

# Méthodes à noyau en bioinformatique

Jean-Philippe Vert

Groupe ‘ ‘Bioinformatique’ ’  
Ecole des Mines de Paris

[Jean-Philippe.Vert@mines.org](mailto:Jean-Philippe.Vert@mines.org)

INRIA Rhône-Alpes, 5 décembre 2002.

# Plan

1. Qu'est-ce qu'un noyau?
2. Que peut-on faire avec un noyau?
3. Quels noyaux pour les gènes?
4. Application: comparaison de réseau protéique et de données d'expression.

## Partie 1

Qu'est-ce qu'un noyau?

## Définition

- Soit un ensemble  $\mathcal{G}$  d'objets à analyser (gènes, composés chimique...)
- Un noyau est une “fonction de similarité”  $K(x, y)$ , définie pour tous objets  $x$  et  $y \in \mathcal{G}$
- Contrainte technique: la fonction  $K(., .)$  doit être:
  - ★ symétrique :  $K(x, y) = K(y, x)$ ,
  - ★ semidéfinie positive:  $\sum_{i,j} a_i a_j K(x_i, x_j) \geq 0$  pour tous  $a_i \in \mathbb{R}$  et  $x_i \in \mathcal{G}$

# Un noyau simple pour vecteurs

- Si les objets à analyser sont des vecteurs:

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

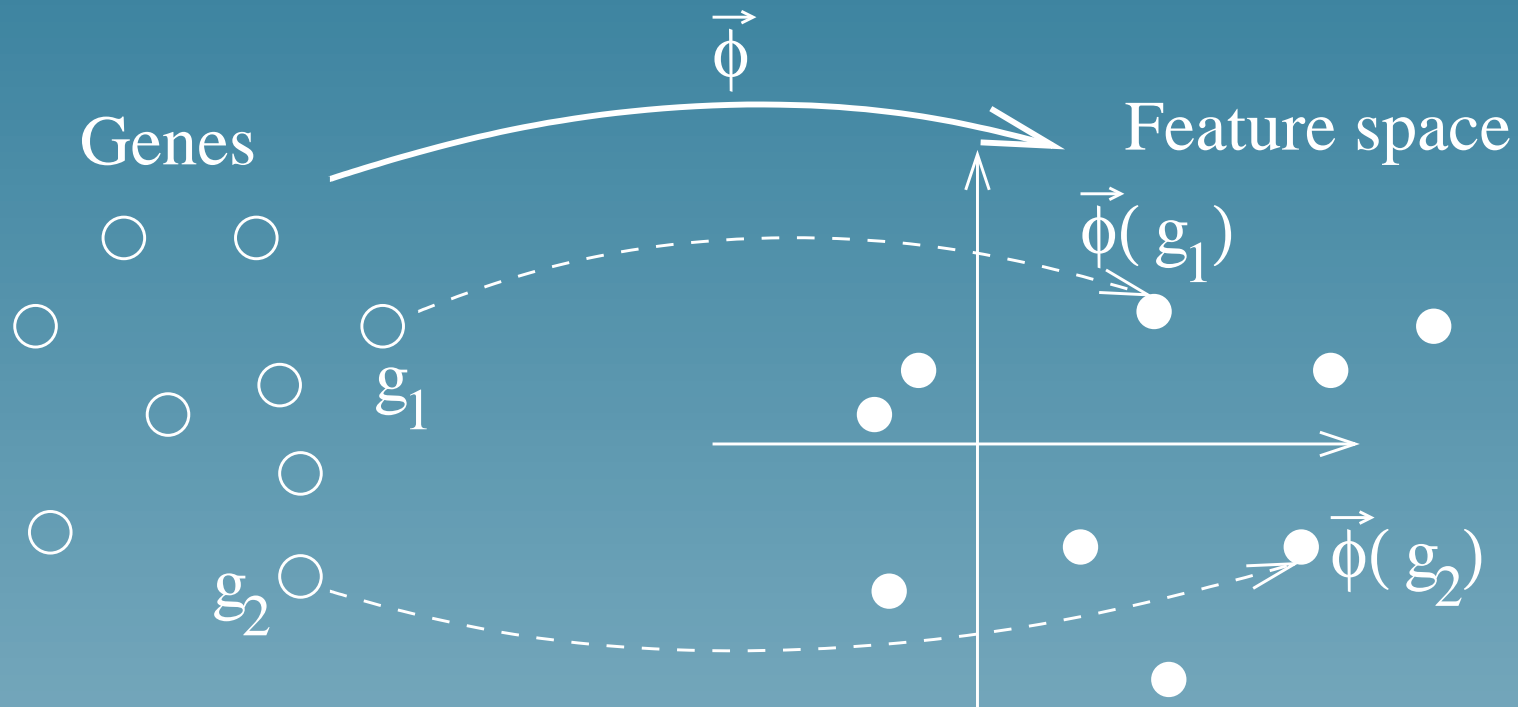
- En effet:

- ★  $\vec{x} \cdot \vec{y} = \vec{y} \cdot \vec{x}$

- ★  $\sum_{i,j} a_i a_j \vec{x}_i \cdot \vec{x}_j = \|\sum_i a_i \vec{x}_i\|^2 \geq 0$

# Comment faire des noyaux en général

$$K(g_i, g_j) \stackrel{def}{=} \vec{\Phi}(g_i) \cdot \vec{\Phi}(g_j)$$



## Exemple: Un noyau simple pour gène

- Représentons chaque gène  $g_1, g_2, \dots$  par un vecteur  $\vec{\Phi}(g_i)$  représentant sa composition en A,T,C,G:

$$\vec{\Phi}(g_1) = \begin{pmatrix} 0.2 \\ 0.3 \\ 0.4 \\ 0.1 \end{pmatrix}, \vec{\Phi}(g_2) = \begin{pmatrix} 0.1 \\ 0.7 \\ 0.1 \\ 0.1 \end{pmatrix}, \dots$$

- Un noyau pour gènes est la fonction  $K(g_i, g_j) = \vec{\Phi}(g_i) \cdot \vec{\Phi}(g_j)$ :

$$K(g_1, g_2) = 0.2 \times 0.1 + 0.3 \times 0.7 + 0.4 \times 0.1 + 0.1 \times 0.1 = 0.28$$

## Noyaux plus généraux

- Une fonction de similitude  $K(x, y)$  est un noyau si la matrice suivante est symétrique semi-définie positive (toutes ses valeurs propres sont positives):

$$K = \begin{pmatrix} K(x_1, x_1) & K(x_1, x_2) & \dots \\ K(x_2, x_1) & K(x_2, x_2) & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

- On peut **vérifier au cas par cas** si une fonction  $K(x, y)$  est un noyau
- Ex: le score d'alignement entre deux séquences calculé par **Smith-Waterman** est un noyau (en général)



## Géométrie implicitement définie

- Si  $K(x, y)$  est un noyau (ex: score de SW), alors on peut montrer qu'il existe une transformation  $\vec{\Phi}(x)$  telle que:

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

- La donnée de  $K(x, y)$  définit donc une structure d'espace vectoriel sur l'espace  $\mathcal{G}$ , de manière implicite.
- Ex: l'algorithme Smith-Waterman définit implicitement une géométrie Euclidienne sur l'espace des séquences biologiques.

