Extracting metabolic pathways activity from gene expression data

Jean-Philippe Vert

Computational biology group
Ecole des Mines de Paris

Jean-Philippe.Vert@mines.org

CEA, Feb 5, 2004.

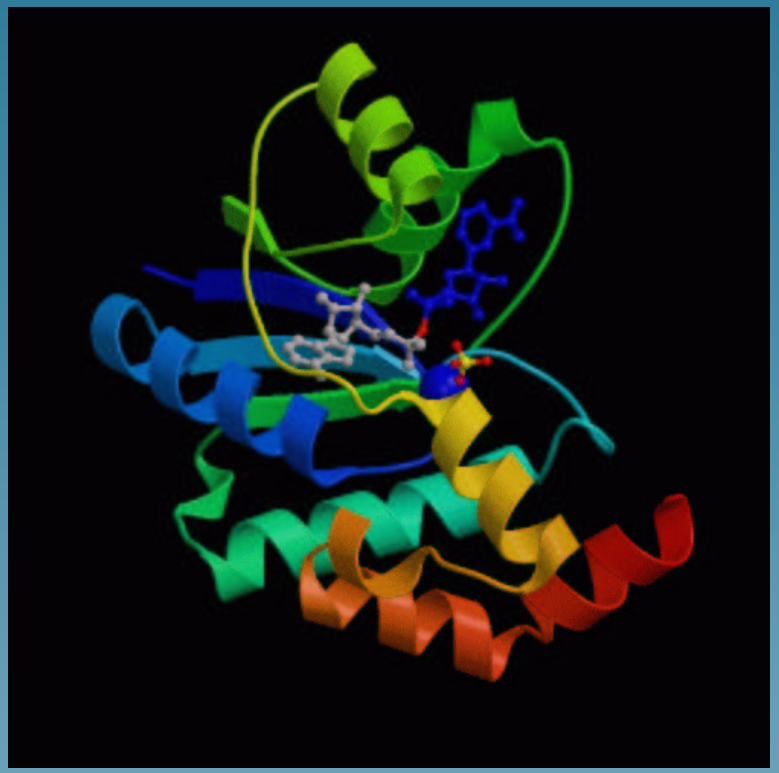
Overview

1. Problem formulation
2. Using expression data only
3. Using a pathway database
4. Combining expression and pathways
5. Experiments

