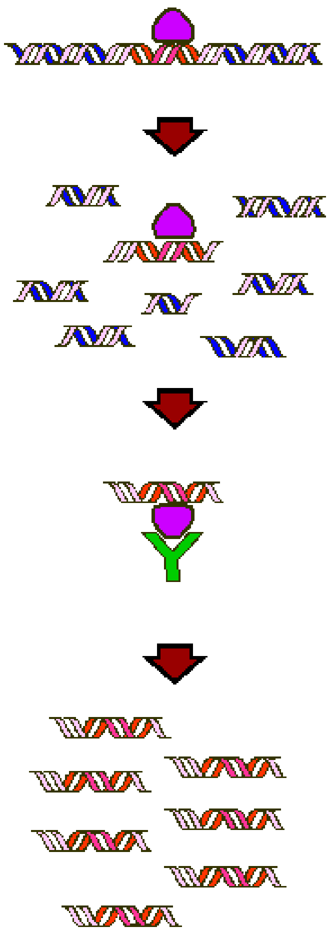
- DNA Sequence
- Microarray
- Proteomics

**Important to Integrate!**




Yeast Genes

# ChIP-chip Data
## (Large-Scale Genomewide Location Data)



- Detecting physical TF binding to DNA on a large scale
- Binding confidence P-value for each TF and DNA region

| | TF1 | TF2 | TF3 |
|---|---|---|---|
| G1 | 0.002 | 0.05 | 0.2 |
| G2 | 0.01 | 0.003 | 0.6 |
| G3 | 0.15 | 0.12 | 0.011 |

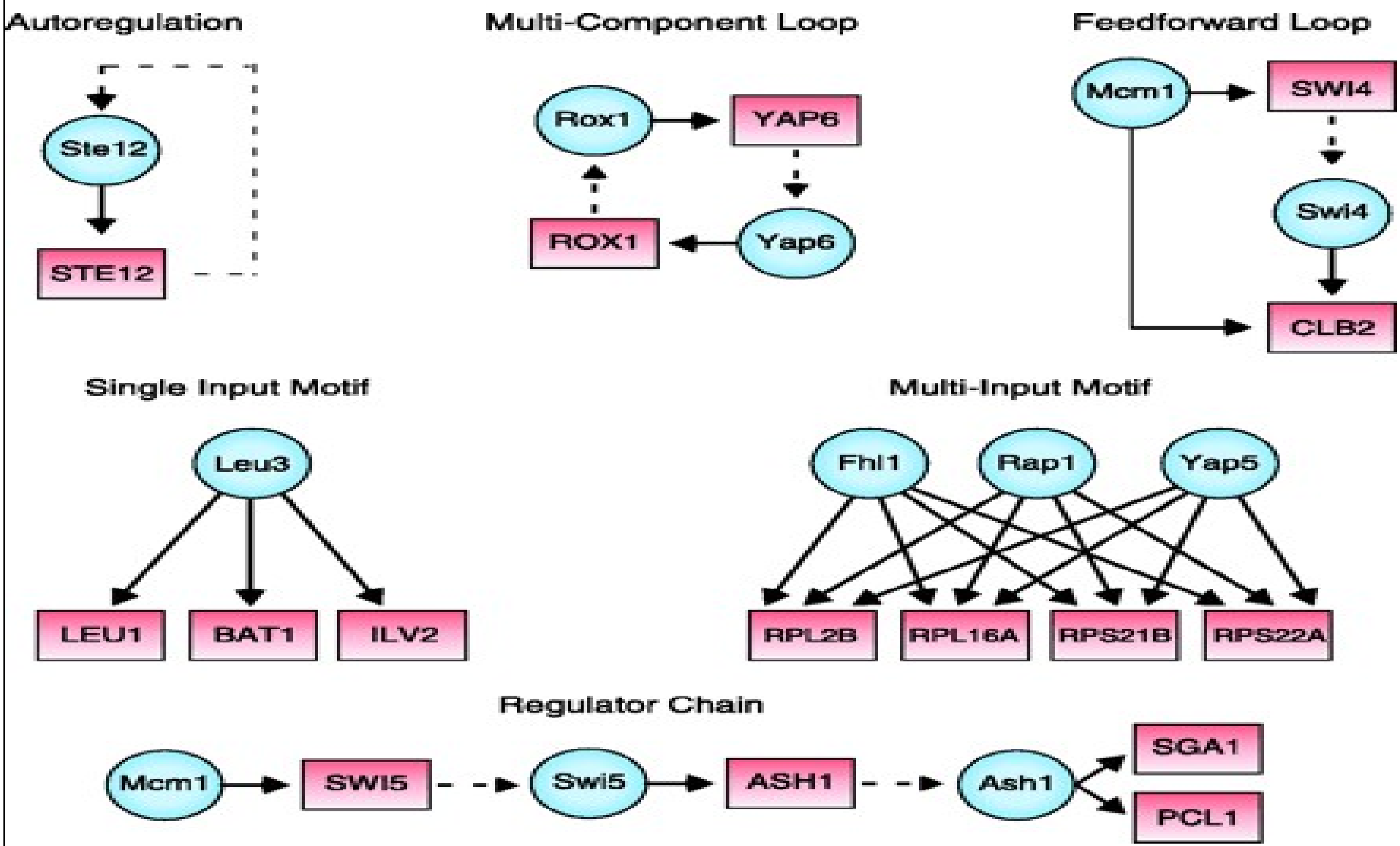
**BUT:**

- Localizes the binding coarsely, to a 1000-2000 bp region
- It doesn't say anything about the nature of the regulation
- It is noisy

