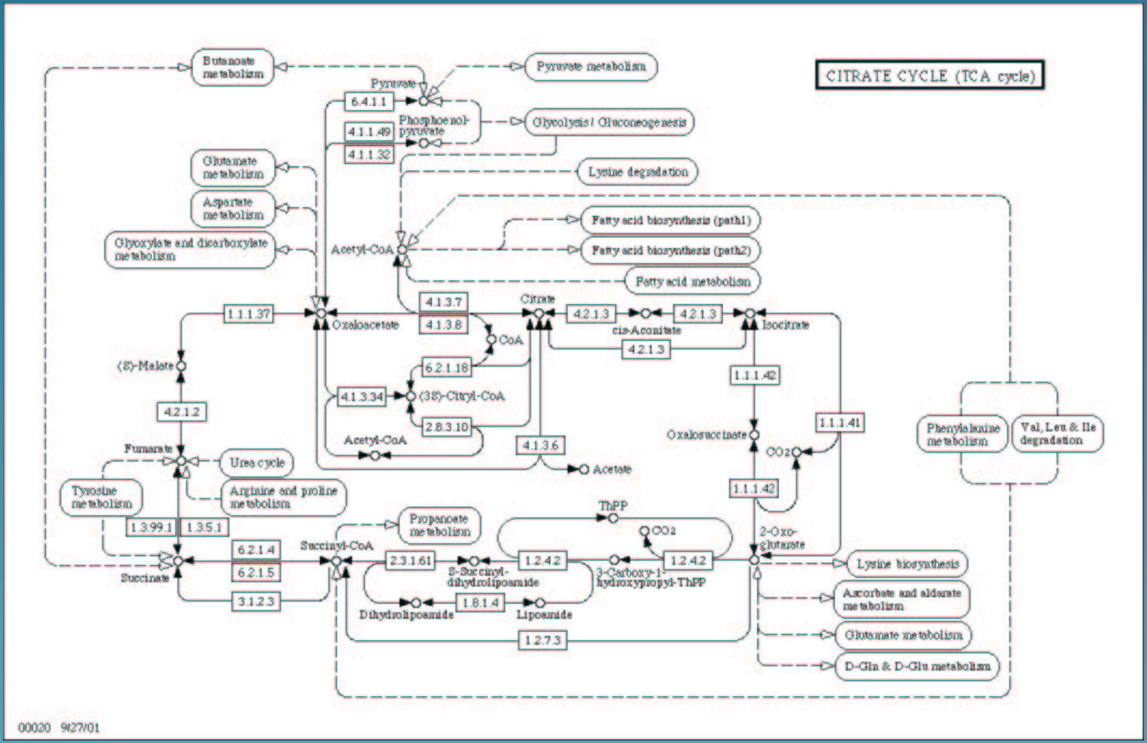
Metabolic networks:
Activity detection and Inference

Jean-Philippe.Vert@mines.org

Ecole des Mines de Paris
Computational Biology group

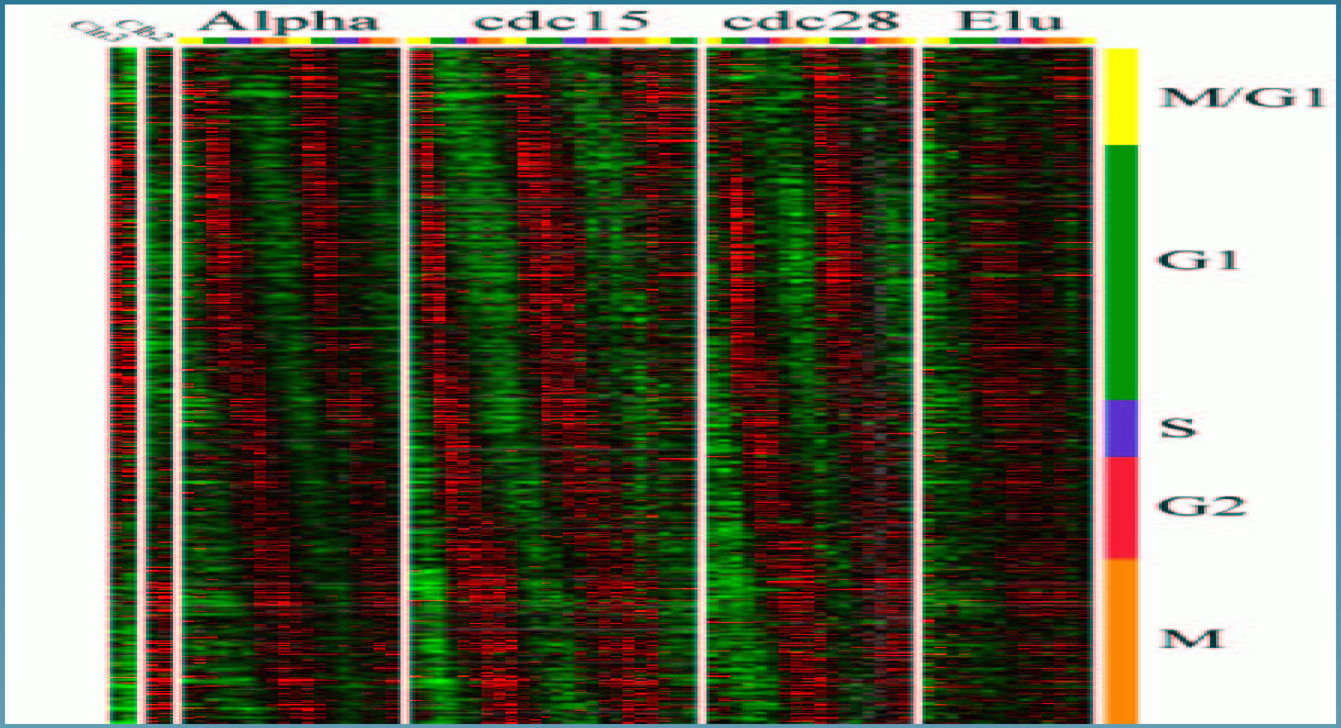
Advanced microarray analysis course, Elsinore, Denmark, May 21th, 2004.

Many metabolic pathways are known



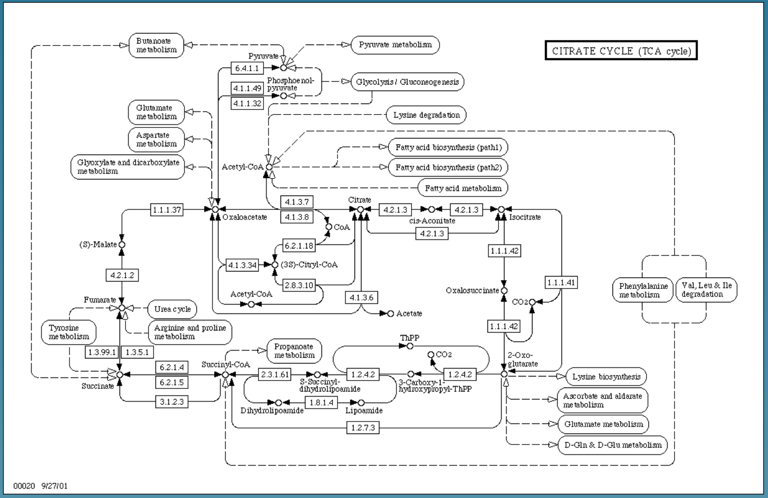
From http://www.genome.ad.jp/kegg/pathway

Microarray technology monitors mRNA quantity

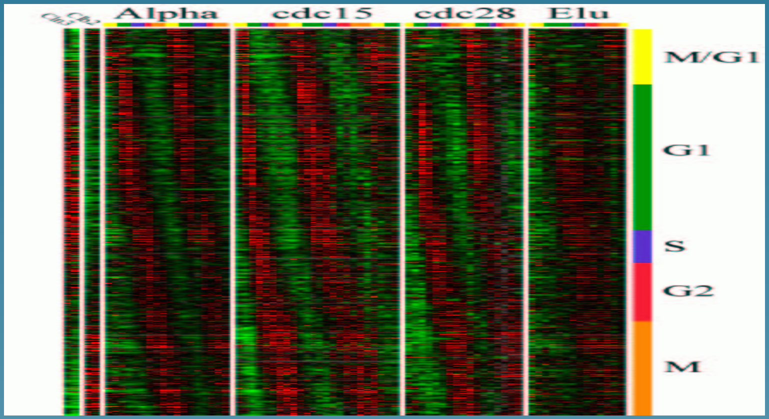


(From Spellman et al., 1998)

Comparing gene expression and pathway databases



VS



Detect active pathways? Denoise expression data?

Denoise pathway database? Find new pathways?

Are there "correlations"?

