Kernel methods in Computational Biology: Two examples

Jean-Philippe.Vert@mines.org

Ecole des Mines de Paris
Computational Biology group

INSA Toulouse, May 20, 2003.

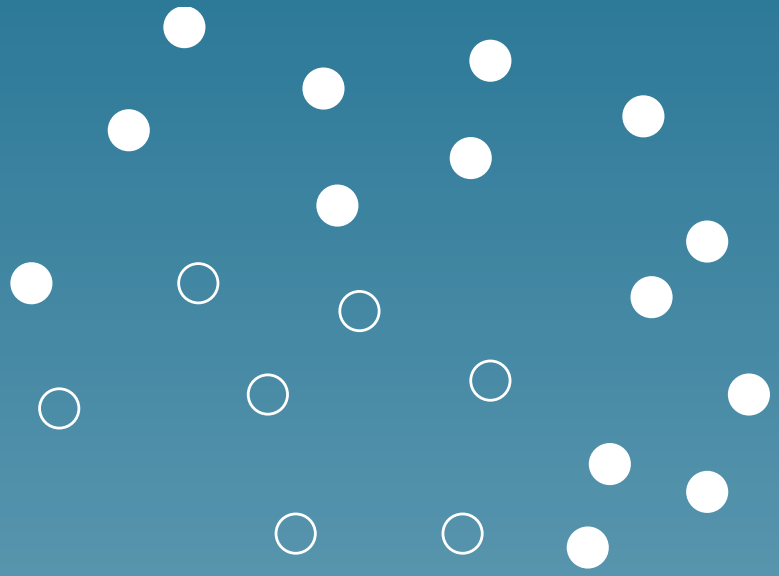
Overview

1. Pattern recognition and Support Vector Machines
2. Remote protein homology detection
3. Analysis of microarray data with pathways information

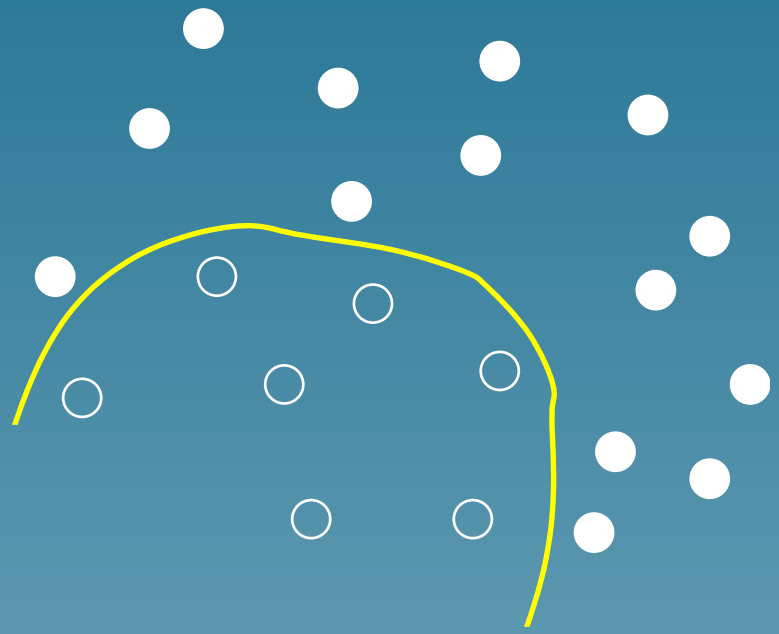
Partie 1

Pattern recognition and Support Vector Machines

The pattern recognition problem

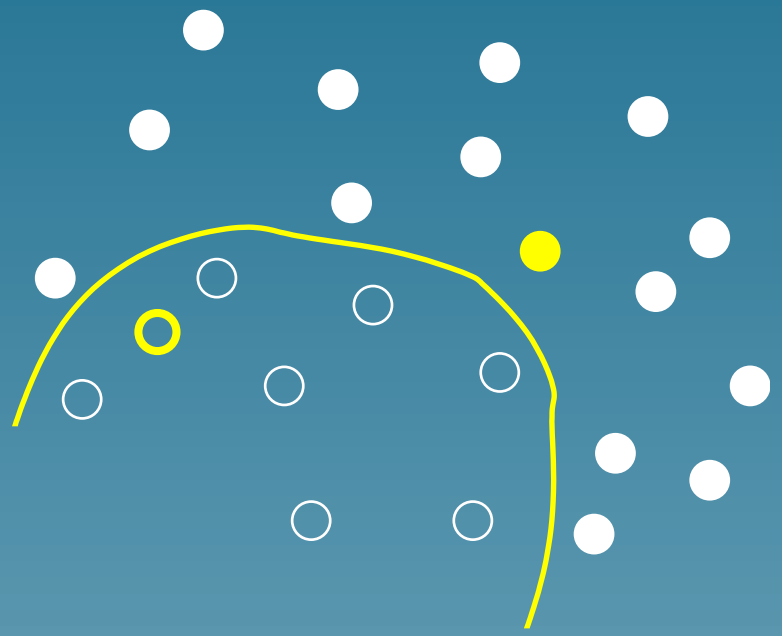


The pattern recognition problem



- Learn from labelled examples a discrimination rule

The pattern recognition problem

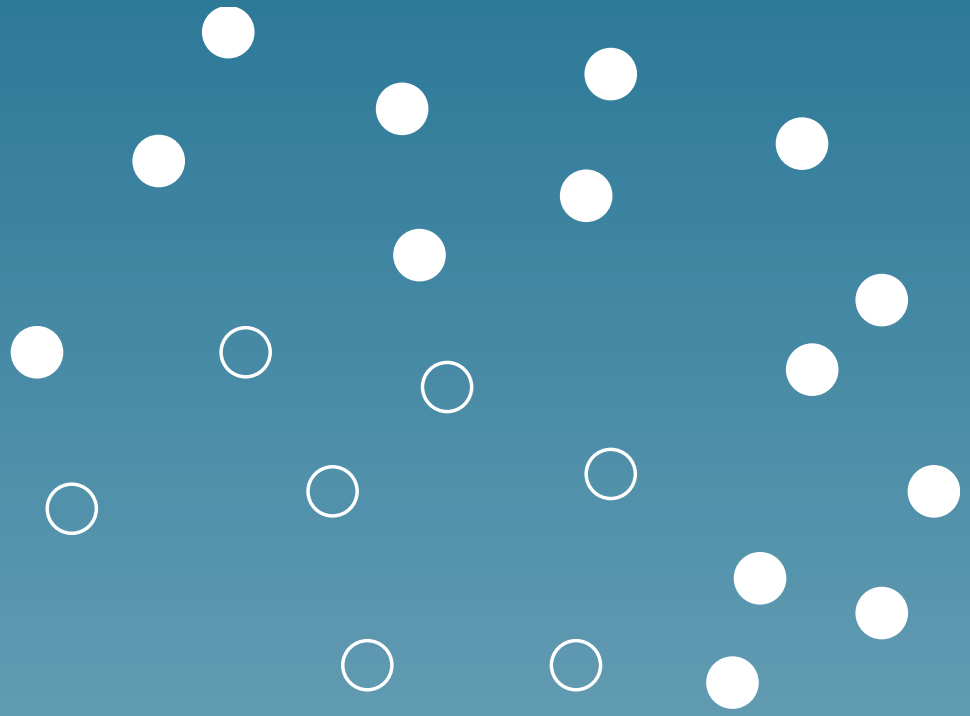


- Learn from labelled examples a discrimination rule
- Use it to predict the class of new points