Part 1

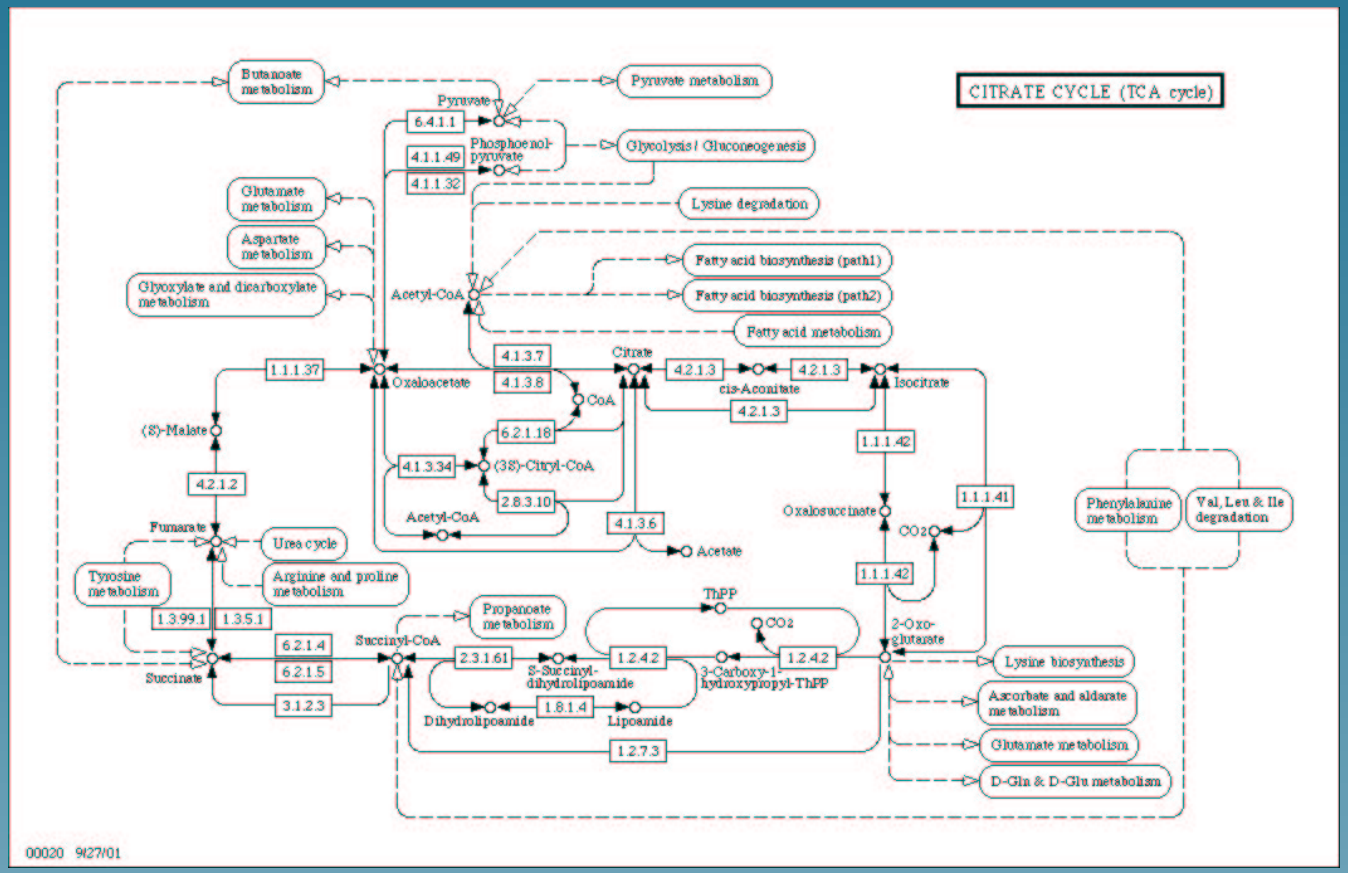
Problem formulation

Genes encode proteins which can catalyse chemical reations



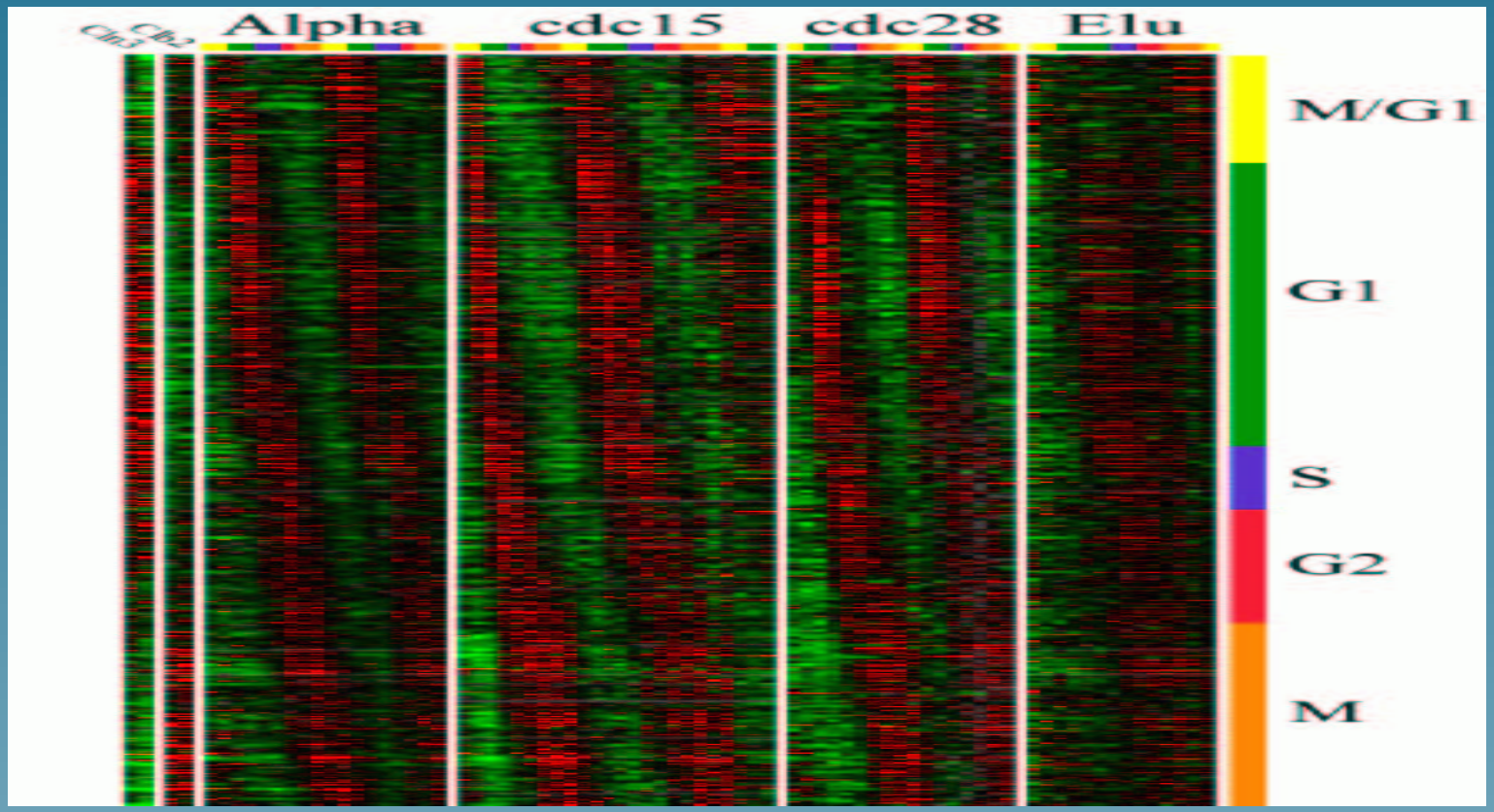
Nicotinamide Mononucleotide Adenylyltransferase With Bound Nad+

Chemical reactions are often parts of pathways



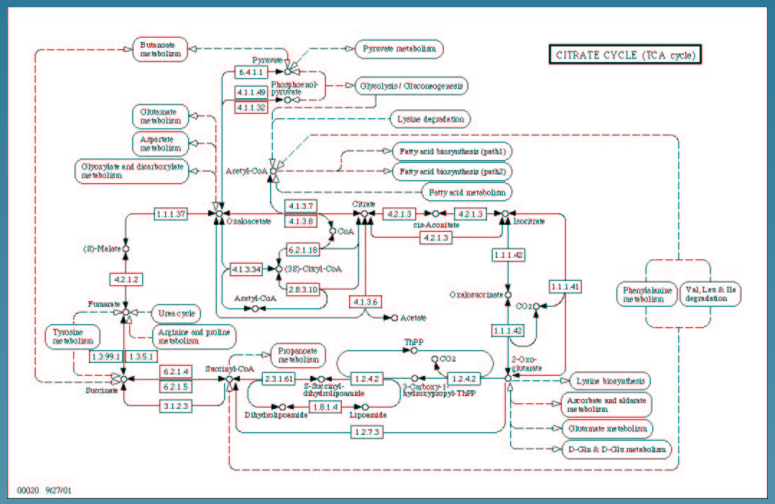
From http://www.genome.ad.jp/kegg/pathway

Microarray technology monitors RNA quantity

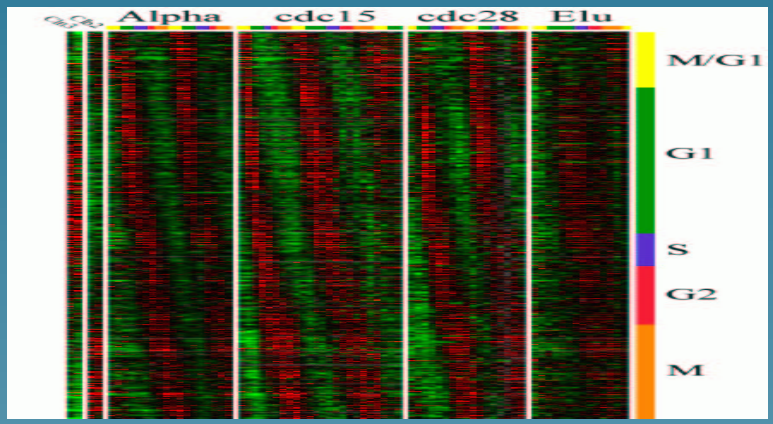


(From Spellman et al., 1998)

