Extracting metabolic pathways from gene expression data using kernel CCA

Jean-Philippe Vert

Bioinformatics group
Ecole des Mines de Paris

Jean-Philippe.Vert@mines.org

Conference Statistical Learning, Theory and Applications, November 14-15, 2002, CNAM Paris.

Overview

1. Motivations
2. Problem Formulation
3. An approach using RKHS
4. Experimental results

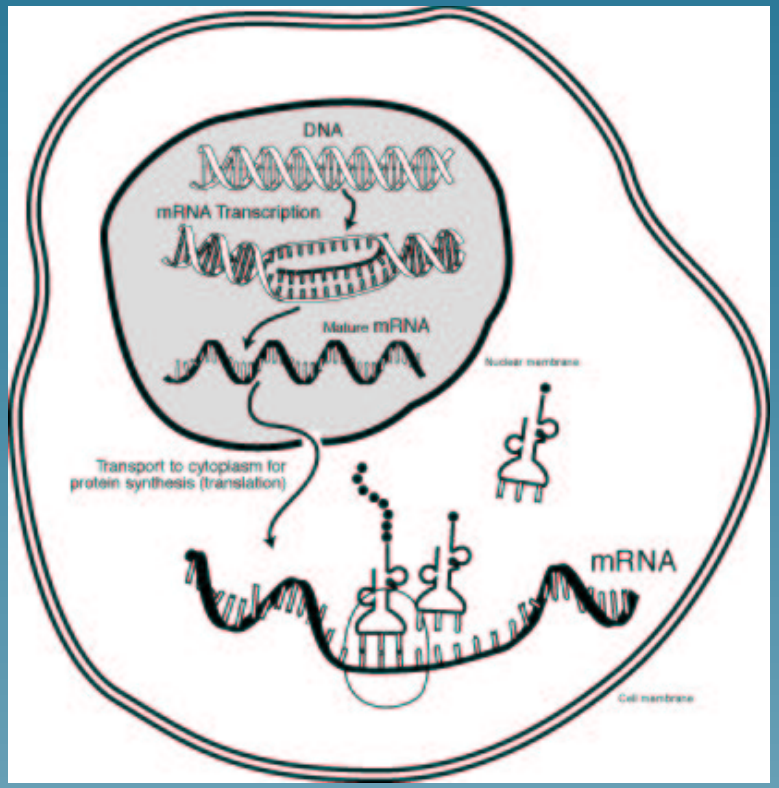
Part 1

Motivations

Context

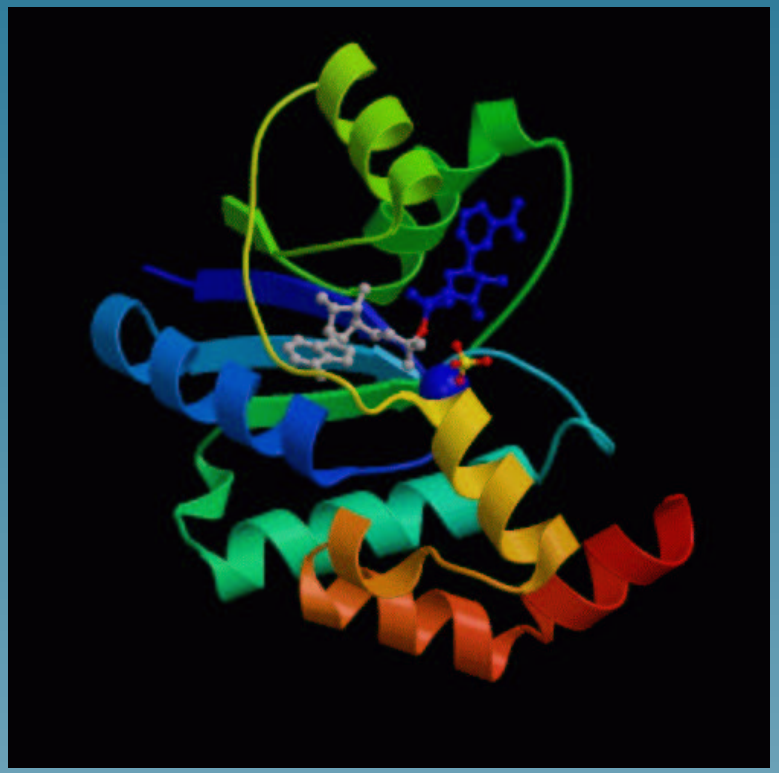
- Data available in bioinformatics: sequences, molecules, graphs, measurements...
  - ⋆ heterogeneous
  - ⋆ large quantity
  - ⋆ noisy.
- Complex biological process still poorly understood