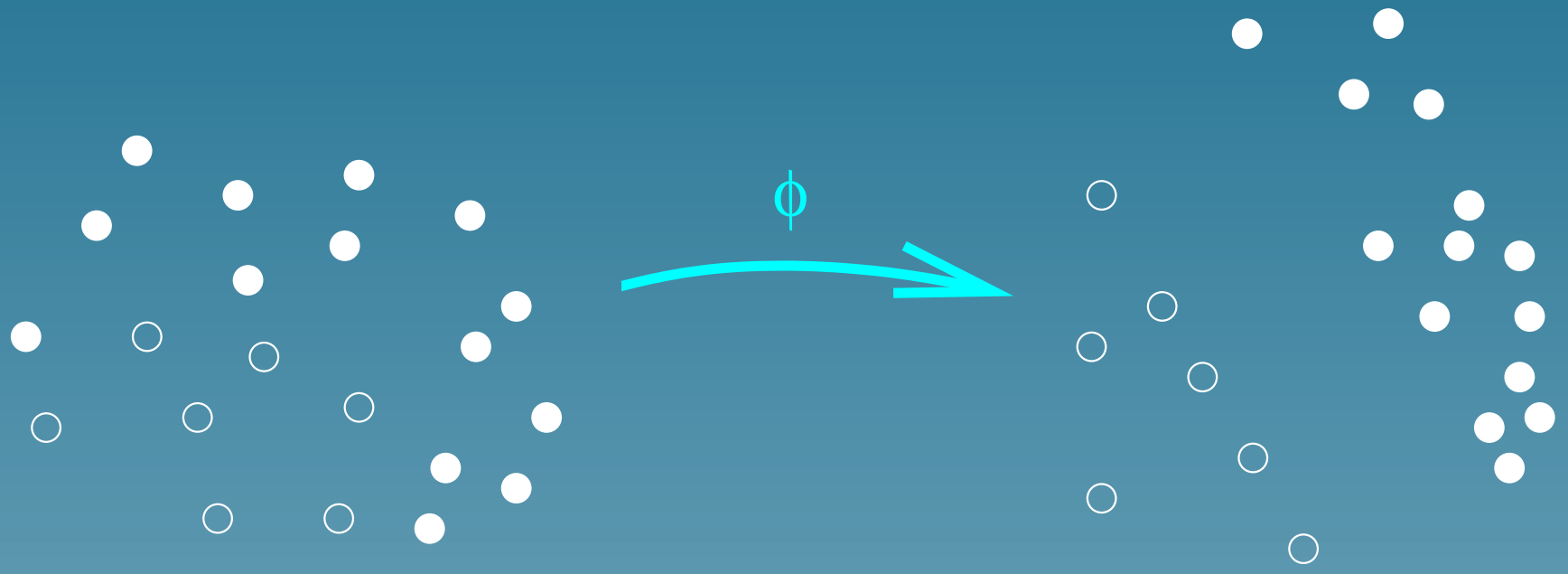
Pattern recognition examples

- Medical diagnosis (e.g., from microarrays)
- Drugability/activity of chemical compouds
- Gene function, structure, localization
- Protein interactions

Support Vector Machines for pattern recognition

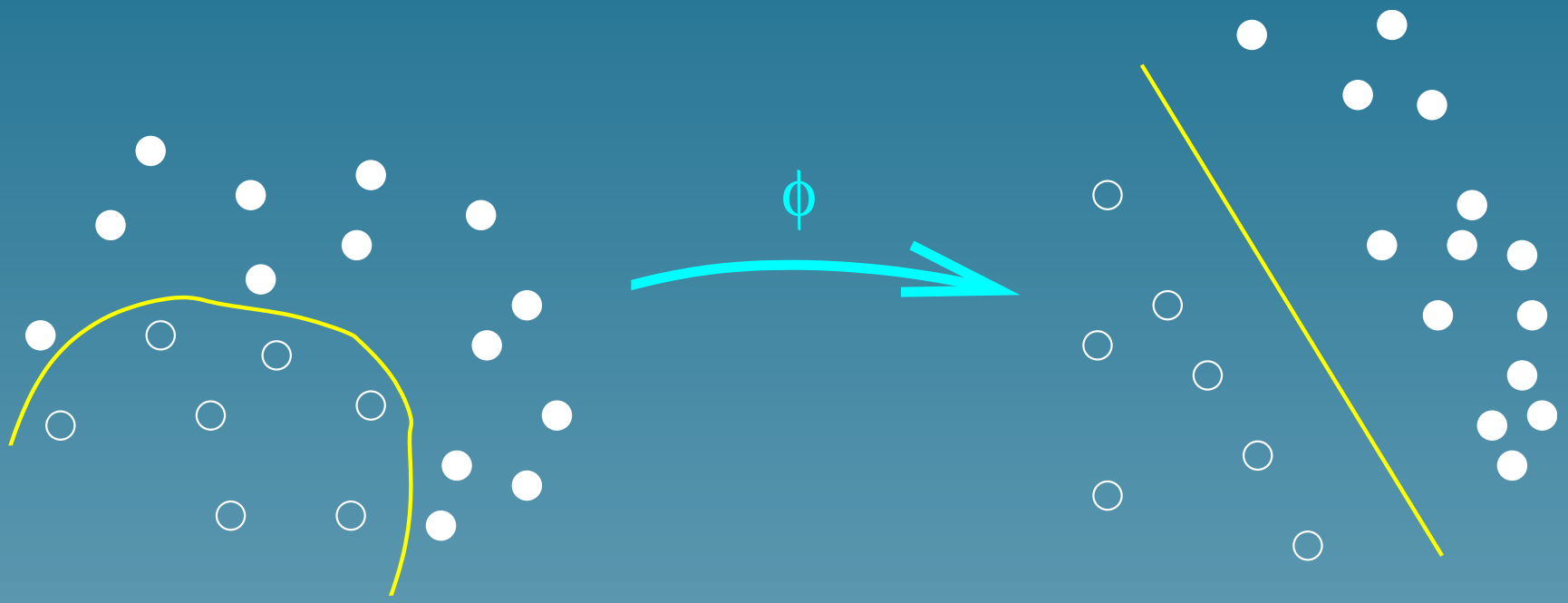


Support Vector Machines for pattern recognition



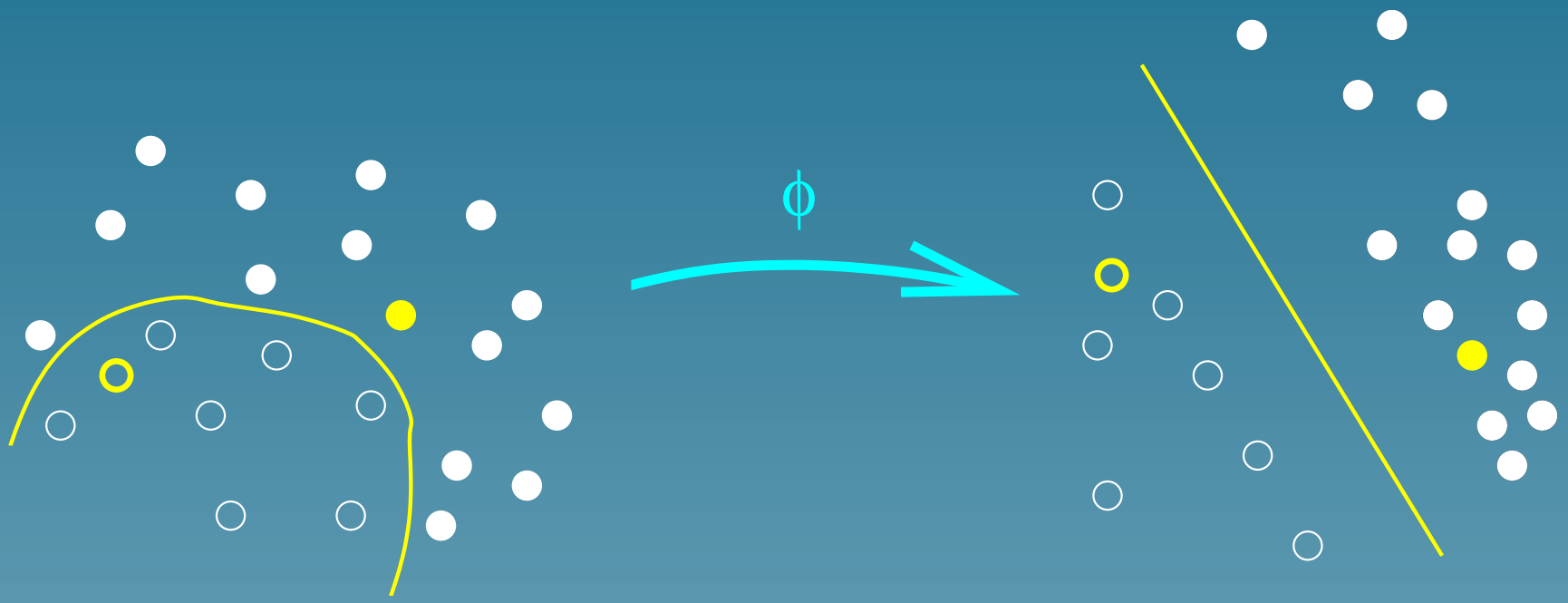
- Object x represented by the vector $\Phi(\vec{x})$ (feature space)

Support Vector Machines for pattern recognition



- Object x represented by the vector $\Phi(\vec{x})$ (feature space)
- Linear separation with large margin in the feature space

Support Vector Machines for pattern recognition



- Object x represented by the vector $\Phi(\vec{x})$ (feature space)
- Linear separation with large margin in the feature space

The kernel trick for SVM

- The separation can be found without knowing $\Phi(x)$. Only the following kernel matters:

$$K(x, y) = \Phi(\vec{x}).\Phi(\vec{y})$$

- Simple kernels $K(x, y)$ can correspond to complex $\vec{\Phi}$
- SVM work with any sort of data as soon as a kernel is defined

