Discovering molecular pathways from protein interaction and gene expression data

José Caldas

9-4-2008

Aim

To have a mechanism for inferring pathways from **gene expression** and **protein interaction** data.

Motivation — Why search for pathways

Pathway

Set of genes that coordinate to achieve a specific task.

Motivation — Why search for pathways

Pathway

Set of genes that coordinate to achieve a specific task.

What do we gain from understanding pathways

1. A coherent global picture of (condition-specific) cellular activity.
2. Application to disease mechanisms.

Motivation — Why use two kinds of data

2 properties of (many) pathways

Motivation — Why use two kinds of data

2 properties of (many) pathways

(A) Genes in the same pathway are activated together $\Rightarrow$ exhibit similar expression profiles.

Motivation — Why use two kinds of data

2 properties of (many) pathways

(A) Genes in the same pathway are activated together $\Rightarrow$ exhibit similar expression profiles.

(B) When genes coordinate to achieve a particular task, their protein products often interact.

Motivation — Why use two kinds of data

2 properties of (many) pathways

(A) Genes in the same pathway are activated together $\Rightarrow$ exhibit similar expression profiles.

(B) When genes coordinate to achieve a particular task, their protein products often interact.

Each data type alone is a **weaker indicator** of pathway activity.

Intuitive Idea

- Detect group of genes that are co-expressed, and whose products interact in the protein data.

Intuitive Idea

- Detect group of genes that are co-expressed, and whose products interact in the protein data.
- Create a model for **gene expression data**.
- Create a model for **protein interaction data**.

Intuitive Idea

- Detect group of genes that are co-expressed, and whose products interact in the protein data.
- Create a model for **gene expression data**.
- Create a model for **protein interaction data**.
- Join them.

Basics

Gene

- Set of genes $G = \{1, \ldots, n\}$.

Basics

Gene

- Set of genes $G = \{1, \ldots, n\}$.
- Each gene g has two attributes:

Basics

Gene

- Set of genes $G = \{1, \ldots, n\}$.
- Each gene g has two attributes:
  - Class (pathway), denoted by $g.C$ (discrete value).

Basics

Gene

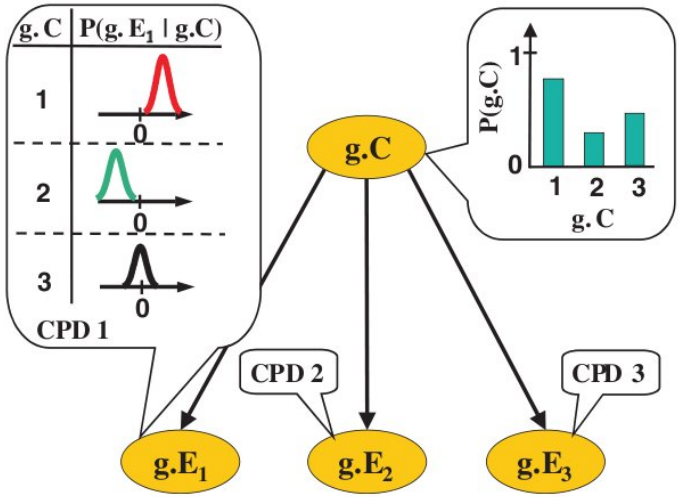
- Set of genes $G = \{1, \ldots, n\}$.
- Each gene g has two attributes:
  - Class (pathway), denoted by $g.C$ (discrete value).
  - Expression in microarray i, denoted by $g.E_i$.