From DNA to proteins



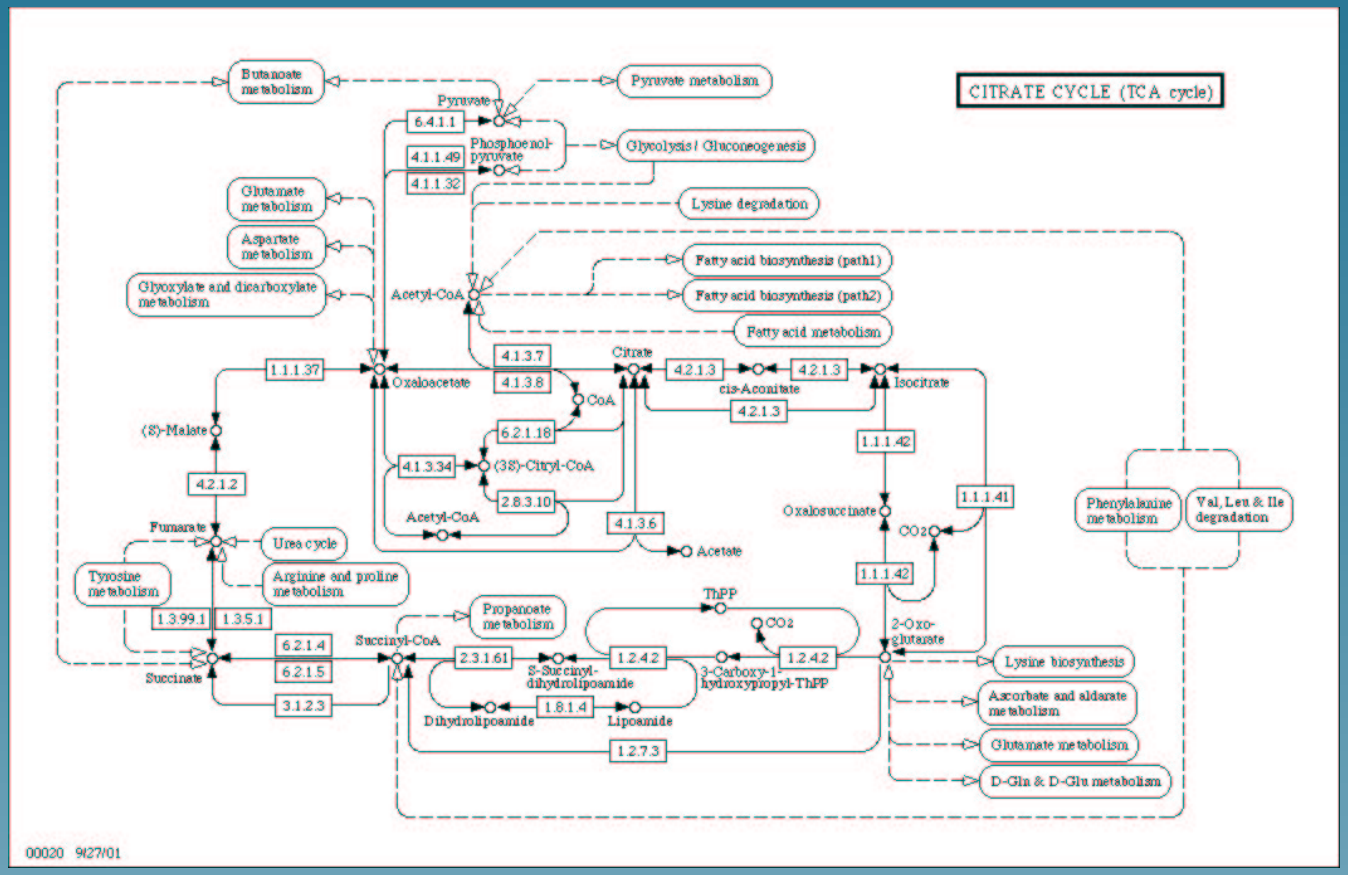
Central dogma: DNA $\longrightarrow$ RNA $\longrightarrow$ Protein

Genes encode proteins which can catalyse chemical reations



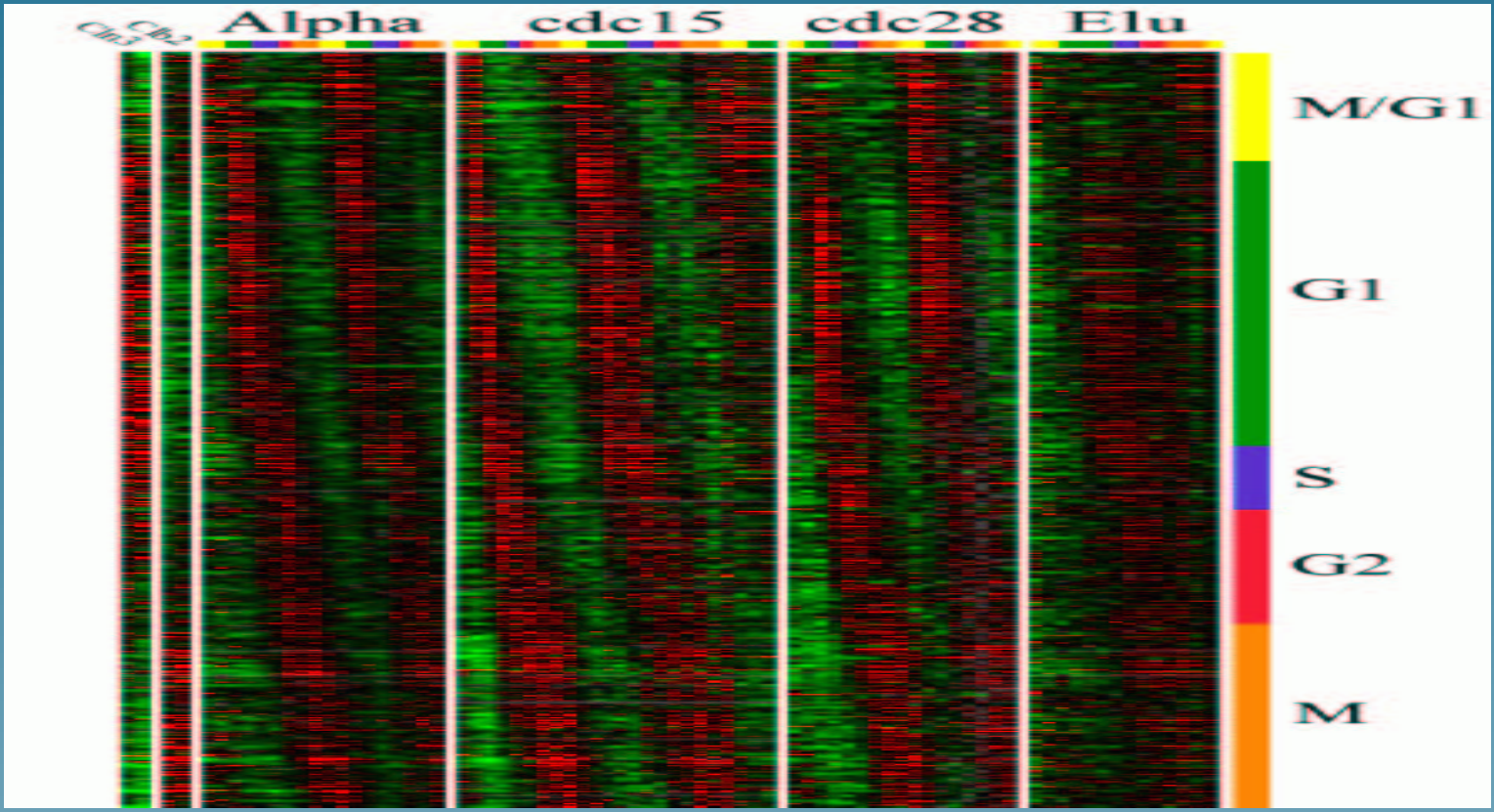
Nicotinamide Mononucleotide Adenylyltransferase With Bound Nad+