## Partie 2

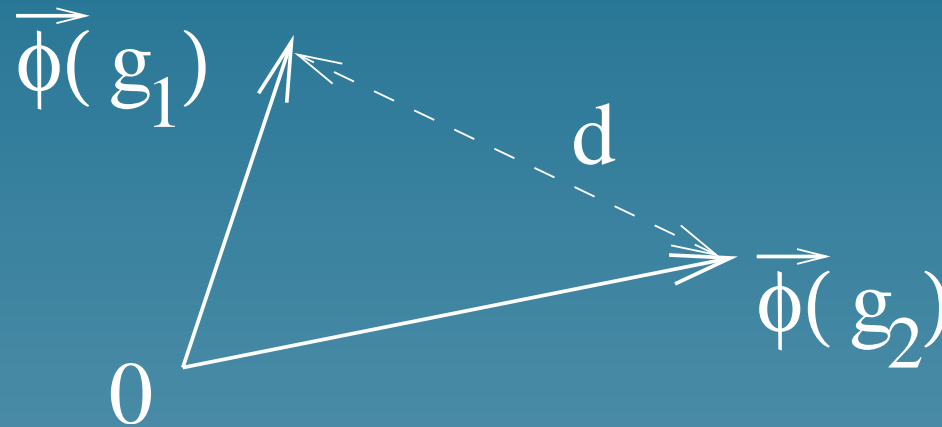
Que peut-on faire avec un  
noyau?

## Résumé

Supposons qu'un noyau  $K(x, y)$  est donné. Alors il est possible de travailler dans l'espace vectoriel associé, **sans jamais calculer l'image  $\vec{\Phi}(x)$  d'un point**. Il est par exemple possible de:

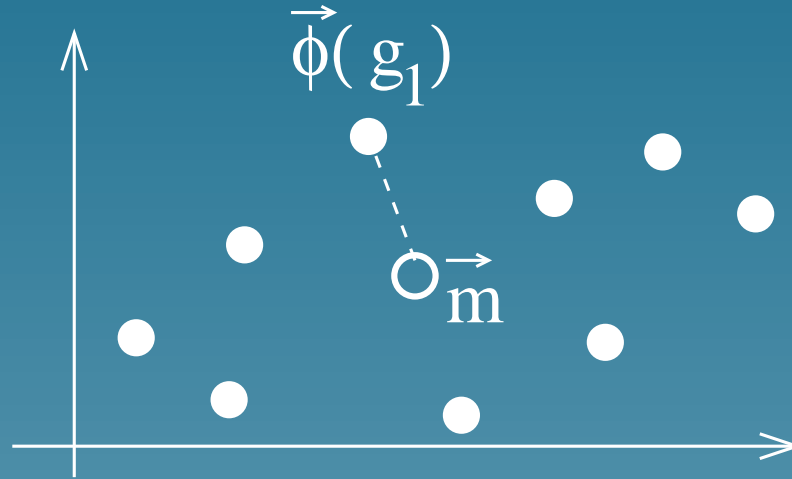
- Calculer la distance entre deux objets, ou entre un objet et le centre de masse d'un ensemble d'objets
- Effectuer une analyse en composantes principales (ACP)
- Effectuer une analyse de corrélation canonique (ACC)
- Classer les objets dans des catégories (Support vector machines)

## Distance entre deux objets



$$\begin{aligned}
 d(g_1, g_2)^2 &= \|\vec{\Phi}(g_1) - \vec{\Phi}(g_2)\|^2 \\
 &= \left( \vec{\Phi}(g_1) - \vec{\Phi}(g_2) \right) \cdot \left( \vec{\Phi}(g_1) - \vec{\Phi}(g_2) \right) \\
 &= \vec{\Phi}(g_1) \cdot \vec{\Phi}(g_1) + \vec{\Phi}(g_2) \cdot \vec{\Phi}(g_2) - 2\vec{\Phi}(g_1) \cdot \vec{\Phi}(g_2) \\
 d(g_1, g_2)^2 &= K(g_1, g_1) + K(g_2, g_2) - 2K(g_1, g_2)
 \end{aligned}$$

## Distance entre un objet et le centre de masse



Center of mass:  $\vec{m} = \frac{1}{N} \sum_{i=1}^N \vec{\Phi}(g_i)$ , hence:

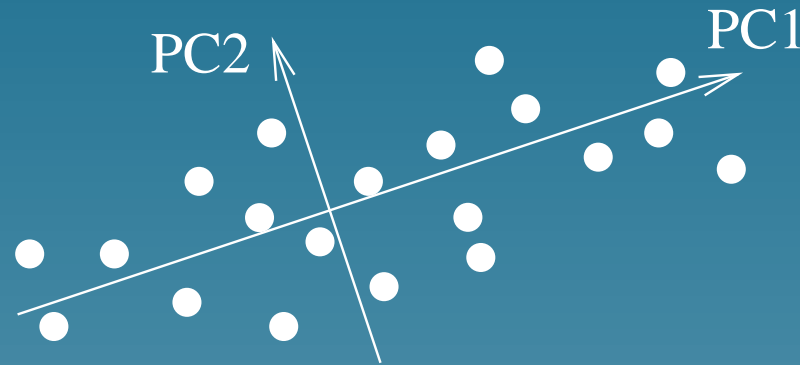
$$\|\vec{\Phi}(g_1) - \vec{m}\|^2 = \vec{\Phi}(g_1) \cdot \vec{\Phi}(g_1) - 2\vec{\Phi}(g_1) \cdot \vec{m} + \vec{m} \cdot \vec{m}$$

$$= K(g_1, g_1) - \frac{2}{N} \sum_{i=1}^N K(g_1, g_i) + \frac{1}{N^2} \sum_{i,j=1}^N K(g_i, g_j)$$

## Exemple: greedy multiple alignment (Gorodkin et al., GIW 2001)

- Utilise le score SW comme noyau pour séquences protéiques
- Calcule la distance entre chaque séquence et le centre de masse
- Commence par aligner les séquences proches du centre de masse
- Ajoute ensuite les séquences une par une à l'alignement multiple, par ordre de proximité au centre de masse