Overview

1. Feature extractions from expression data only

2. Detecting correlations with the metabolic databse

3. Experiments

4. Inferring new pathways

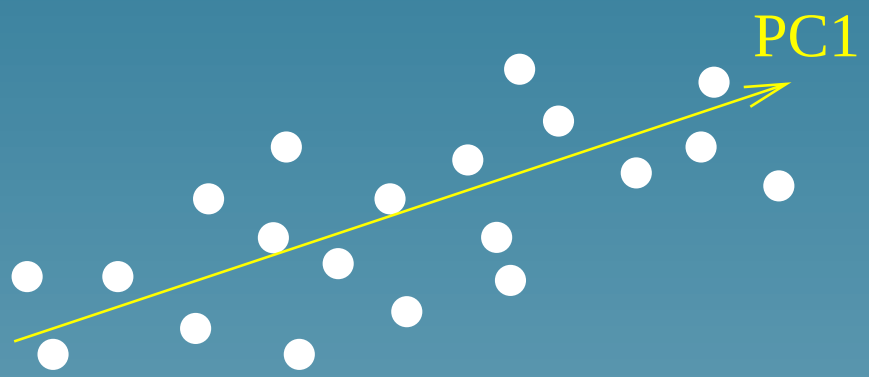
Part 1

Feature extraction from expression data only

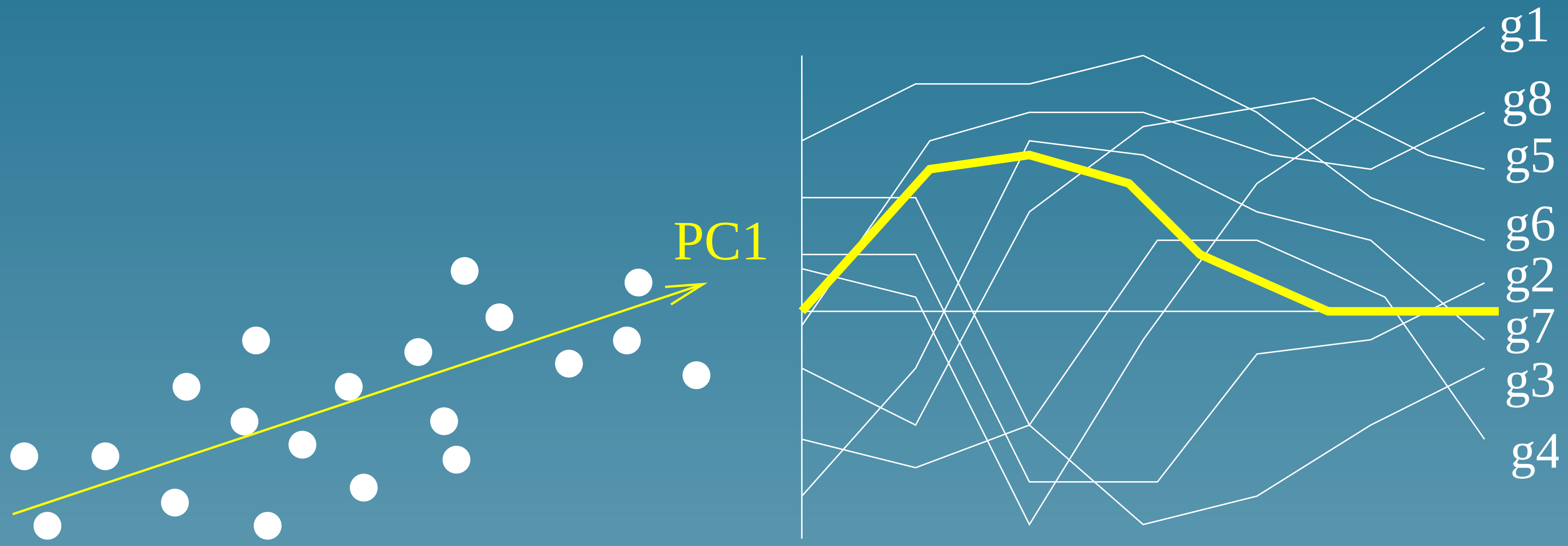
Motivation

- Pathways and biological events involve the coordinated action of several genes
- Co-regulation is an important way to coordinate the action of several genes
- Systematic variations in the set of gene expression profiles might be an indicator of an underlying biological phenomenon

Using microarray only



Using microarray only



PCA finds the directions (*profiles*) explaining the largest amount of variations among expression profiles.

PCA formulation

- Let $f_v(i)$ be the projection of the i-th profile onto v.
- The amount of variation captured by f_v is:

$$h_1(v)^{-1} = \sum_{i=1}^{N} f_v(i)^2$$

- PCA finds an orthonormal basis by solving successively:

$$\min_{||v||=1} h_1(v)$$

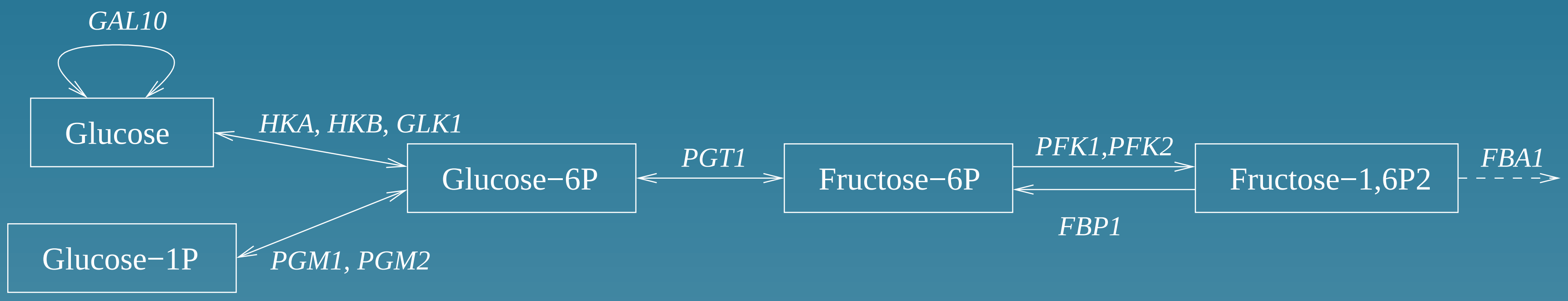
Part 2

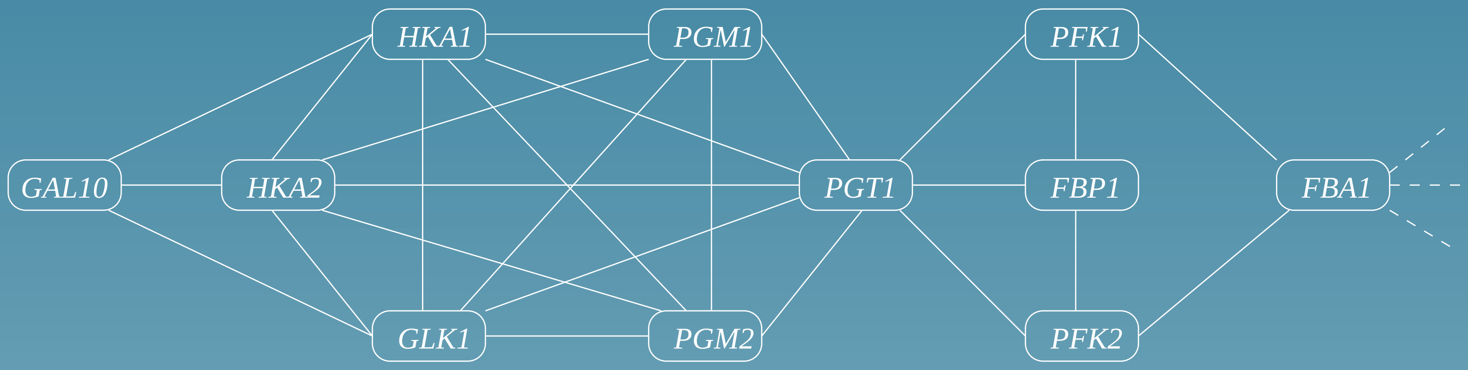
Detecting correlations with the metabolic database

Motivation

- PCA is useful if there is a small number of strong signal
- In concrete applications, we observe a noisy superposition of many events
- Using a prior knowledge of metabolic networks can help denoising the information detected by PCA

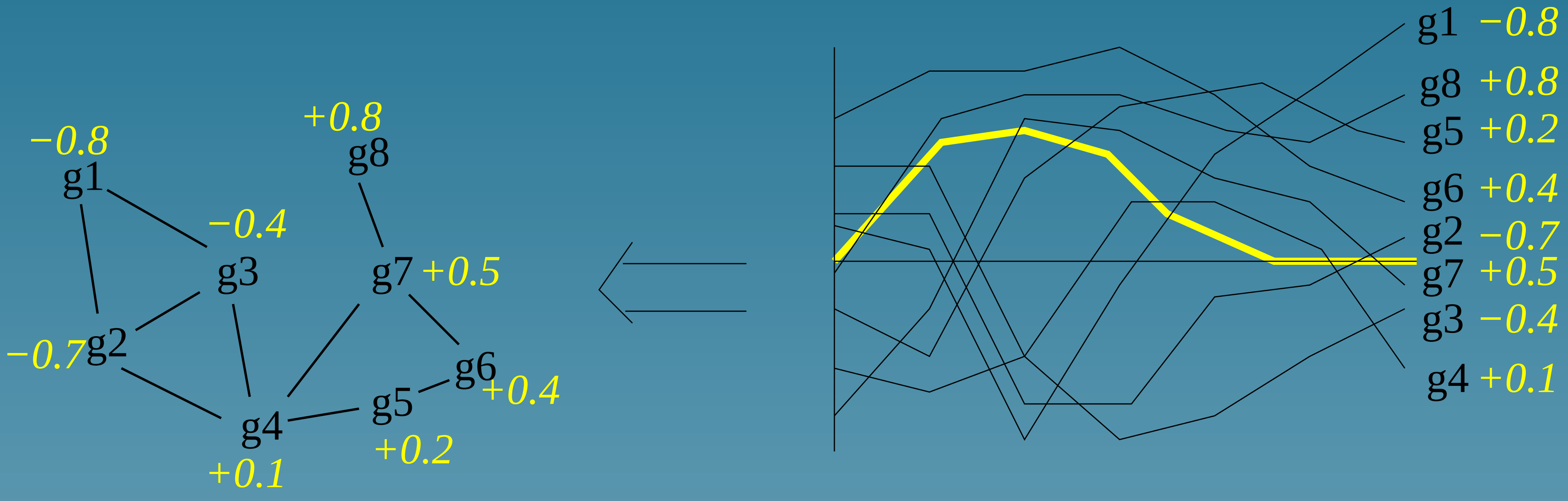
The metabolic gene network





Link two genes when they can catalyze two successive reactions

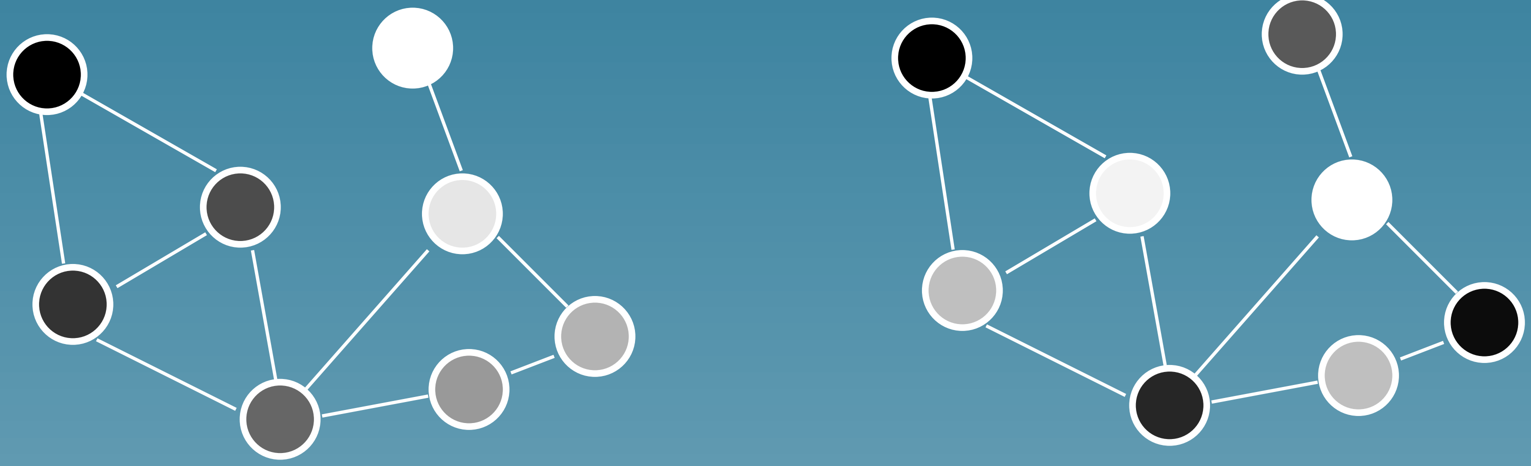
Mapping f_v to the metabolic gene network



Does it look interesting or not?