Comparing gene expression and pathway databases

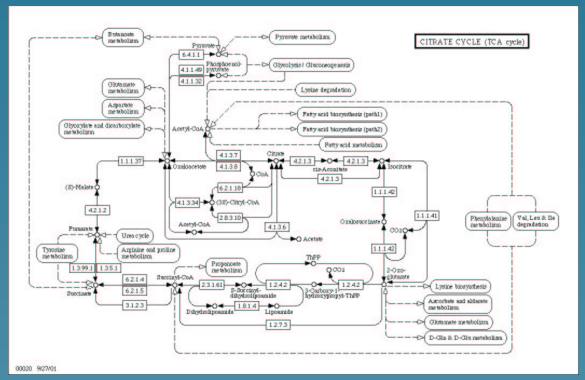


VS

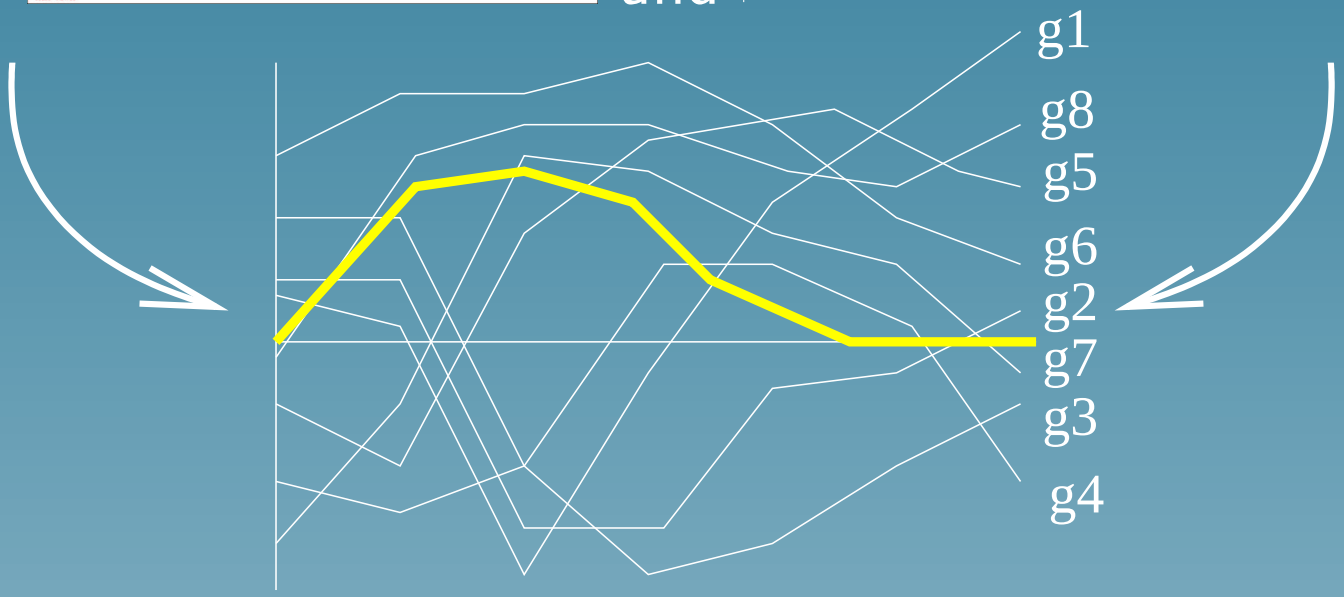


Detect active pathways? Denoise expression data?
Denoise pathway database? Find new pathways?
Are there "correlations"?

A useful first step



and



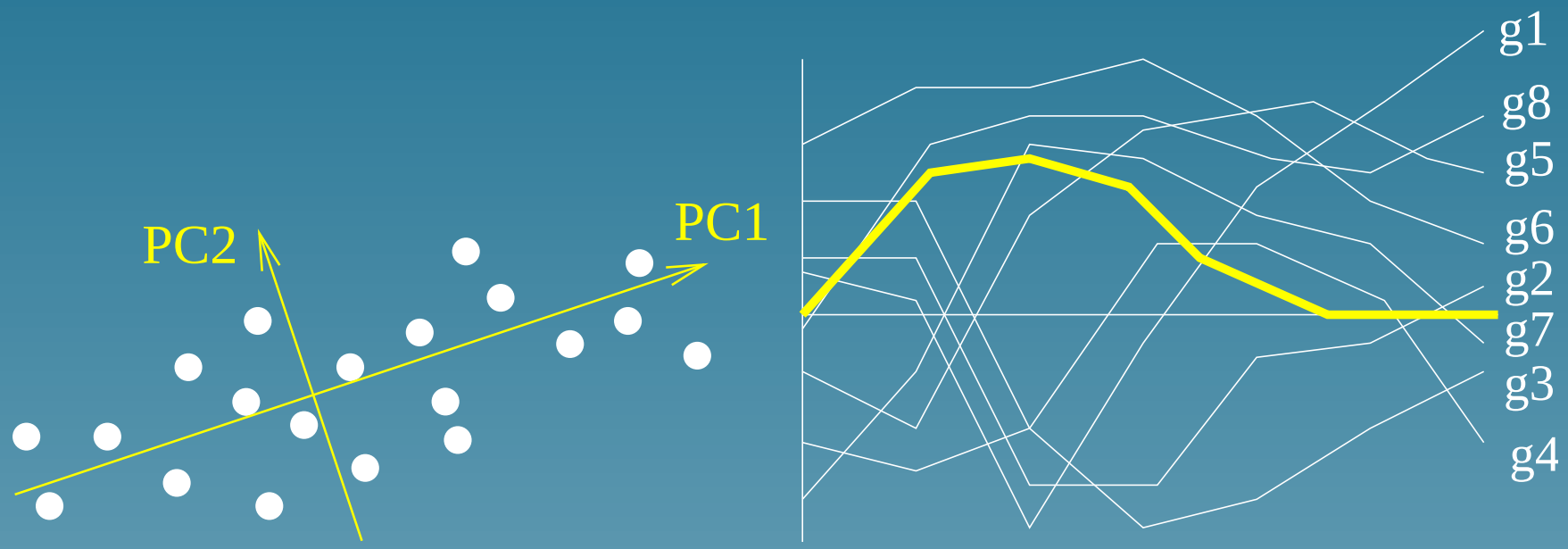
Part 1

Using 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

Principal component analysis (PCA)



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

PCA notations

- N genes, P experimental conditions
- $e_i \in \mathbb{R}^P$ the expression profile of gene $i = 1, \ldots, N$.
- The expression profiles are centered: $\sum_{i=1}^{N} e_i = 0$
- For a candidate profile $v \in \mathbb{R}^p$, $f_v(i) = v^\top e_i$ the projection of e_i onto v

PCA classical formulation

- The amount of variation captured by f_v is:

$$||f_v||^2_{L_2} = \sum_{i=1}^{N} f_v(i)^2$$

- The norm of v is

$$||f_v||^2_{H_1} = \sum_{i=1}^{P} v_i^2$$

- PCA solves:

$$\max_{||f_v||_{H_1}=1} ||f_v||^2_{L_2} = \max_{f_v} \frac{||f_v||^2_{L_2}}{||f_v||^2_{H_1}}$$

