Support vector machines and Applications in bioinformatics

Jean-Philippe.Vert@mines.org

Ecole des Mines de Paris
Computational Biology group

Sanofi, May 19, 2003.

Ecole des Mines de Paris

- 1770 persons (250 academics, 400 PhD students, 670 undergraduates/M.S.)
- Excellent formation (undergraduate and graduate)
- 19 research centers (earth science, energy, mechanics, applied maths, economics)
- 21.5 Million euros of research contracts (through Armines)

Computational biology at the Ecole des Mines de Paris

- Started 11/2002, 5 persons in 9/2003
- Expertise in statistics, machine learning, data mining...
- Projects: functional genomics, learning from heterogeneous data, virtual screening of chemical compouds, microarray data and pathway analysis...

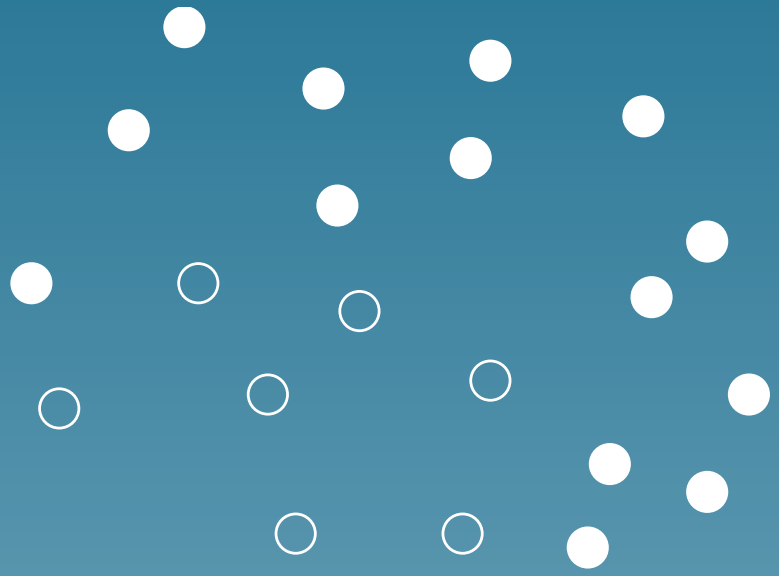
Overview

1. Pattern recognition and Support Vector Machines
2. Signal peptide detection
3. Virtual screening of small molecules
4. Analysis of microarray data with pathways information

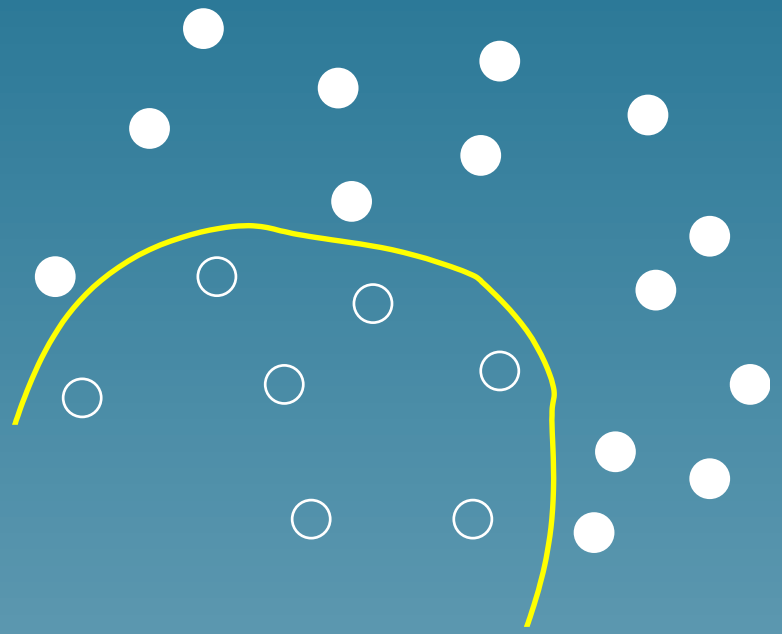
Partie 1

Pattern recognition and Support Vector Machines

The pattern recognition problem

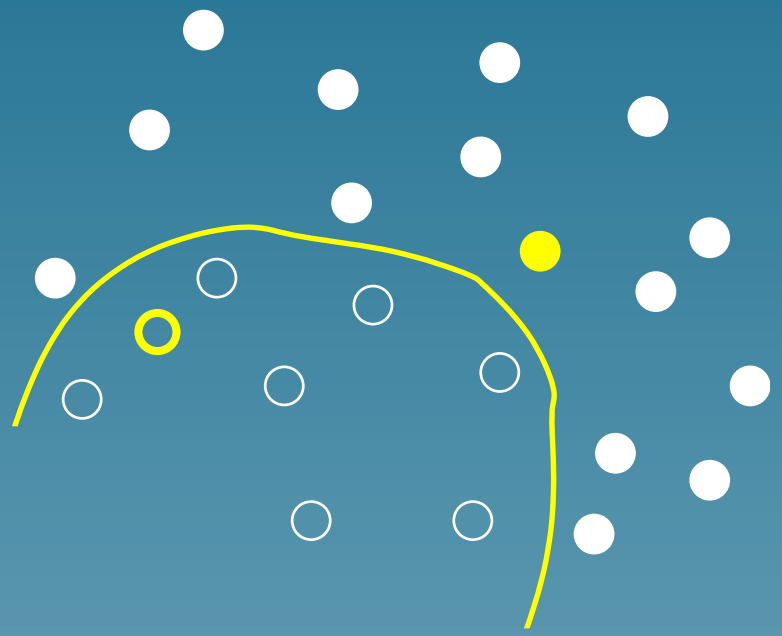


The pattern recognition problem



- Learn from labelled examples a discrimination rule

The pattern recognition problem

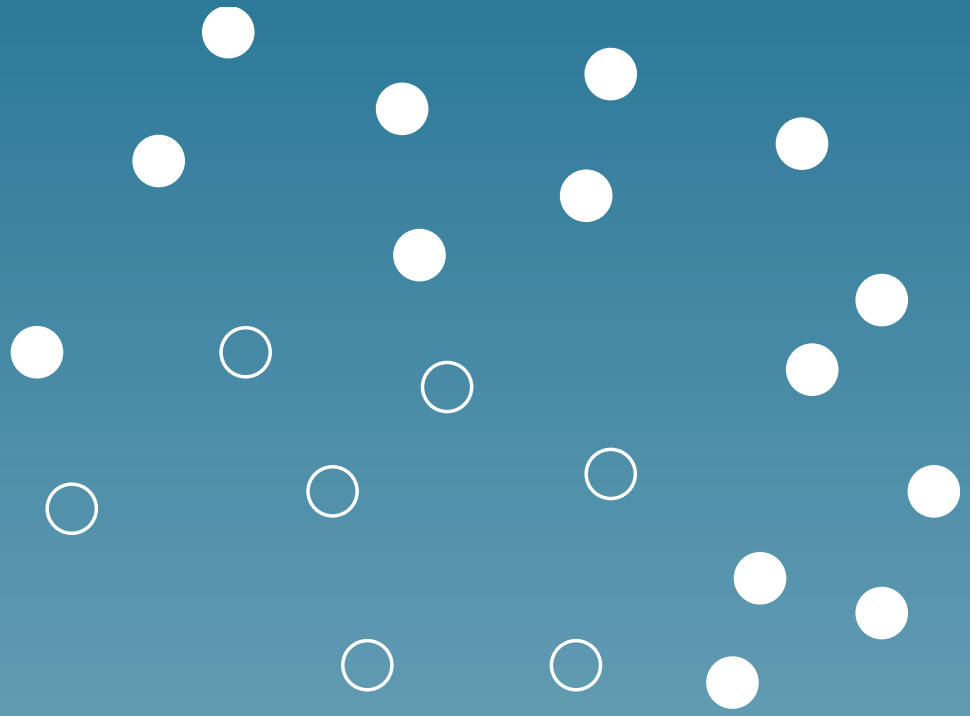


- Learn from labelled examples a discrimination rule
- Use it to predict the class of new points