Kernels

- A kernel can be thought of as a measure of similarity.
- There are mathematical conditions to ensure that a function $K(x, y)$ is a valid kernel (it must be symmetric positive semidefinite).
- As soon as $K(.,.)$ is a valid kernel, SVM can be used for pattern recognition

Advantages of SVM

- Works well on real-world applications
- Large dimensions, noise OK
- Can be applied to any kind of data as soon as a kernel is available

Advantages of kernels

- Kernels can be engineered for a particular type of data (vectors, strings, graphs, structure, ...)
- Prior knownledge of the problem can be included in the kernel
- Not restricted to SVM: the kernel trick works with many other kernel methods

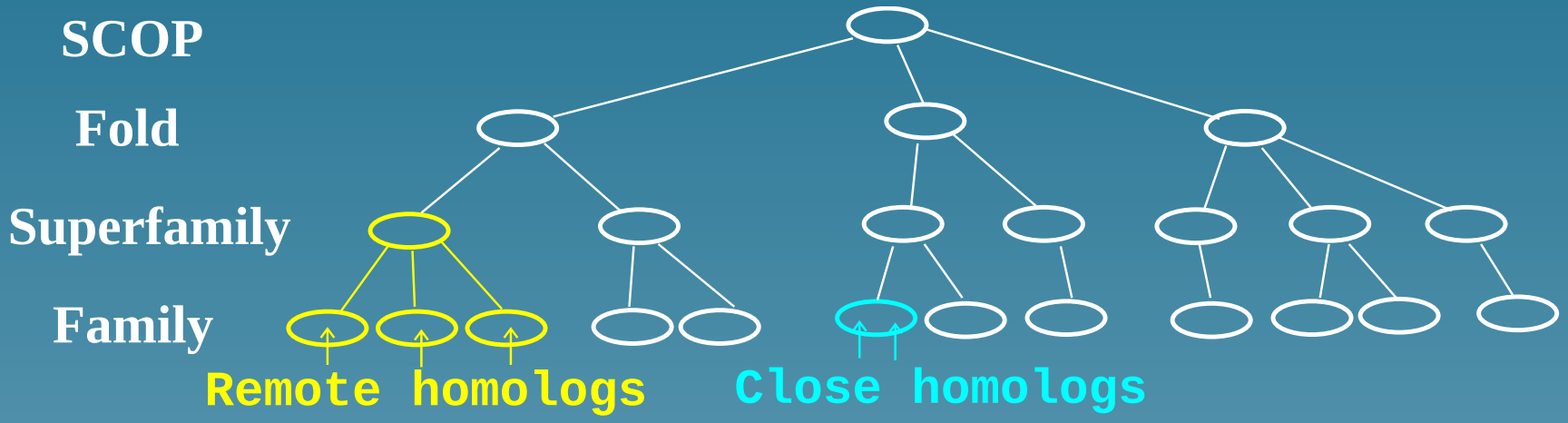
Partie 2

Application: remote protein homology detection

The problem

- Same structure/function but sequence diverged
- Remote homology can not be found by direct sequence similarity

SCOP database



A benchmark experiment

- Can we predict the superfamily of a domain if we have not seen any member of its family before?

A benchmark experiment

- Can we predict the superfamily of a domain if we have not seen any member of its family before?
- During learning: remove a family and learn the difference between the superfamily from the rest

A benchmark experiment

- Can we predict the superfamily of a domain if we have not seen any member of its family before?
- During learning: remove a family and learn the difference between the superfamily from the rest
- Then, use the model to test each domain of the family removed

Using SVM

- We need a kernel $K(s_1, s_2)$ between any two sequences

Using SVM

- We need a kernel $K(s_1, s_2)$ between any two sequences
- Idea: can we use classical sequence similarity scores as kernels?

Local alignment kernel

- For two strings x and y, a local alignment π with gaps is:

Local alignment kernel

- For two strings x and y, a local alignment π with gaps is:

- The score is:

$$s(x, y, \pi) = s(E, E) + s(F, E) + s(G, G) + s(I, I) - s(gaps)$$

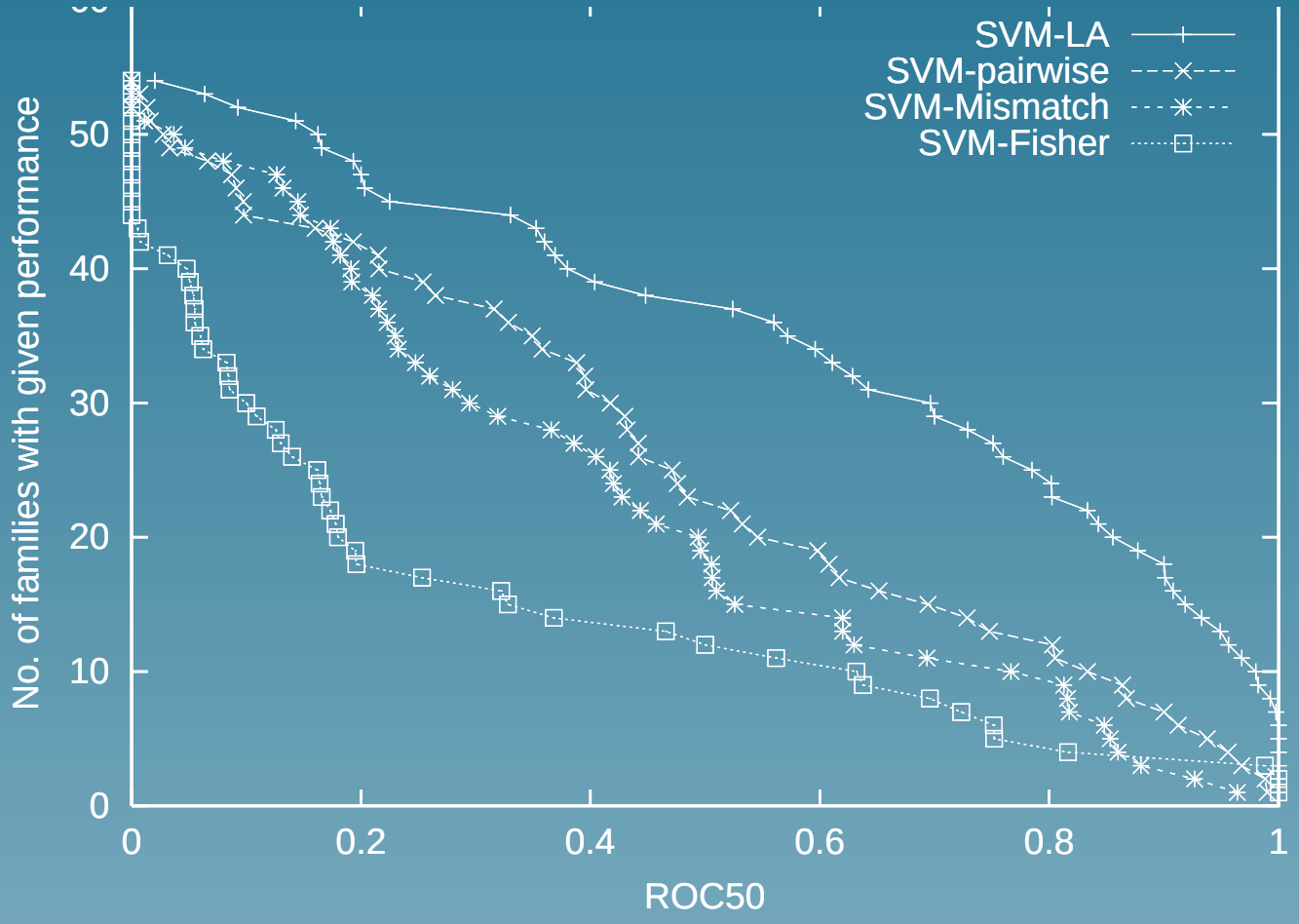
Smith-Waterman (SW) score

$$SW(x,y) = \max_{\pi \in \Pi(x,y)} s(x,y,\pi)$$

- This is not a kernel in general
- But the following is a valid kernel:

$$K_{LA}^{(\beta)}(x,y) = \sum_{\pi \in \Pi(x,y)} \exp\left(\beta s(x,y,\pi)\right),$$