PCA conclusion

- For any candidate profile $v \in \mathbb{R}^p$,

$$h_1(v) = \frac{||f_v||^2_{L_2}}{||f_v||^2_{H_1}}$$

  is a first indicator of how relevant v is: the larger the better

- In the absence of other information, maximizing $h(v)$ is natural: this is PCA

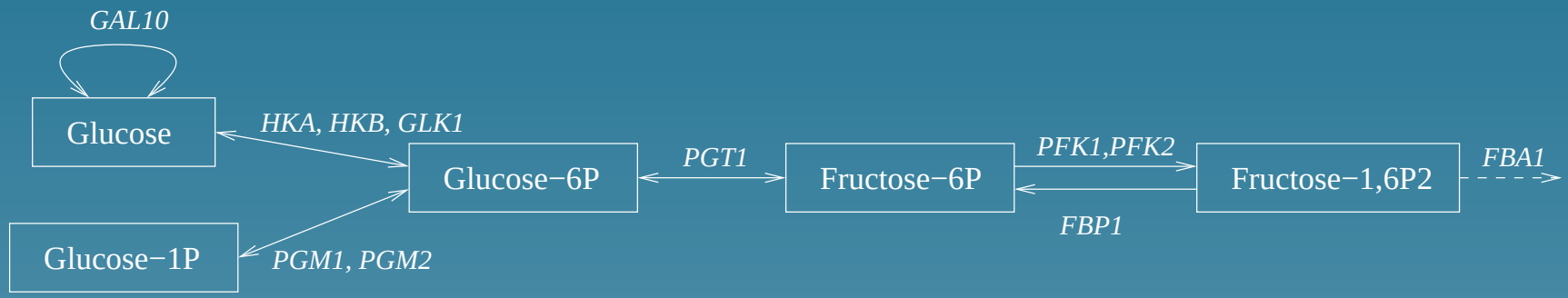
Part 3

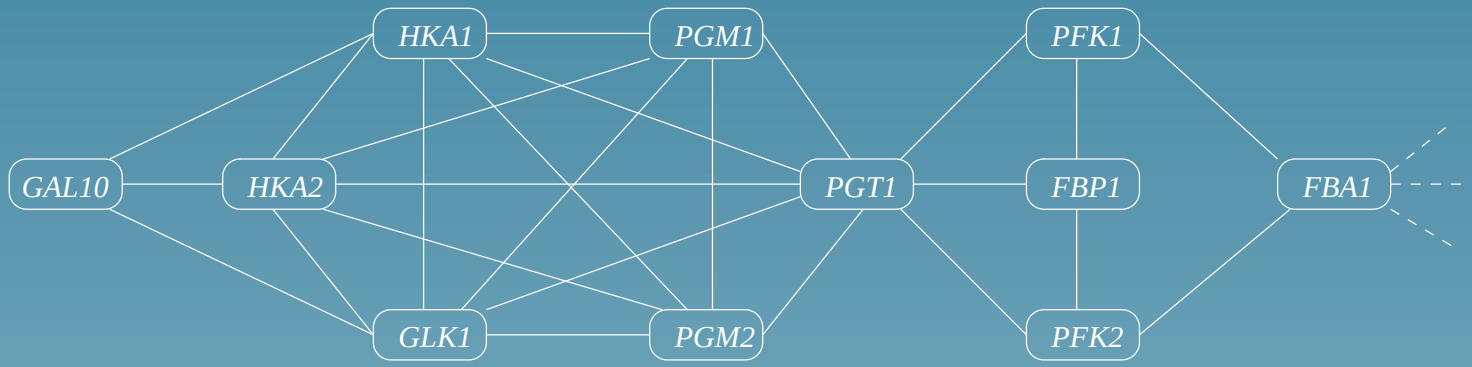
Using the metabolic database

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

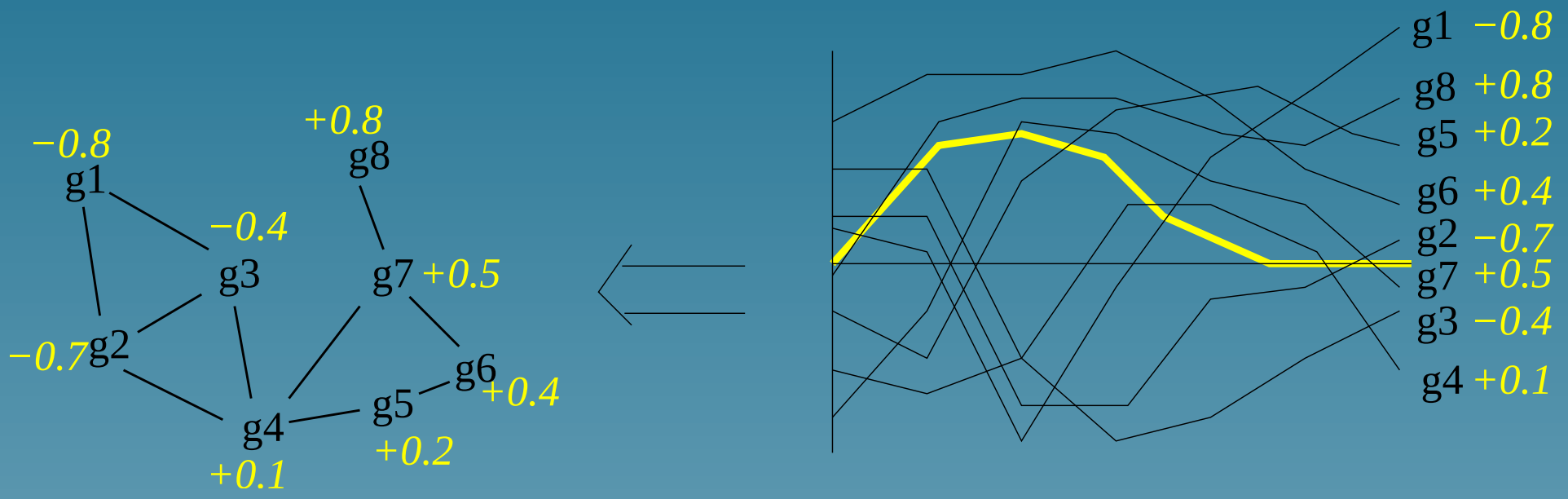
The metabolic gene network





Link two genes when they can catalyze two successive reactions

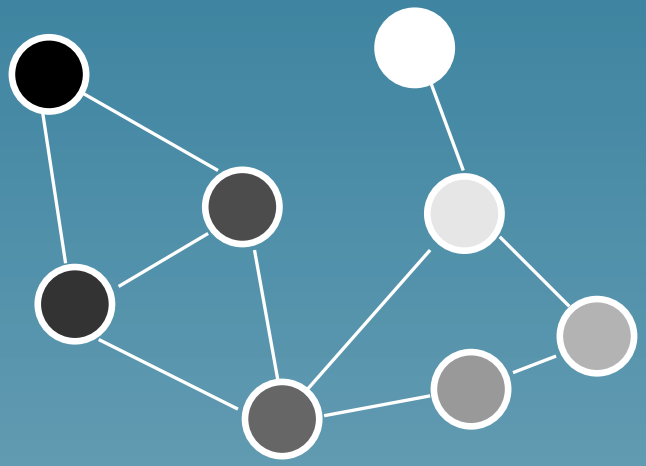
Mapping f_v to the metabolic gene network



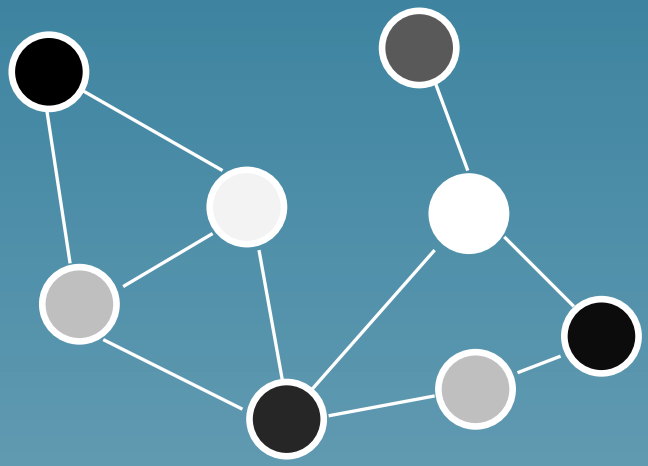
Does it look interesting or not?