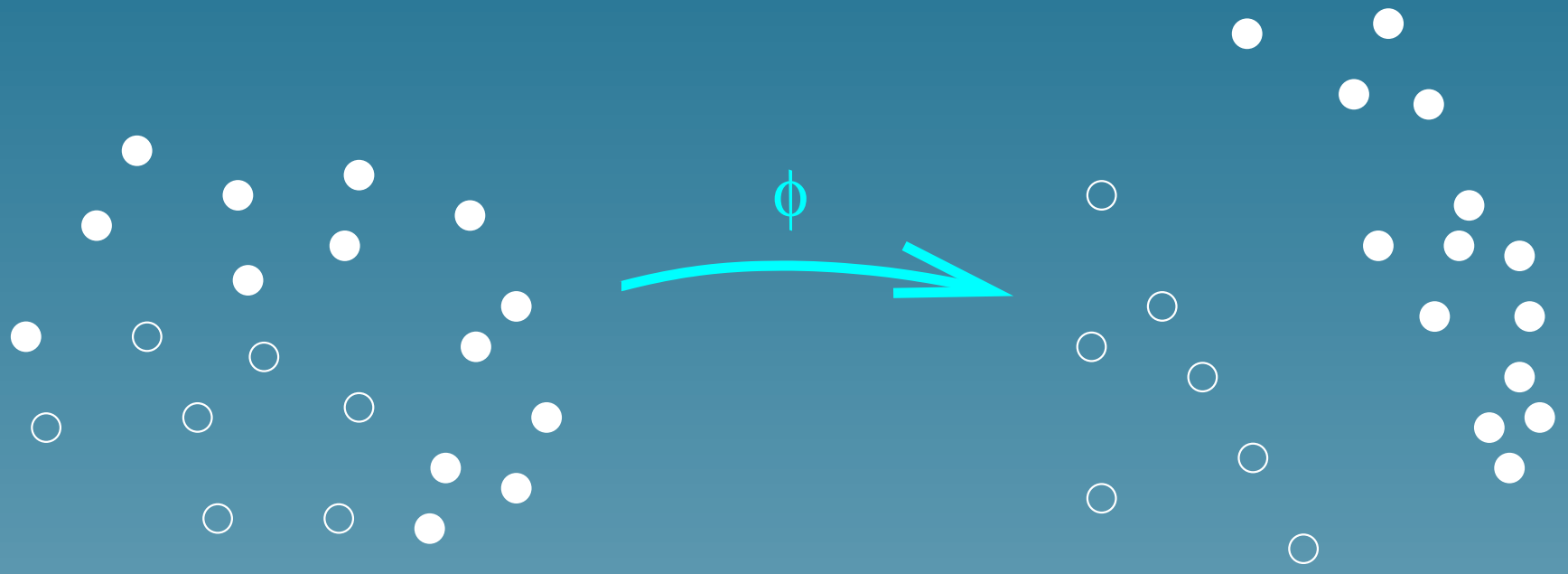
Pattern recognition examples

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

Support Vector Machines for pattern recognition

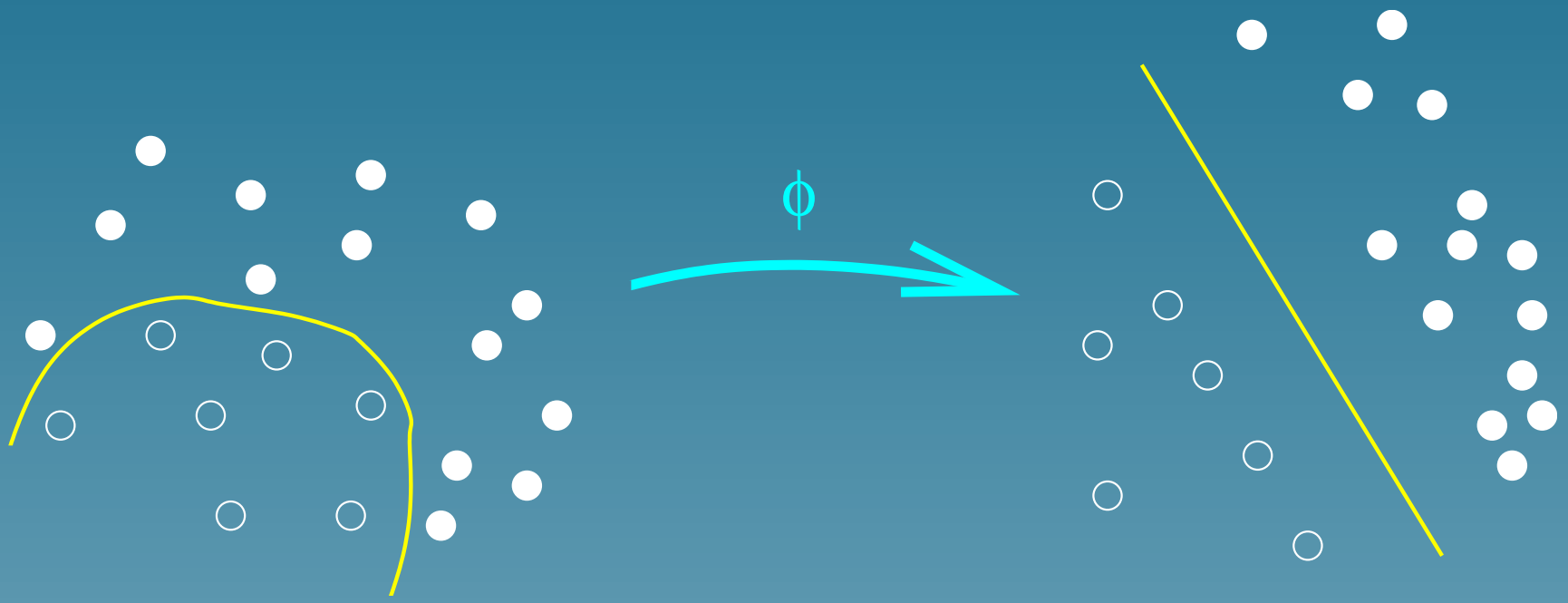


Support Vector Machines for pattern recognition



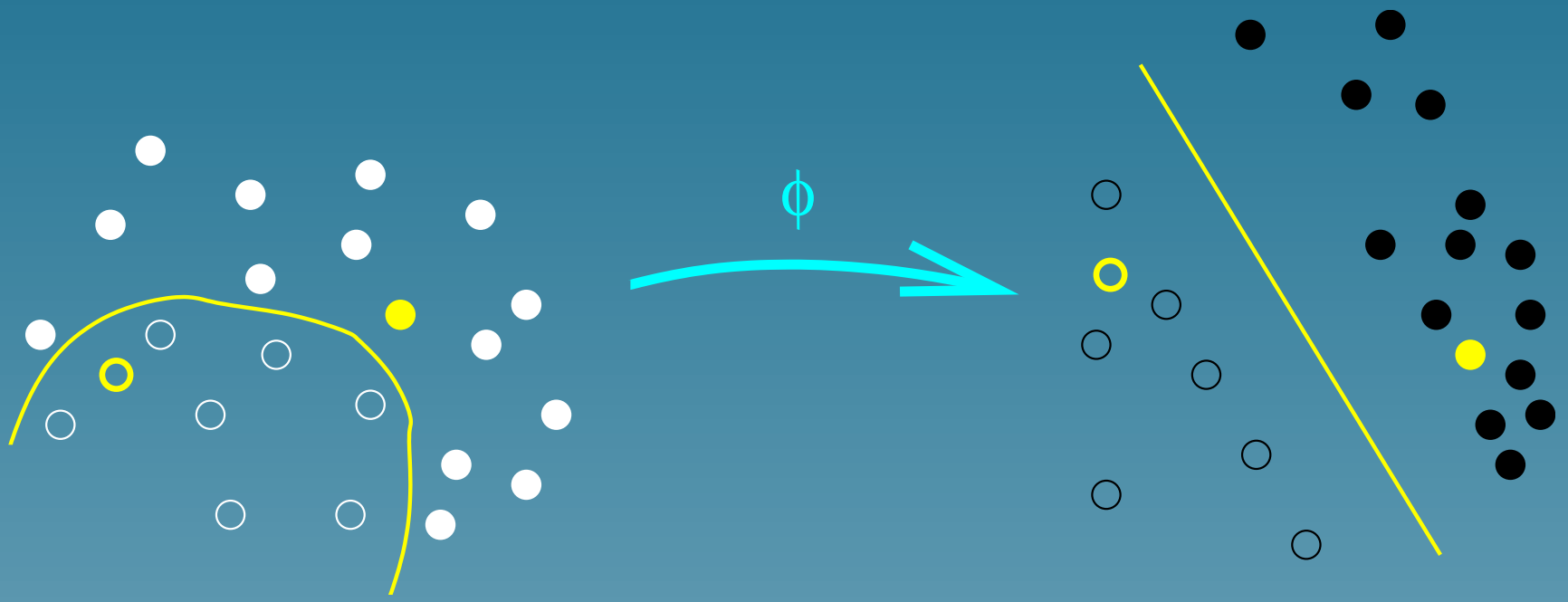
- Object x represented by the vector $\Phi(\vec{x})$ (feature space)

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

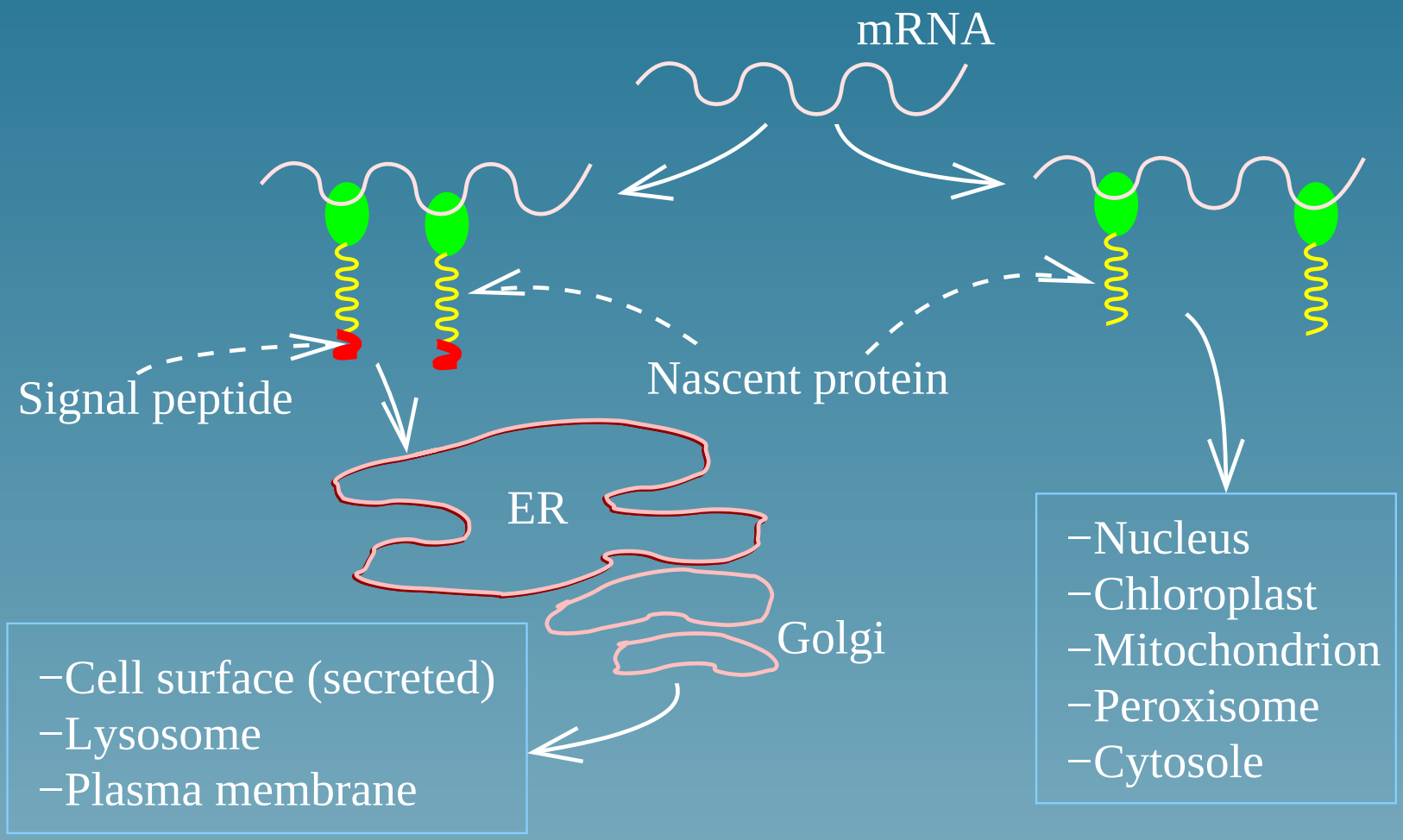
Advantages of SVM

- 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

Partie 2

Signal peptide cleavage site detection

Secretory pathway



Signal peptides

Protein	-1 +1
(1)	MKANAKTIIAGMIALAISHTAMA EE...
(2)	MKQSTIALALLPLLFTPVTKA RT...
(3)	MKATKLVLGAVILGSTLLAG CS...