Basics

Gene

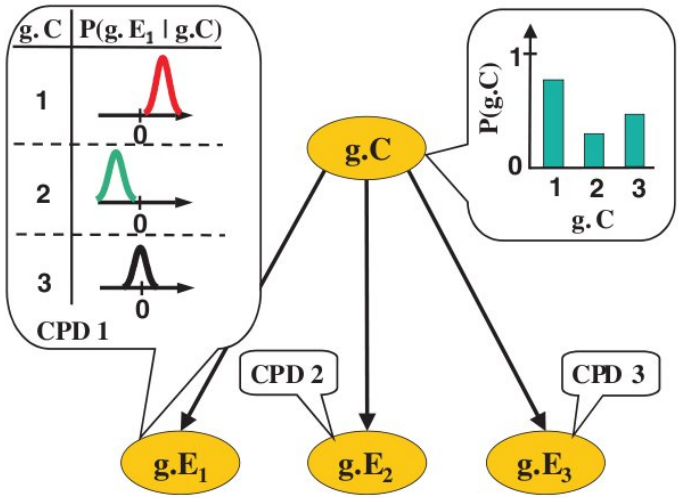
- Set of genes $G = \{1, \ldots, n\}$.
- Each gene g has two attributes:
  - Class (pathway), denoted by $g.C$ (discrete value).
  - Expression in microarray i, denoted by $g.E_i$.
  - If there are m microarrays $\Rightarrow g.\mathbf{E} = \{g.E_1, \ldots, g.Em\}$.

Model for expression profiles — Naive Bayes



Naive Bayes — given the class label $g.C$, $g.E_i$ and $G.E_j$ are independent.

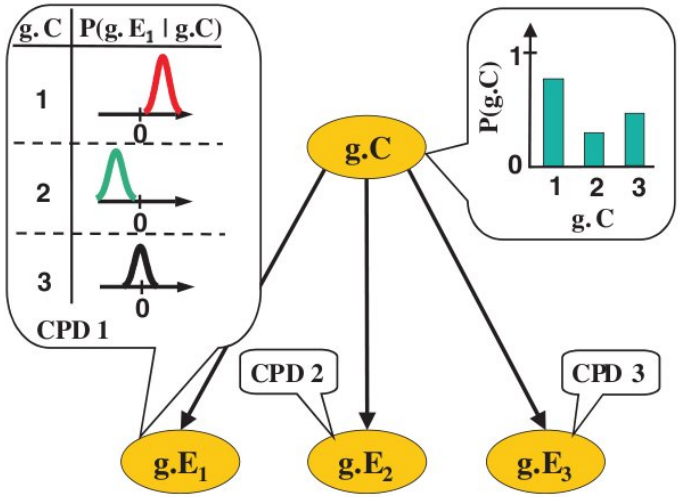
Model for expression profiles — Naive Bayes



Class probability

- $g.C$ follows a multinomial probability distribution
- $p(g.C = k) = \theta_k$
- $\sum_{i=1}^{K} \theta_i = 1$

Model for expression profiles — Naive Bayes



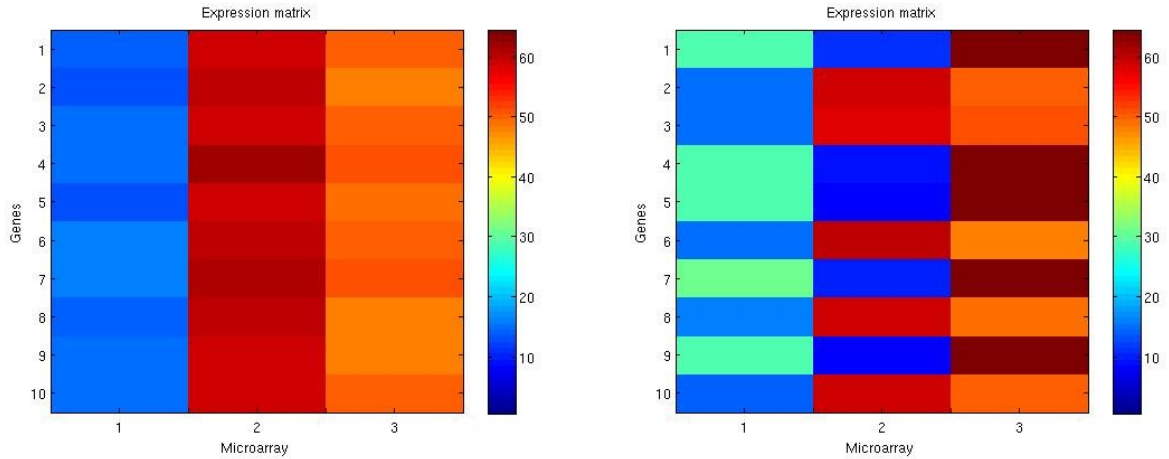
Expression profiles

- $g.E_i|g.C = k \sim N(\mu_{ki}, \sigma^2_{ki})$
- A pathway i specifies the **average** expression level for each microarray and also the variance.