# Analyse en composantes principales

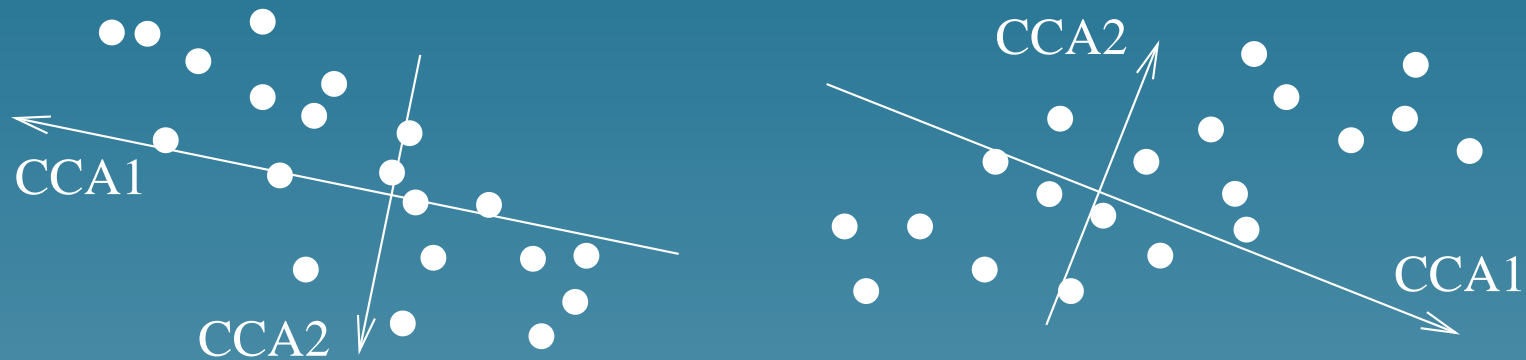


Il suffit de trouver les vecteurs propres de la matrice:

$$\begin{aligned} K &= \left( \vec{\Phi}(g_i) \cdot \vec{\Phi}(g_j) \right)_{i,j=1\dots N} \\ &= \left( K(g_i, g_j) \right)_{i,j=1\dots N} \end{aligned}$$

Utile pour projeter les objets sur un espace de petite dimension (feature extraction).

## Analyse de corrélation canonique (ACC)



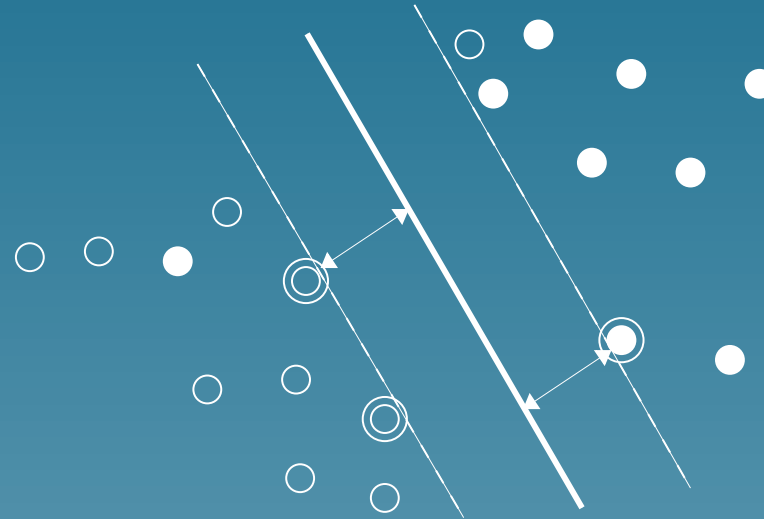
$K_1$  et  $K_2$  sont deux noyaux différents pour les mêmes objets. L'ACC est obtenue en résolvant le problème de valeurs propres généralisé suivant:

$$\begin{pmatrix} 0 & K_1 K_2 \\ K_2 K_1 & 0 \end{pmatrix} \vec{\xi} = \rho \begin{pmatrix} K_1^2 & 0 \\ 0 & K_2^2 \end{pmatrix} \vec{\xi}$$

Utile pour trouver des corrélations entre différentes représentations d'un même type d'objet (ex: genes, ...)



# Classification: support vector machines (SVM)



Trouve une frontière linéaire avec une grande marge en résolvant:

$$\begin{cases} \max_{\vec{\alpha}} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(g_i, g_j) \\ \forall i = 1, \dots, n \quad 0 \leq \alpha_i \leq C \\ \sum_{i=1}^n \alpha_i y_i = 0 \end{cases}$$

## Exemples: SVM en bioinformatique

- Gene functional classification from microarray: Brown et al. (2000), Pavlidis et al. (2001)
- Tissue classification from microarray: Mukherje et al. (1999), Furey et al. (2000), Guyon et al. (2001)
- Protein family prediction from sequence: Jaakkola et al. (1998)
- Protein secondary structure prediction: Hua et al. (2001)
- Protein subcellular localization prediction from sequence: Hua et al. (2001)

## Résumé

- Une fois qu'un noyau  $K(x, y)$  est donné, beaucoup d'opérations peuvent être effectuées de manière implicite dans l'espace vectoriel associé
- Ces méthodes (ex: SVM) sont **extrêmement performantes** sur des problèmes réels
- Modularité: **chaque noyau fonctionne avec chaque méthode**

## Part 3

Quels noyaux pour les gènes?

## Qu'est-ce qu'un gène

- une séquence d'ADN?
- une structure primaire, secondaire ou 3D de protéine?
- un profil d'expression?
- un noeud dans un réseau d'interaction ou de régulation?
- une séquence promotrice?
- un profile phylogénétique?
- ...

# Noyaux pour séquences protéiques

- spectrum kernel (Eskin et al., 2002)
- Fisher kernel (Jaakkola et al., 1999)
- Smith-Waterman score (Vert et al., submitted)
- Résultats très encourageants pour la détection d'homologie lointaine ou la prédiction de fonction et de localisation

## Noyaux pour profils d'expression

Un profil d'expression est un vecteur  $\vec{\Phi}(x)$ , pour lequel il existe de nombreux noyaux "classiques" :

- Linéaire:  $K(x, y) = \vec{\Phi}(x) \cdot \vec{\Phi}(y)$ .
- Polynomial:  $K(x, y) = \left( \vec{\Phi}(x) \cdot \vec{\Phi}(y) + 1 \right)^d$ .
- Gaussien:  $K(x, y) = \exp \left( -\frac{\|\vec{\Phi}(x) - \vec{\Phi}(y)\|^2}{2\sigma^2} \right)$ .