(1):Leucine-binding protein, (2):Pre-alkaline phosphatase, (3)Pre-lipoprotein

Signal peptides

Protein	-1 +1
(1)	MKANAKTIIAGMIALAISHTAMA EE...
(2)	MKQSTIALALLPLLFTPVTKA RT...
(3)	MKATKLVLGAVILGSTLLAG CS...

(1):Leucine-binding protein, (2):Pre-alkaline phosphatase, (3)Pre-lipoprotein

- 6-12 hydrophobic residues (in yellow)
- (-3,-1) : small uncharged residues

The classification problem(s)

- Problem 1 :
  Given an aminoacids windows:

$$[x_{-8}, x_{-7}, \ldots, x_{-1}, x_1, x_2] = \text{ILGSTLLACS}$$

  is there a cleavage site between x_{-1} and x_1 ?

The classification problem(s)

- Problem 1 :
  Given an aminoacids windows:

$$[x_{-8}, x_{-7}, \ldots, x_{-1}, x_1, x_2] = \text{ILGSTLLACS}$$

  is there a cleavage site between x_{-1} and x_1 ?

- Problem 2 :
  Given an protein sequence, does it contain a signal peptide?

Current methods : Problem 1

- Weight matrix method: compute the score of a window by:

$$s(ILGSTLLACS) = s_{-8}(I) + s_{-7}(L) + \ldots + s_2(S)$$

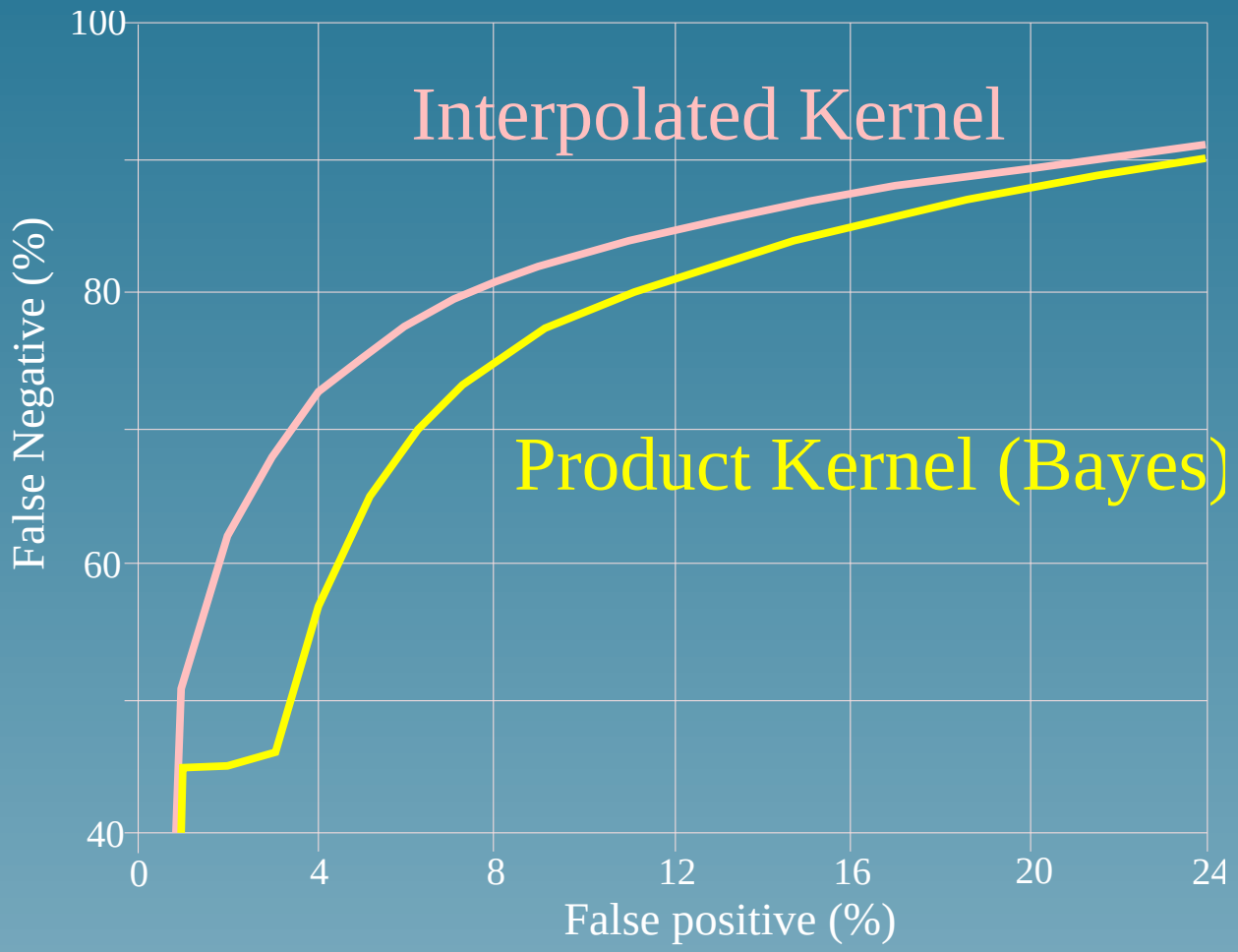
  where s_i have been trained from example to discriminate between windows with or without cleavage site (Von Heijne)

- Neural networks (Brunak et al.)

SVM approach (*PSB 2002*)

- We need a kernel $K(w_1, w_2)$ between 2 windows
- It is possible to transform a weight matrix into a kernel (technical, see paper)
- Experiment : 1,418 positive examples, 65,216 negative examples, cross-validation

Result: ROC curves



Remarks

- The weight matrix is used to define the geometry of the feature space (through the kernel)
- The SVM algorithm learns a linear discrimination in this space

Problem 2: signal peptide detection

- Classical approach: move a window along the sequence, check whether it looks like a typical signal peptide
- SVM approach: we need a string kernel $K(p_1, p_2)$ for variable-length protein sequences
- String kernel examples: Fisher kernel (Jaakkola et al. 99), spectrum and mismatch kernels (Leslie et al. 02), local alignment kernel (Vert et al. 03)...

Local alignment kernel

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

- The score is:

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

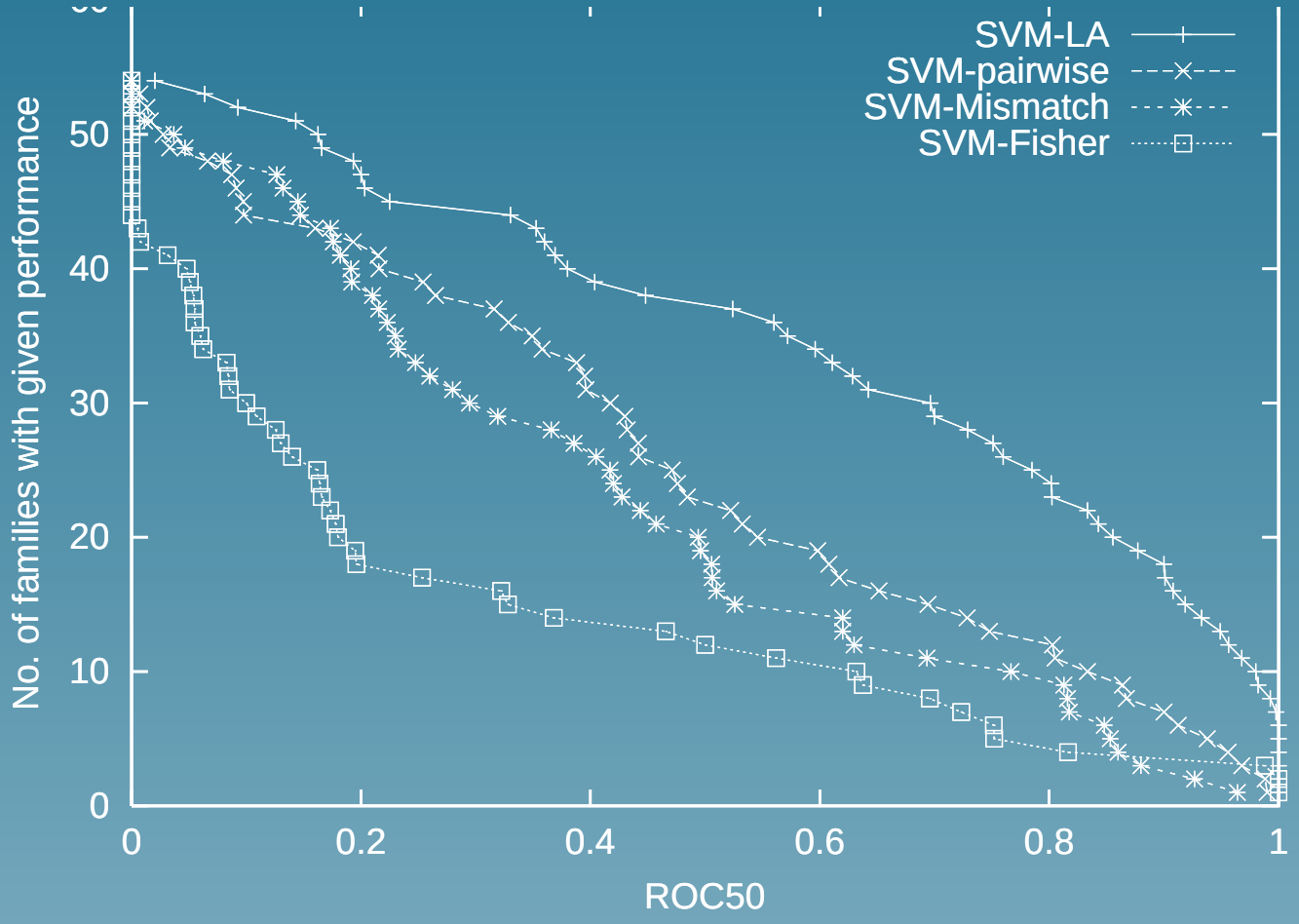
Smith-Waterman (SW) score

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

- 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

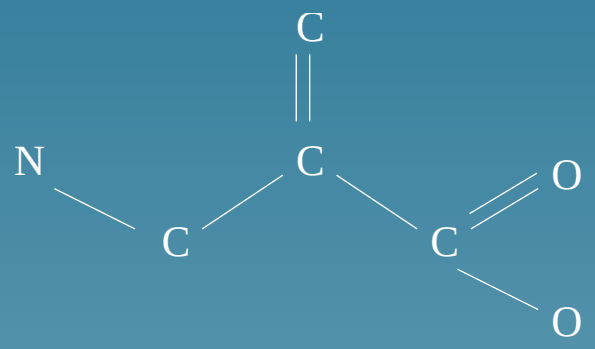


Partie 3

Virtual screening of small molecules

The problem

- Objects = chemical compounds (formula, structure..)



- Problem = predict their:
  - ⋆ drugability
  - ⋆ pharmacocinetics
  - ⋆ activity on a target etc...

Classical approaches

- Use molecular descriptors to represent the compouds as vectors
- Select a limited numbers of relevant descriptors
- Use linear regression, NN, nearest neighbour etc...

SVM approach

- We need a kernel $K(c_1, c_2)$ between compounds

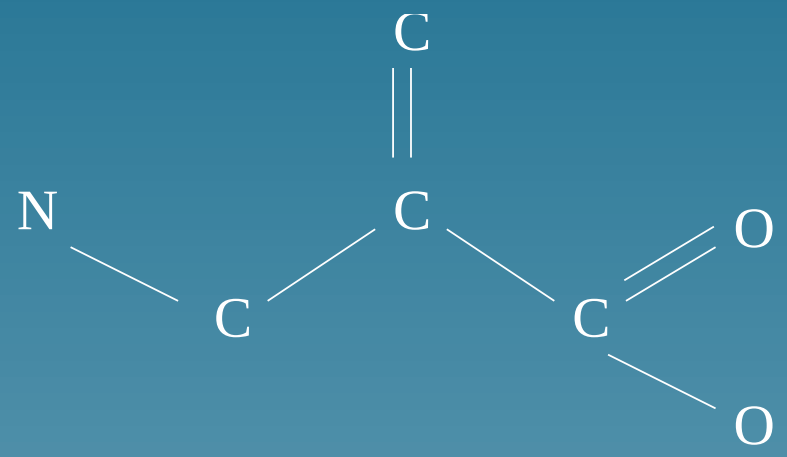
SVM approach

- We need a kernel $K(c_1, c_2)$ between compounds
- One solution: inner product between vectors

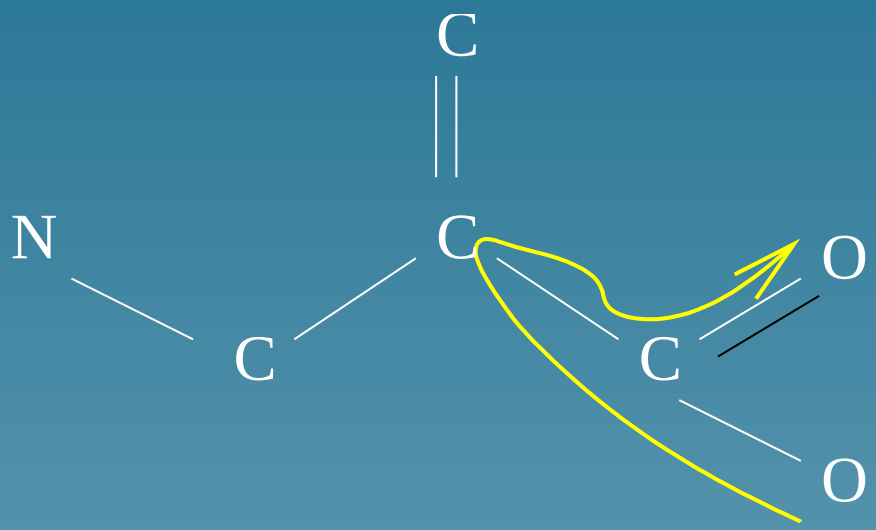
SVM approach

- We need a kernel $K(c_1, c_2)$ between compounds
- One solution: inner product between vectors
- Alternative solution: define a kernel directly using graph comparison tools

Example: graph kernel

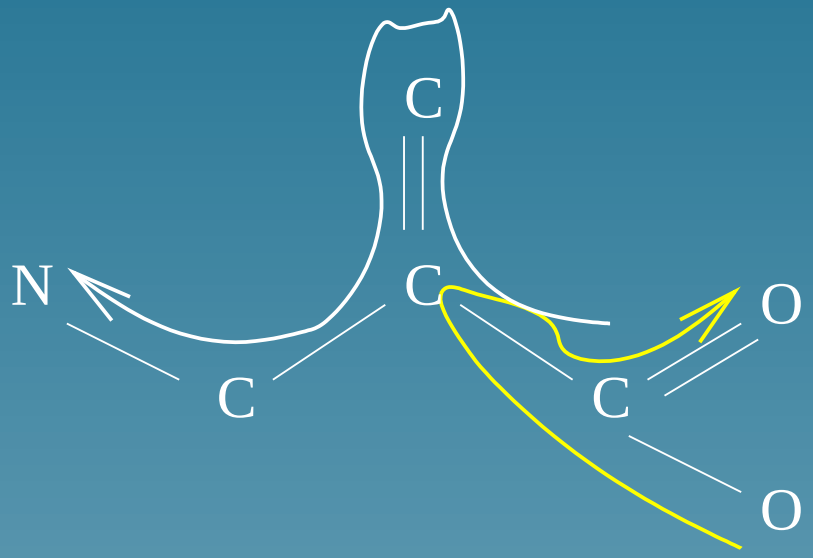


Example: graph kernel



Extract random paths

Example: graph kernel



Extract random paths

Example: graph kernel

- Let H_1 be a random path of a compound c_1
- Let H_2 be a random path of a compound c_2
- The following is a valid kernel:

$$K(c_1, c_2) = \mathsf{Prob}(H_1 = H_2).$$

Remarks

- Interesting preliminary results in mutagenesis prediction (benchmark dataset)

Remarks

- Interesting preliminary results in mutagenesis prediction (benchmark dataset)
- Two compounds are compared in terms of their common substructures

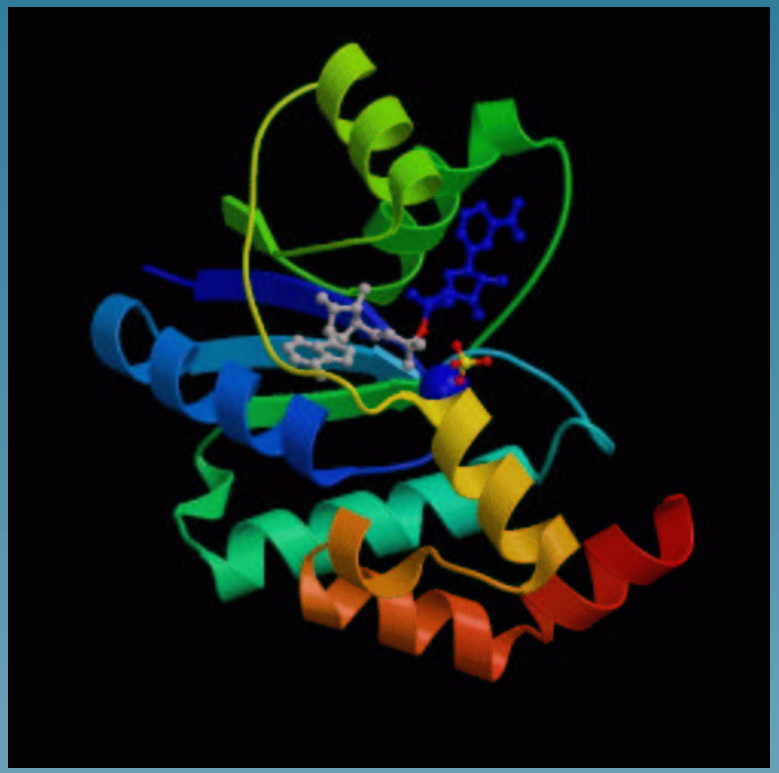
Remarks

- Interesting preliminary results in mutagenesis prediction (benchmark dataset)
- Two compounds are compared in terms of their common substructures
- What about kernels for the 3D structure?