Model for expression profiles — Naive Bayes

Example:

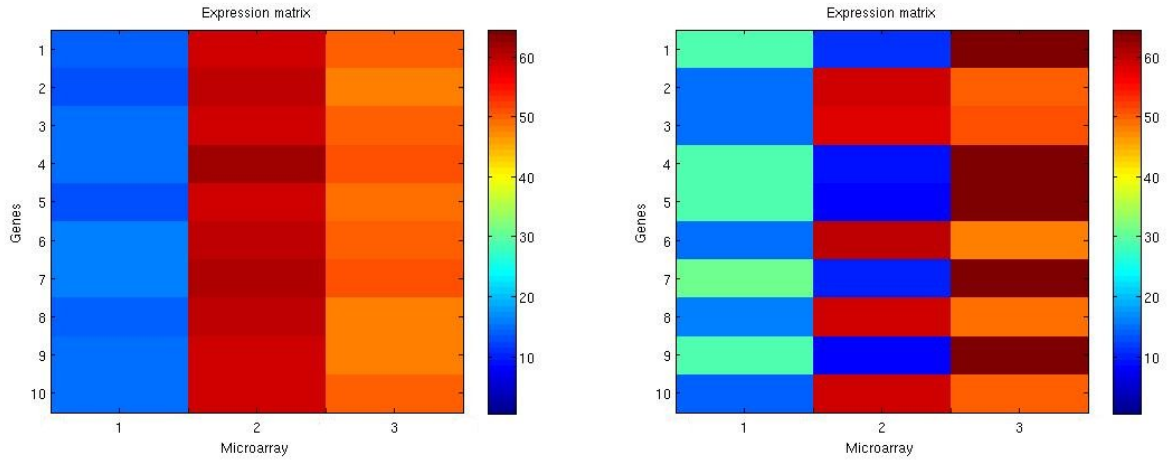
- 1 pathway, 10 genes, 3 microarrays
- Pathway specifies the averages $\mu = (15, 60, 50)$
- What is the most likely expression matrix?



Model for expression profiles — Naive Bayes

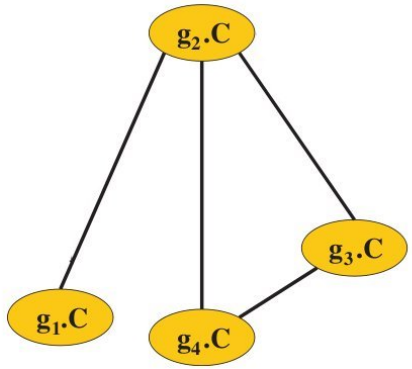
Example:

- 1 pathway, 10 genes, 3 microarrays
- Pathway specifies the averages $\mu = (15, 60, 50)$
- What is the most likely expression matrix?



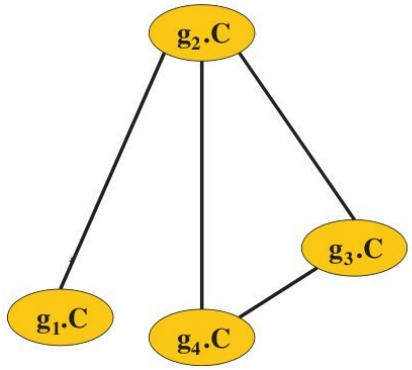
- (The matrix on the left)

Model for protein interaction — Markov random field



Undirected graph, $V = \{g_1.C, \ldots, g_n.C\}$, E =set of protein interactions.

Model for protein interaction — Markov random field



Undirected graph, $V = \{g_1.C, \ldots, g_n.C\}$, E =set of protein interactions.

Assumption

Interacting proteins are more likely to be in the same pathway.

Intuitive idea

If a pair of nodes share the same class $\Rightarrow$ likelihood is higher

Markov random field — Formalism

- Each $g_i.C$ is associated with a *potential* $\phi_i(g_i.C)$.