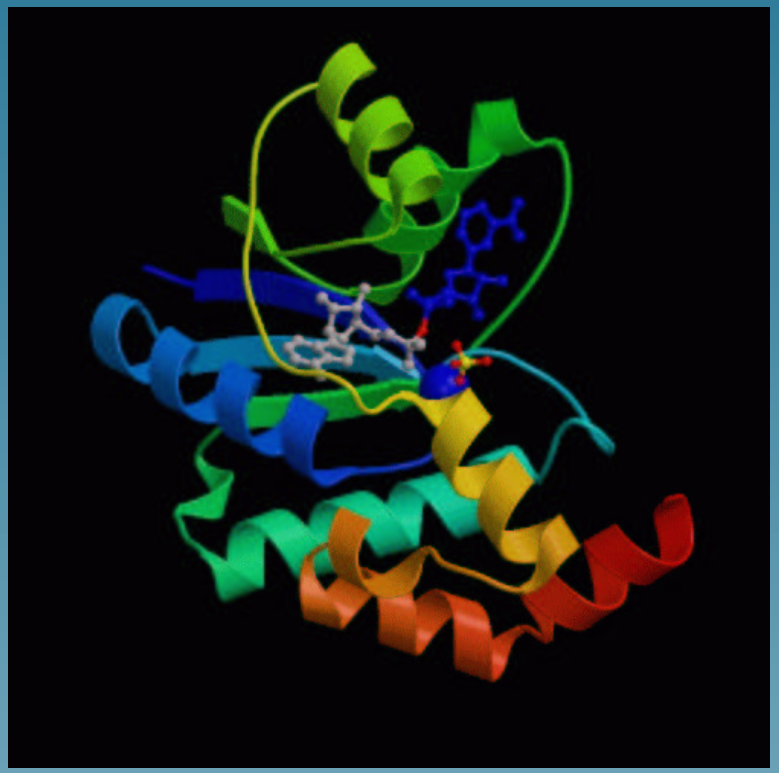
SCOP superfamily recognition benchmark



Partie 3

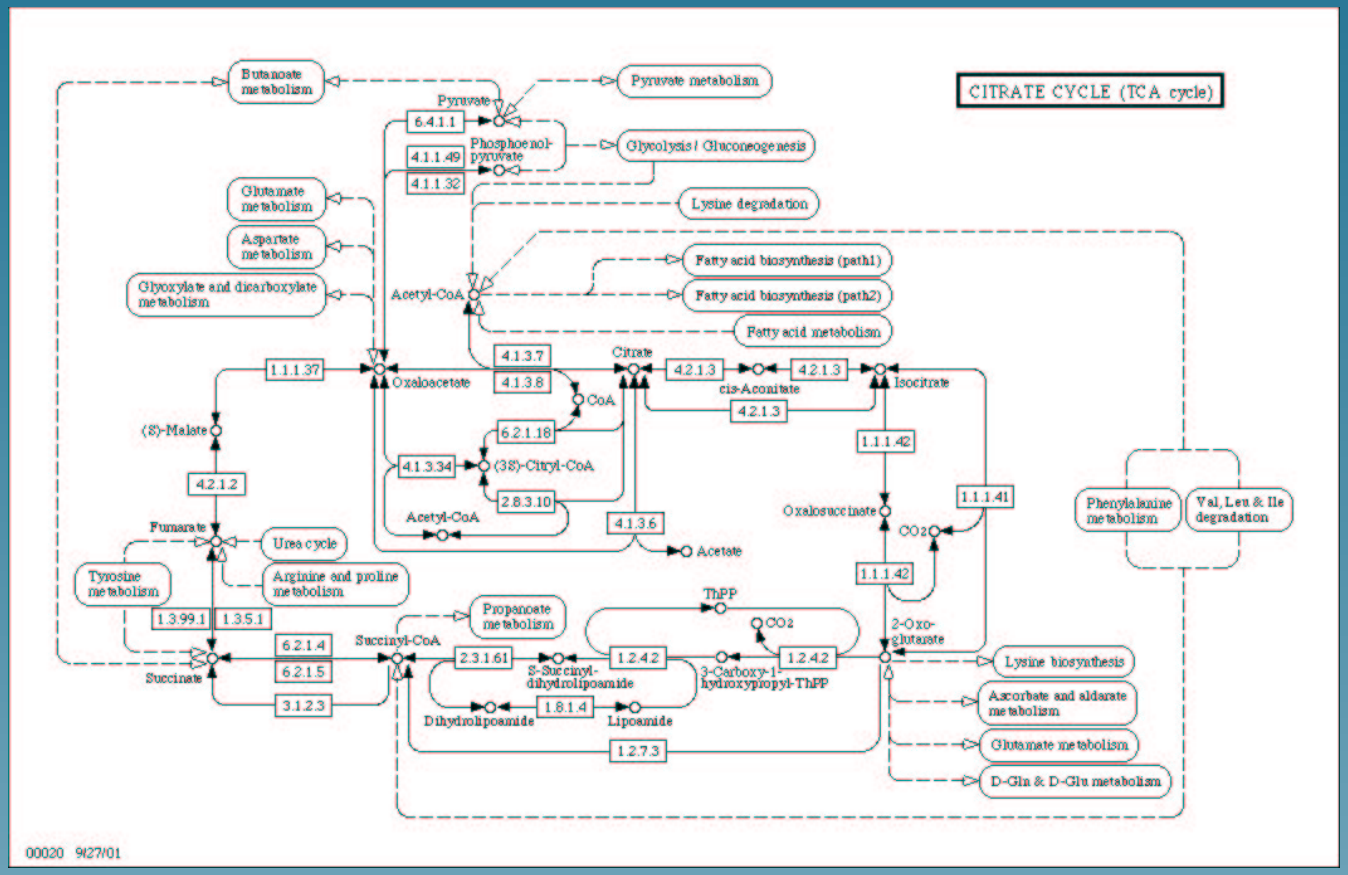
Analysis of microarray data with pathways information

Genes encode proteins which can catalyse chemical reations



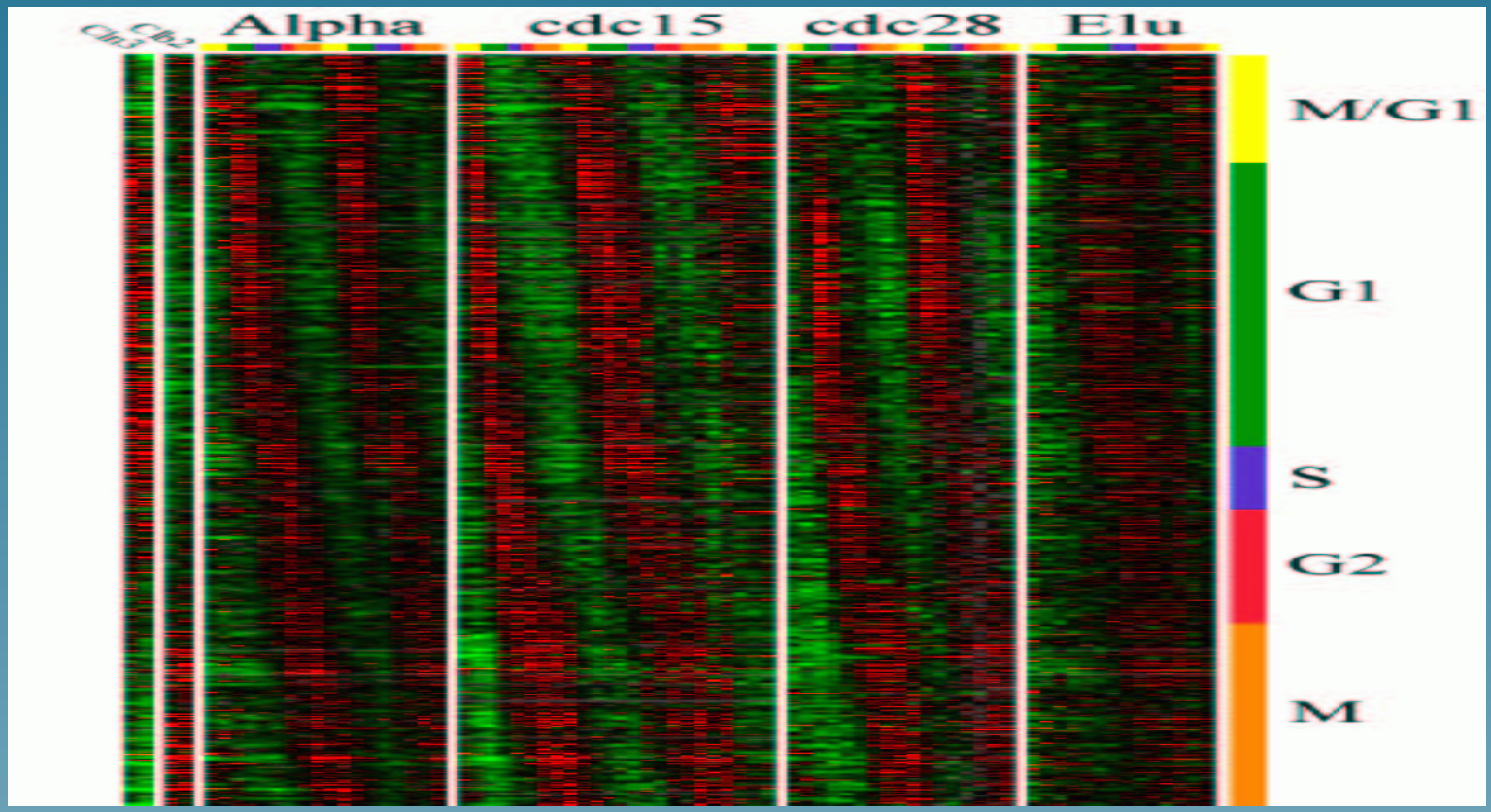
Nicotinamide Mononucleotide Adenylyltransferase With Bound Nad+