Chemical reactions are often parts of pathways



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

Microarray technology monitors RNA quantity

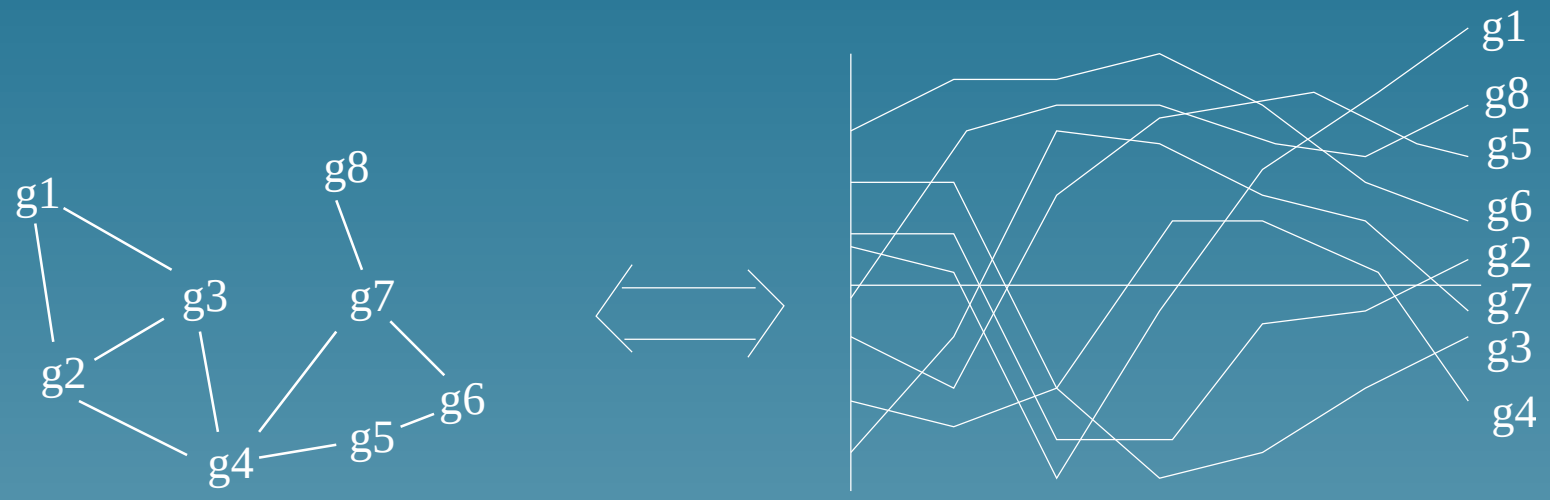


(From Spellman et al., 1998)

Part 2

Problem formulation

Comparing gene expression and protein network



Gene network

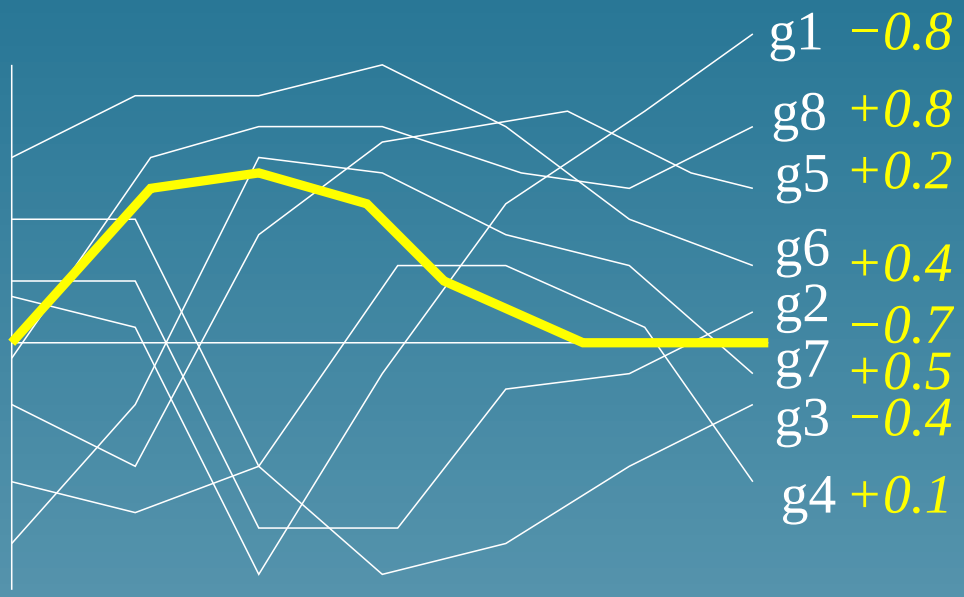
Expression profiles

Are there "correlations"?

What is a correlation?

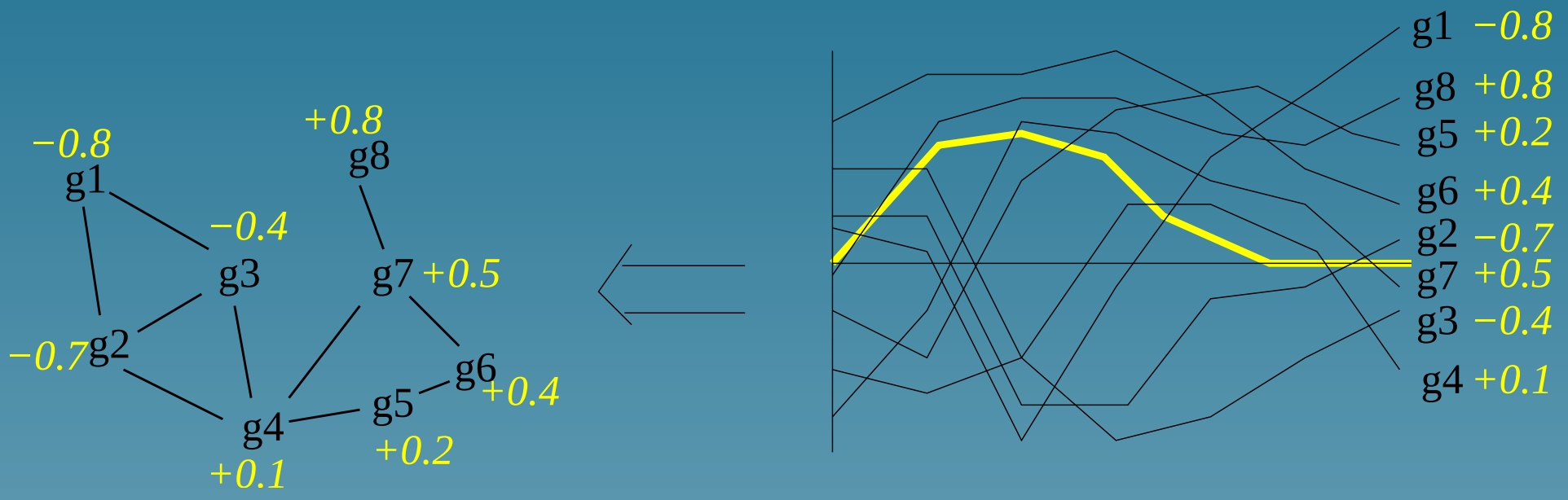
- A pattern of expression shared by genes close to each other on the graph
  - ⋆ activity level of a metabolic pathway
  - ⋆ environmental change