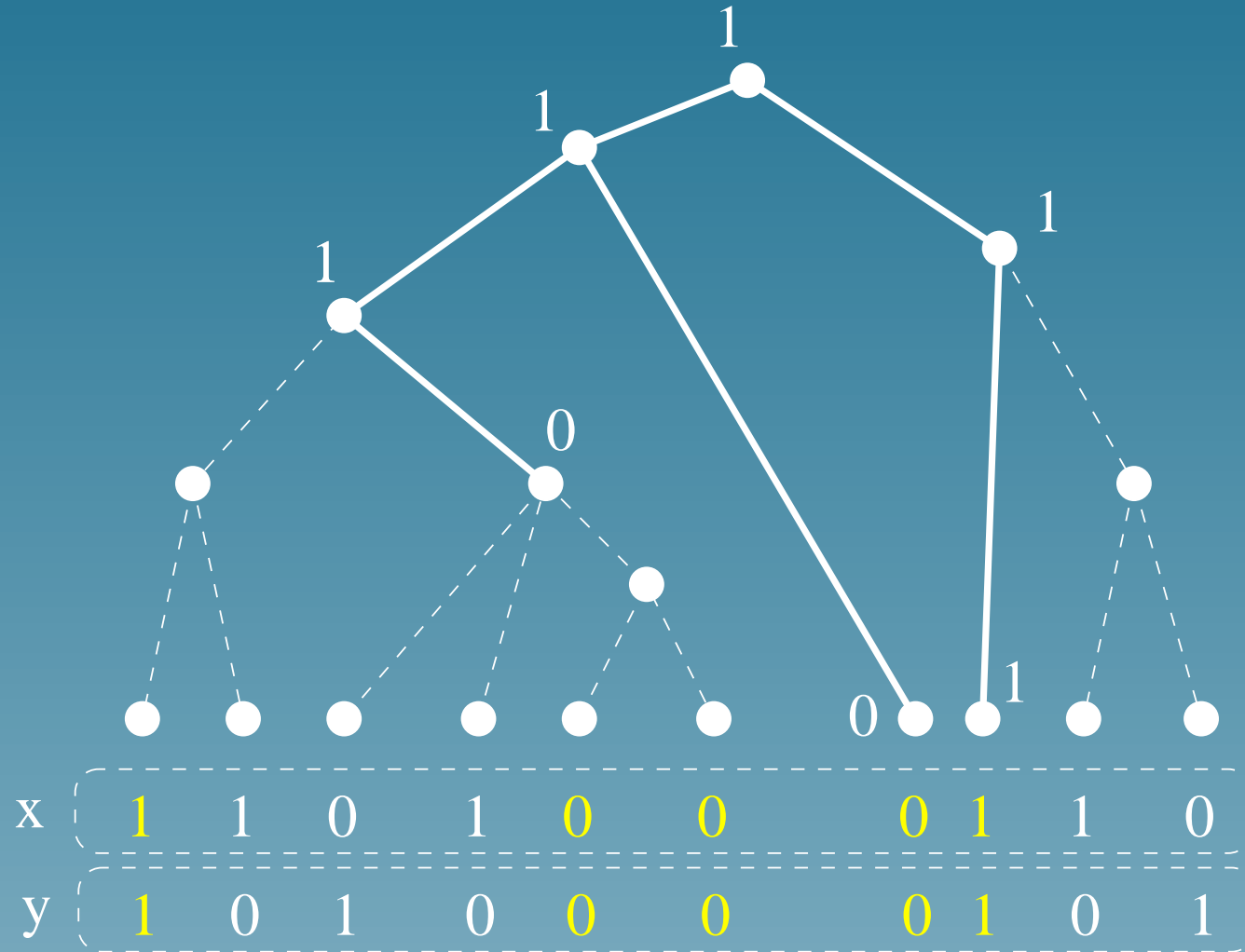
## Noyaux pour profils phylogénétiques (ISMB02)

- **Profile phylogénétique:** une suite de bits (0/1) qui indiquent la présence ou l'absence d'un homologue dans chaque génome séquencé.
- **Intuition:** deux gènes sont similaires si ils partagent des “patterns” d'évolution
- **Solution:** définir un modèle probabiliste simple de transmission de gènes entre espèces au cours de l'évolution, et:

$$K(x, y) = \sum_{e \text{ evolution pattern}} p(e)p(x|e)p(y|e)$$



# Evolution patterns



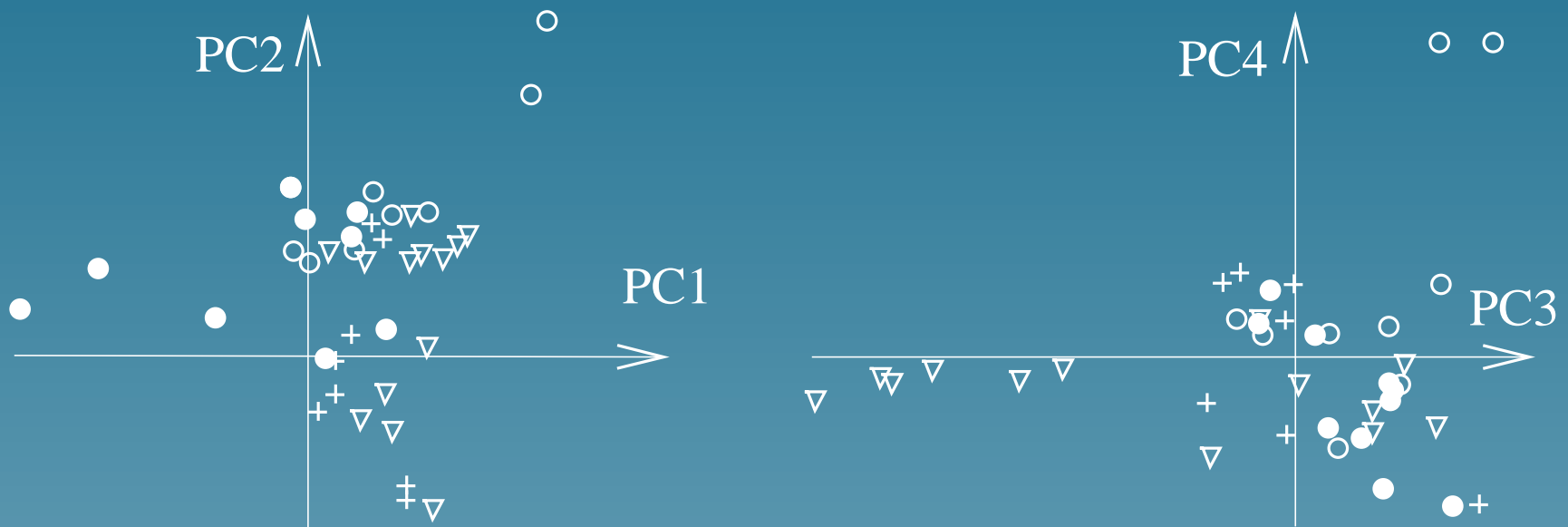
## Propriétés de $K(x, y)$

- $K(., .)$  est un noyau
- Deux gènes sont similaires quand ils ont partagé des “pattern” d'évolution avec une grande probabilité
- Implémentation: pour des profiles de longueur  $n$ ,  $K(x, y)$  se calcule une complexité linéaire en  $O(n)$  (bien qu'il y ait un nombre exponentiel de patterns à additionner)