Important hypothesis

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

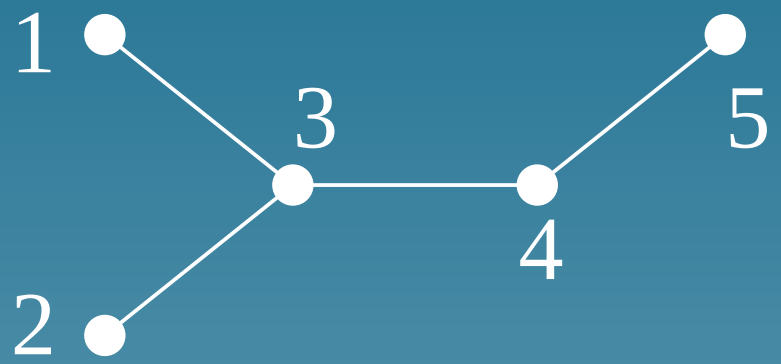


Smooth



Rugged

Graoh Laplacian



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

How smooth is f?

- Local quantification:

$$f^\top L f = \sum_{i \sim j} (f_i - f_j)^2 \left(= \int \frac{\partial f}{\partial x}^2 dx \right)$$

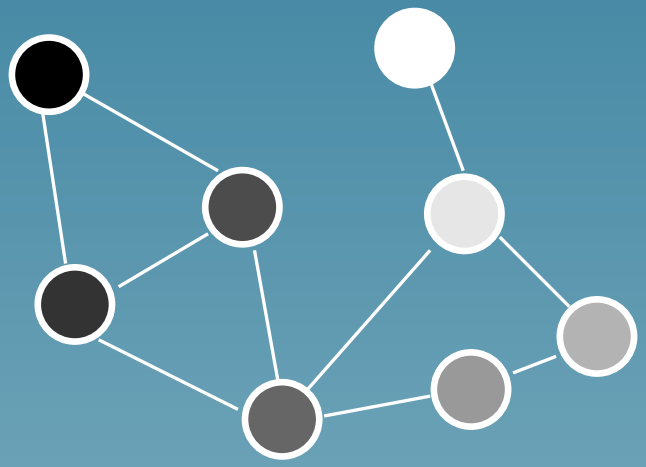
- Spectral quantification:

$$||f||_{H_2}^2 = f^\top \exp(L) f = \sum_{j=1}^{N} \hat{f}_j e^{\lambda_j} \left(= \int \hat{f}(\omega) e^{\omega^2} d\omega \right)$$

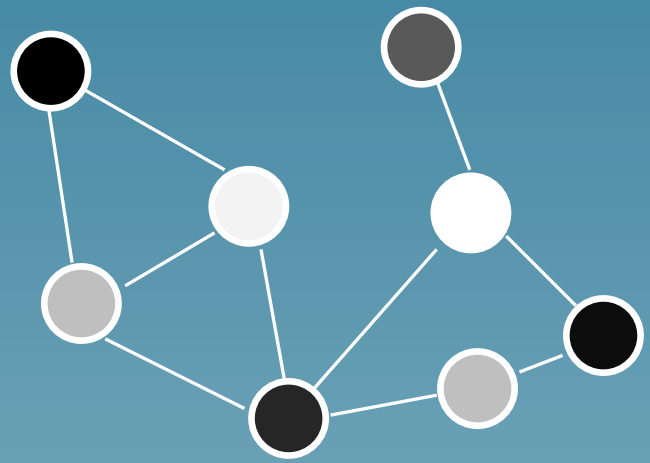
Smoothness quantification

$$h_2(f) = \frac{||f||^2_{L_2}}{||f||^2_{H_2}}$$

is large when f is smooth



h(f) = 2.5



h(f) = 34.2

Part 3

Combining expression and metabolic pathways

Motivation

For a candidate profile v,

- $h_1(f_v)$ is large when v captures a lot of natural variation among profiles
- $h_2(f_v)$ is large when f_v is smooth on the graph

Try to maximize both terms in the same time

Problem reformulation

Find a function f_v and a function f_2 such that:

- $h_1(f_v) = ||f_v||_{L^2}/||f_v||_{H_1}$ be large
- $h_2(f_2) = ||f_2||_{L^2}/||f_2||_{H_2}$ be large
- f_v and f_2 be correlated :

$$\frac{f_v^\top f_2}{||f_v||_{L^2}||f_2||_{L^2}}$$

be large

Problem reformulation (2)

The three goals can be combined in the following problem:

$$\max_{f_v, f_2} \frac{f_v^\top f_2}{\left(||f_v||^2_{L^2} + \delta||f_v||^2_{H_1}\right)^{\frac{1}{2}} \left(||f_2||^2_{L^2} + \delta||f_2||^2_{H_2}\right)^{\frac{1}{2}}}$$

where the parameter δ controls the trade-off between relevance/smoothness on the one hand, correlation on the other hand.

Solving the problem

This formultation is equivalent to a generalized form of CCA (Kernel-CCA, Bach and Jordan, 2002), which is equivalent to the following generalized eigenvector problem

$$\begin{pmatrix} 0 & K_1K_2 \\ K_2K_1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho \begin{pmatrix} K_1^2 + \delta K_1 & 0 \\ 0 & K_2^2 + \delta K_2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

where $[K_1]_{i,j} = e_i^\top e_j$ and $K_2 = \exp(-L)$.
Then, $f_v = K_1\alpha$ and $f_2 = K_2\beta$.