Important hypothesis

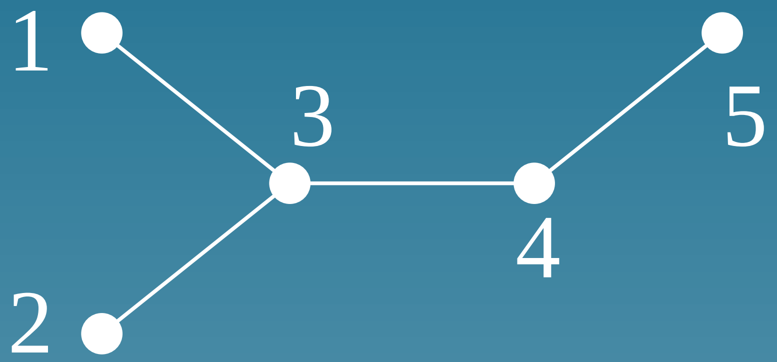
If v is related to a metabolic activity, then f_v should vary ”smoothly” on the graph



Smooth

Rugged

Graph Laplacian $L = D - A$



$$L = \begin{pmatrix} 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ -1 & -1 & 3 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 1 \end{pmatrix}$$

Smoothness quantification

$$h_2(f) = \sum_{i \sim j} (f(i) - f(j))^2 = f^\top L f$$

or

$$h_2(f) = \sum_i \hat{f}_i^2 e^{\beta \omega_i} = f^\top \exp(\beta L) f$$

is small when f is smooth

Where we are now...

For a candidate profile v,

- $h_1(f_v)$ is small when v captures a lot of natural variation among profiles

- $h_2(f_v)$ is small when f_v is smooth on the graph

Try to minimize both terms in the same time

Problem reformulation

Find a function f_v (and therefore a profile v) that solves:

$$\min_{v} \{h_1(f_v) + \lambda h_2(f_v)\}$$

λ is a parameter that controls the trade-off.

Solving the problem

- By the representer theorem, v can be expanded as:

$$v = \sum_{i=1}^{n} \alpha_i e(x_i).$$

Solving the problem (cont.)

- The problem can then be rewritten:

$$\min_{\alpha \in \mathbb{R}^n} \left\{ \alpha^\top K_0 K_2 K_0 \alpha + \lambda \alpha^\top K_0 \alpha \right\}$$

under the constraint $\alpha^\top K_0^2 \alpha = 1$, where:

⋆ $K_2 = \exp(-\beta L)$ is the $n \times n$ diffusion kernel

⋆ K_0 is the centered $n \times n$ Gram matrix ($[K_0]_{i,j} = e_i^\top e_j$)

Solving the problem (cont.)

- The problem can then be rewritten:

$$\min_{\alpha\in\mathbb{R}^n}\left\{\alpha^\top K_0K_2K_0\alpha+\lambda\alpha^\top K_0\alpha\right\}$$

under the constraint $\alpha^\top K_0^2\alpha=1$, where:

⋆ $K_2=\exp(-\beta L)$ is the $n\times n$ diffusion kernel
⋆ K_0 is the centered $n\times n$ Gram matrix $([K_0]_{i,j}=e_i^\top e_j\ )$

- It is equivalent to solving the generalized eigenvalue problem:

$$(K_2K_0+\lambda I)\alpha=\mu K_0\alpha.$$

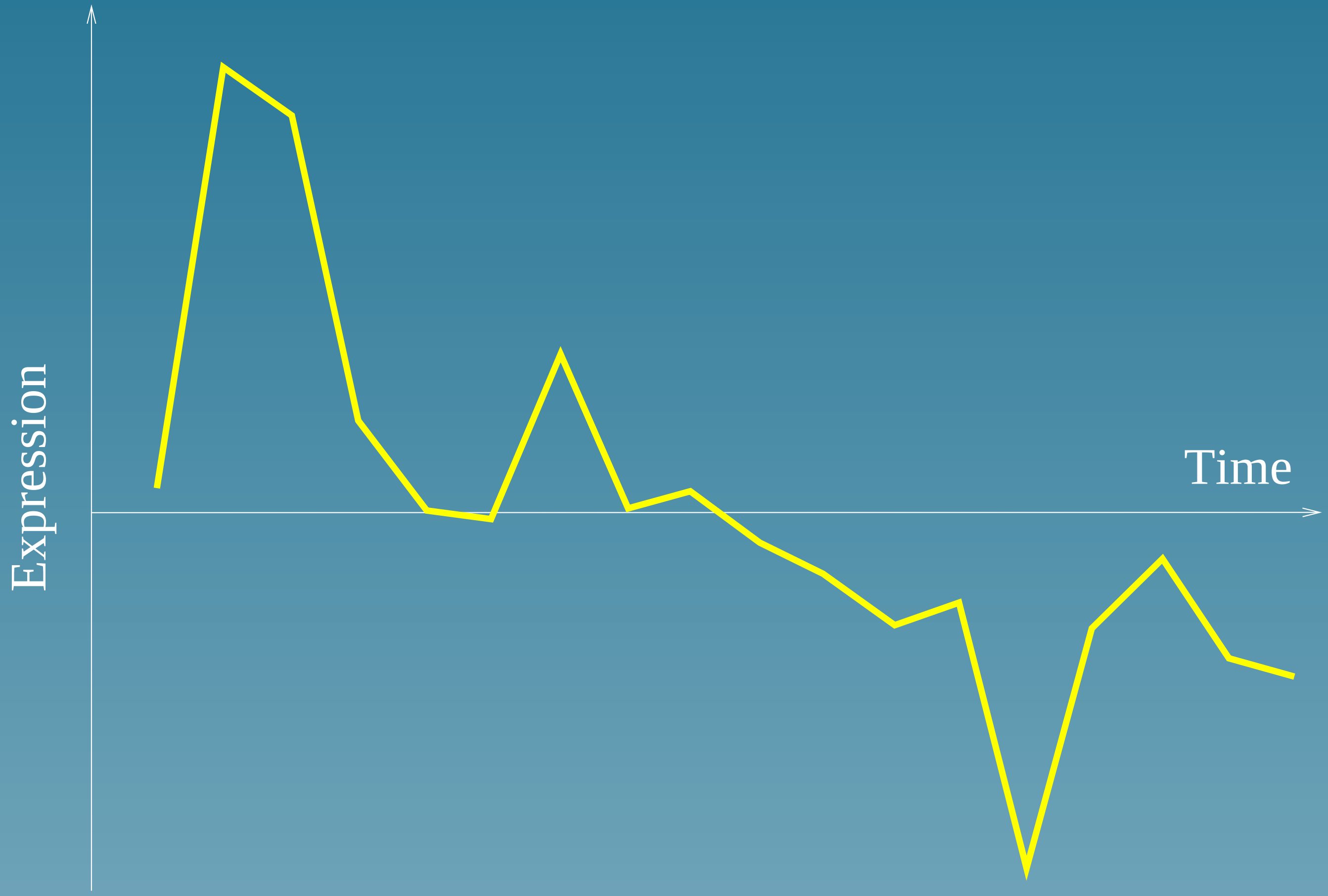
Part 3

Experiments

Data

- Gene network: two genes are linked if the catalyze successive reactions in the KEGG database (669 yeast genes)
- Expression profiles: 18 time series measures for the 6,000 genes of yeast, during two cell cycles