Markov random field — Formalism

- Each $g_i.C$ is associated with a *potential* $\phi_i(g_i.C)$.
- Each edge $g_i.C - g_j.C$ is associated with a *compatibility potential* $\phi_{i,j}(g_i.C, g_j.C)$.

Markov random field — Formalism

- Each $g_i.C$ is associated with a *potential* $\phi_i(g_i.C)$.
- Each edge $g_i.C - g_j.C$ is associated with a *compatibility potential* $\phi_{i,j}(g_i.C, g_j.C)$.

Joint distribution is

$$P(g_1.C, \ldots, g_n.C) = \frac{1}{Z} \prod_{i=1}^{n} \phi_i(g_i.C) \prod_{\{g_i.C - g_j.C\} \in \mathrm{E}} \phi_{i,j}(g_i.C, g_j.C) \tag{1}$$

Z is a normalization constant.

Markov random field — Formalism

$$\phi_{i,j}(g_i.C = p,\, g_j.C = q) = \begin{cases} \alpha & p = q \\ 1 & \text{otherwise} \end{cases}$$

Markov random field — Formalism

$$\phi_{i,j}(g_i.C = p,\, g_j.C = q) = \begin{cases} \alpha & p = q \\ 1 & \text{otherwise} \end{cases}$$

$(\alpha \geq 1)$.

Unified Model

What we already have

- Model for expression data (Naive Bayes)
- Model for class probability (Markov random field)

Unified Model

What we already have

- Model for expression data (Naive Bayes)
- Model for class probability (Markov random field)

What we want

Probability distribution $P(\mathbf{G}.C, \mathbf{G}.E)$, using expression and protein data.

Unified Model

What we already have

- Model for expression data (Naive Bayes)
- Model for class probability (Markov random field)

What we want

Probability distribution $P(\mathbf{G}.C, \mathbf{G}.E)$, using expression and protein data.

What we are missing

- Naive Bayes provides that prob. distribution, but does not use protein data.

Unified Model

What we already have

- Model for expression data (Naive Bayes)
- Model for class probability (Markov random field)

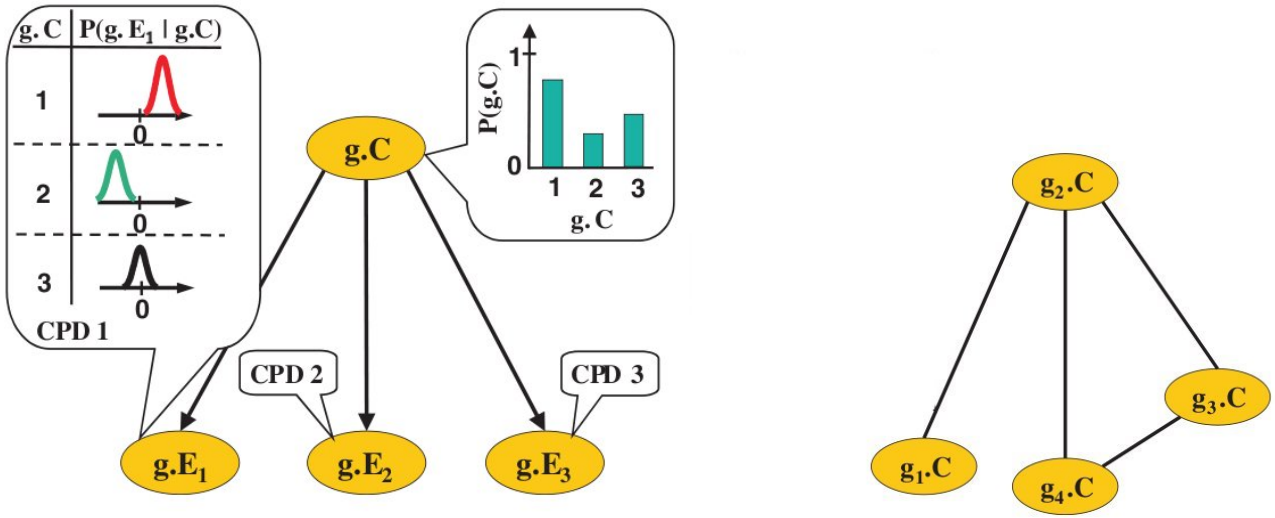
What we want

Probability distribution $P(\mathbf{G}.C, \mathbf{G}.E)$, using expression and protein data.