## Application: Prédiction de fonction à partir des profiles phylogénétiques (indice $ROC_{50}$ )

Functional class	Dot kernel	Tree kernel	Difference
Amino-acid transporters	0.74	0.81	+ 9%
Fermentation	0.68	0.73	+ 7%
ABC transporters	0.64	0.87	+ 36%
C-compound transport	0.59	0.68	+ 15%
Amino-acid biosynthesis	0.37	0.46	+ 24%
Amino-acid metabolism	0.35	0.32	- 9%
Tricarboxylic-acid pathway	0.33	0.48	+ 45%
Transport Facilitation	0.33	0.28	- 15%

# Application: kernel PCA of phylogenetic profiles



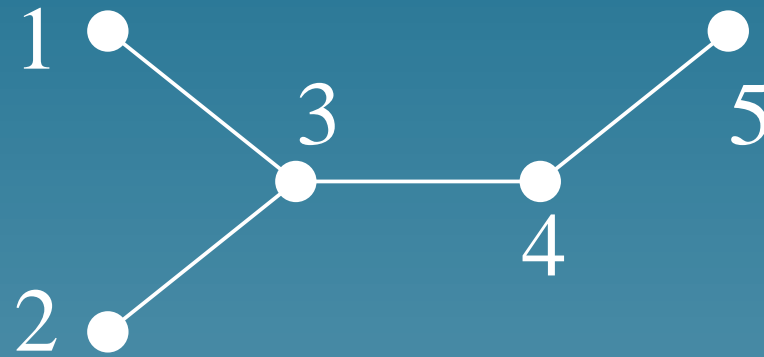
- Amino-acid transporters
- Fermentation
- ▽ ABC transporters
- + C-compound, carbohydrate transport

## Noyau pour les noeuds d'un graphe (Kandor, 2001)

- **But:** Supposons que l'on puisse définir des relations binaires entre gènes (ex: interaction des protéines). Comment construire un noyau qui représente la topologie du graphe ainsi construit?
- **Intuition :** Deux gènes doivent être d'autant plus similaires qu'il y a de nombreux chemins courts entre eux dans le graphe.
- **Solution** Soit  $L$  la matrice Laplacienne du graphe. Pour chaque  $\lambda > 0$ , on obtient un noyau par:

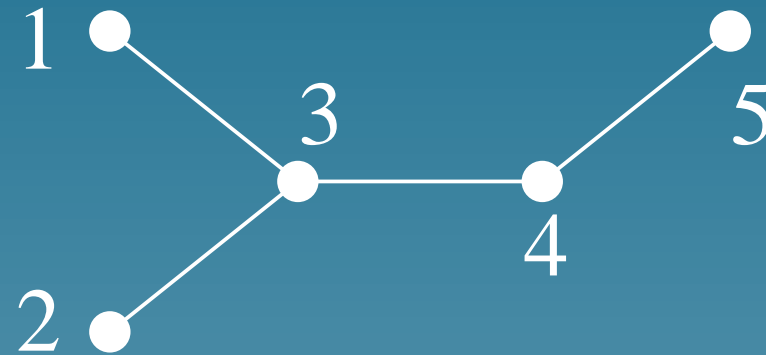
$$K = \exp(-\lambda L)$$

## Exemple d'un graphe (1)



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

## Exemple d'un graphe (2)



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

## Résumé

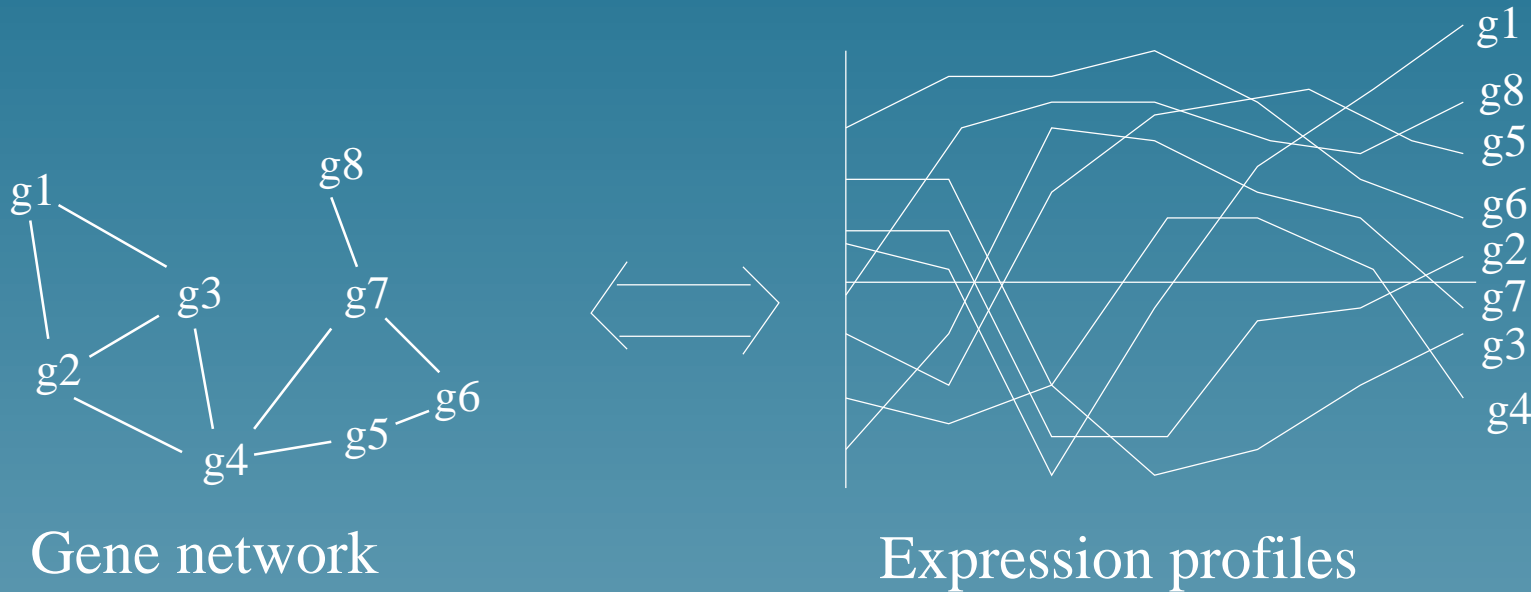
- Des noyaux peuvent être construits à partir de **connaissances biologiques a priori**.
- La connaissance biologique est une information sur “quand deux gènes doivent être considérés comme similaires”
- De nombreux noyaux ont été imaginés récemment