First pattern of expression

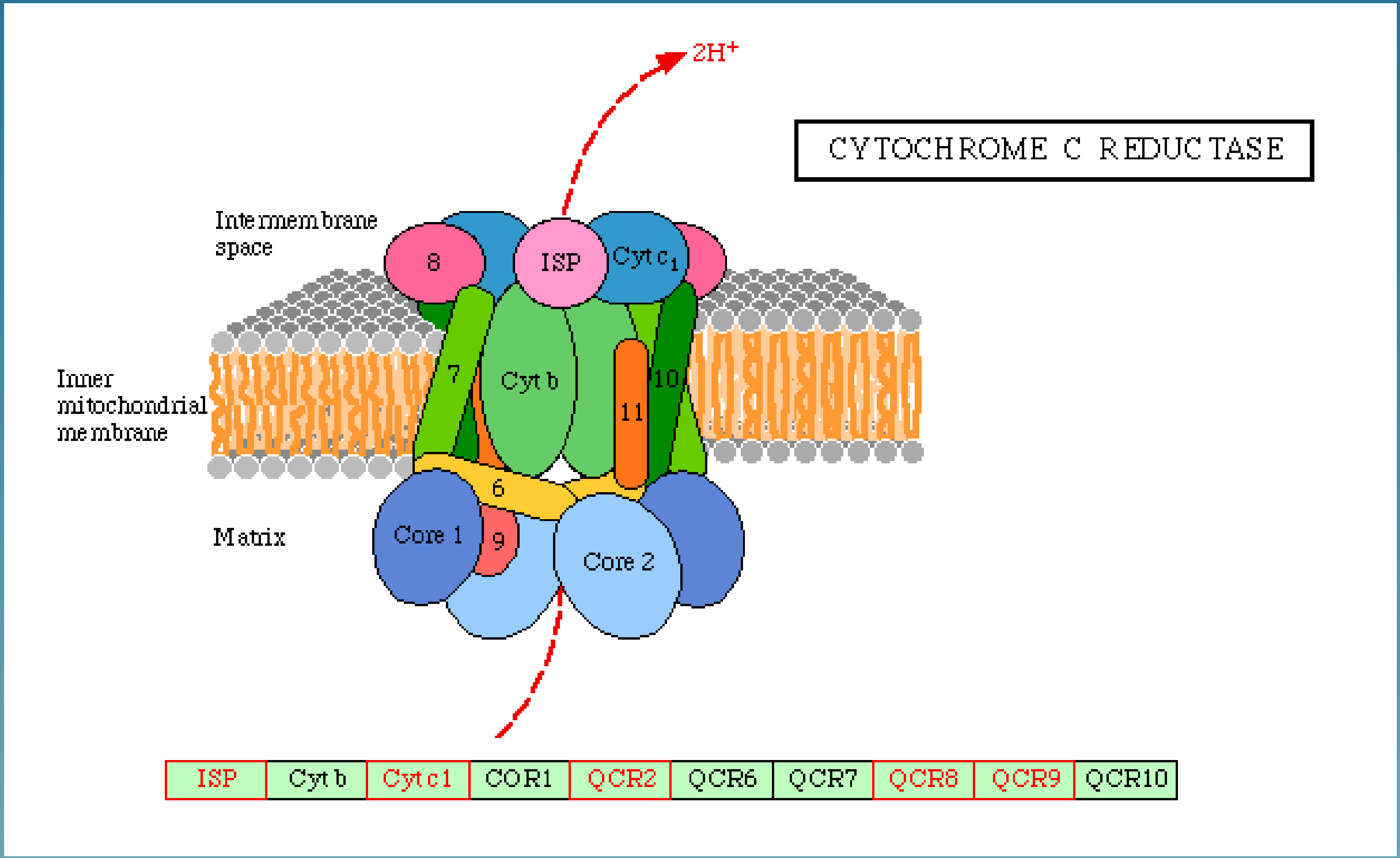


Related metabolic pathways

50 genes with highest $s_2 - s_1$ belong to:

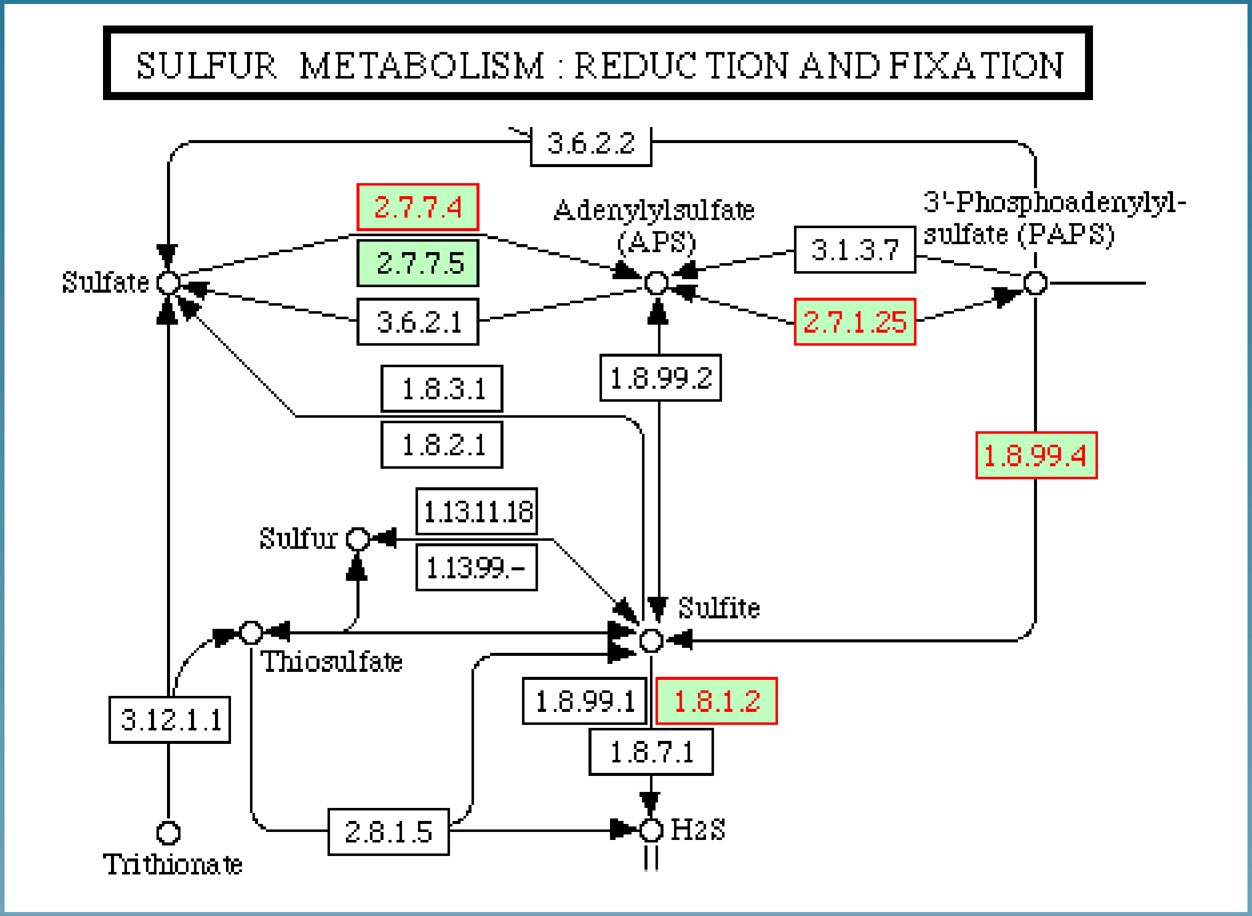
- Oxidative phosphorylation (10 genes)
- Citrate cycle (7)
- Purine metabolism (6)
- Glycerolipid metabolism (6)
- Sulfur metabolism (5), etc...