# Using ChIP-chip Data (Lee et al. 2002)

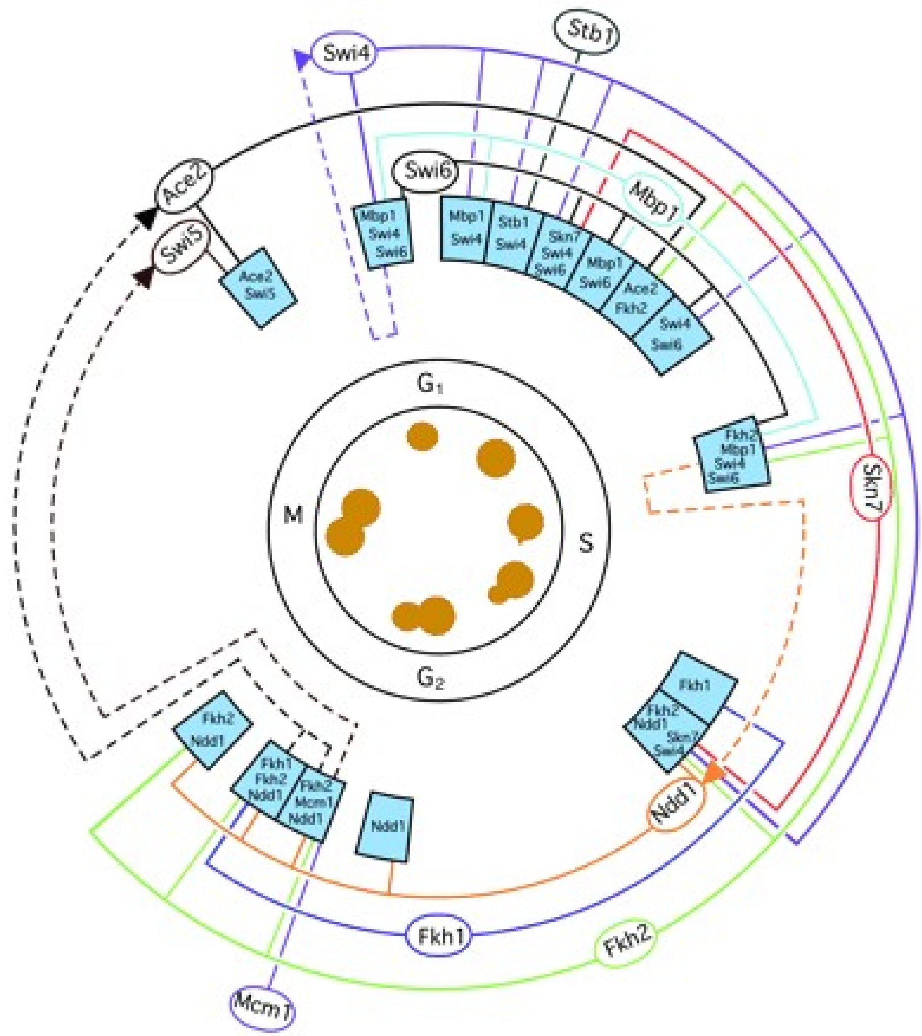
- Transcriptional Regulatory Networks in S. Cerevisiae (yeast)
- 106 TFs, 6000+ genes of yeast
- P-value: 0.001 (1/10 % by chance)
- Connect the TFs to the genes they regulate above that threshold
- The result is a regulatory network