Pattern of expression



- A pattern is by definition a profile.
- The correlation between a candidate pattern and a gene quantifies how much the gene shares the pattern

Pattern smoothness



- The correlation function with interesting patterns should vary smoothly on the graph

Pattern relevance

- Interesting patterns involve many genes
- The projection of profiles onto an interesting pattern should capture a lot of variations among profiles

Problem

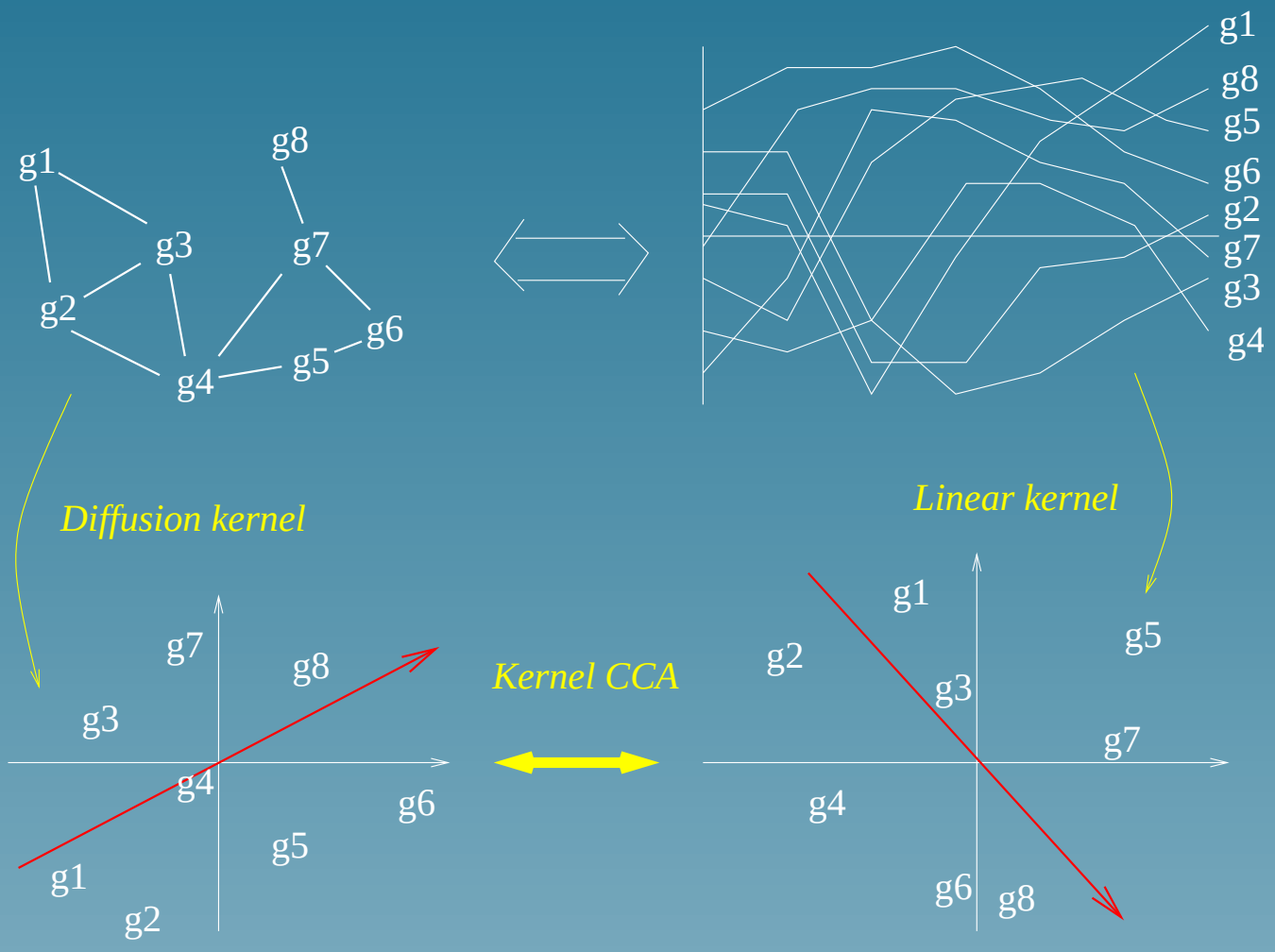
Find patterns of expression which are simultaneously

- smooth
- relevant

Part 3

An approach using RKHS

The idea


g1
g8
g3
g7
g2
g6
g4
g5
g1
g8
g5
g6
g2
g7
g3
g4
Diffusion kernel
Linear kernel
Kernel CCA
g7
g8
g3
g4
g6
g5
g1
g2
g1
g5
g2
g3
g7
g4
g6
g8

Pattern relevance

- Let $e(x)$ the profile of gene x, and $v = \sum_x \alpha_x e(x)$ a candidate pattern.
- Let $K_1(x, y) = e(x).e(y)$ be the linear kernel matrix on the space of genes
- The relevance of a pattern is quantified as:

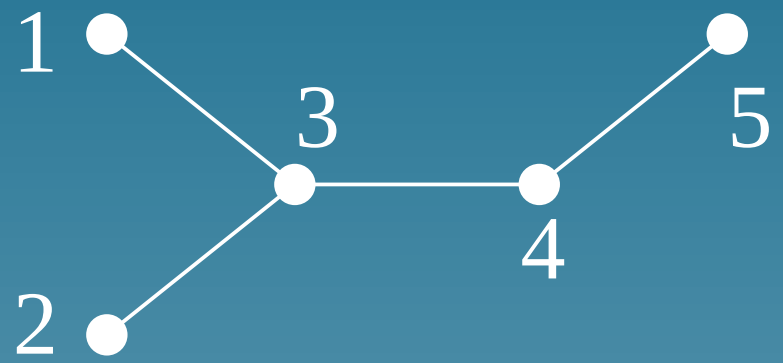
$$R(v) \triangleq \frac{\sum_x (v.e(x))^2}{||v||^2} = \frac{\alpha' K_1^2 \alpha}{\alpha' K_1 \alpha} = \frac{||K_1\alpha||_{L^2}}{||K_1\alpha||_{H_1}}$$