Partie 4

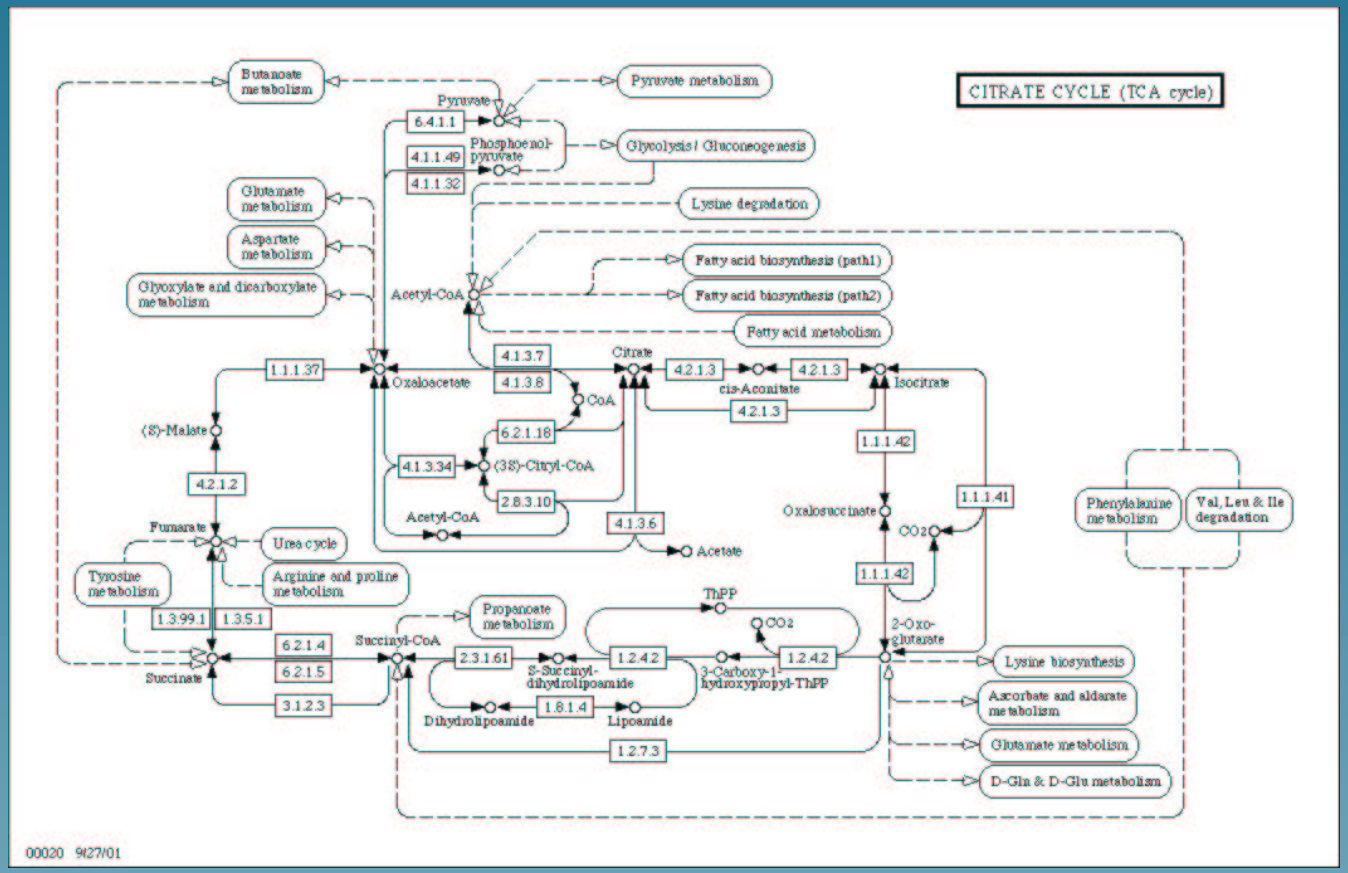
Analysis of microarray data with pathways information

Genes encode proteins which can catalyse chemical reations



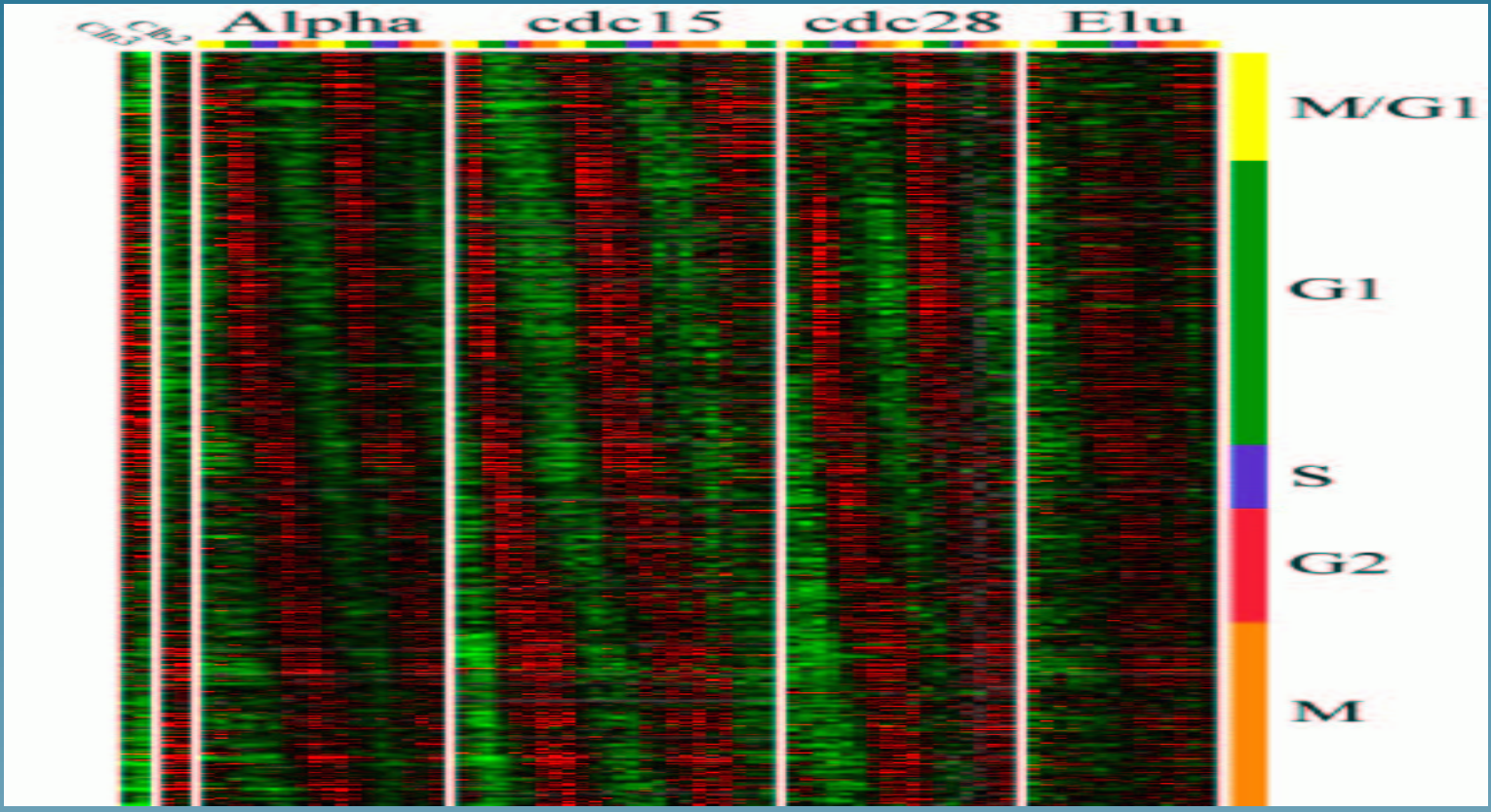
Nicotinamide Mononucleotide Adenylyltransferase With Bound Nad+

Chemical reactions are often parts of pathways



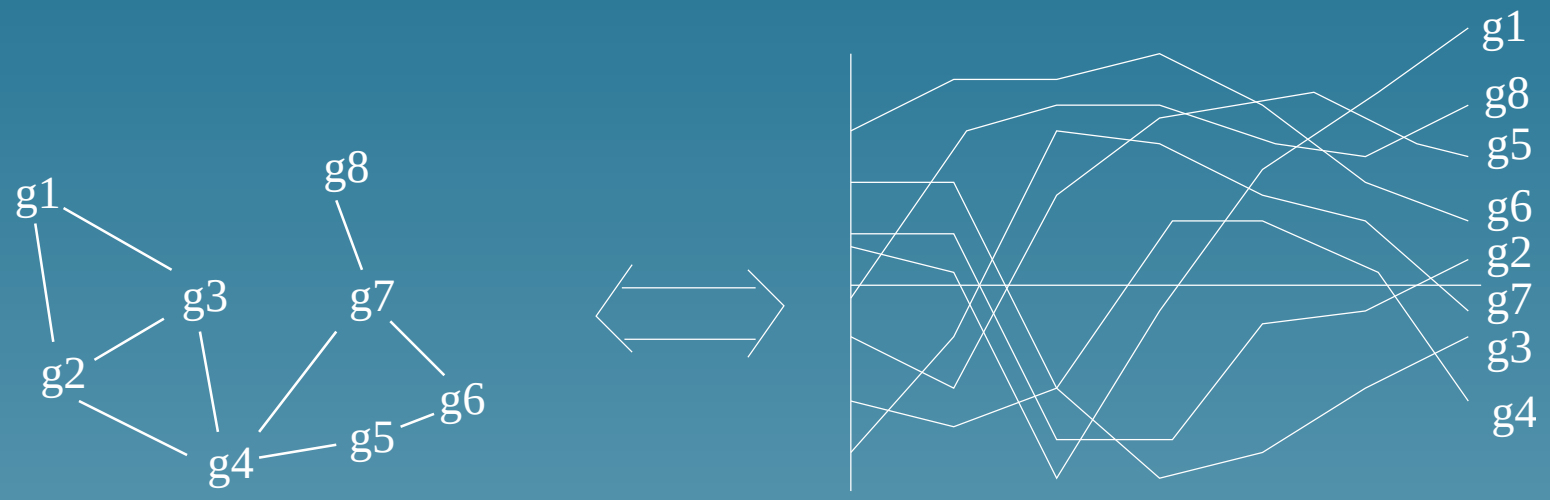
From http://www.genome.ad.jp/kegg/pathway

Microarray technology monitors RNA quantity



(From Spellman et al., 1998)