## Part 4

Application: comparer un réseau  
de protéines et des données  
d'expression

# Le problème

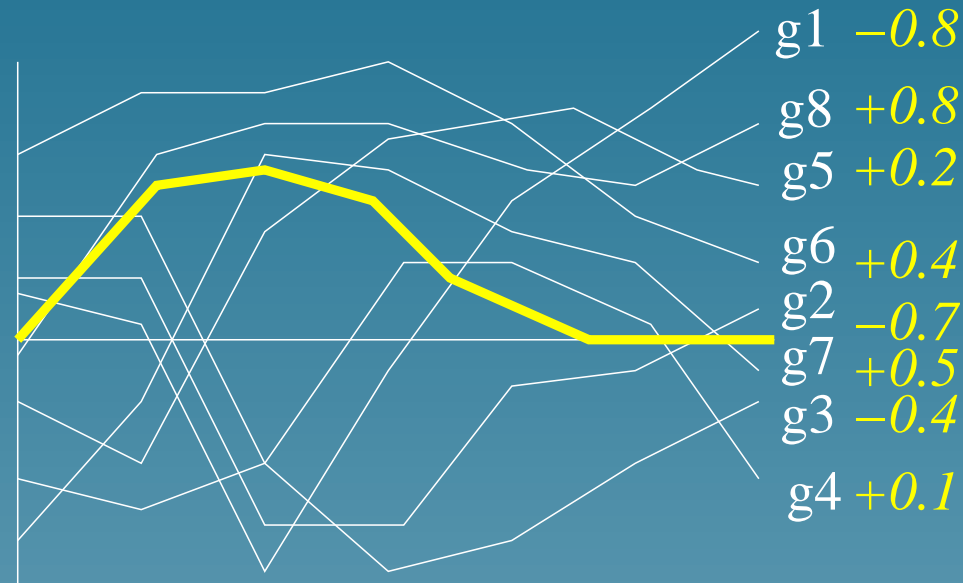


Y a-t-il des "corrélations" ?

## Qu'est-ce qu'une corrélation?

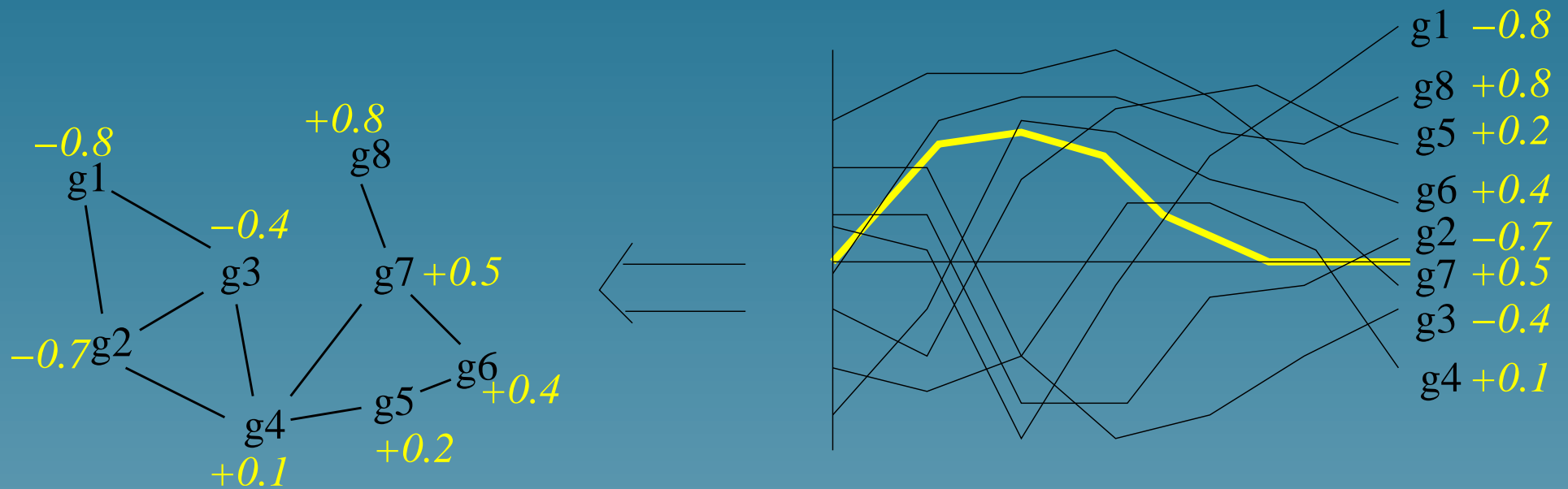
- Une “tendance” d'expression partagée par des gènes proches les uns des autres dans le réseau
- Exemples:
  - ★ **Activation d'une voie métabolique:** les enzymes qui catalyzent des réactions successives ont une expression qui est liée à l'activation de la voie
  - ★ **Formation d'un complexe de protéines:** Les protéines qui le composent doivent avoir une expression coordonnée