What we are missing

- Naive Bayes provides that prob. distribution, but does not use protein data.
- We haven't specified the potentials $\phi_i(g_i.C)$.

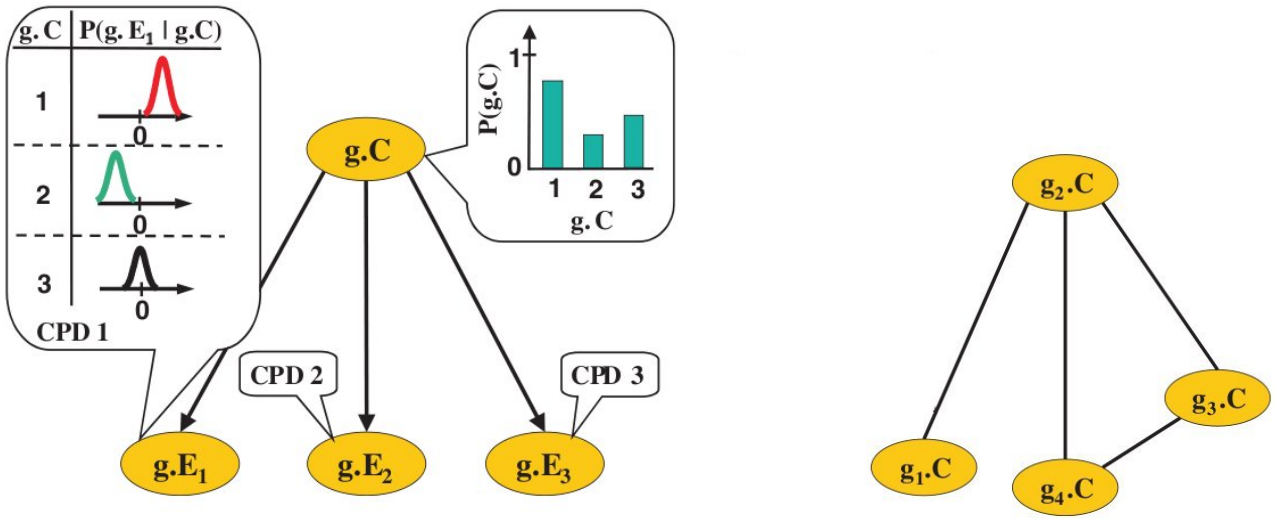
Unified Model



Solution

- Use Markov random field as $P(\mathbf{G}.C)$.

Unified Model



Solution

- Use Markov random field as $P(\mathbf{G}.C)$.
- Use multinomial dist. $P(g_i.C)$ from Naive Bayes as potential $\phi_i(g_i.C)$.
- Call it $P^*(g_i.C)$.

Unified Model

$$P(\mathbf{G}.C, \mathbf{G}.\mathbf{E}) =$$

$$\frac{1}{Z} \prod_{i=1}^{n} P^*(g_i.C) \prod_{\{g_i.C - g_j.C\} \in \mathrm{E}} \phi_{i,j}(g_i.C, g_j.C) \cdot$$

$$\prod_{i=1}^{n} \prod_{j=1}^{m} P(g_i.E_j | g_i.C)$$

$P(\mathbf{G}.C) \rightarrow$ Markov random field.
$P(\mathbf{G}.\mathbf{E}) \rightarrow$ Gaussian distributions.

Learning Algorithm

EM algorithm

Parameters to be estimated

- Multinomial distribution $\rightarrow (\theta_1, \ldots, \theta_K)$.
- Mean and variance for gaussian distributions

Datasets

Gene Expression

- 173 arrays (Gasch *et al.* 03)
- 77 arrays (Spellman *et al.* 98)

Protein Interaction

10705 interactions (Xenarios *et al.* 05)

After preprocessing $\longrightarrow$ 3589 genes.