Comparing gene expression and protein network

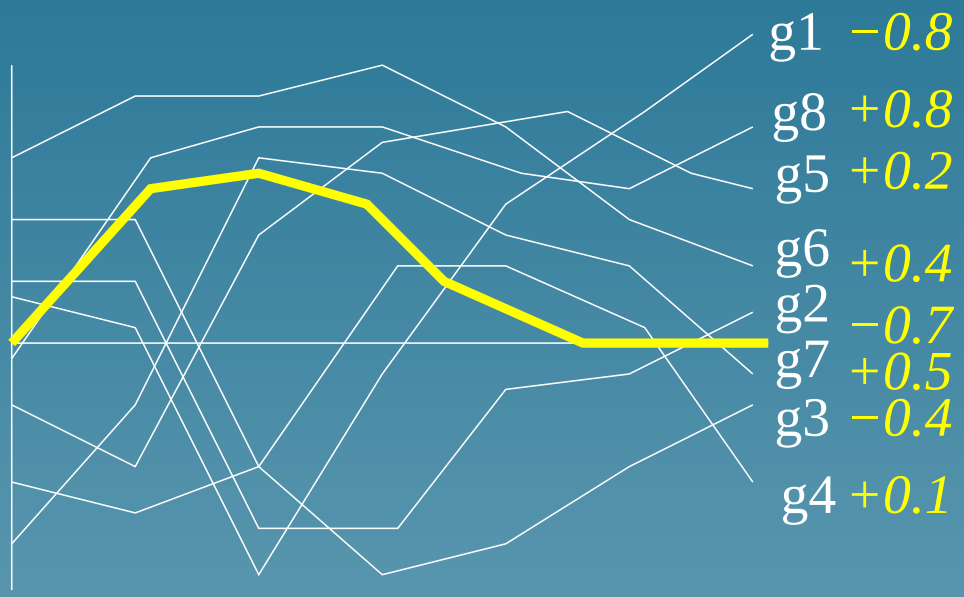


Gene network

Expression profiles

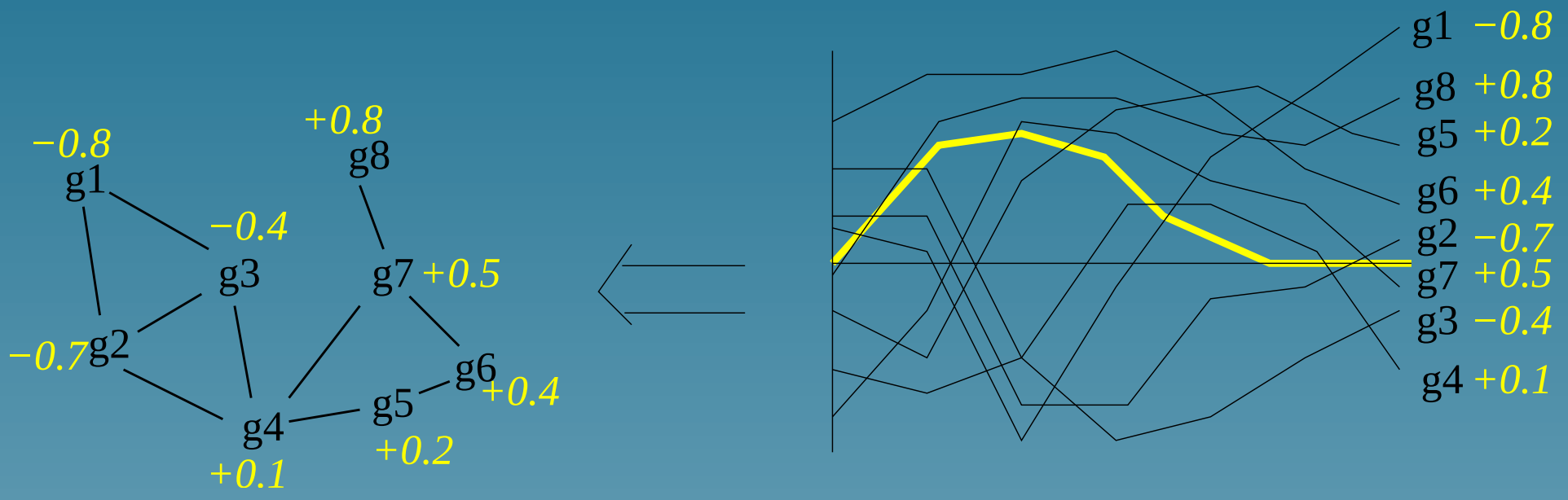
Are there "correlations"?

Pattern of expression



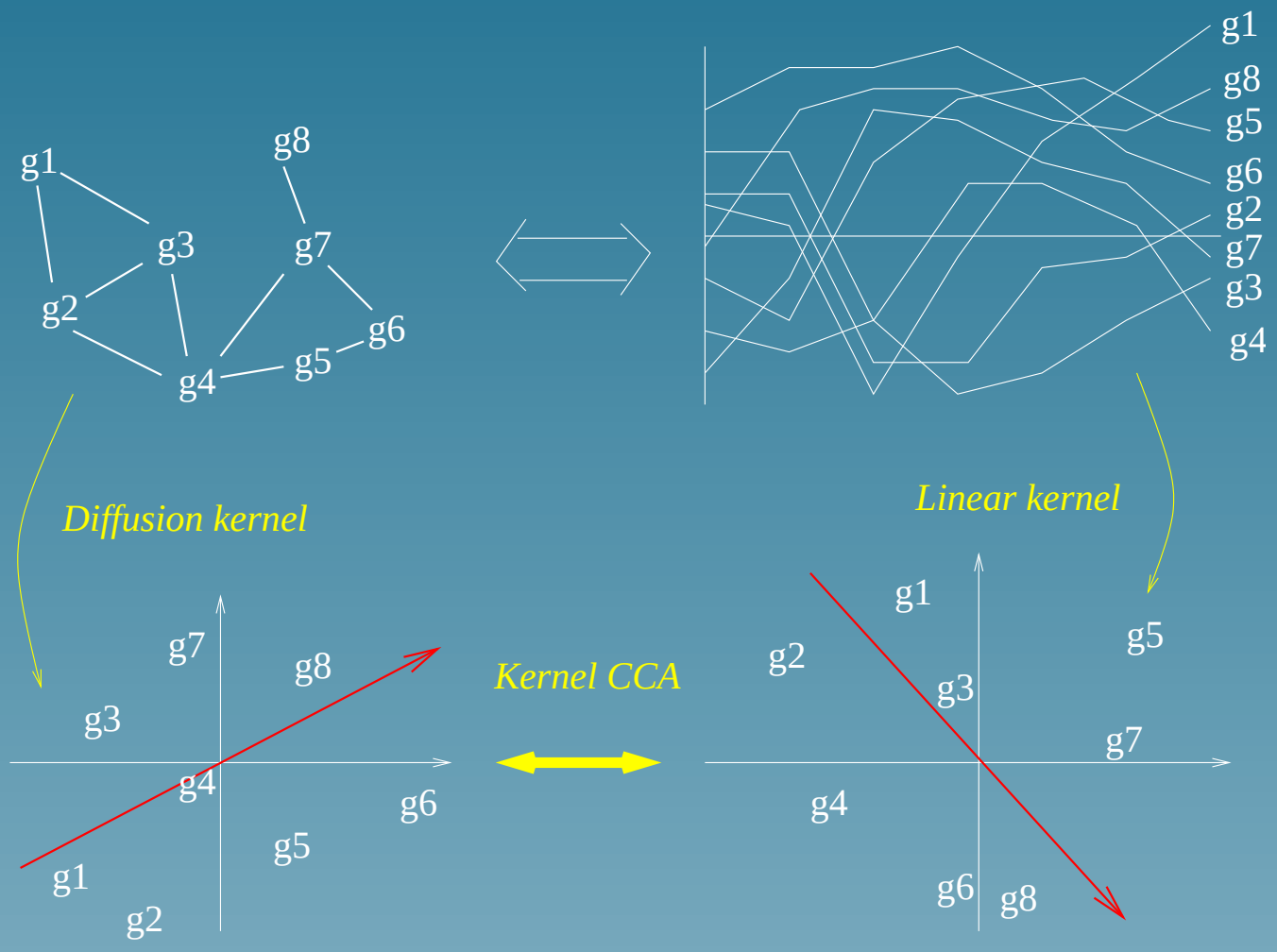
- In yellow: a candidate pattern , and the correlation coefficient with each gene profile