## Tendance d'expression



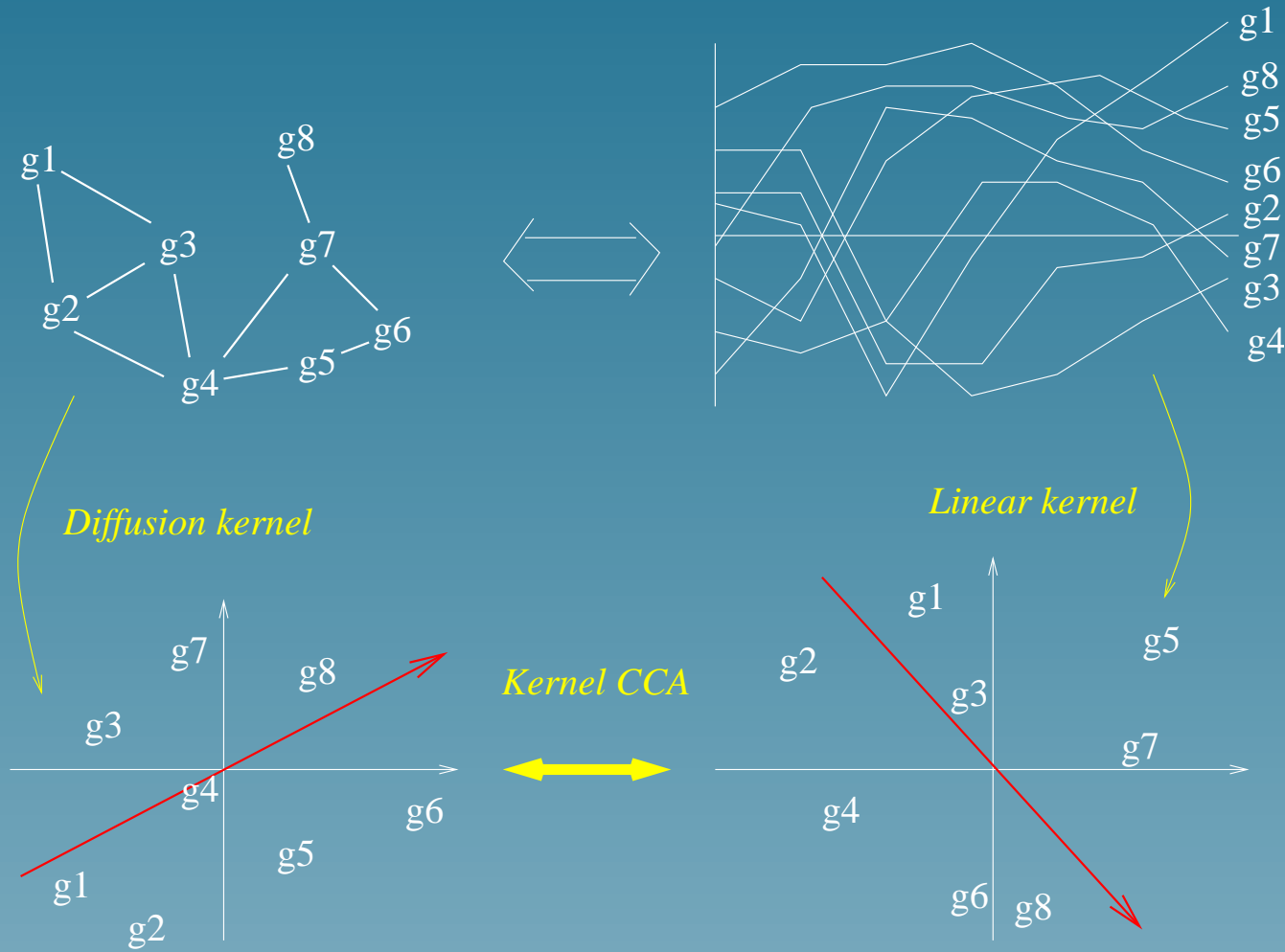
- Une **tendance d'expression** est n'importe quel profile.
- La **correlation** entre une tendance et le profile d'expression d'un gène quantifie comment le gène partage la tendance.

# Régularité d'une tendance sur le graphe



- Les tendances dont la corrélation varie **lentement** sur le graphe sont intéressantes.

# L'idée



# Noyaux utilisés

- Graphe: un noyau de diffusion
- Expression: un noyau linéaire

## Kernel CCA (Bach and Jordan, 2002)

- Soit  $K_1$  le noyau de diffusion et  $K_2$  le noyau linéaire (correspondant à deux espaces Euclidiens différents).
- Les directions de grandes corrélations peuvent être trouvées en résolvant l'équation:

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



## Pourquoi ça marche

On peut montrer que avec ces choix de noyaux, les directions de grandes corrélation dans les espaces Euclidiens associés correspondent à des tendances d'expression qui varient régulièrement sur le graphe!

# Données

- **Réseau de gène**: deux gènes sont liés si ils catalysent deux réactions successives (extrait de la base de données KEGG)
- **Profiles d'expression**: 18 mesures pour tous les genes (6,000) de la levure *S. Cerevisiae* par Spellman et al., correspondant à deux cycles cellulaires.