# Network Analysis Detects Component Reuse (Network Motifs)



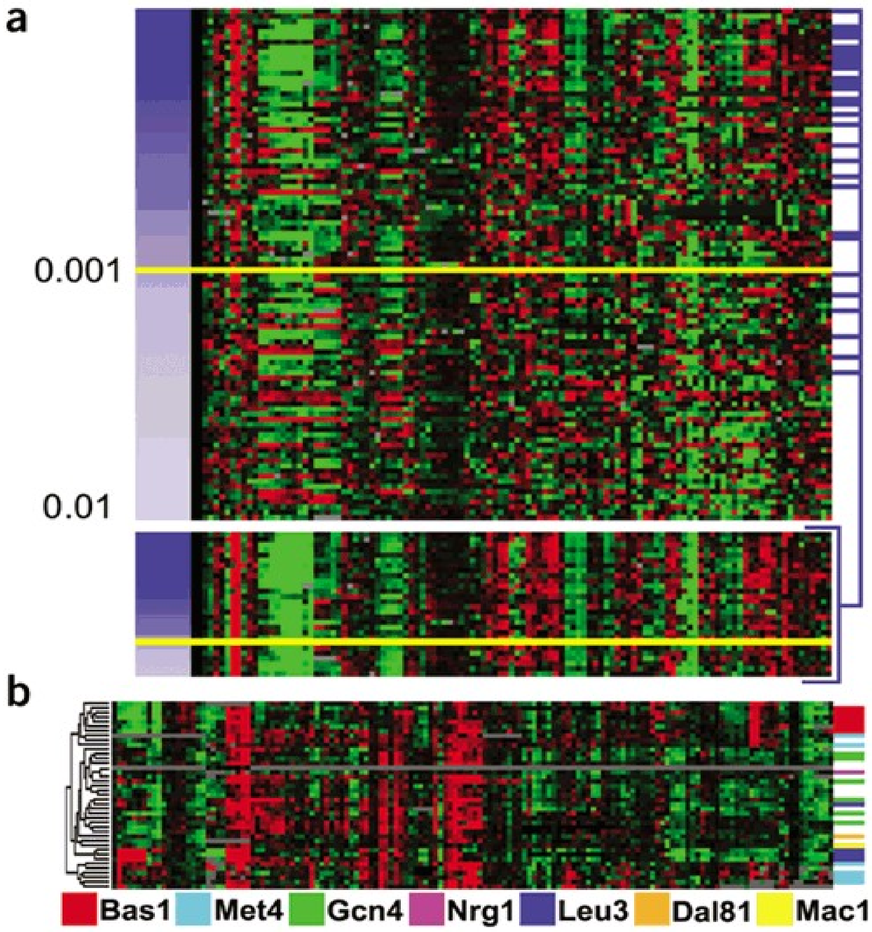
# Gene Modules and The Cell Cycle

- TF-DNA data combined with Expression Data from 500+ experiments
- Multi-Input Network Motifs + Coexpressed genes = MIM-CE (coexpressed and coregulated gene modules)
- MIM-CE aligned with known genes expressed in different parts of the cell-cycle
- Result: cell cycle modular regulatory network

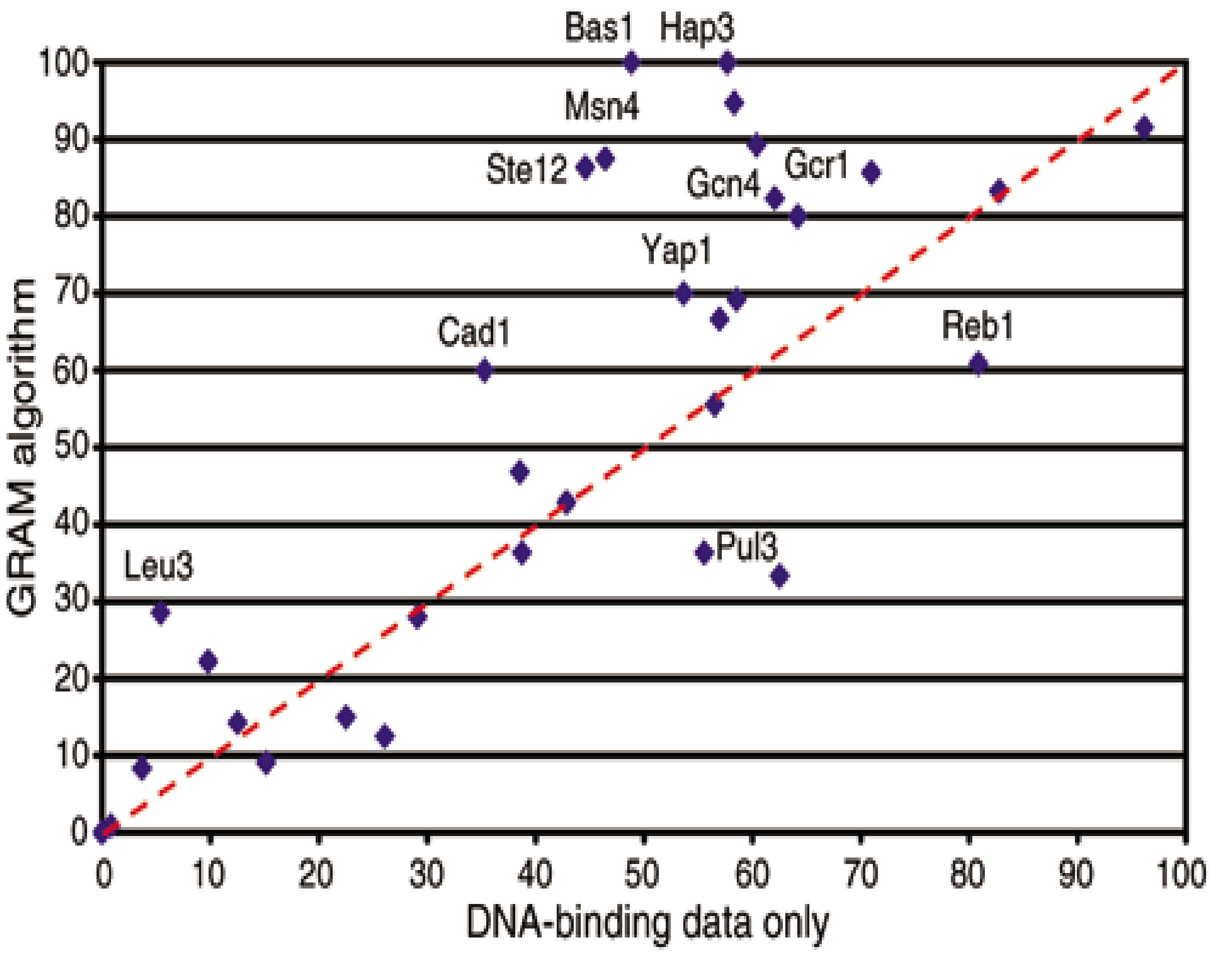


# Improved Expression+TF-DNA Integration: Gene Modules (Bar-Joseph et al. 2003)

- GRAM algorithm: eliminate the strict significance threshold (P-value) for TF-DNA binding by considering expression data
- Find coexpressed genes and their regulators (ie gene modules)