Running the algorithm

- EM for optimizing parameters
- Number of pathways fixed as 60
- Starting point for parameters $\longrightarrow$ use hierarchical clustering

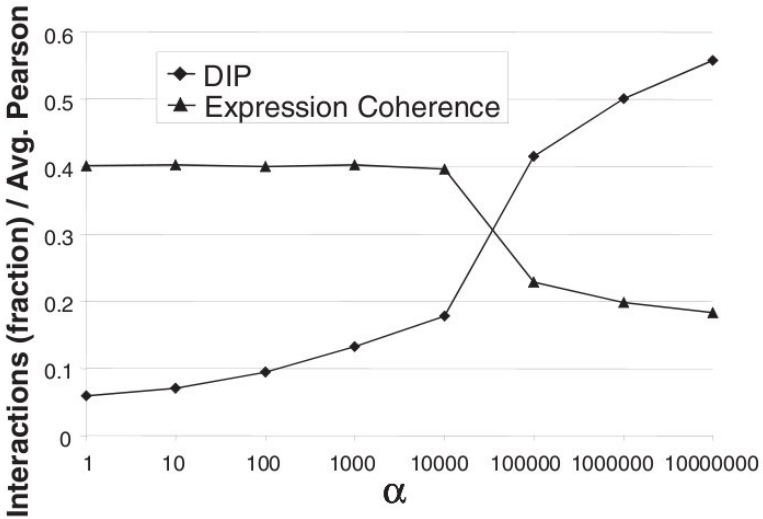
How to set the α parameter?

Setting α

- Recall: α is the *compatibility potential* when two proteins interact and belong to the same pathway.

Setting α

- Recall: α is the *compatibility potential* when two proteins interact and belong to the same pathway.



Comparisons with other methods

Methods that use only one type of data

- Markov Cluster (Enright *et al.* 02)
- Hierarchical clustering (Eisen *et al.* 98)

Tests

- Prediction of held-out interactions.
- Functional enrichment in Gene Ontology.
- Coverage of protein complexes.
- Assigning new roles to unknown proteins.

Prediction of held-out interactions

- Cross-validation — divide protein data into 5 disjoint sets (4 for training, 1 for testing)

Prediction of held-out interactions

- Cross-validation — divide protein data into 5 disjoint sets (4 for training, 1 for testing)
- Get average number of held-out interactions between genes in the same pathway

Prediction of held-out interactions

- Cross-validation — divide protein data into 5 disjoint sets (4 for training, 1 for testing)
- Get average number of held-out interactions between genes in the same pathway
- Result: 222.4 ± 13.2
- (MCL) 383.2 ± 29.1

Biological coherence of the inferred pathways

General result

More functionally coherent than when using standard clustering or MCL

Biological coherence of the inferred pathways

General result

More functionally coherent than when using standard clustering or MCL

Example — Pathways related to translation, protein degradation, transcription, and DNA replication

- Genes in these pathways interact with many genes from other categories.
- They are also co-expressed.

Biological coherence of the inferred pathways

General result

More functionally coherent than when using standard clustering or MCL

Example — Pathways related to translation, protein degradation, transcription, and DNA replication

- Genes in these pathways interact with many genes from other categories.
- They are also co-expressed.
- MCL cannot isolate them.

Protein Complexes

Motivation

The components of many pathways are protein complexes. Thus, a good pathway model should assign the member genes of many of these complexes to the same pathway.

Protein Complexes

Motivation

The components of many pathways are protein complexes. Thus, a good pathway model should assign the member genes of many of these complexes to the same pathway.

Procedure

- Use experimental assays (Gavin *et al.* 02) and (Ho *et al.* 02)
- Associate each gene to the complexes to which it belongs.
- Measure enrichment in pathways.