Chemical reactions are often parts of pathways



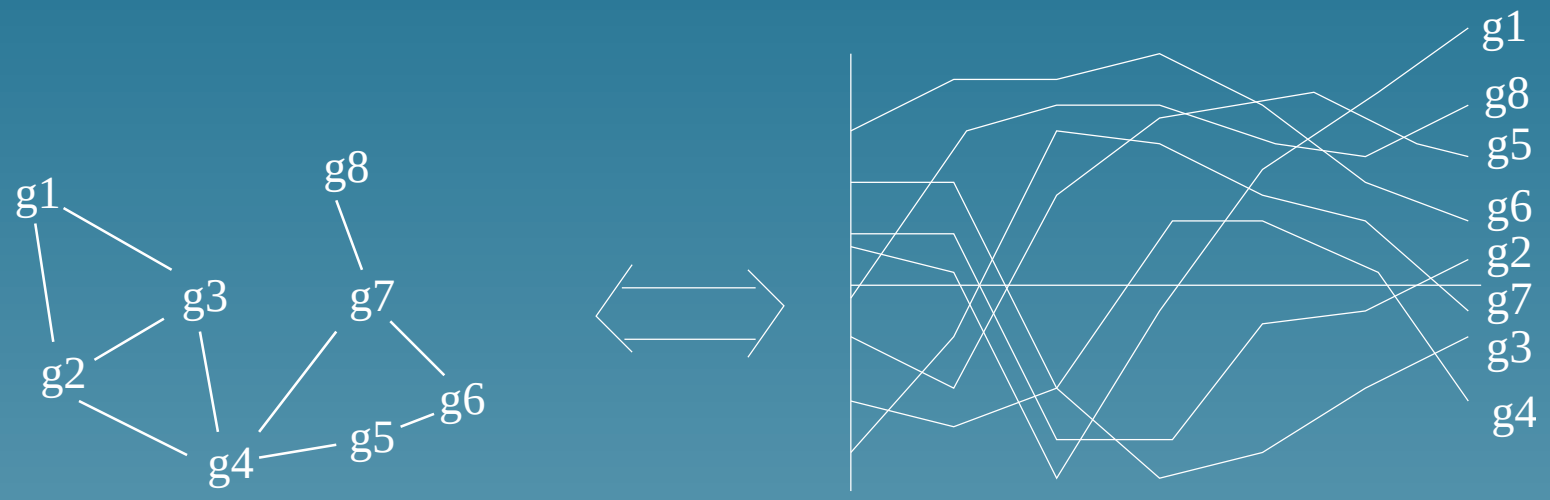
From http://www.genome.ad.jp/kegg/pathway

Microarray technology monitors RNA quantity



(From Spellman et al., 1998)

Comparing gene expression and protein network

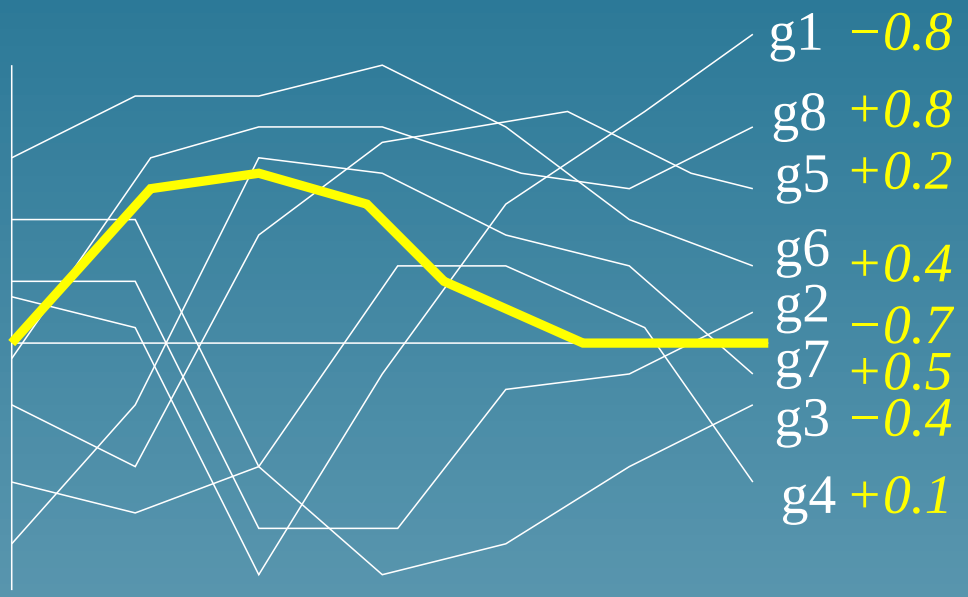


Gene network

Expression profiles

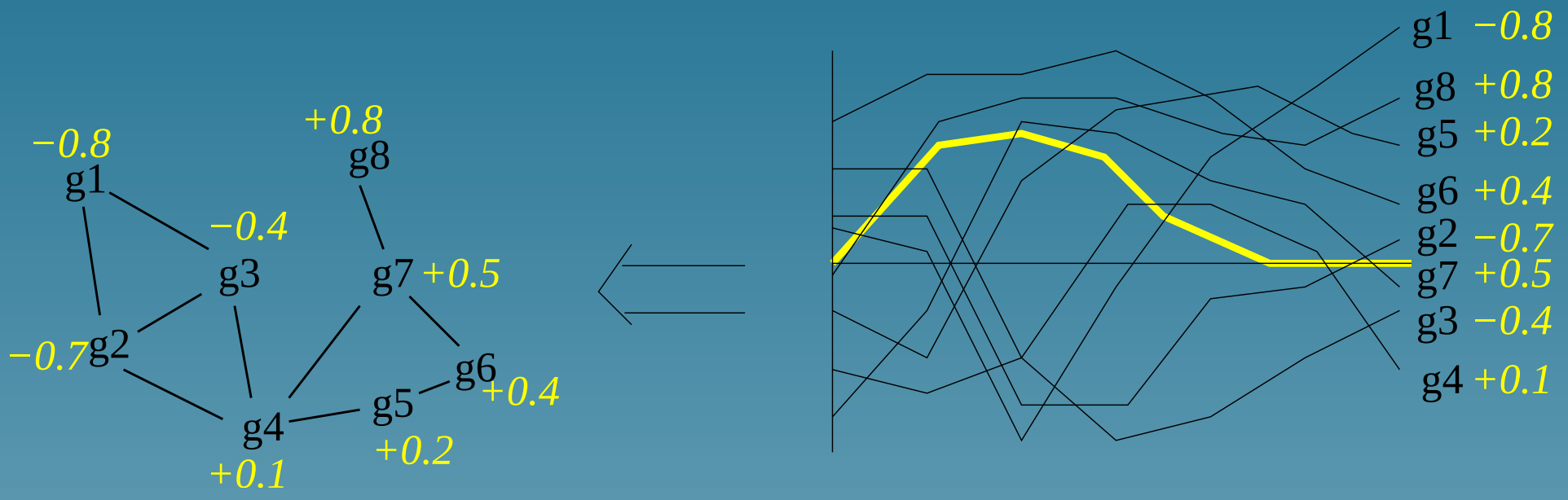
Are there "correlations"?