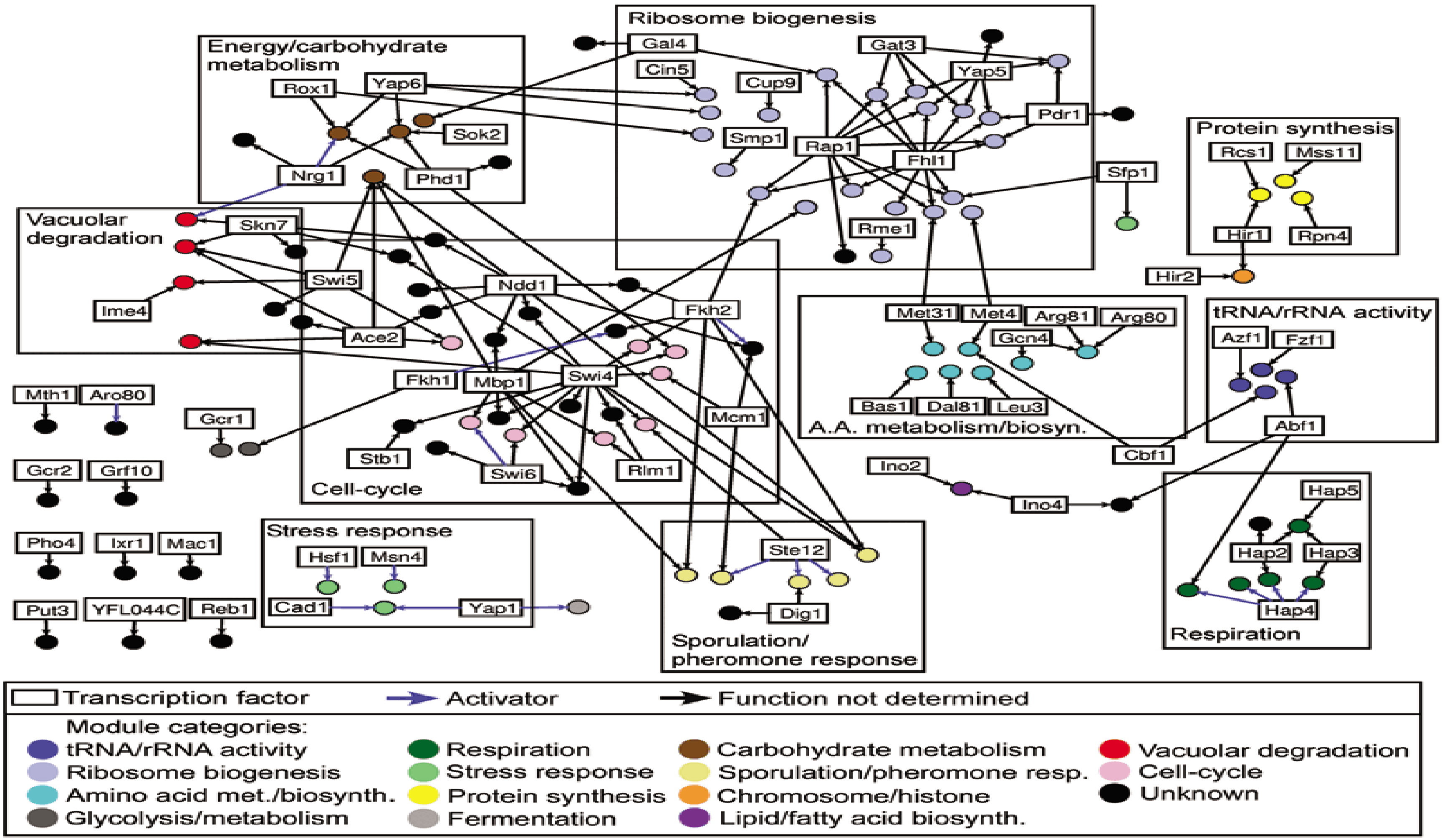


# Benefits of Integration




GRAM detected gene modules are more likely to contain upstream the appropriate binding sites (from TRANSFAC) then the modules detected by using TF-DNA data only

# Results From Integration



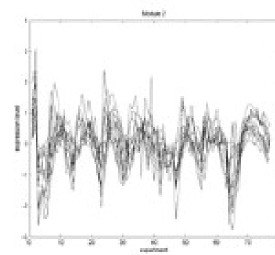
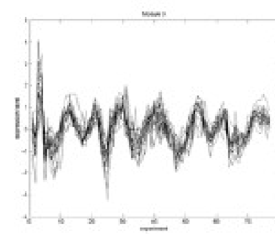
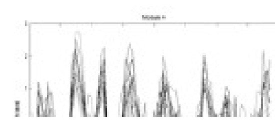
# ChIP-chip + Promoter Sequence (Liu et al. 2002)