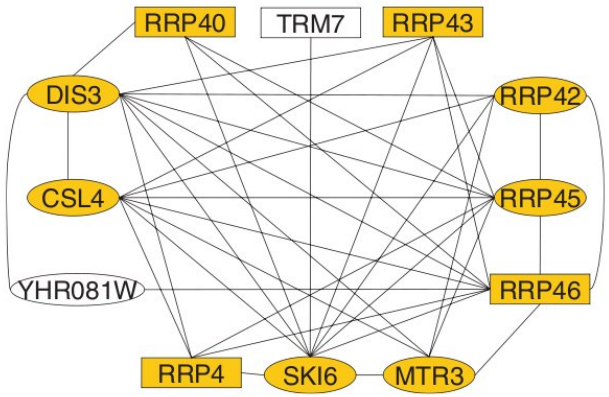
Protein Complexes — Results

In general, better than clustering:

- 374 complexes significantly enriched (higher than in clustering).
- Stress data $\rightarrow$ 124 complexes in which more than 50% of members appear in the same pathway.
- Clustering $\rightarrow$ only 46 complexes that verify that condition.

Assigning New Roles to Unknown Proteins

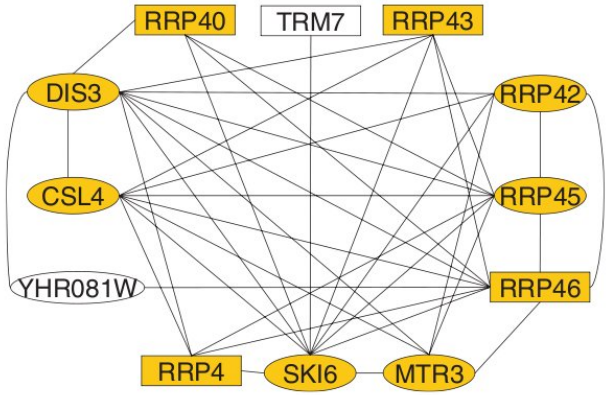
Largest connected component of pathway 1 (cytoplasmic exosome):



- YHR081W is uncharacterized

Assigning New Roles to Unknown Proteins

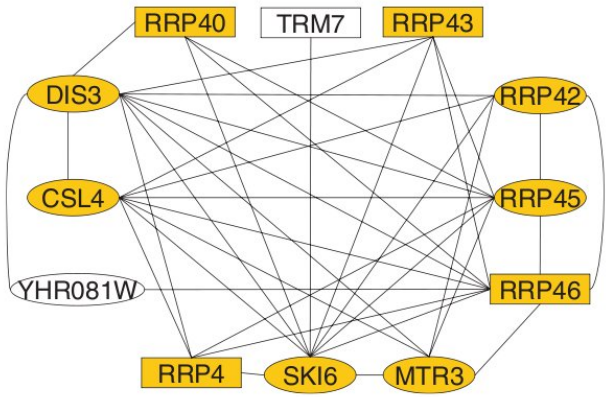
Largest connected component of pathway 1 (cytoplasmic exosome):



- YHR081W is uncharacterized
- Clustering — Only 4 genes in pathway

Assigning New Roles to Unknown Proteins

Largest connected component of pathway 1 (cytoplasmic exosome):



- YHR081W is uncharacterized
- Clustering — Only 4 genes in pathway
- MCL — Includes 114 additional genes in connected component

Conclusion

Summary

- Probabilistic model for integrating gene expression and protein interaction data

Conclusion

Summary

- Probabilistic model for integrating gene expression and protein interaction data
- Method aims at finding co-expressed and connected genes (pathways)

Conclusion

Summary

- Probabilistic model for integrating gene expression and protein interaction data
- Method aims at finding co-expressed and connected genes (pathways)

Comparison with single-source methods

Conclusion

Summary

- Probabilistic model for integrating gene expression and protein interaction data
- Method aims at finding co-expressed and connected genes (pathways)

Comparison with single-source methods

Some pathways are only obtainable by combining both types of data

Conclusion

Limitations

Conclusion

Limitations

- Model for co-expression is too restrictive

Conclusion

Limitations

- Model for co-expression is too restrictive
- Assignment of each gene to a *single* pathway

Conclusion

Limitations

- Model for co-expression is too restrictive
- Assignment of each gene to a *single* pathway
- Pathways should be condition-specific (same goes for protein interaction)

Questions

(1) On which two assumptions about pathways is the model based?

(2) Map each of the previous assumptions into a property of the model

(3) Why must the α parameter in the markov random field be greater than one?

(4) What happens when (a) $\alpha = 1$ or when (b) α is close to infinity?