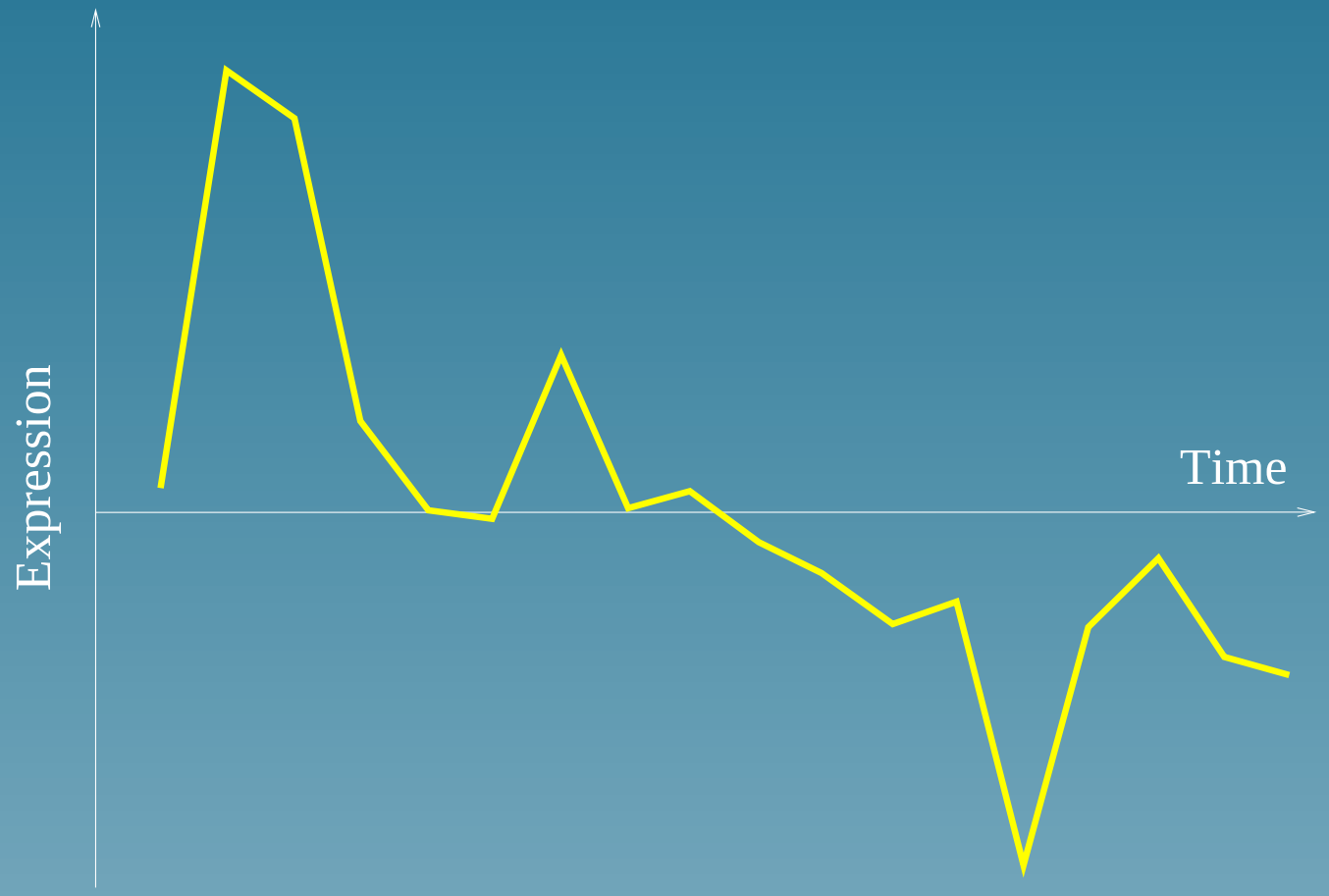
Part 4

Experimental results

Data

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

First pattern of expression

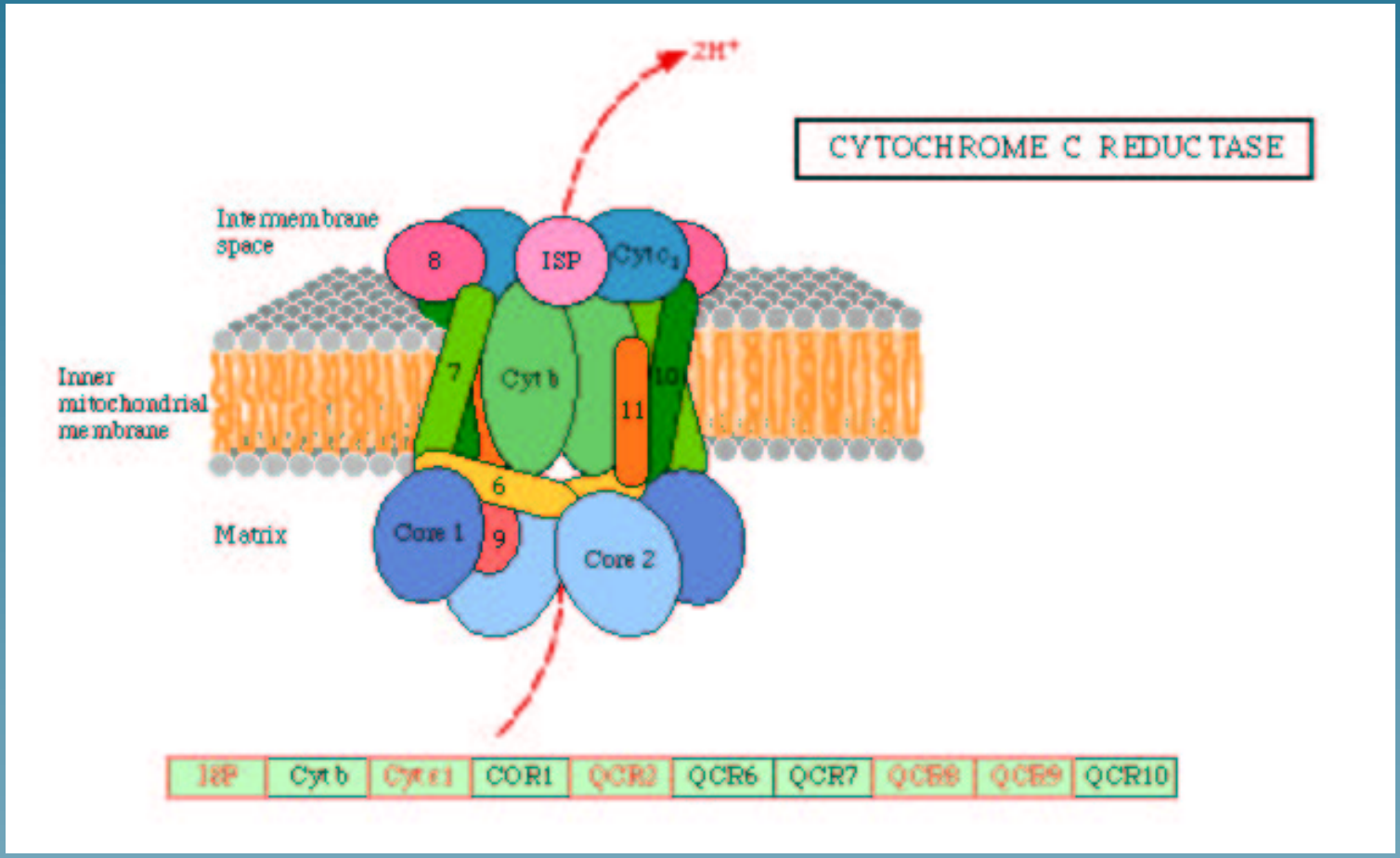


Related metabolic pathways