Pattern smoothness

- Let $K_2(x, y)$ be the diffusion kernel obtained from the gene network.
- It can be considered as a discretized version of a Gaussian kernel (solving the heat equation with the graph Laplacian)
- The norm in the RKHS defined by K_2 is a smoothness functional: the smoother a function $K_2\beta$, the larger the function:

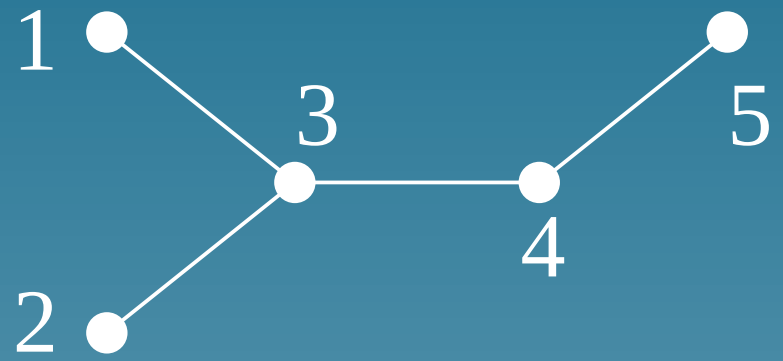
$$S(K_2\beta) = \frac{||K_2\beta||_{L^2}}{||K_2\beta||_{H_2}} = \frac{\beta' K_2^2 \beta}{\beta' K_2 \beta}$$

Diffusion kernel (Kondor and Lafferty, 2002)



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

Diffusion kernel (Kondor and Lafferty, 2002)



$$K = \exp(-L) = \begin{pmatrix} 0.49 & 0.12 & 0.23 & 0.10 & 0.03 \\ 0.12 & 0.49 & 0.23 & 0.10 & 0.03 \\ 0.23 & 0.23 & 0.24 & 0.17 & 0.10 \\ 0.10 & 0.10 & 0.17 & 0.31 & 0.30 \\ 0.03 & 0.03 & 0.10 & 0.30 & 0.52 \end{pmatrix}$$

Problem reformulation

Find a linear function $K_1\alpha$ and a function $K_2\beta$ such that:

- $K_1\alpha$ be relevant : $||K_1\alpha||_{L^2}/||K_1\alpha||_{H_1}$ be large
- $K_2\beta$ be smooth : $||K_2\beta||_{L^2}/||K_2\beta||_{H_2}$ be large
- $K_1\alpha$ and $K_2\beta$ be correlated :

$$\frac{\alpha' K_1 K_2 \beta}{||K_1\alpha||_{L^2}||K_2\beta||_{L^2}}$$

be large

Problem reformulation (2)

The three goals can be combined in the following problem:

$$\max_{\alpha,\beta} \frac{\alpha' K_1 K_2 \beta}{\left(||K_1\alpha||^2_{L^2} + \delta||K_1\alpha||^2_{H_1}\right)^{\frac{1}{2}} \left(||K_2\beta||^2_{L^2} + \delta||K_2\beta||^2_{H_2}\right)^{\frac{1}{2}}}$$