Pattern smoothness



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

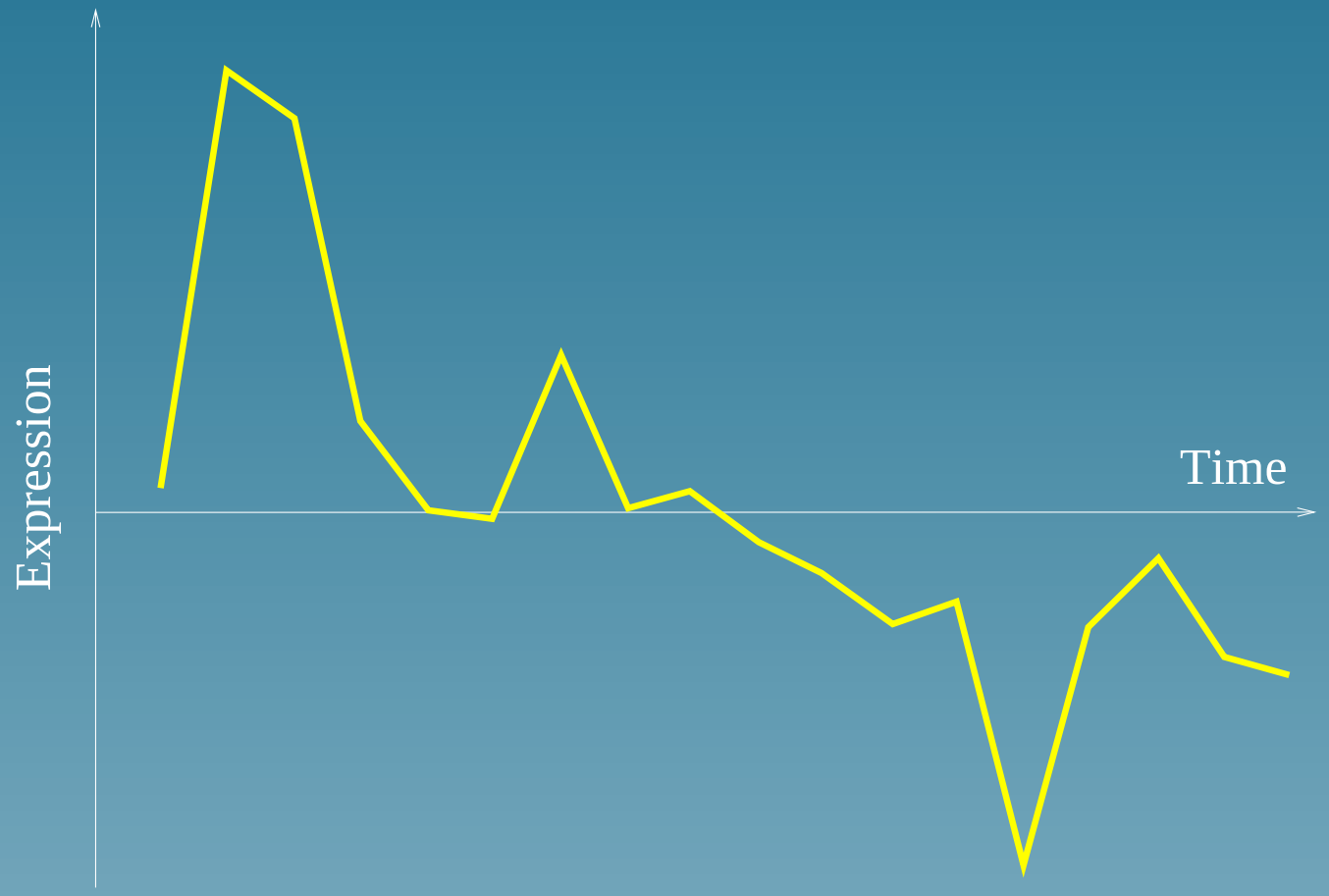
Summary


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

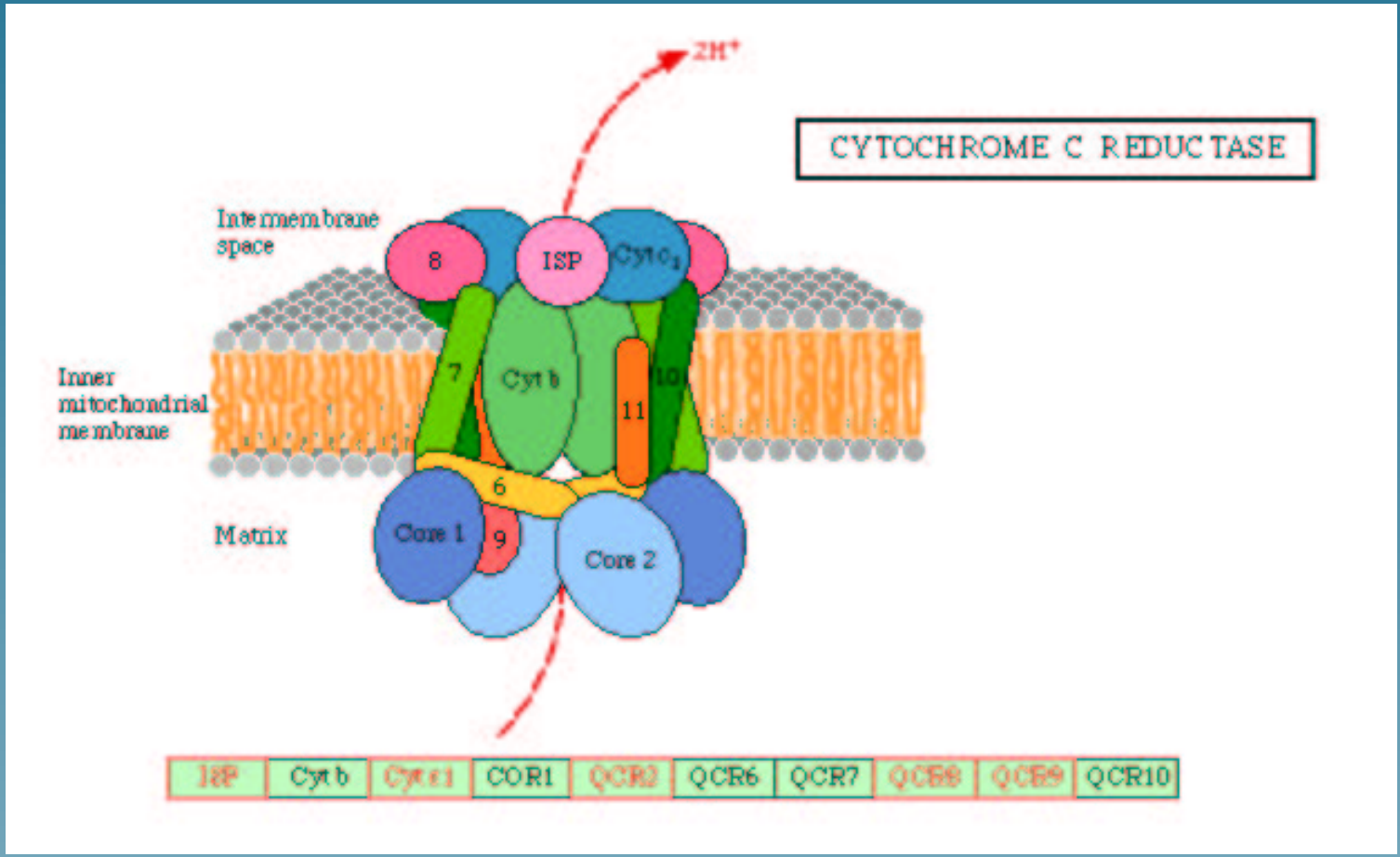


Related metabolic pathways

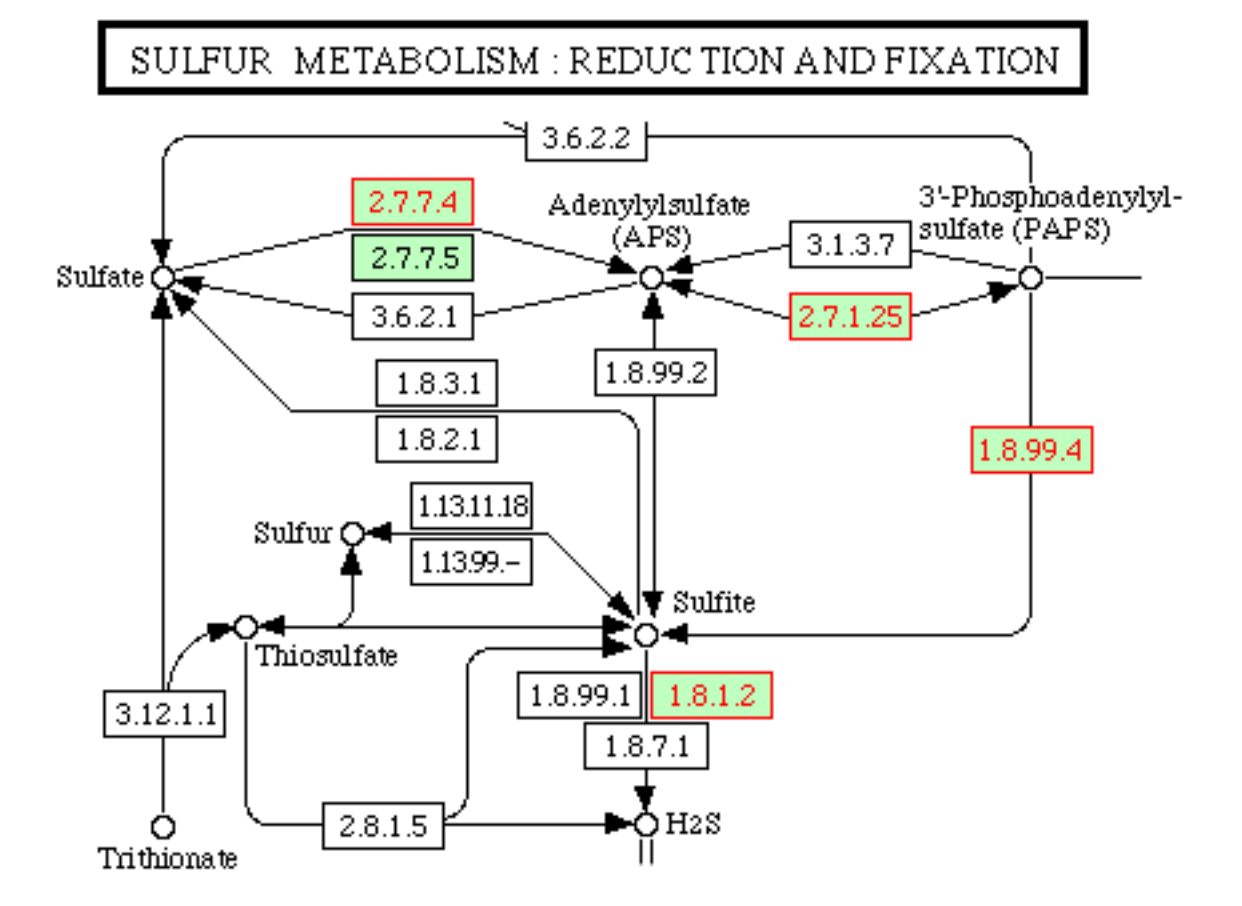
50 genes with highest $s_2 - s_1$ belong to:

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

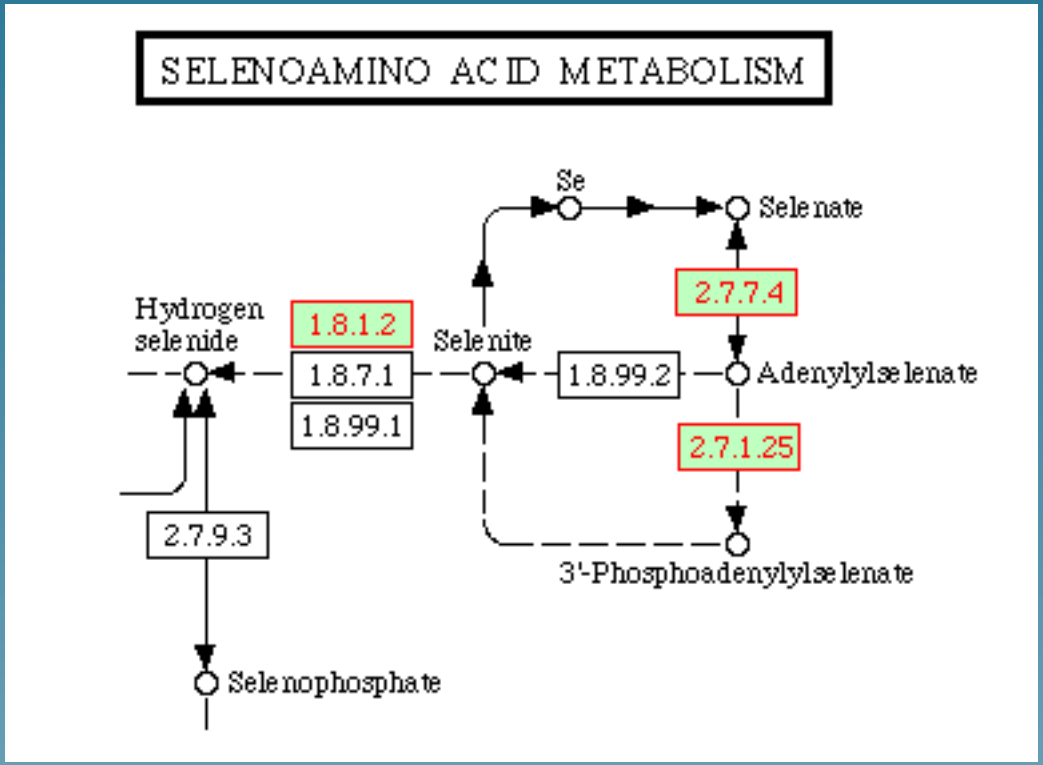
Related genes



Related genes



Related genes



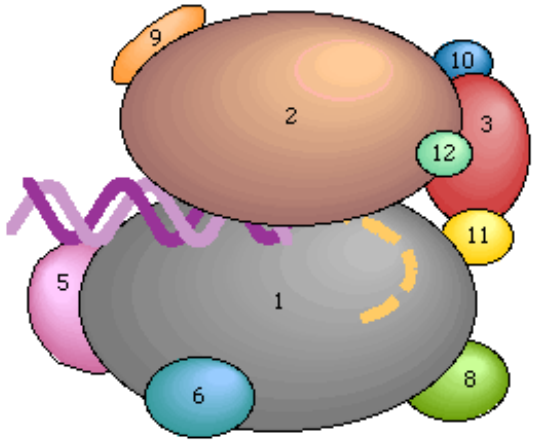
Opposite pattern

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

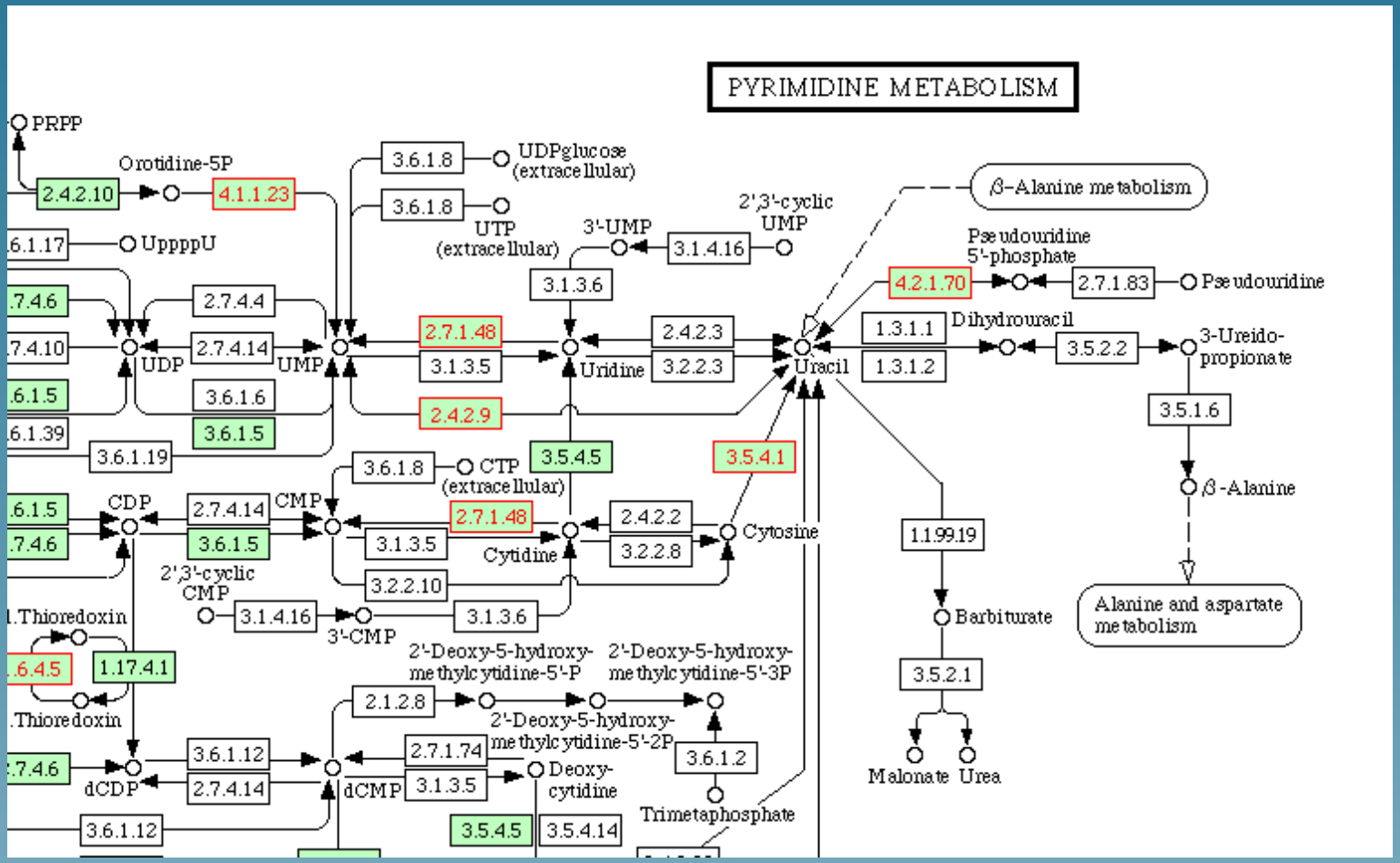
Eukaryotic Pol III

C2	C3	C4	C5	C11
C1	C19	C25	C31	C34

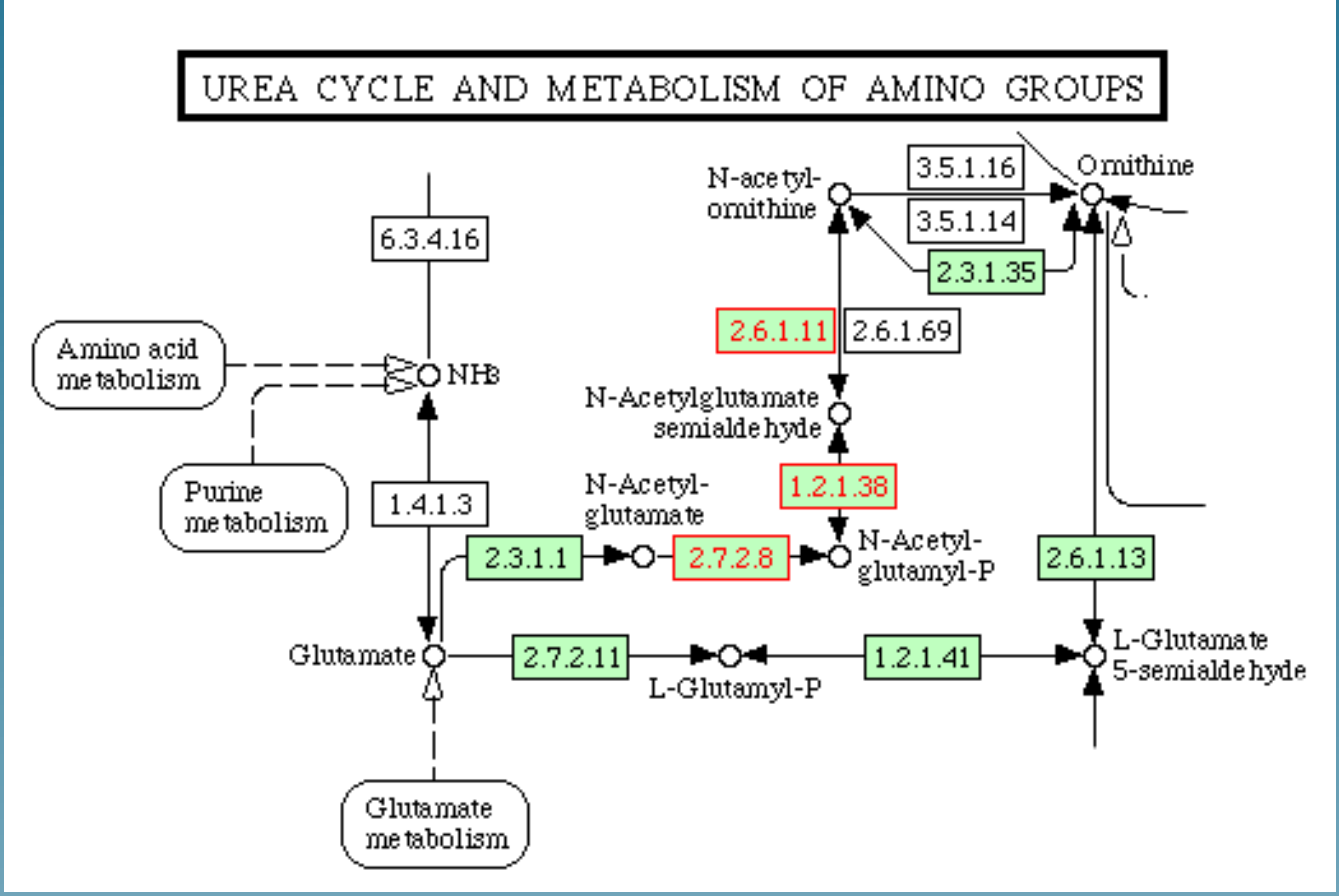
Eukaryotic Pol I

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
- Encouraging results on real-world applications
- Many opportunities in developping kernels for particular applications