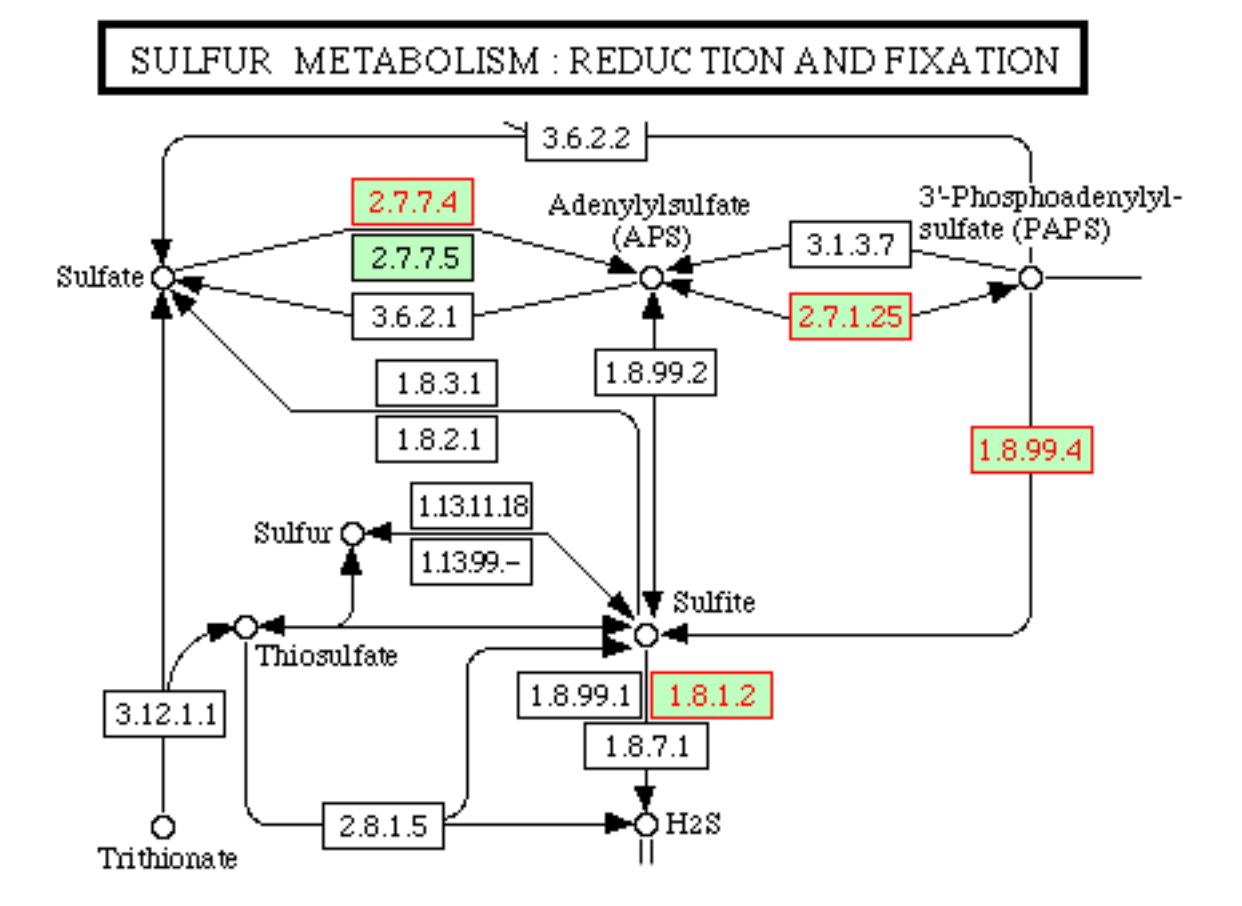
50 genes with highest $s_2 - s_1$ belong to:

- Oxidative phosphorylation (10 genes)
- Citrate cycle (7)
- Purine metabolism (6)
- Glycerolipid metabolism (6)
- Sulfur metabolism (5)
- Selenoaminoacid metabolism (4) , etc...