- TF-DNA binding is very coarsely localized with ChIP-chip data
- To find the actual binding sites use the TF-DNA binding evidence to narrow down the search space for motif finding
- MDscan: rank the DNA sequences and give more weight to sequences more enriched for TF binding
- Works better than motif finders that do not take TF-DNA binding into account

# Sequence + ChIP-chip + Expression Data (de Bie et al, 2004)

- Uses 3 independent data sources of yeast:
  - M, Sequence motifs (obtained by comparative genomics)
  - R, ChIP-chip binding (from Young's lab)
  - A, Microarray expression data experiments
- A module is:
  - A set of regulators
  - A set of genes it regulates
  - A set of sequence motif where the regulators bind
- Algorithm: simple threshold based procedure for each data source, made efficient by observing hierarchical properties of modules

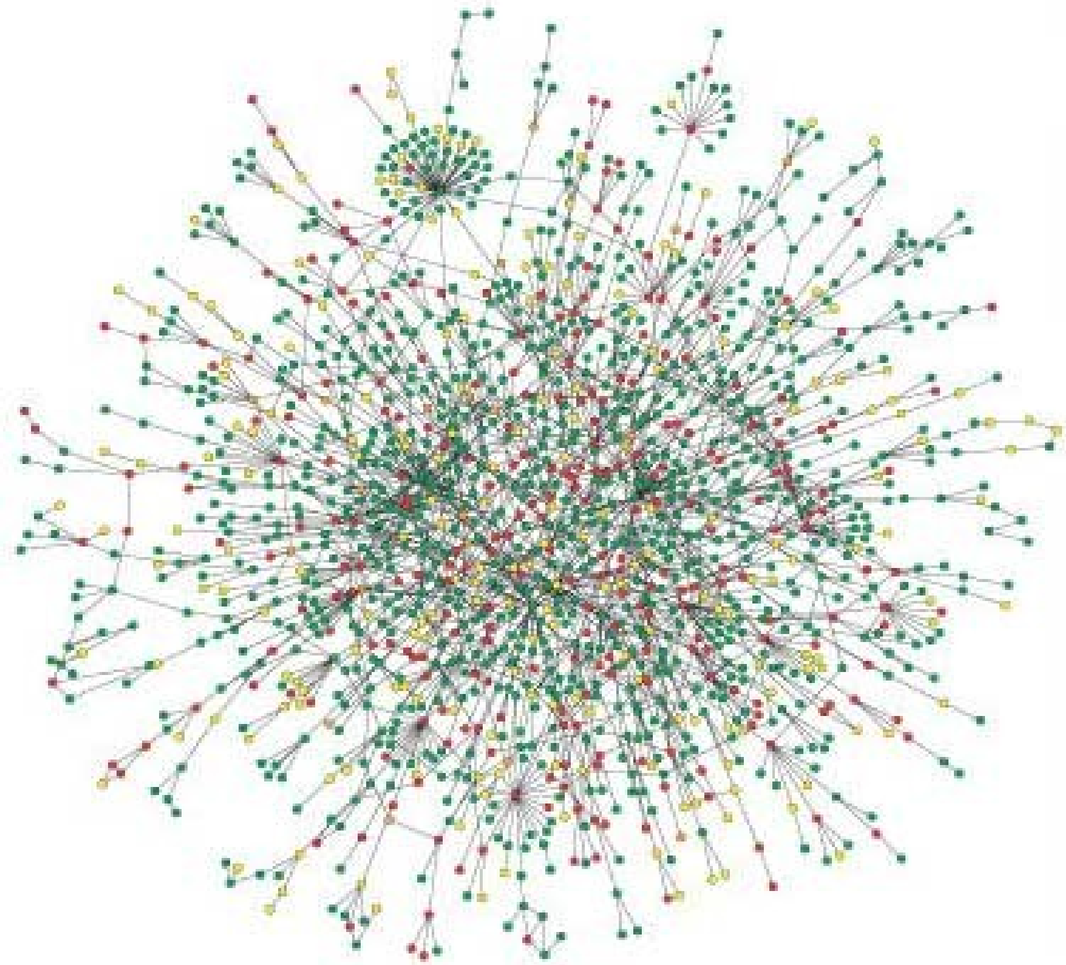
# Results

| | | | | |
|---|---|---|---|---|
| 1 | | (Swi4) | | |
| Module 2 | Swi4 Mbp1 Swi6 FKH2 | M_18 (Mbp1) M_12 (Mbp1) M_11 (Swi4) M_8 (Mcm) | 40 CELL FATE: 5.2e-4<br>**40.01** cell growth / morphogenesis: 2.6e-3<br><br>43 CELL TYPE DIFFERENTIATION: 5.2e-3<br>**43.01** f ungal/microorganismic cell type differentiation: 5.2e-3<br><br>**34.11** cellular sensing and response:5.3e-3<br>**01.05.01** C-compound and carbohydrate utilization: 6.8e-3<br>**10.03.04.03** chromosome condensation: 9.4e-3 |  |
| Module 3 | NDD1 FKH2 Mcm1 | M_8 (Mcm) M_30 (Mcm) | 43 CELL TYPE DIFFERENTIATION: 3.6e-3<br>**43.01** fungal/microorganismic cell type differentiation: 3.6e-3<br><br>**10.03.03** cytokinesis (cell division) /septum formation: 4.8e-3 |  |
| Mod | | | |  |