Related genes

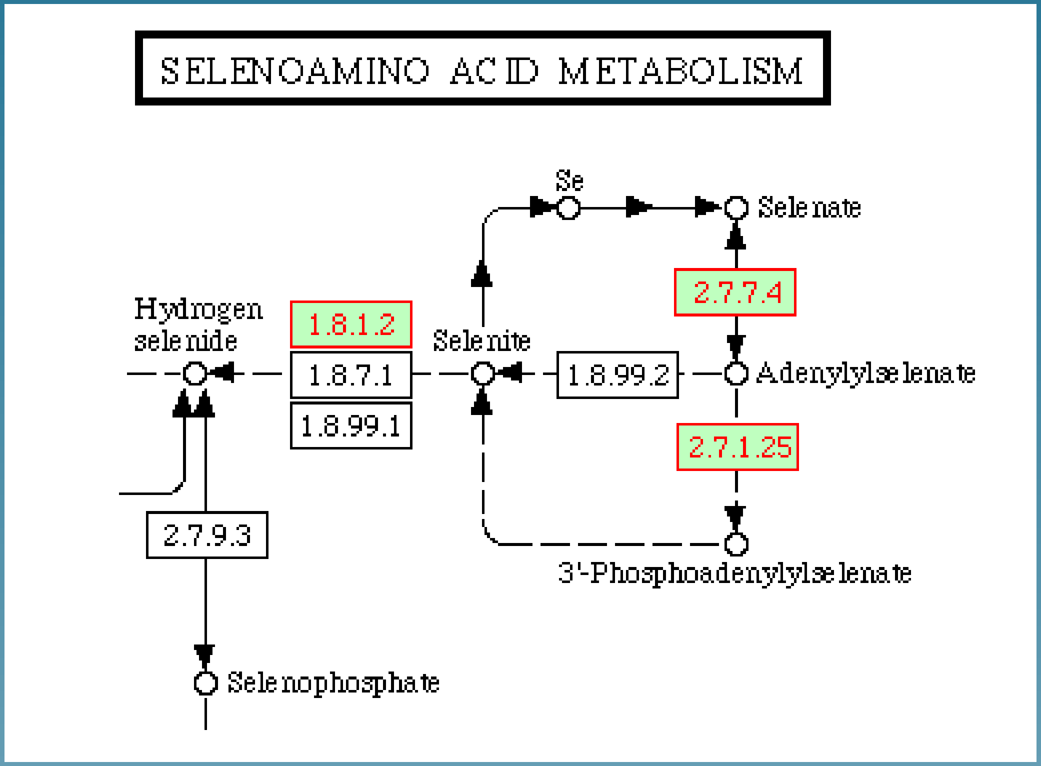


ISP	Cyt b	Cyt c1	COR1	QCR2	QCR6	QCR7	QCR8	QCR9	QCR10

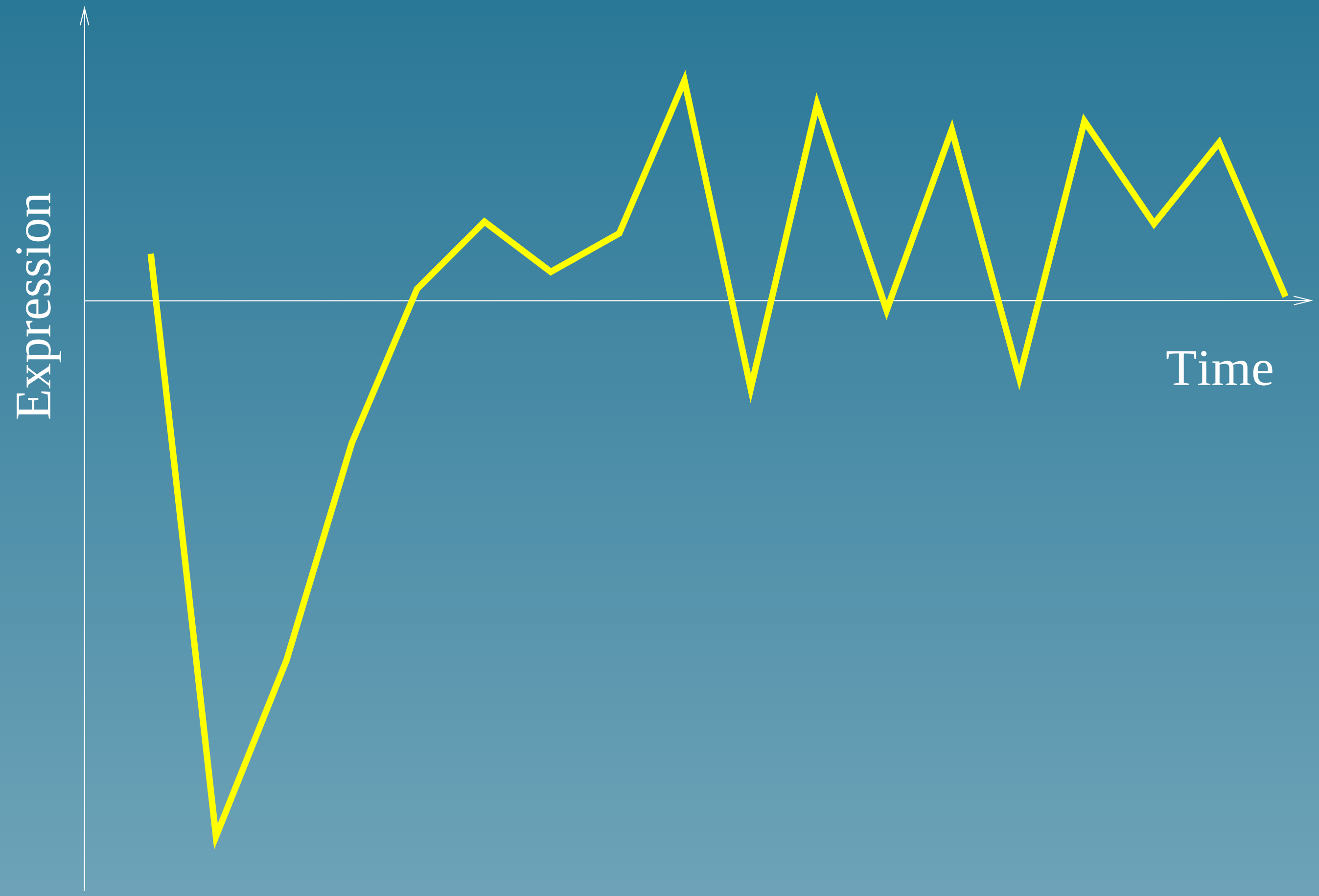
Related genes



Related genes



Opposite pattern

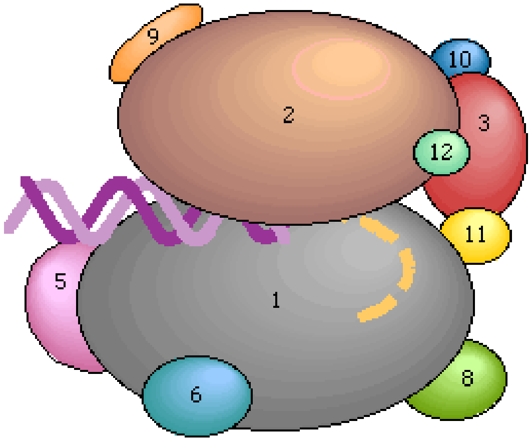


Related genes

- RNA polymerase (11 genes)
- Pyrimidine metabolism (10)
- Aminoacyl-tRNA biosynthesis (7)
- Urea cycle and metabolism of amino groups (3)
- Oxidative phosphorlation (3)
- ATP synthesis(3) , etc...

Related genes

RNA POLYMERASE



RNA polymerase II (Saccharomyces cerevisiae)