Pattern of expression



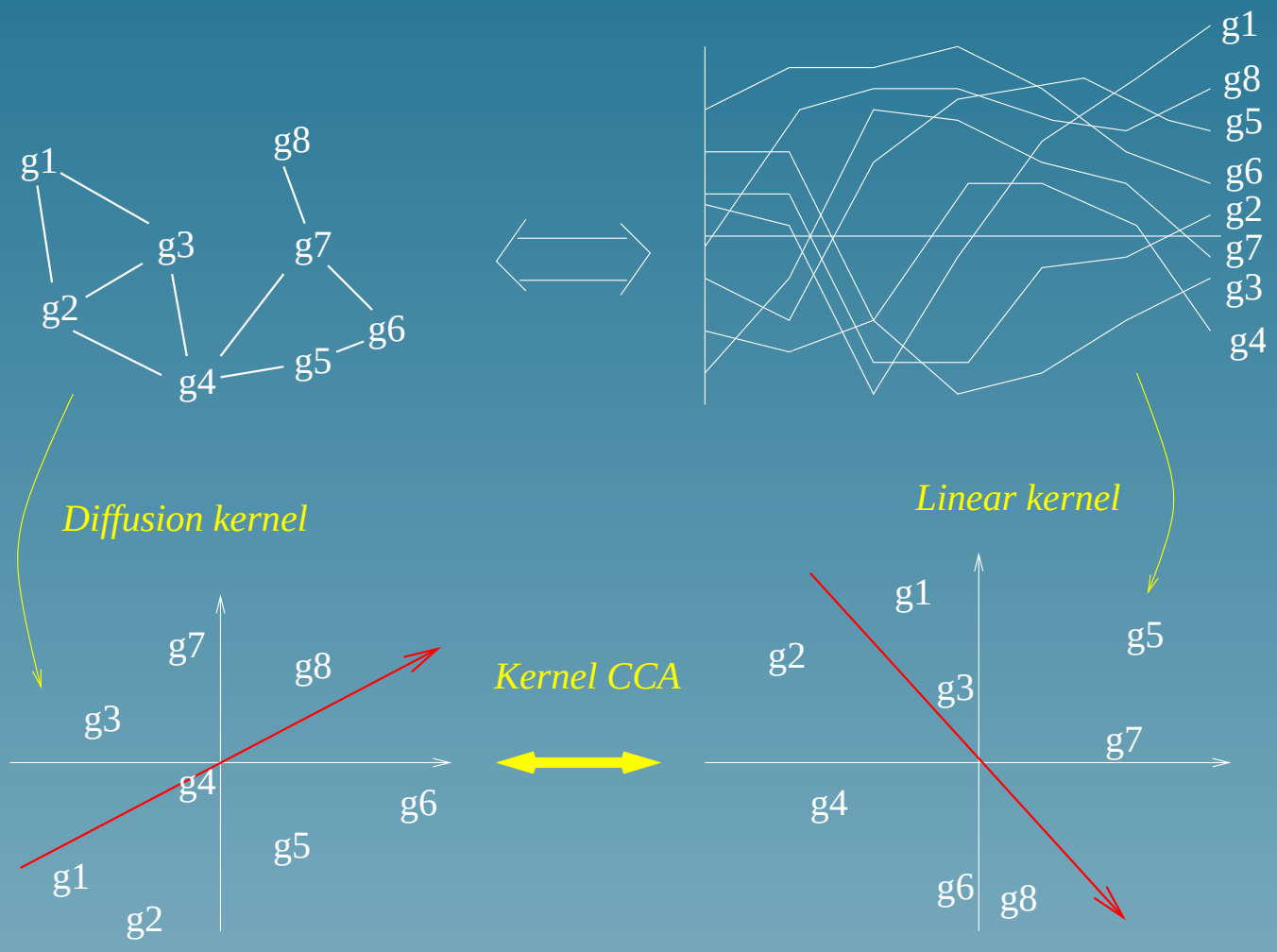
- In yellow: a candidate pattern , and the correlation coefficient with each gene profile