Known functional gene modules in yeast identified

# Protein-Protein Interaction Data (PPI)

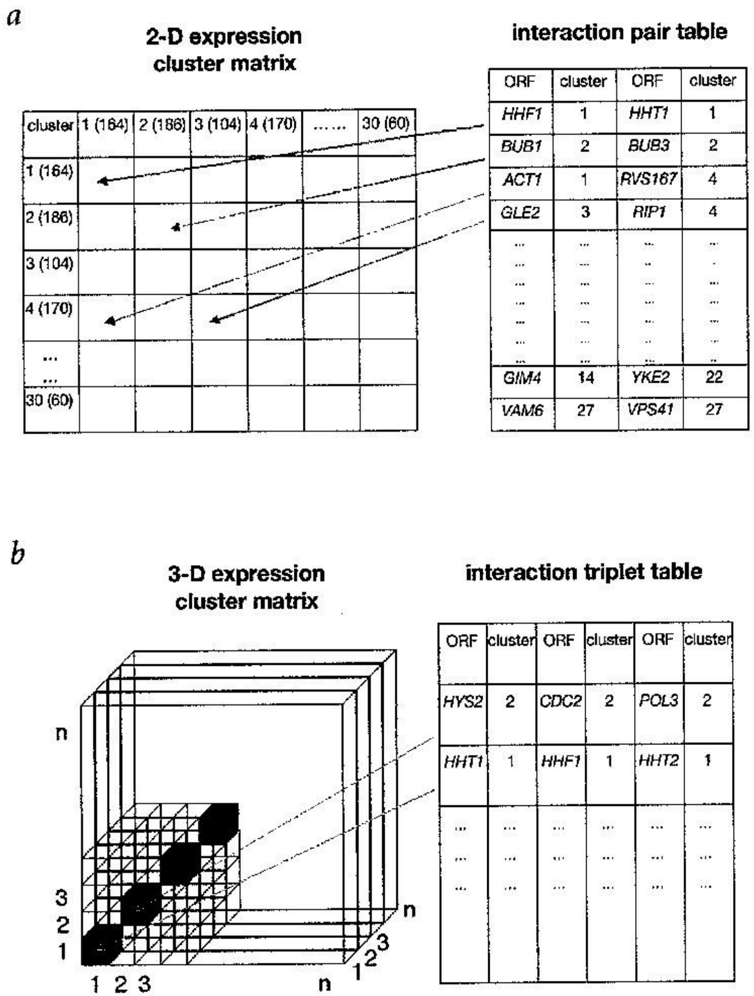
- Yeast-two-Hybrid technology
- Large-scale
- Available for yeast, drosophila



Yeast PPI Network (largest cluster): nodes are proteins, edges PPIs (red, lethal; green, non-lethal; orange, slow growth; yellow, unknown)

# Gene Expression + PPI Data

(Ge et al., 2001)