Eukaryotic Pol II

B2	B3	B4	B5	B6	B7
B1	B8	B9	B10	B11	B12

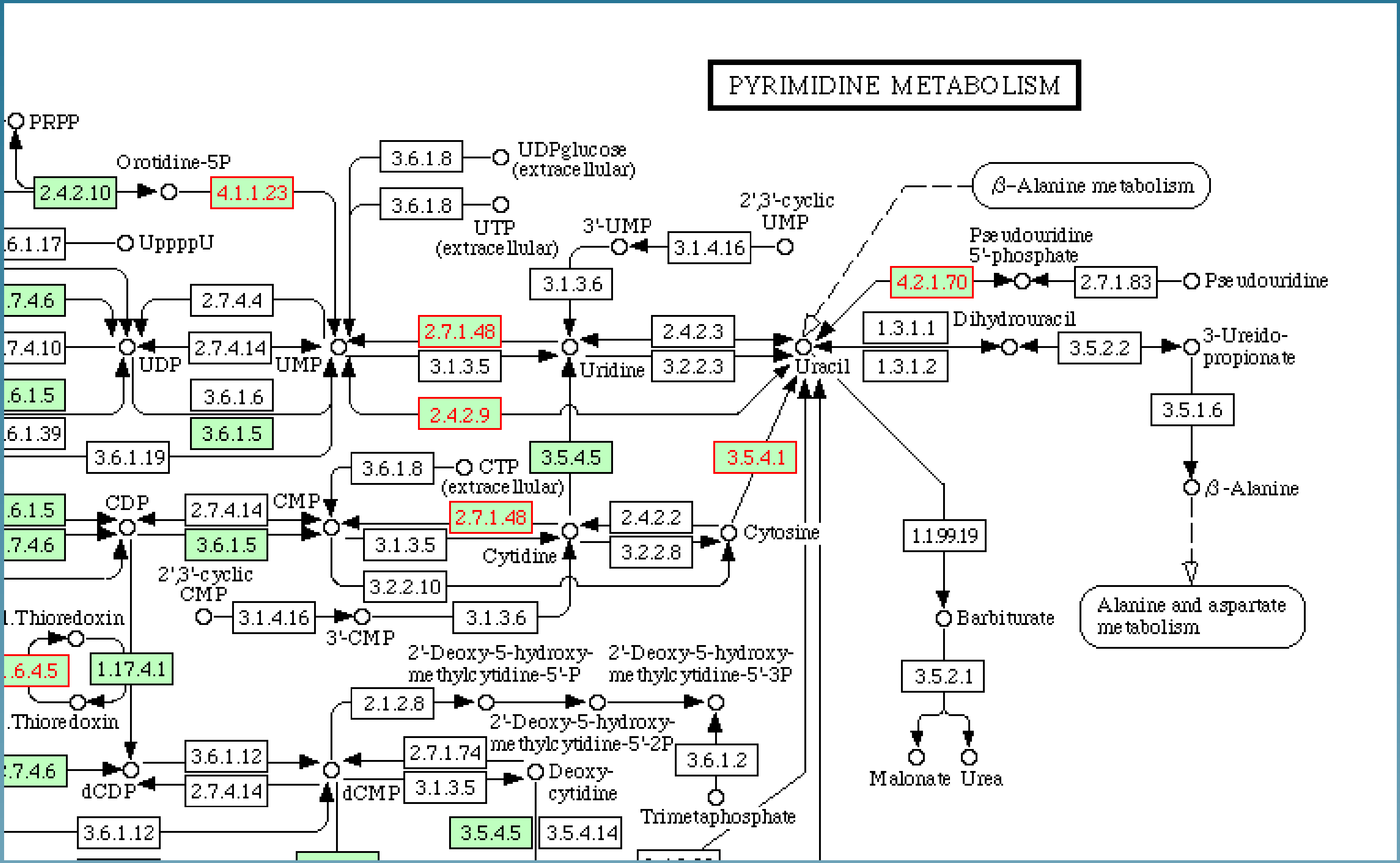
Eukaryotic Pol III

C2	C3	C4	C5	C11
C1	C19	C25	C31	C34

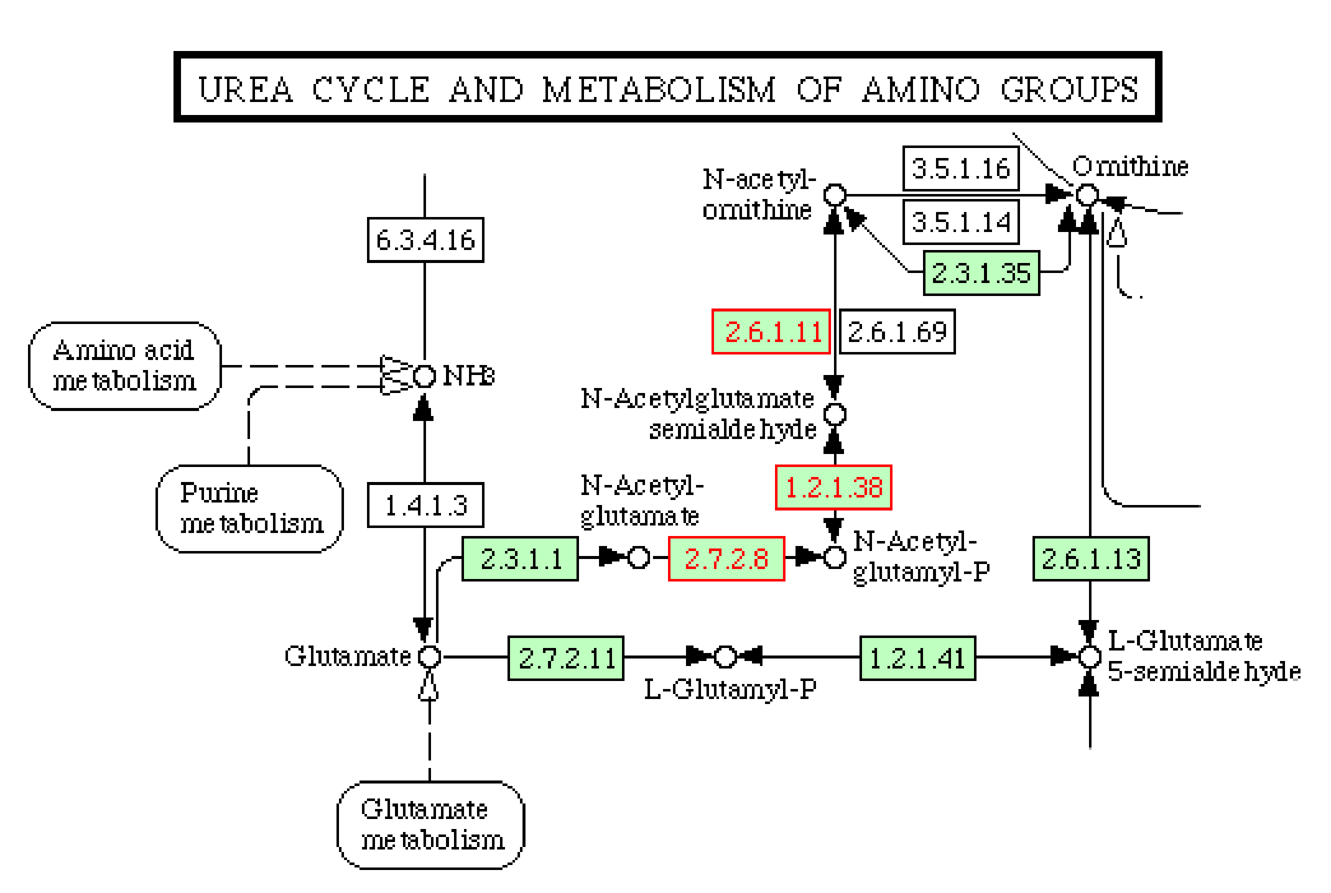
Eukaryotic Pol I

A2	A12	A14	A34	A43	A49
A1					

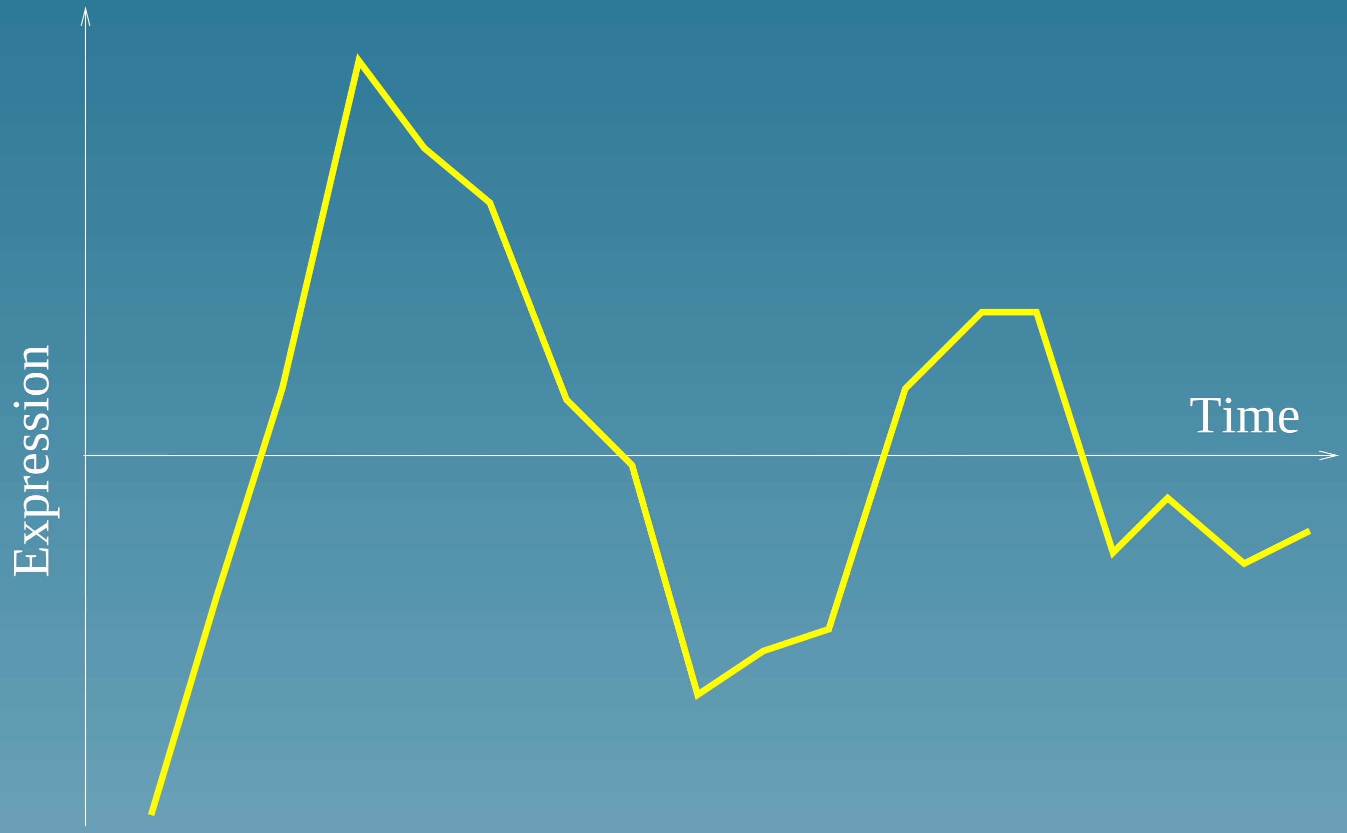
Related genes



Related genes



Second pattern



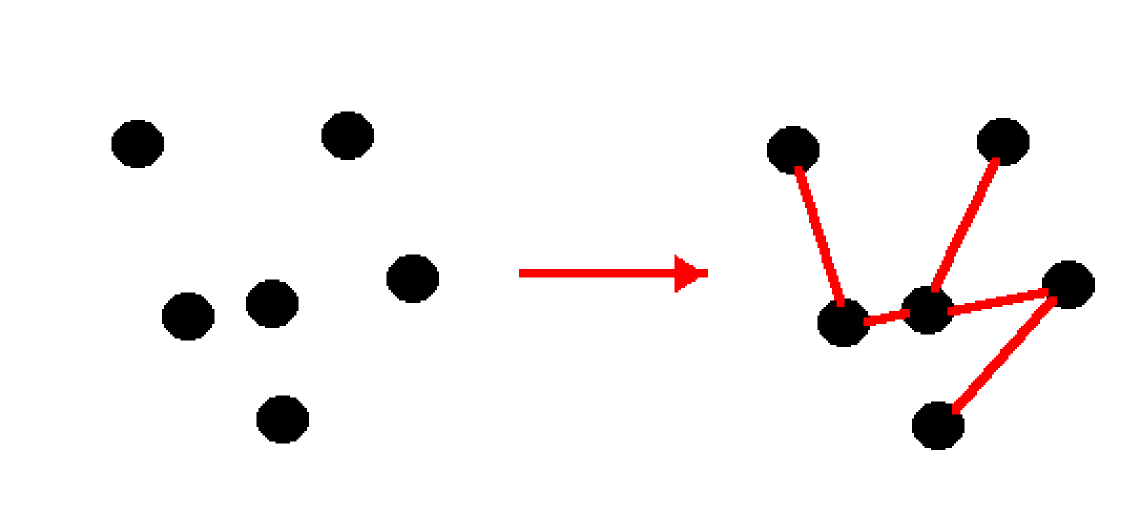
Part 4

Inferring new pathways

(with Y.Yamanishi)

The network inference problem

Given some measurement/observation about the genes (sequences, structure, expression, ...), infer "the" gene network



Related approaches

- Bayesian nets for regulatory networks (Friedman et al. 2000)
- Boolean networks (Akutsu, 2000)
- Joint graph method (Marcotte et al, 1999)

A direct (unsupervised) approach

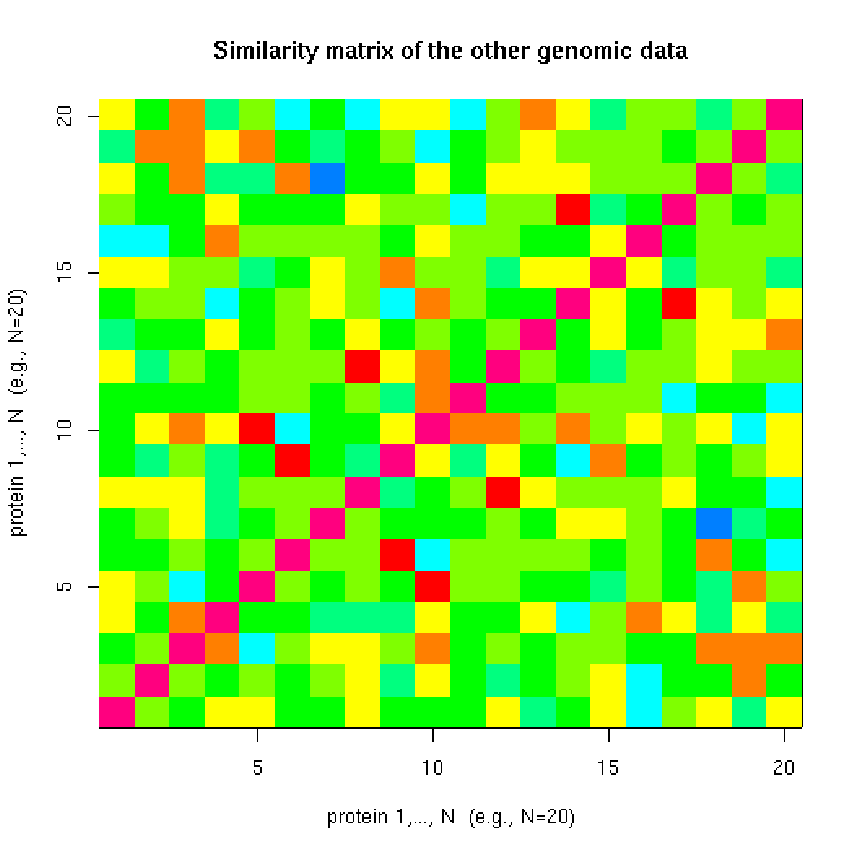
- Let $K(x,y)$ be a measure of similarity (a kernel) between genes x and y based on available measurements, e.g.,

$$K(x,y) = \exp\left(-\frac{||e(x) - e(y)||^2}{2\sigma^2}\right)$$

- For a set of n genes $\{x_1,\ldots,x_n\}$, let K be the $n \times n$ matrix of pairwise similarity (Gram matrix)

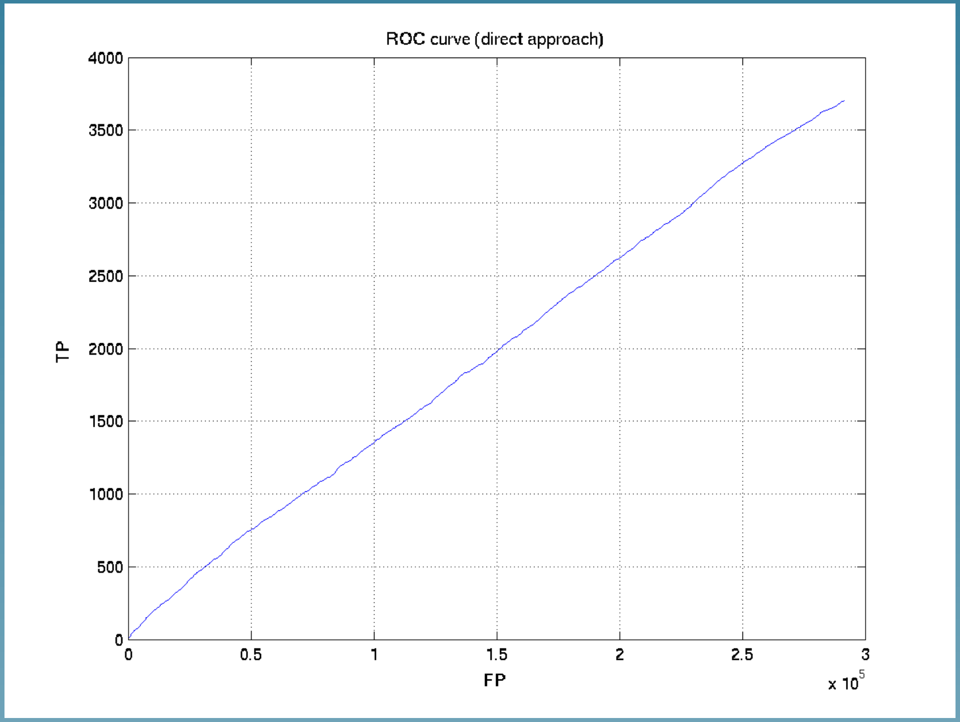
- Direct strategy: add edges between genes by decreasing similarity.