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

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

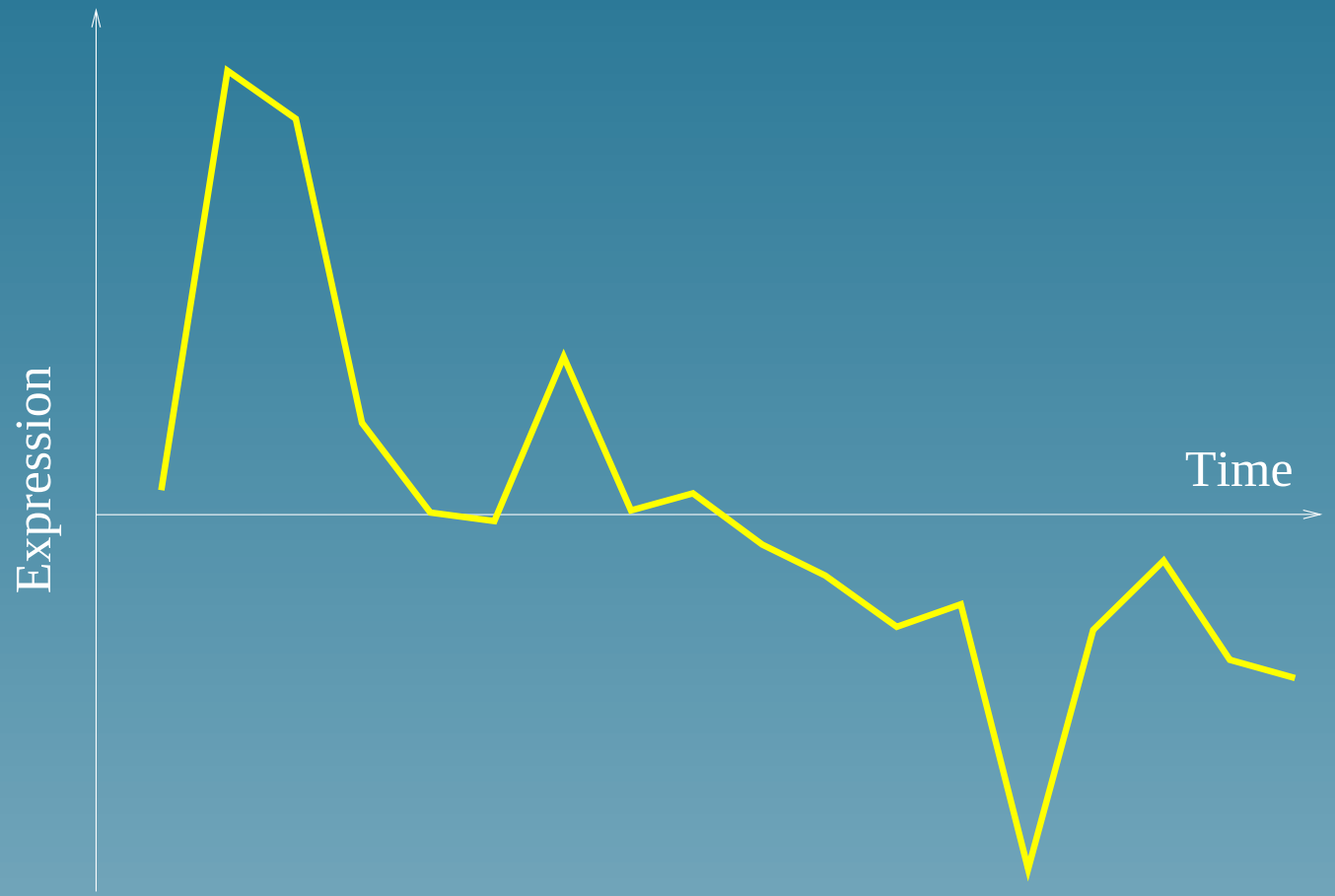
Part 4

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

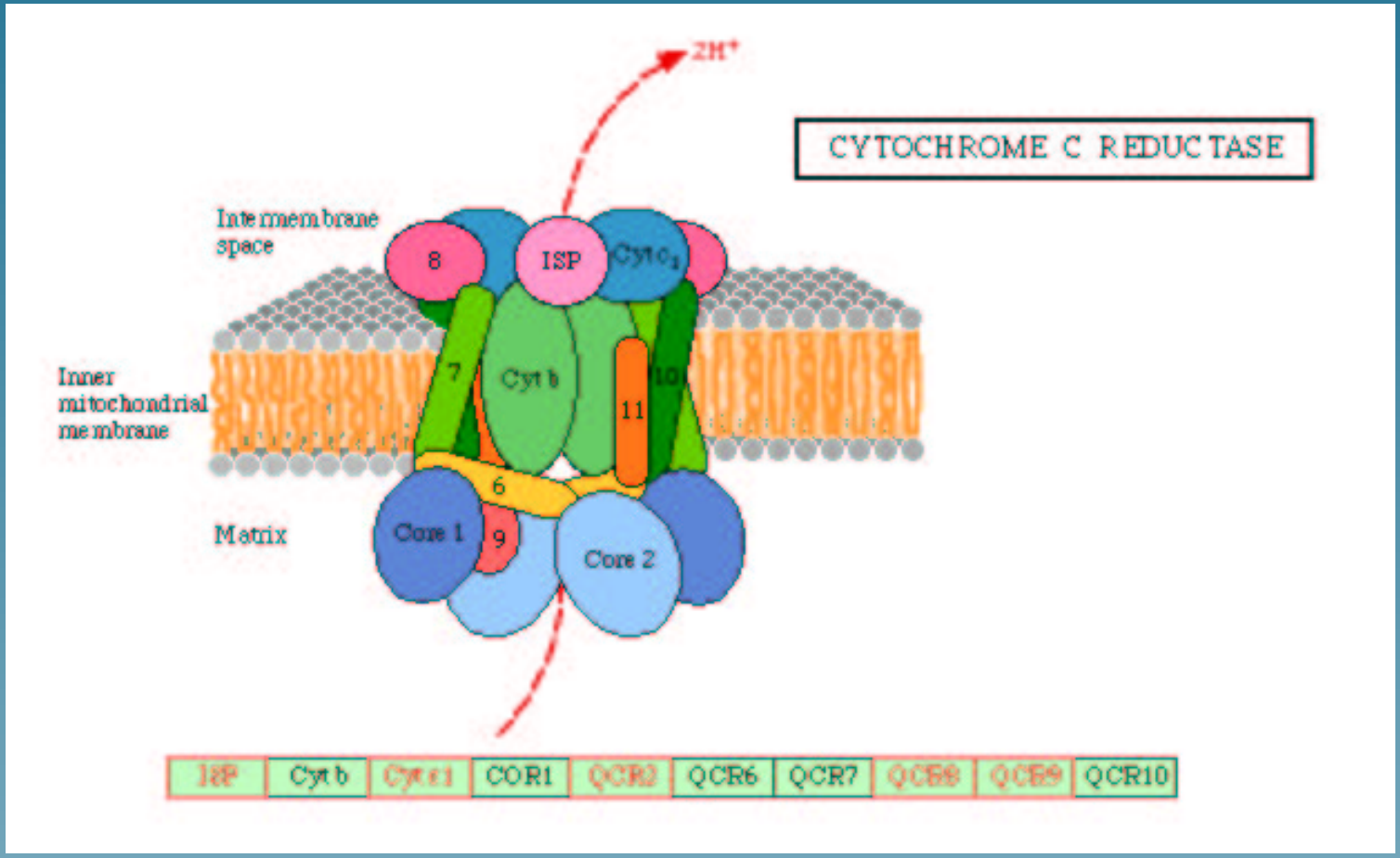


Related metabolic pathways

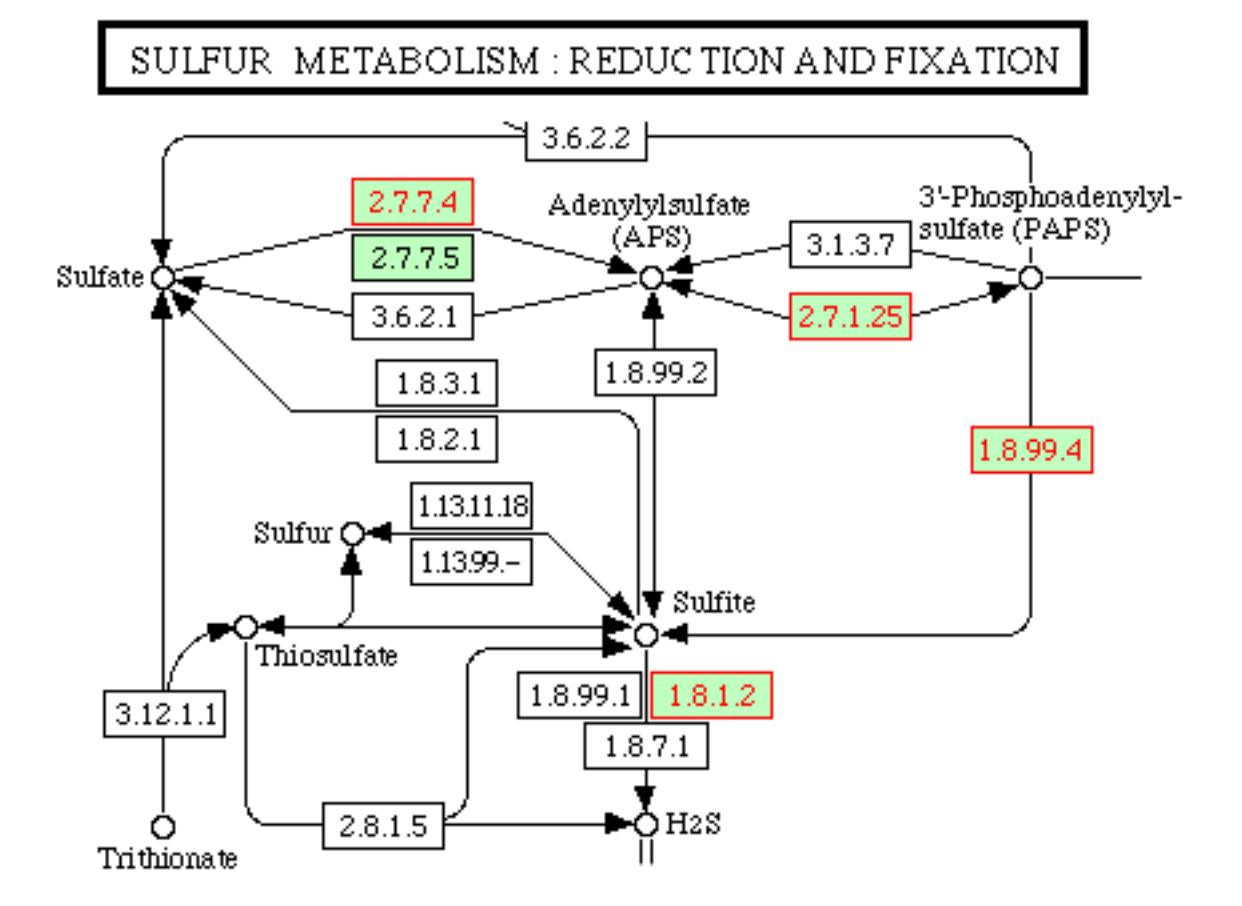
50 genes with highest $s_2 - s_1$ belong to:

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

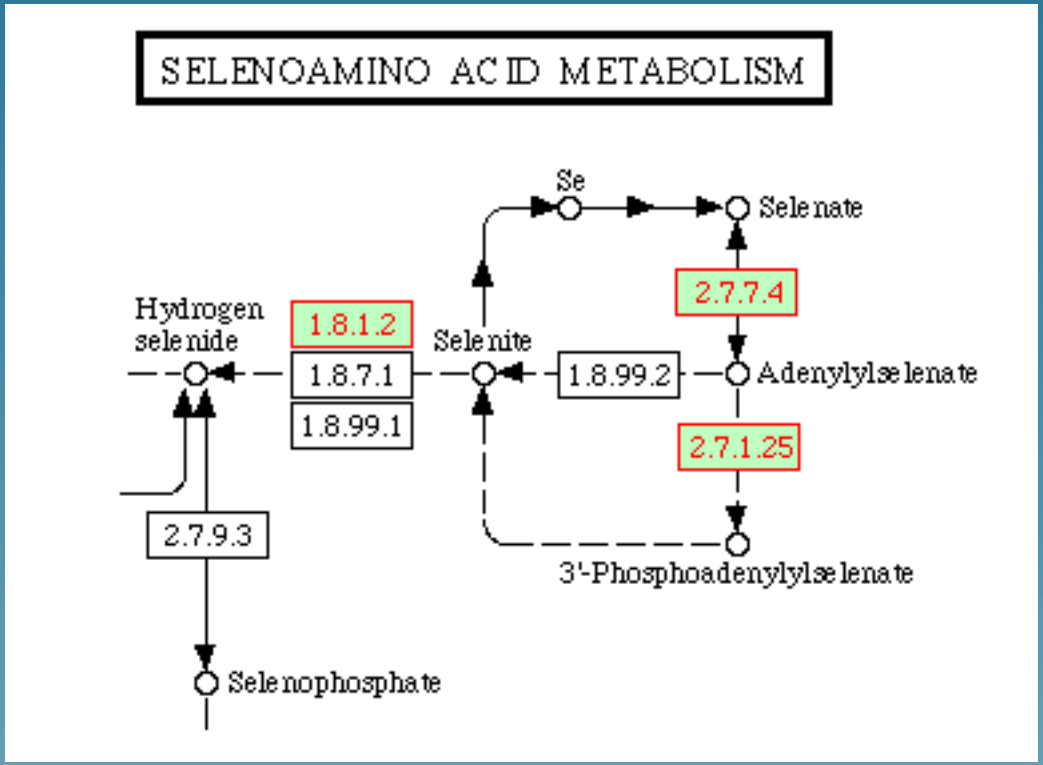
Related genes



Related genes



Related genes



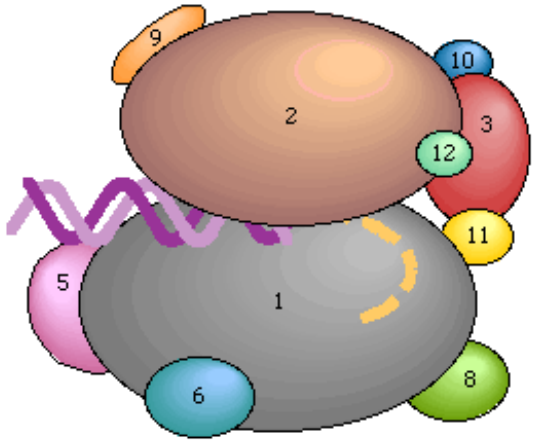
Opposite pattern

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

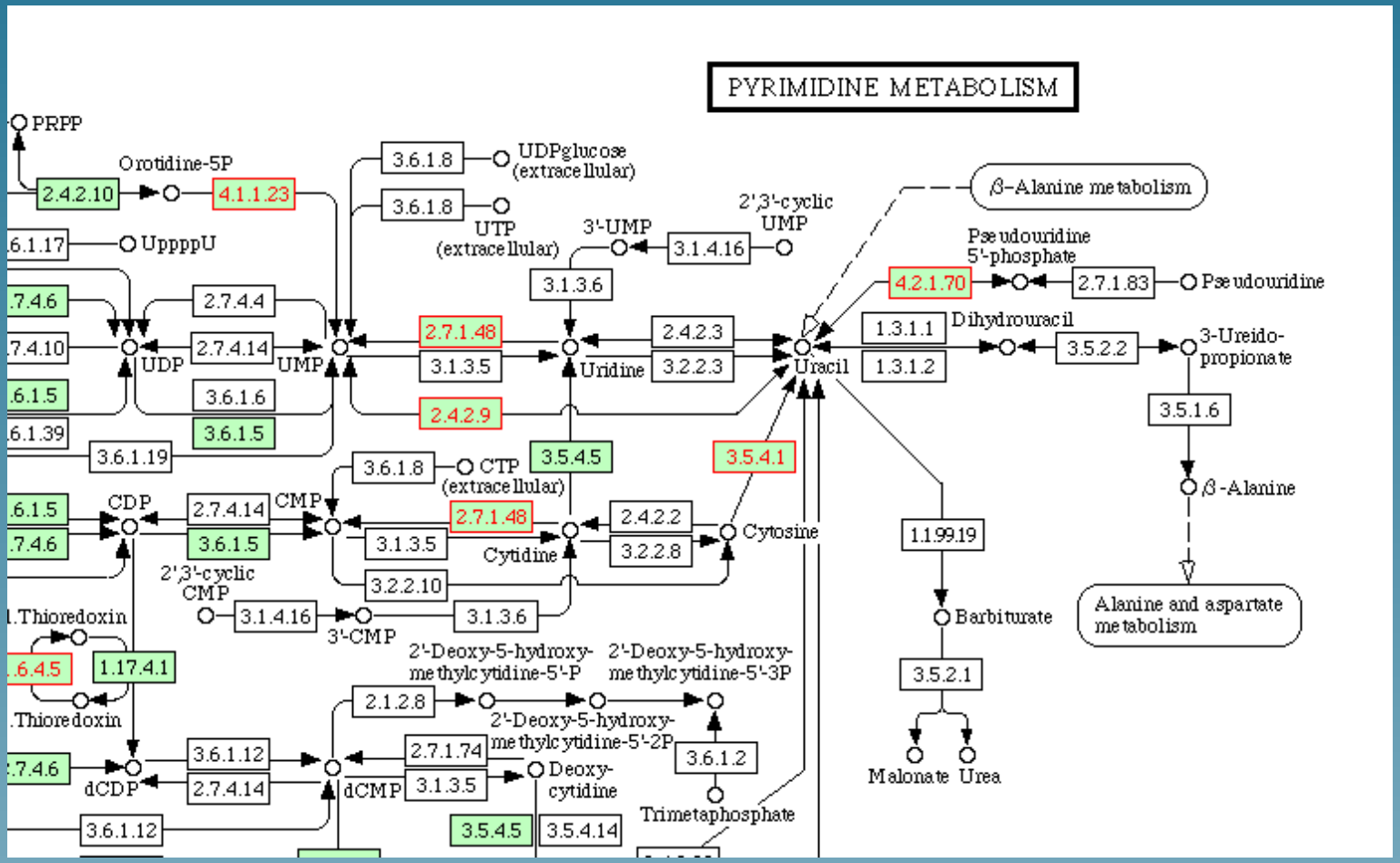
Eukaryotic Pol III

C2	C3	C4	C5	C11
C1	C19	C25	C31	C34

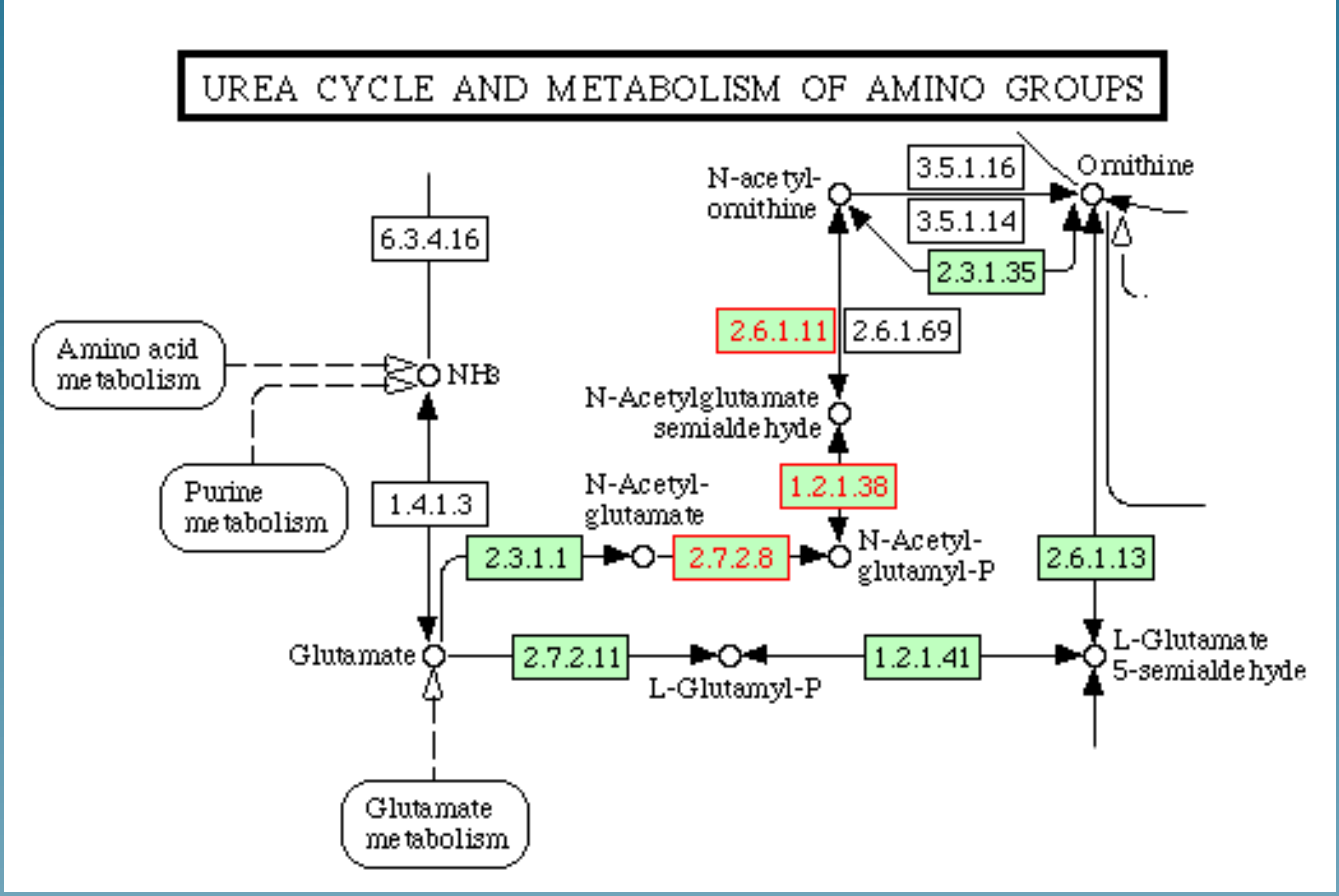
Eukaryotic Pol I

A2	A12	A14	A34	A43	A49
A1					

Related genes



Related genes



Conclusion

Conclusion

- Heterogeneous data can be integrated with kernels
- The approach can be generalized (non-linear kernel for gene expression, string kernels...)
- Working in RKHS can help solve real-world problems

Workshop

Kernel Methods in Bioinformatics

Harnack-Haus, Berlin, April 14, 2003

http://www.cg.ensmp.fr/vert/kmb03