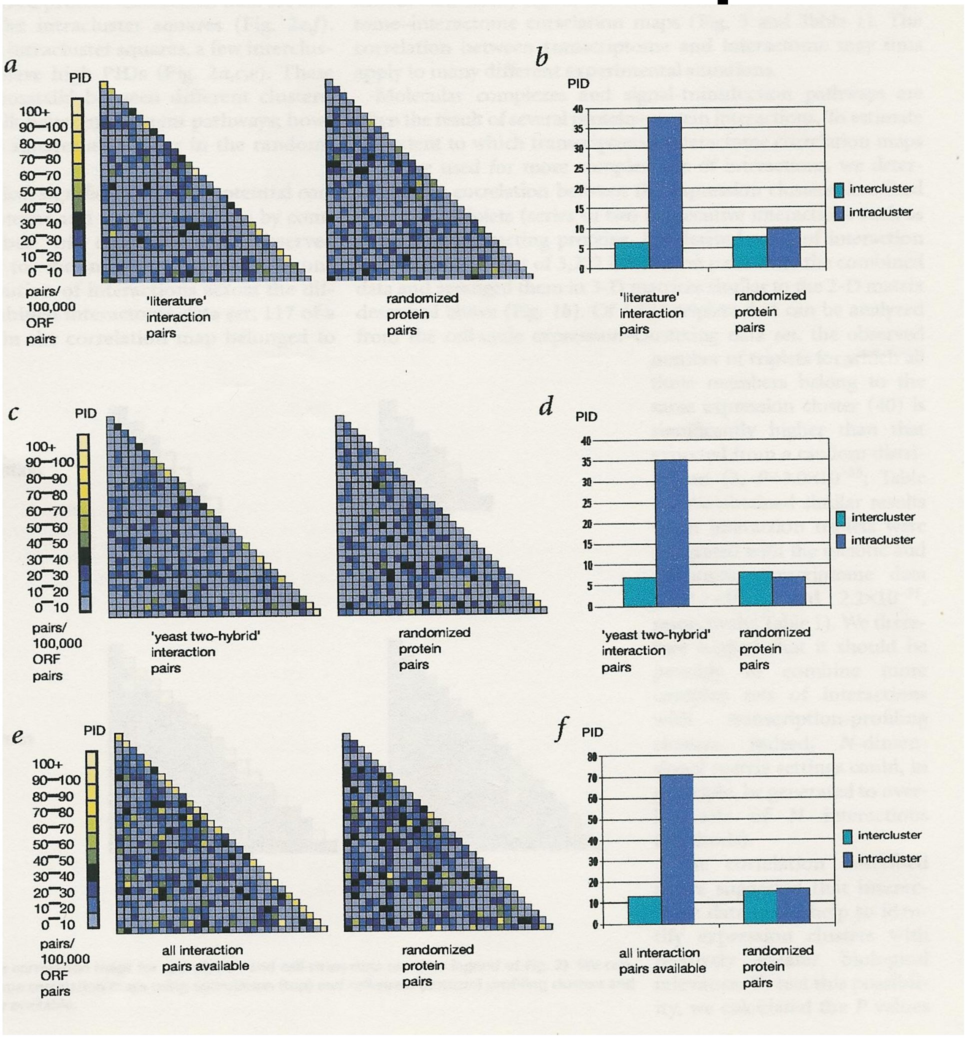

Goals

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

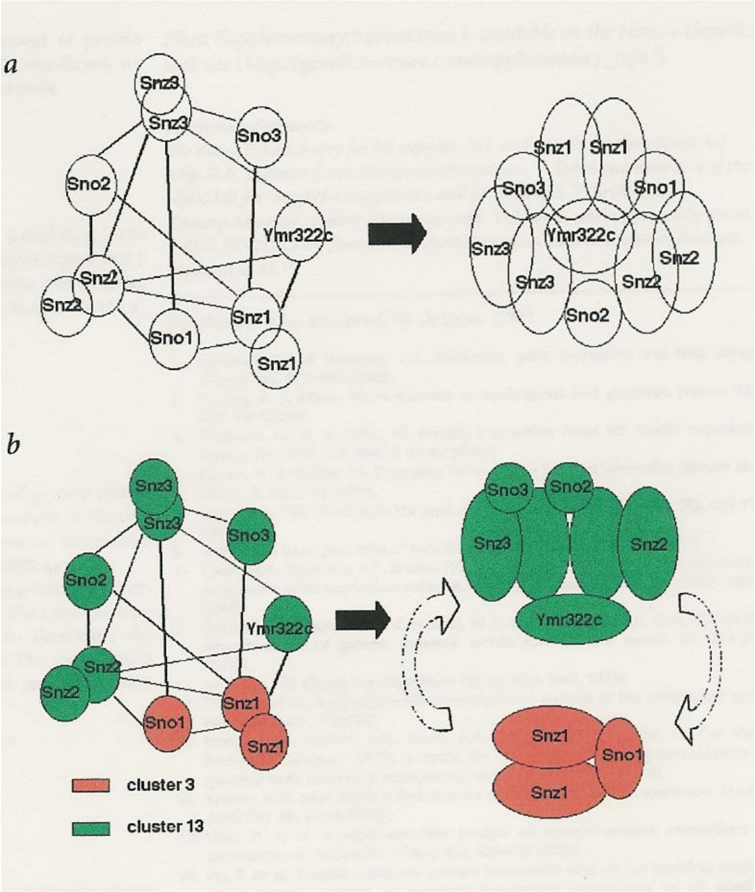
# Transcriptome – Interactome Correlation Maps



# More Knowledge Yields Better Models

a) Protein-protein interaction data

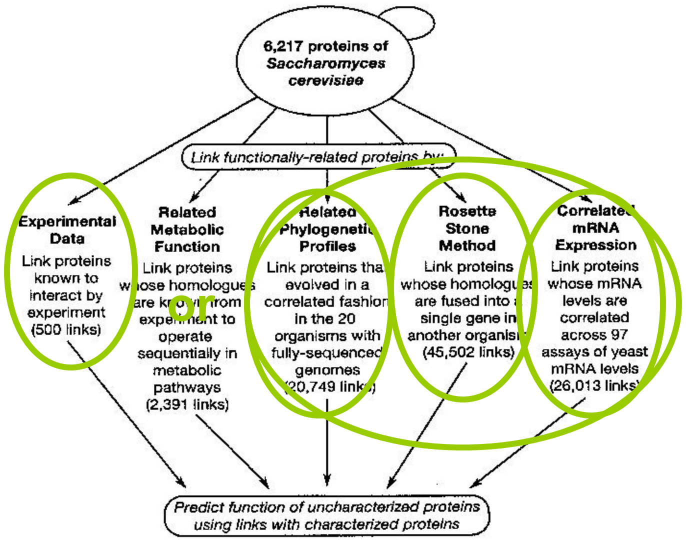
b) Protein Interaction + Gene Expression Data




Stress response proteins

# Protein Function Prediction

(Marcotte et al., 1999)