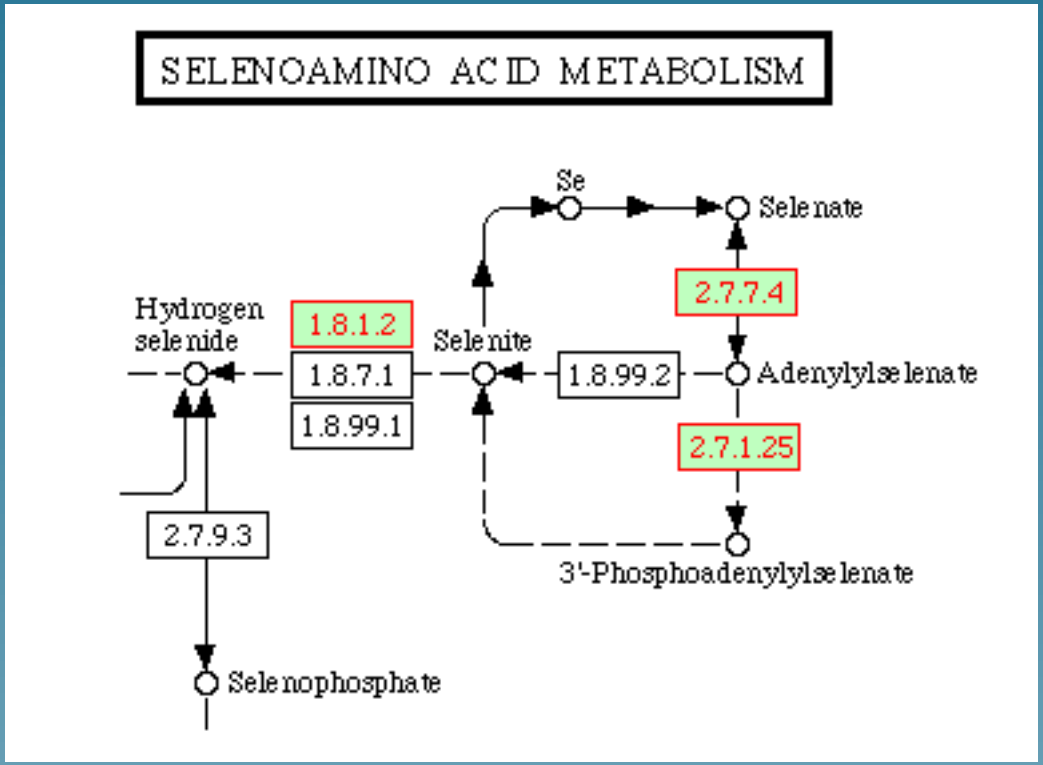
Related genes



Related genes



Related genes



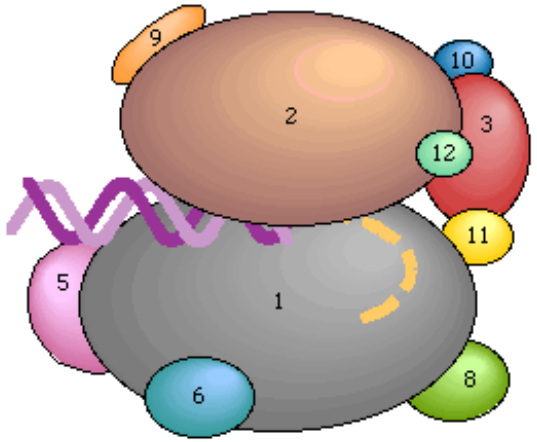
Opposite pattern

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

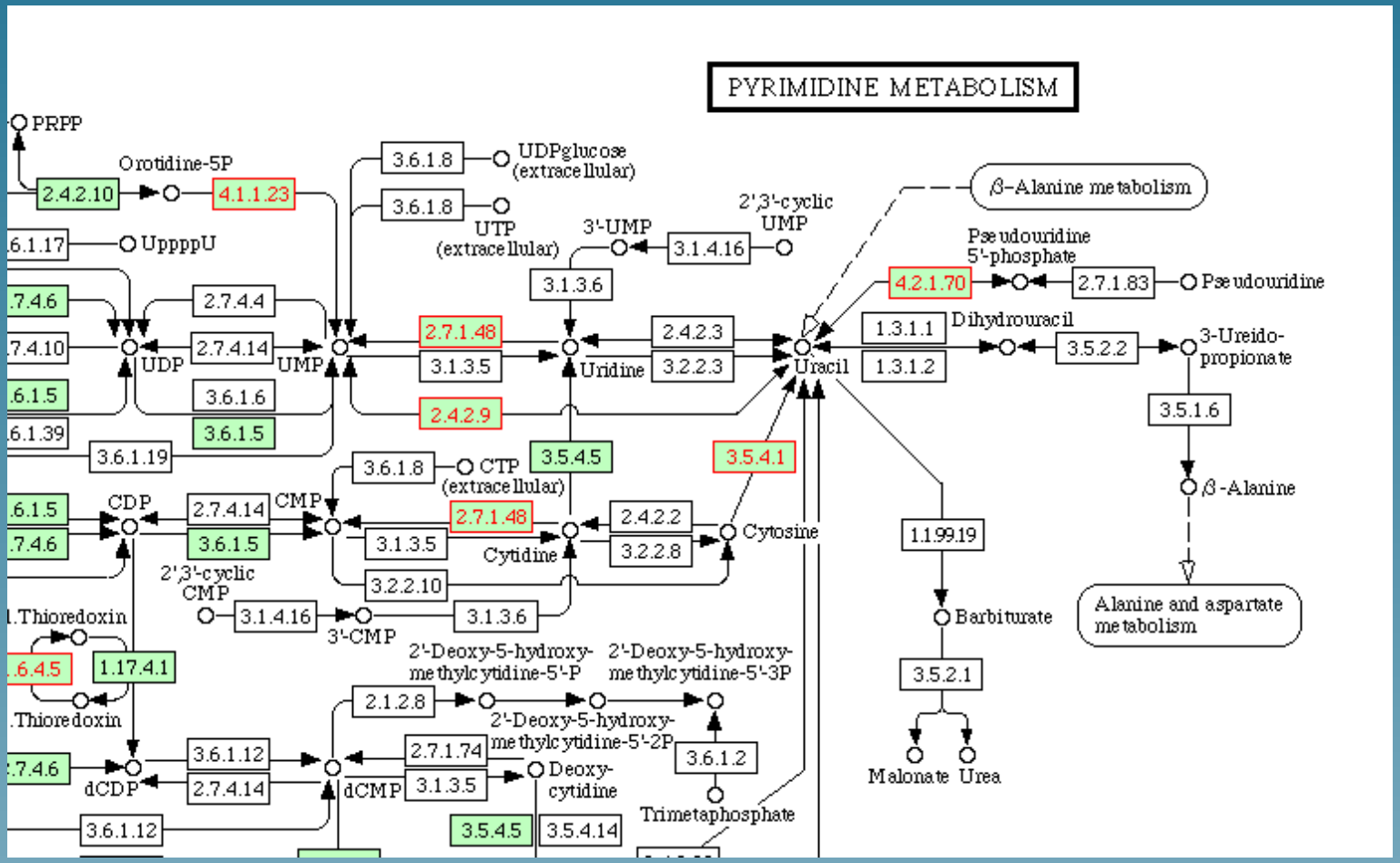
Eukaryotic Pol III

C2	C3	C4	C5	C11
C1	C19	C25	C31	C34

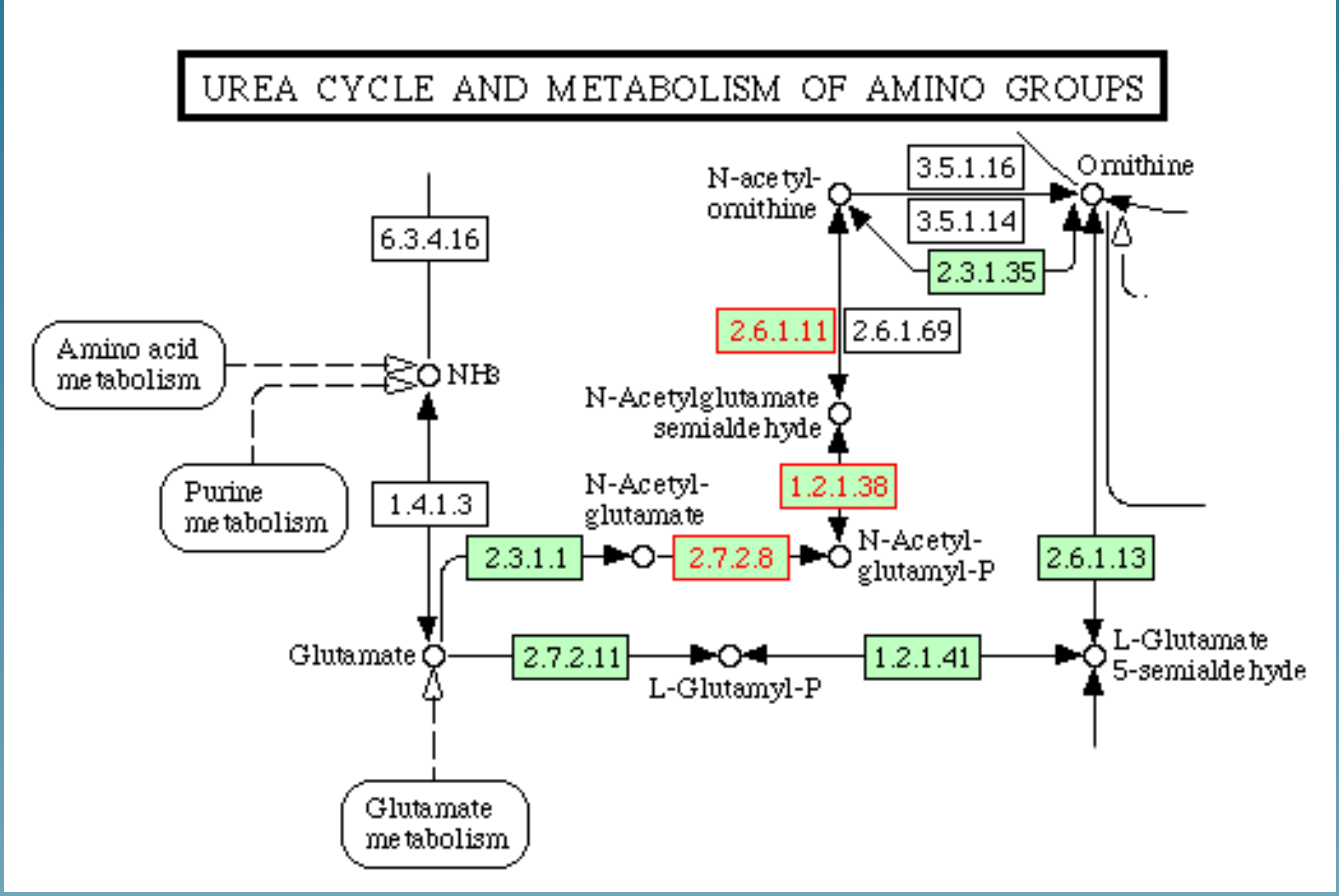
Eukaryotic Pol I

A2	A12	A14	A34	A43	A49
A1					

Related genes



Related genes



Conclusion

Conclusion

- An approach to integrate heterogeneous data (expression profiles and network)
- A particular case of more generic methods (kernel methods)
- Generalization to other types of data and more than two datasets is possible (see ISMB's paper with Yamanishi)