Example of similarity matrix

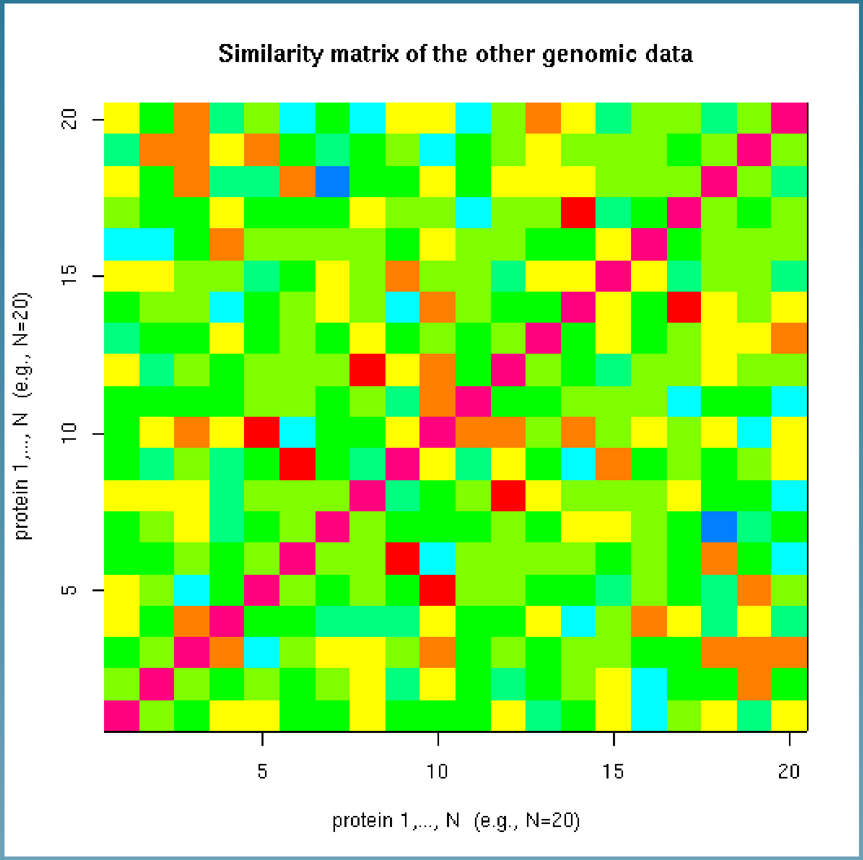


Evaluation of the direct approach

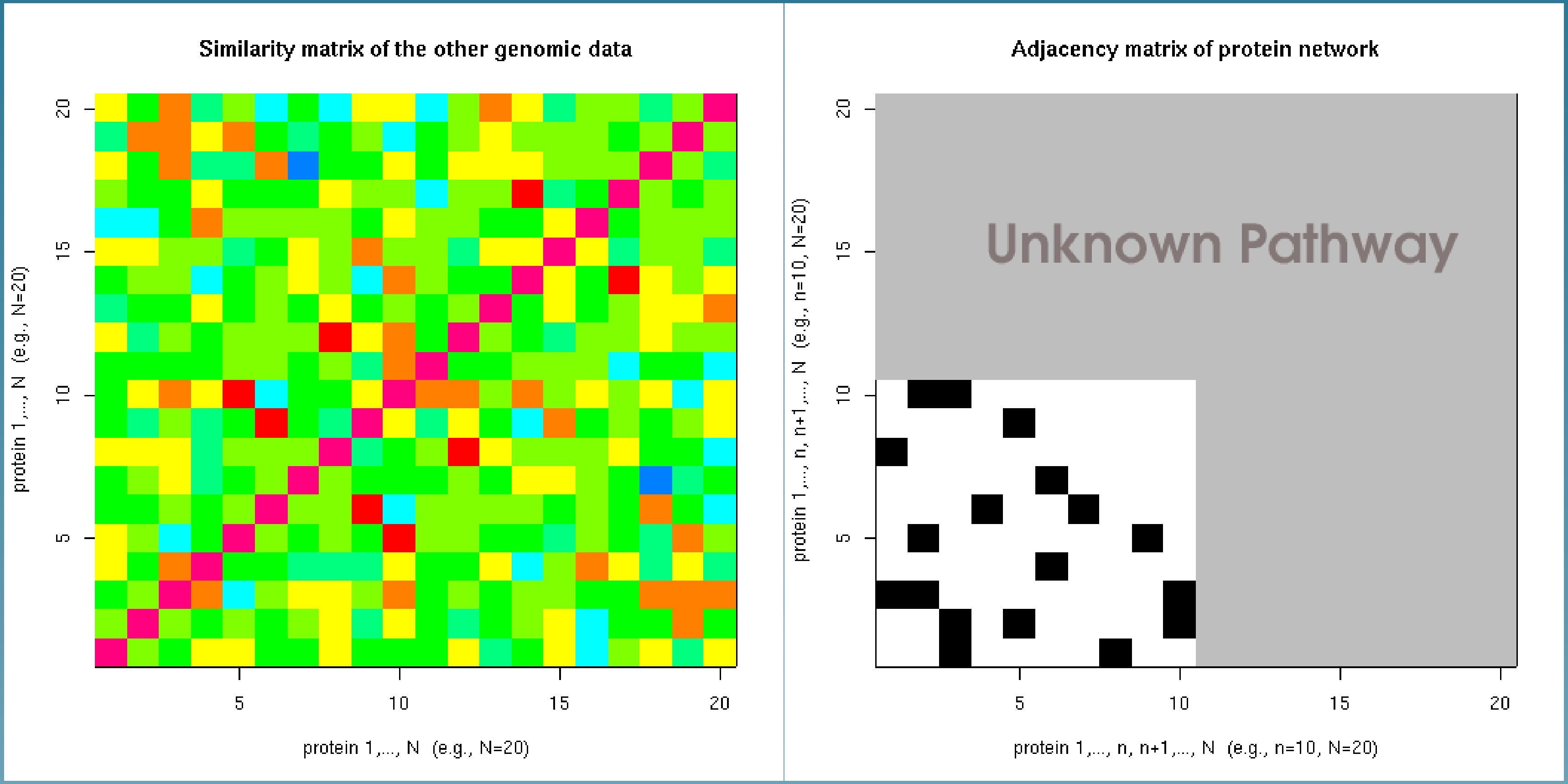
The metabolic network of the yeast involves 769 genes. Each gene is represented by 157 expression measurements. (ROC=0.52)



The supervised gene inference problem



The supervised gene inference problem



The idea in a nutshell

- Use the known network to define a more relevant measure of similarity
- For any positive definite similarity $n \times n$ matrix, there exists a representation as n-dimensional vectors such that the matrix similarity is exactly the similarity between vectors.
- In this space, look for projections onto small-dimensional spaces that better fit the known network.

A two-step strategy

- First map any gene x onto a vector

$$\Phi(x) = (f_1(x), \ldots, f_d(x))' \in \mathbb{R}^d$$

A two-step strategy

- First map any gene x onto a vector

$$\Phi(x) = (f_1(x), \ldots, f_d(x))' \in \mathbb{R}^d$$

- Then apply the direct strategy to reconstruct the graph from the images $\{\Phi(x_1), \ldots, \Phi(x_n)\}$

A two-step strategy

- First map any gene x onto a vector

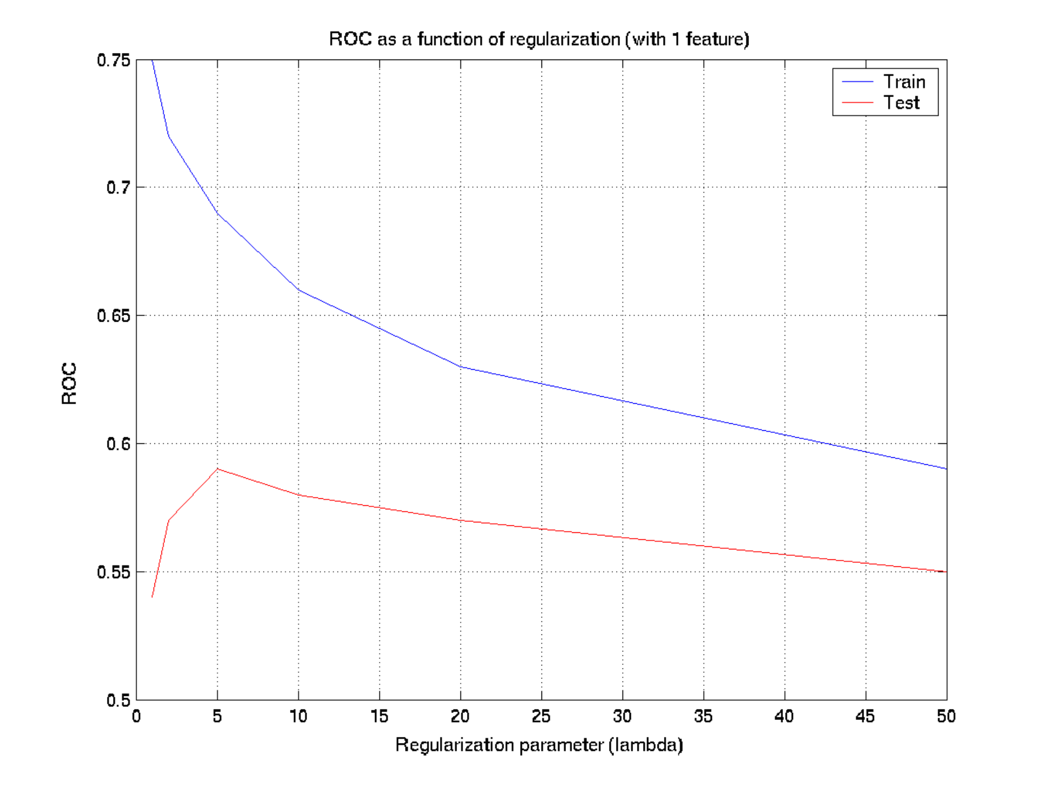
$$\Phi(x) = (f_1(x), \ldots, f_d(x))' \in \mathbb{R}^d$$