Pattern smoothness



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

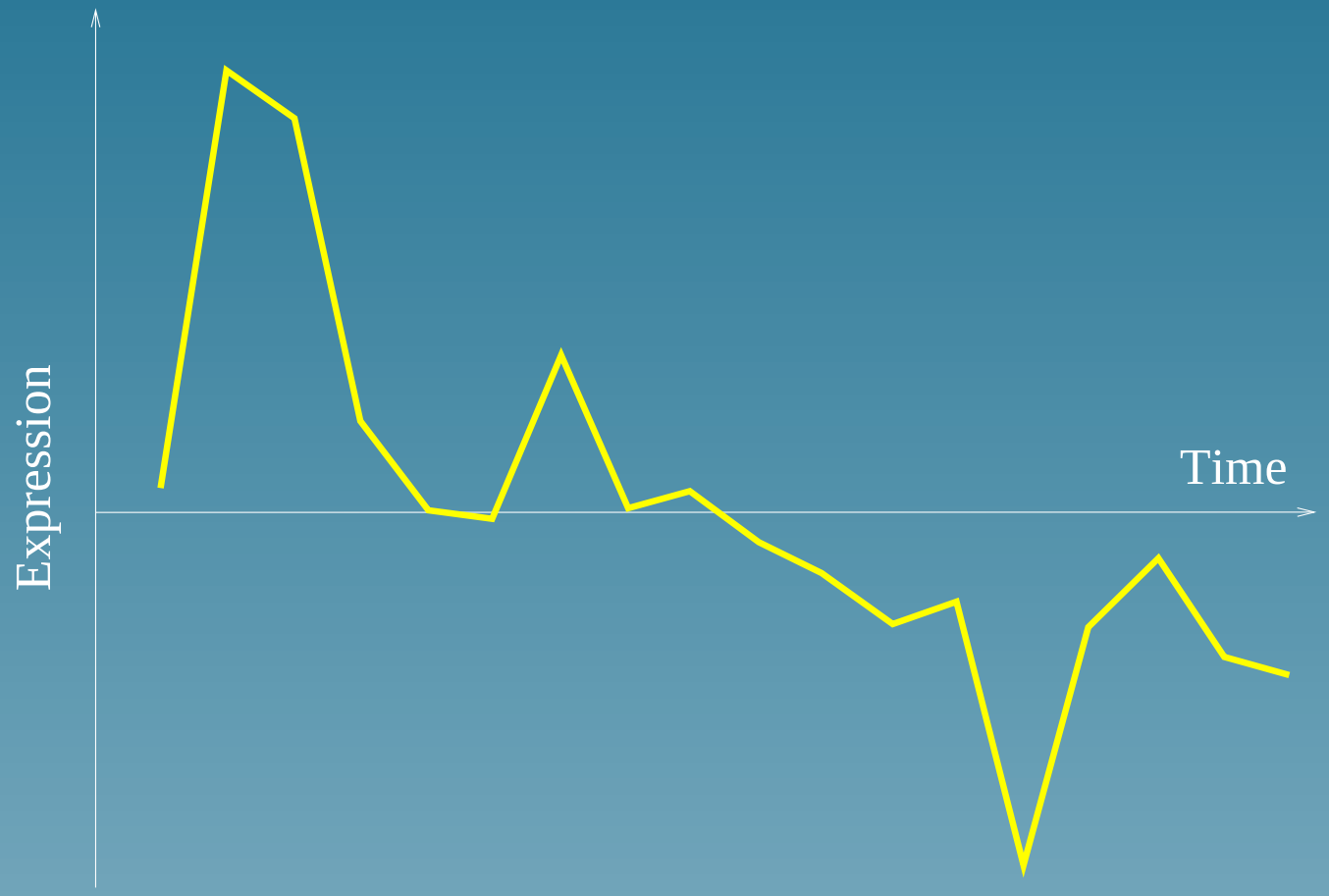
Summary


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

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

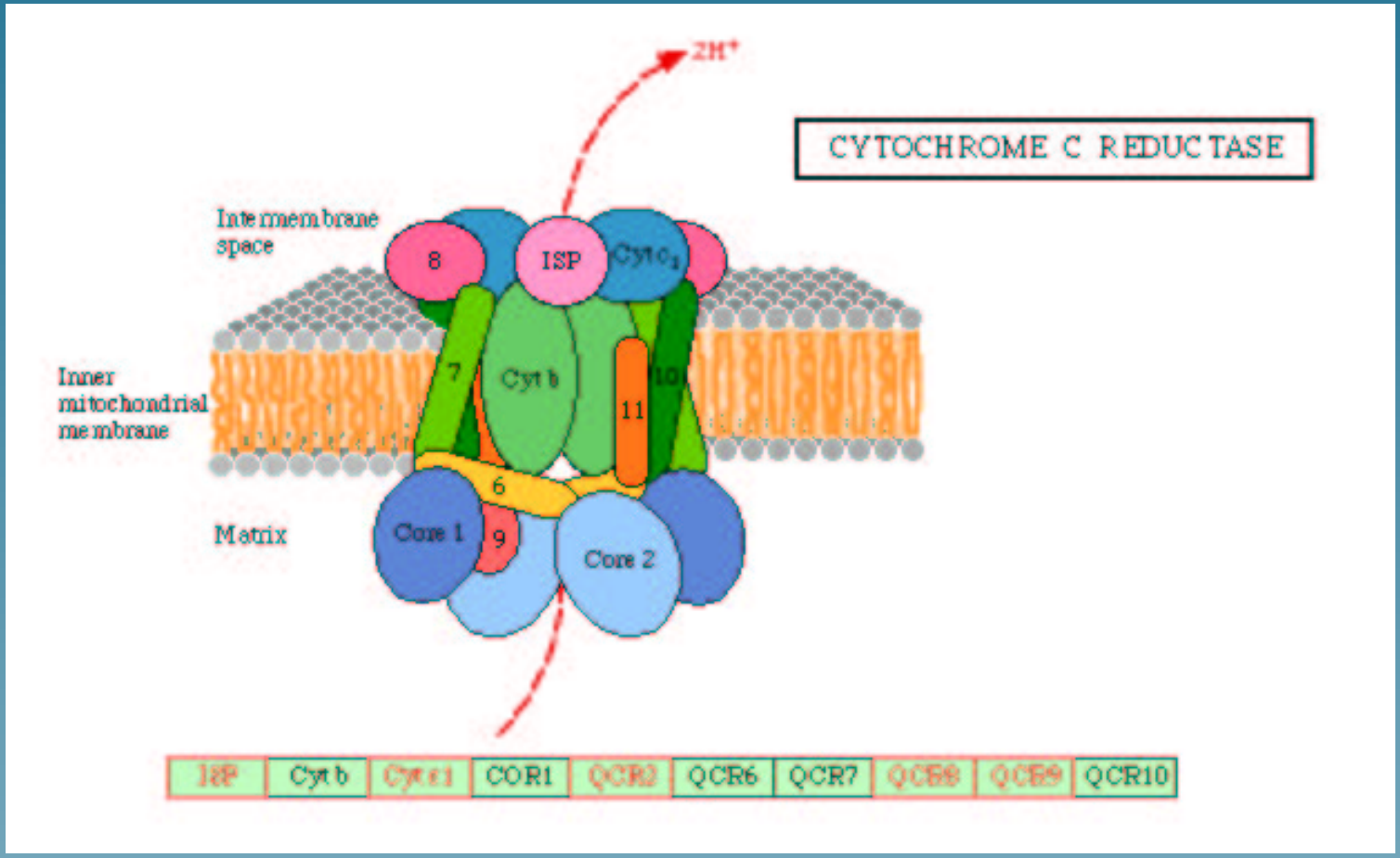


Related metabolic pathways

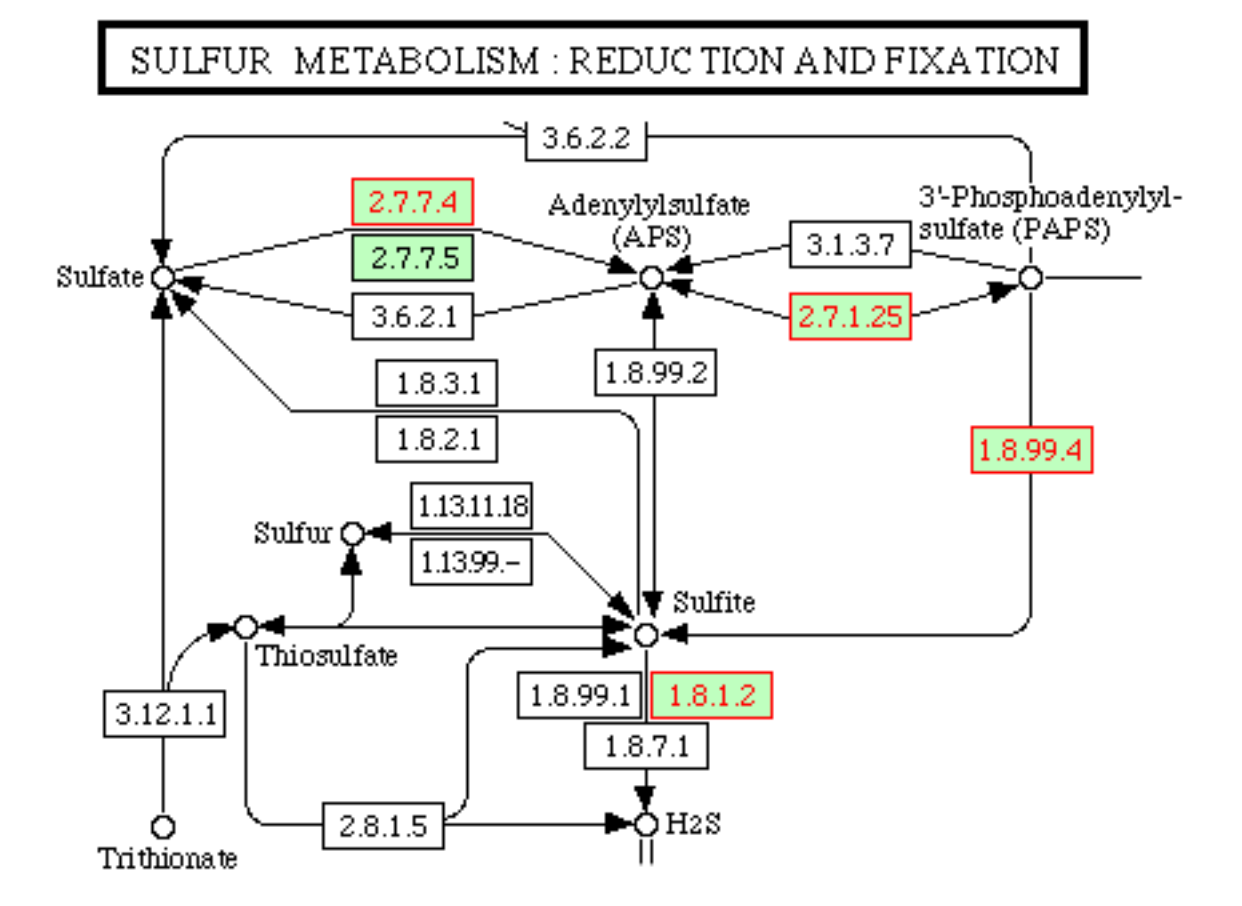
50 genes with highest $s_2 - s_1$ belong to:

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

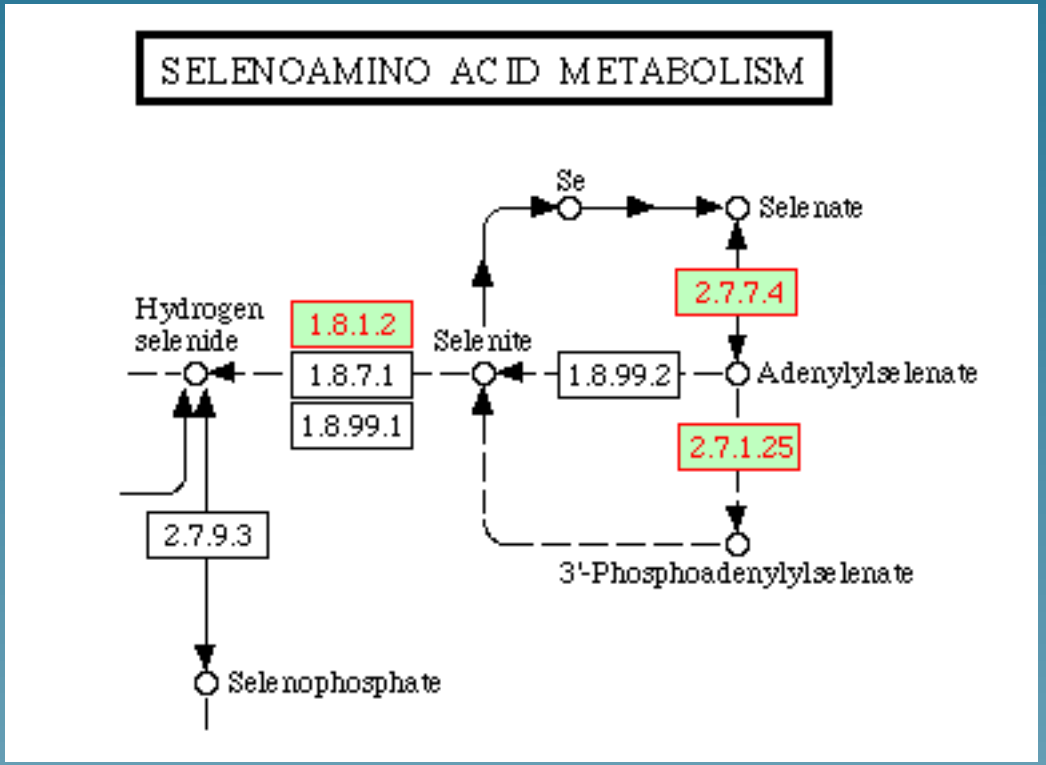
Related genes



Related genes



Related genes



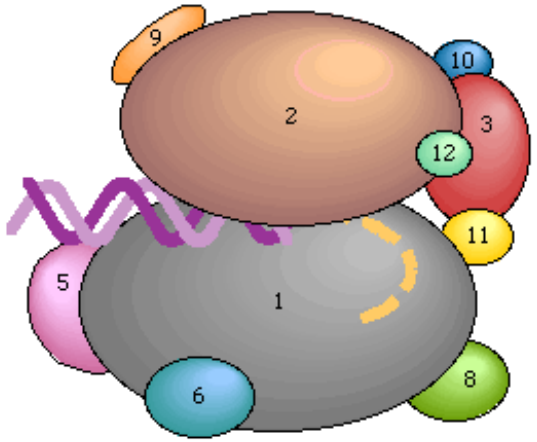
Opposite pattern

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

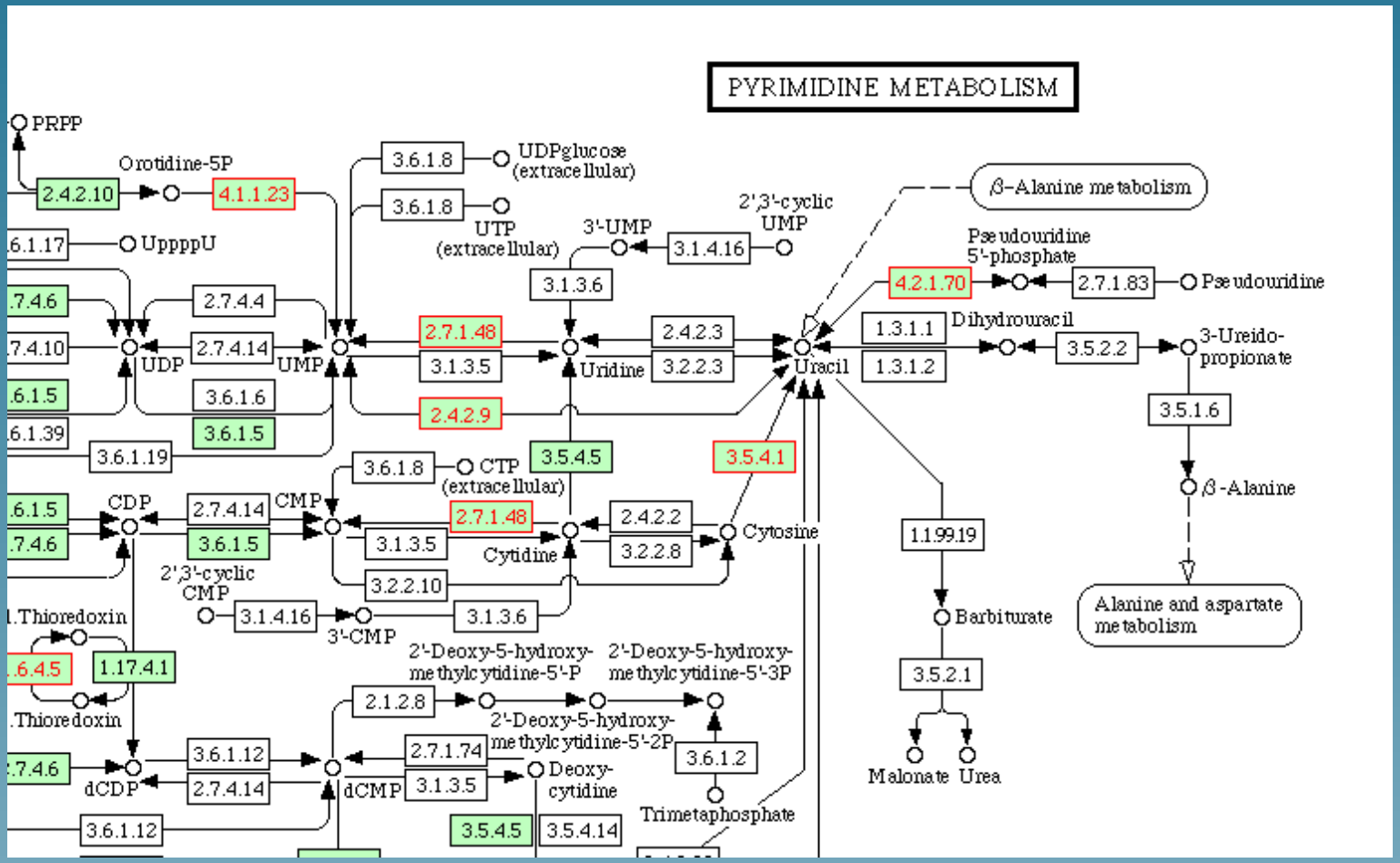
Eukaryotic Pol III

C2	C3	C4	C5	C11
C1	C19	C25	C31	C34

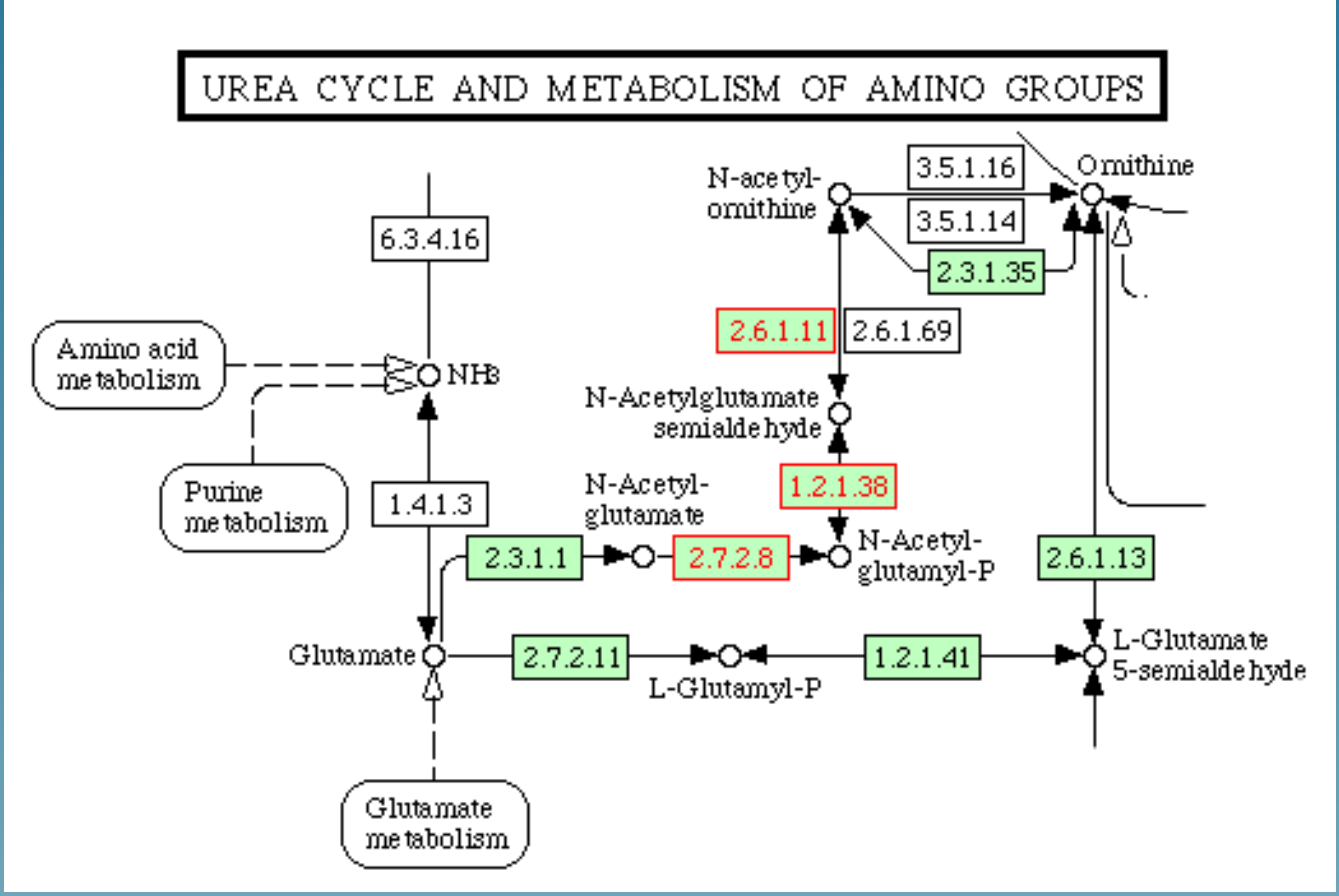
Eukaryotic Pol I

A2	A12	A14	A34	A43	A49
A1					

Related genes



Related genes



Extensions

- Can be used to extract features from expression profiles (preprint 2002)
- Can be generalized to more than 2 datasets and other kernels
- Can be used to extract clusters of genes (e.g., operon detection, *ISMB 03* with Y. Yamanishi, A. Nakaya and M. Kanehisa)

Conclusion

Conclusion

- Kernels offer a versatile framework to represent biological data
- SVM and kernel methods work well on real-life problems, in particular in high dimension and with noise
- Data integration with kernel CCA is possible