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 |

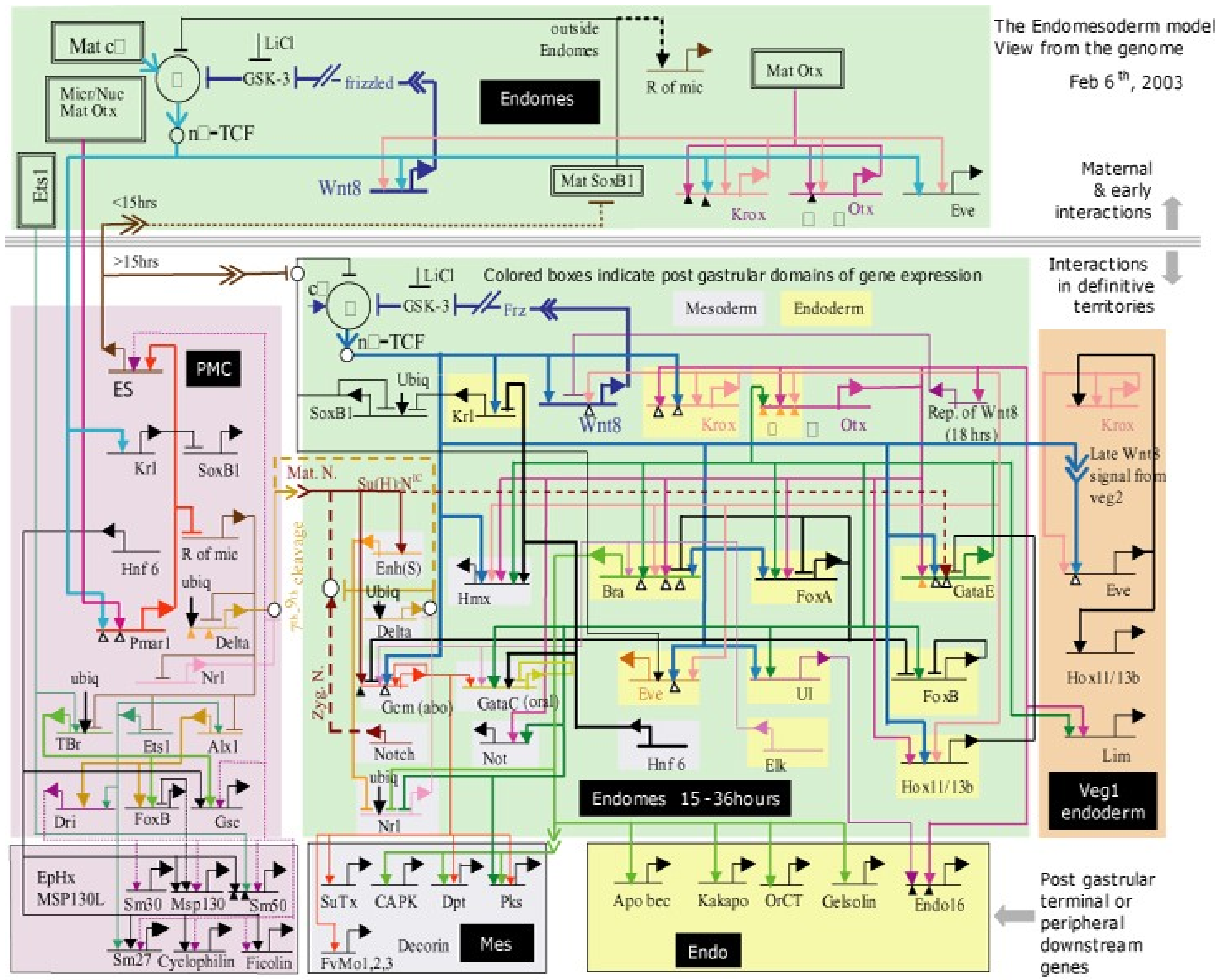
* The reliability of individual links was calculated as the percentage of pairwise links found between proteins of known function but having no functional categories in common (as tabulated in the MIPS database[4], ignoring the functional categories 'unclassified' and 'classification not clear cut'). 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:

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

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.

# 4. Putting It All Together?

Davidson et al., 2002



# Bibliography


- Lee et al., Transcriptional regulatory networks in S. Cerevisiae, Science 2002
- De Bie et al., Discovering transcriptional modules from motif, ChIP-chip and microarray data, PSB 2004
- Liu et al., An algorithm for finding protein-DNA binding sites with applications to chromatin immunoprecipitation microarray experiments, Nat. Biotech, 2002
- Bar-Joseph et al., Computational discovery of gene modules and regulatory networks, Nat. Biotech. 2003
- Marcotte et al., *A Combined Algorithm for Genome-wide Prediction of Protein Function*, Nature, v. 402, 1999, 83-86.
- Ge et al., *Correlation Between Transcriptome and Interactome Mapping Data from Saccharomyces Cerevisiae,* Nature Genetics, v. 29, 2001, 482-486.

OTHER:

- 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)
- Hartemnik et al., *Combining Location and Expression Data for Principled Discovery of Genetic Regulatory Network Models*, Pacific Symposium on Biocomputing 2002.
- Pavlidis et al., *Learning Gene Functional Classification from Multiple Data Types*, Journal of Computational Biology, v. 9, 2002, 401-411.