- Then apply the direct strategy to reconstruct the graph from the images $\{\Phi(x_1), \ldots, \Phi(x_n)\}$
- The functions $f_1, \ldots, f_d$ can be learned from the knowledge of the graph on the first n genes

Choice of f

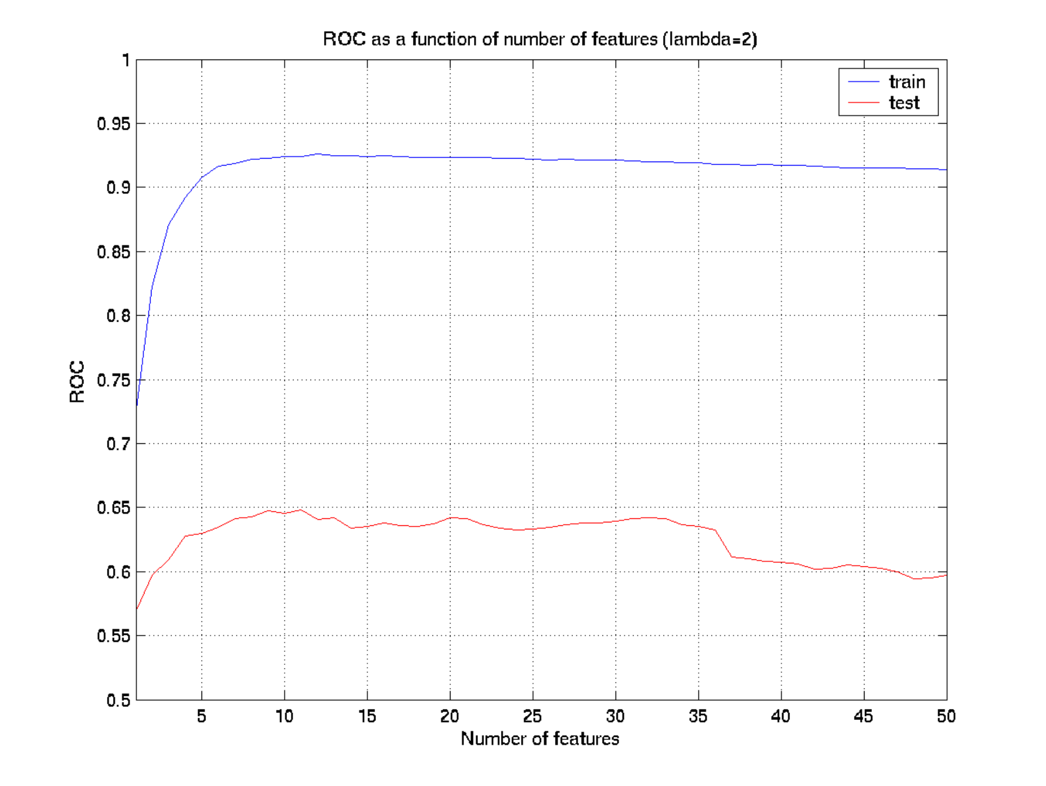
- A feature $f : \mathcal{X} \rightarrow \mathbb{R}$ is good on the training set if connected genes have similar value.
- This is exactly what we did in the previous part!
- So use the features already extracted to map new genes onto a vector space by projection

Evaluation of the supervised approach: effect of λ



Metabolic network, 10-fold cross-validation, 1 feature

Evaluation of the supervised approach: number of features ($\lambda = 2$)



Learning from heterogeneous data

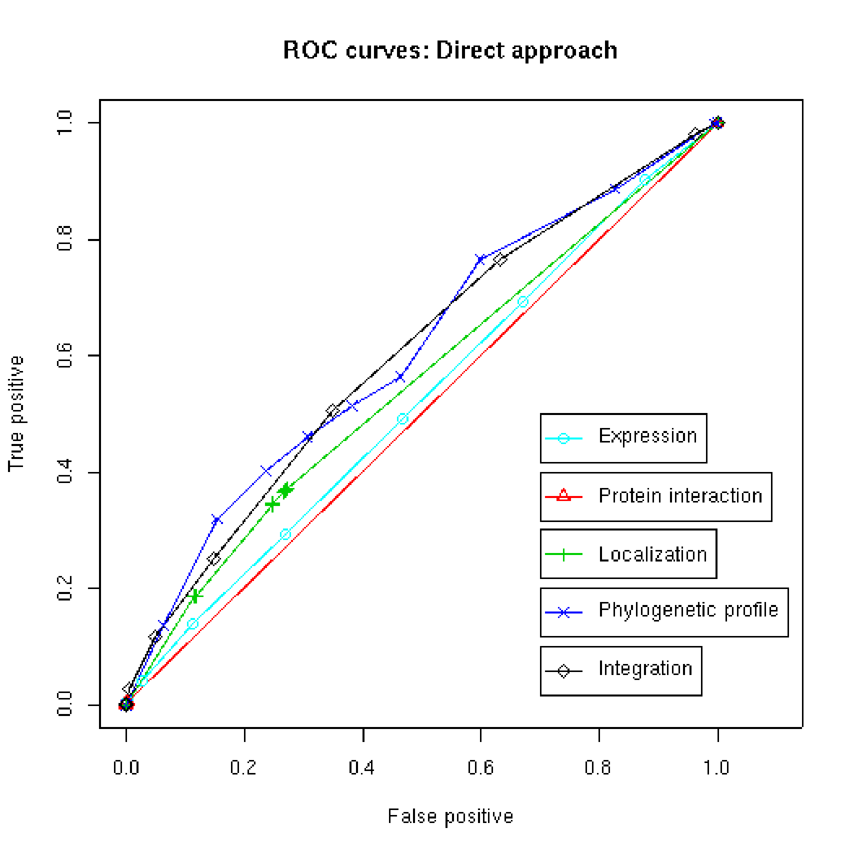
- Suppose several data are available about the genes, e.g., expression, localization, struture, predicted interaction etc...

- Each data can be represented by a positive definite similarity matrix $K_1, \ldots, K_p$ called kernels

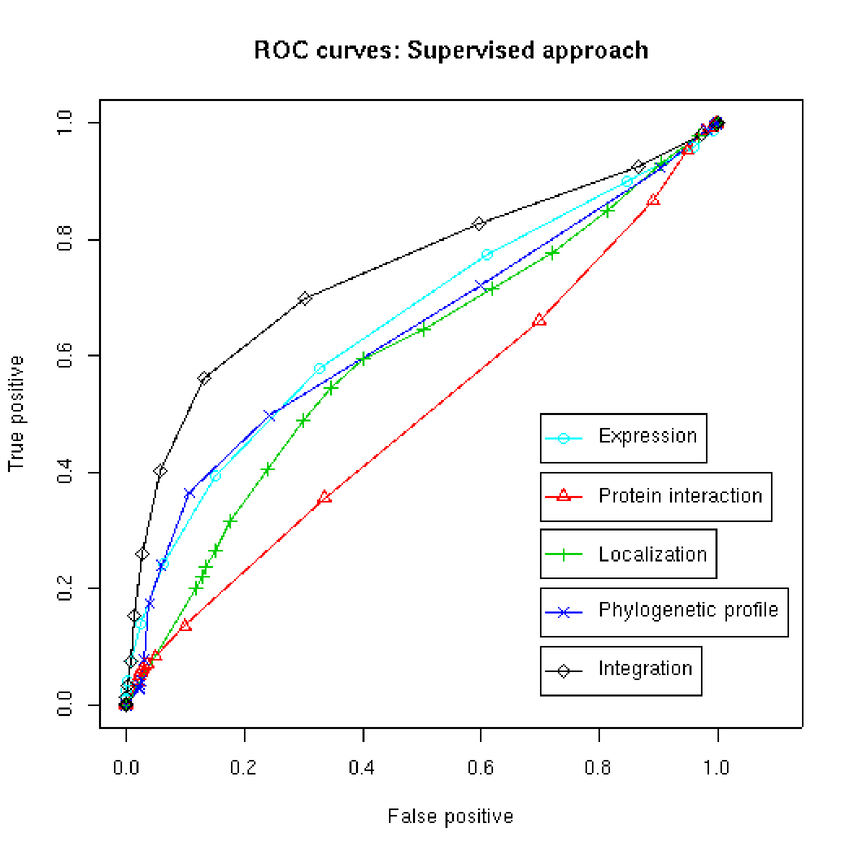
- Kernel can be combined by various operations, e.g., addition:

$$K = \sum_{i=1}^{p} K_i$$

Learning from heterogeneous data (unsupervised)



Learning from heterogeneous data (supervised)



Extensions

- The diffusion kernel can be replaced by another graph kernel
- Other formulations can lead to kernel CCA (NIPS 02)

Open questions / Ongoing work

- What should be the number of features (problem of embedding a graph in low dimension)
- Other cost functions
- How to better integrate several similarities? (semi-definite programming?)

Conclusion

Conclusion

- A new approach to feature extractions and supervised network inference, many possible variants and extensions
- Straightforward generalization to any network (e.g., interactome): the same data can be used to infer different networks
- Possible connections with other algorithms (SVM, kernel CCA..)