# Première tendance d'expression

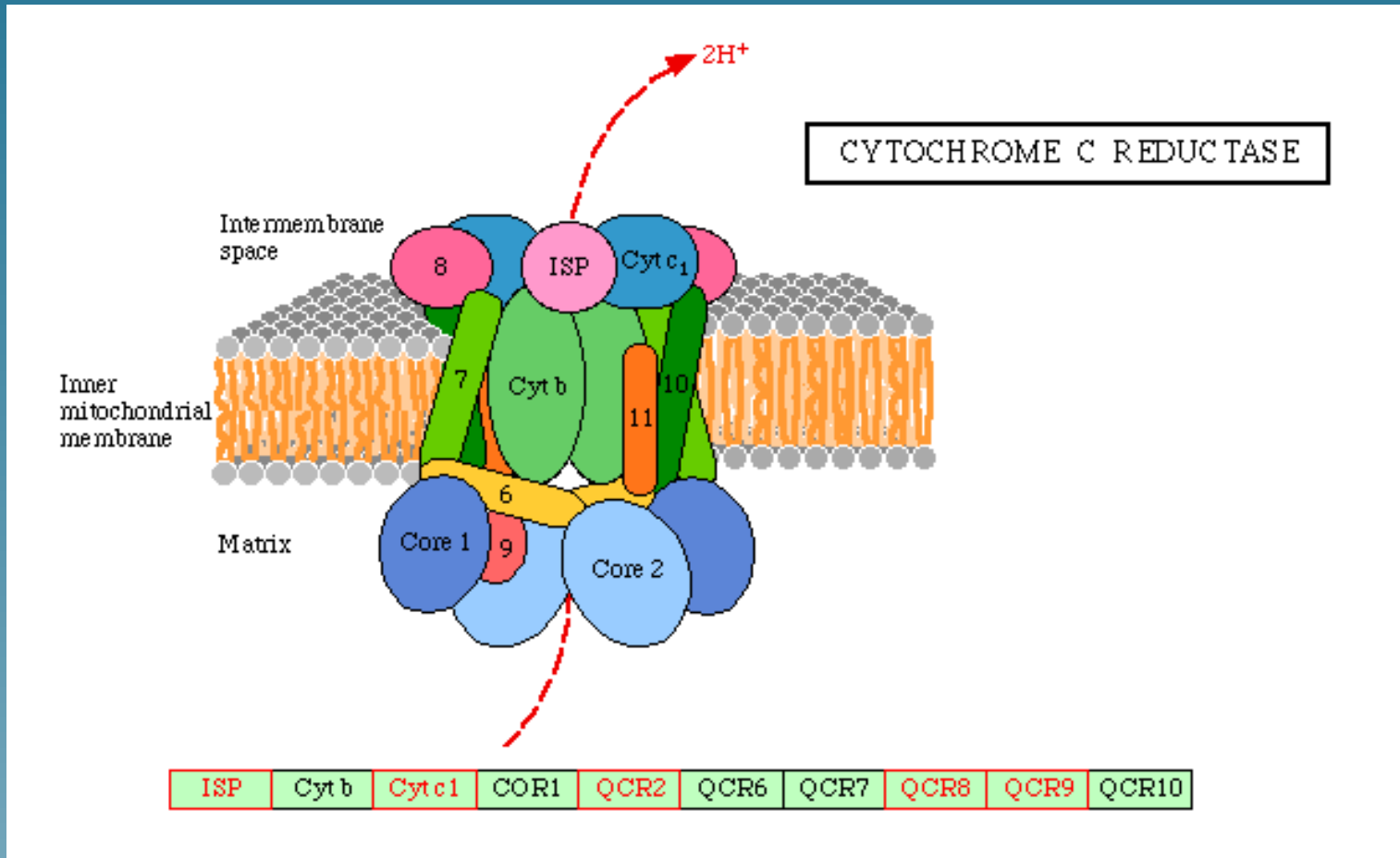


## Voies métaboliques associées

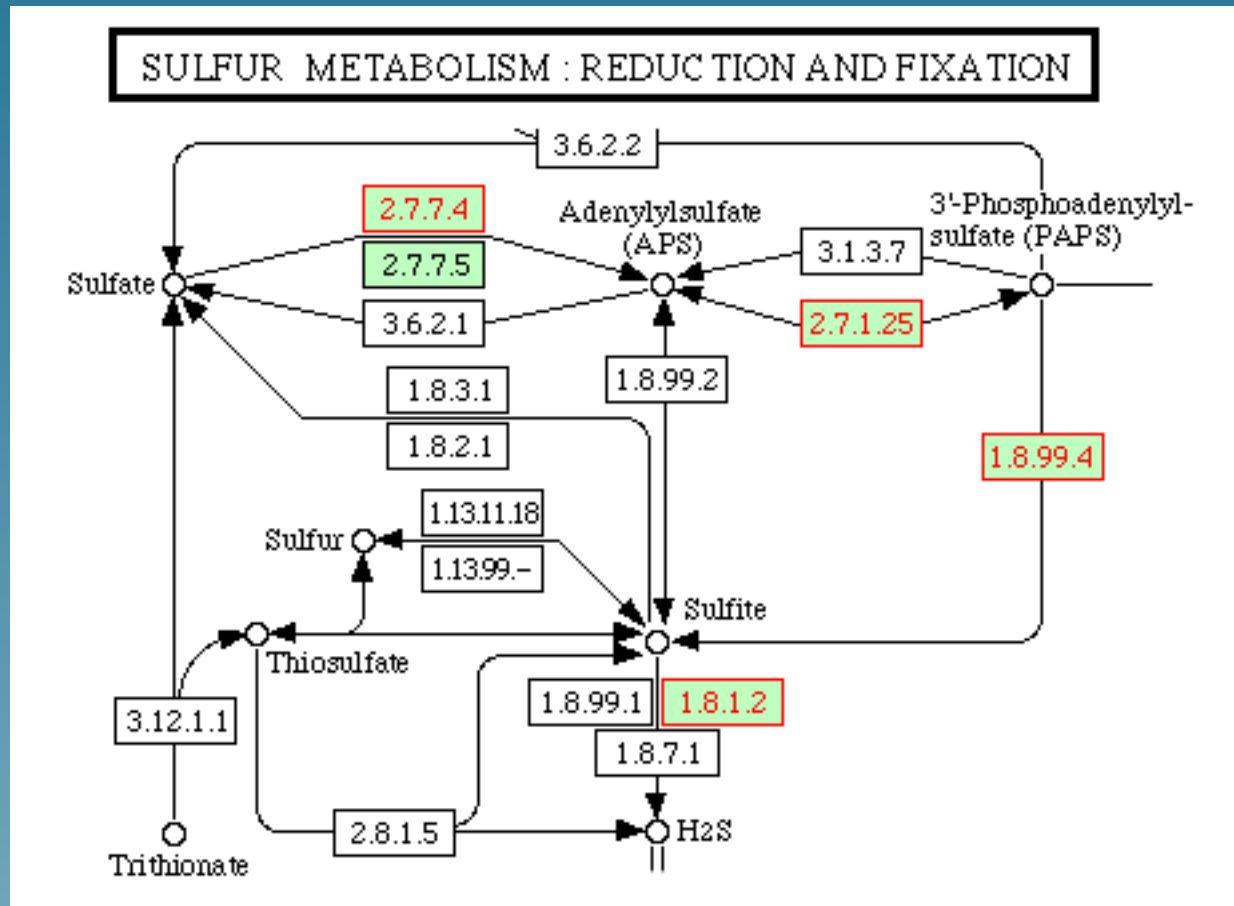
50 genes with highest  $s_2 - s_1$  belong to:

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

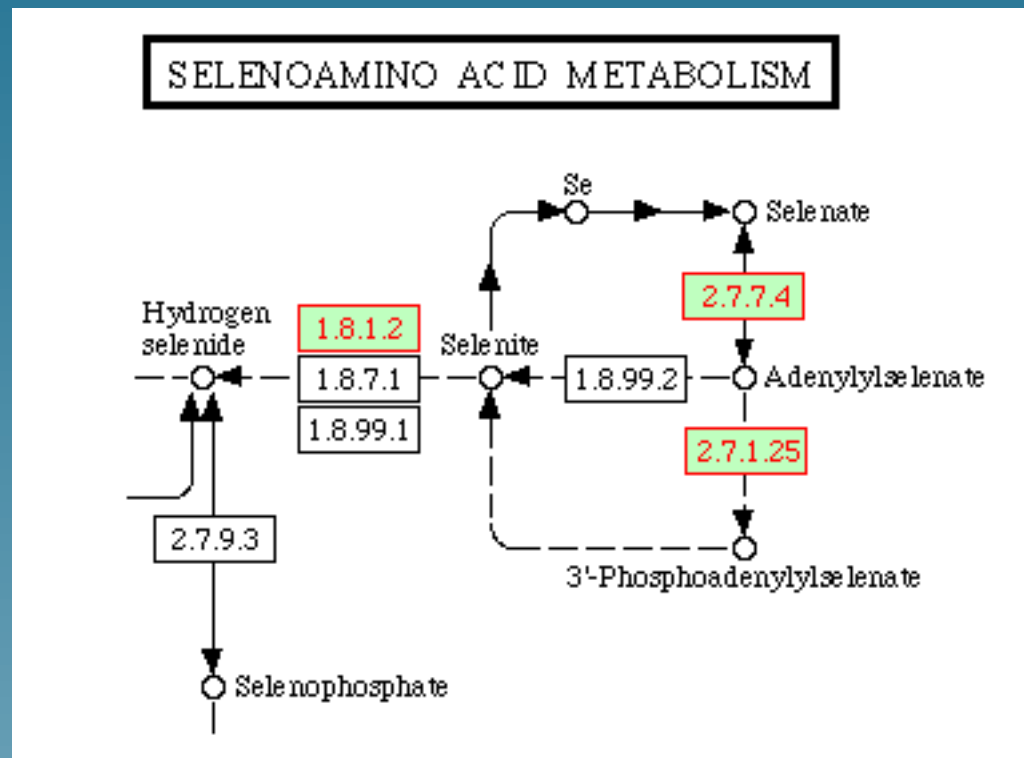
# Gènes associés



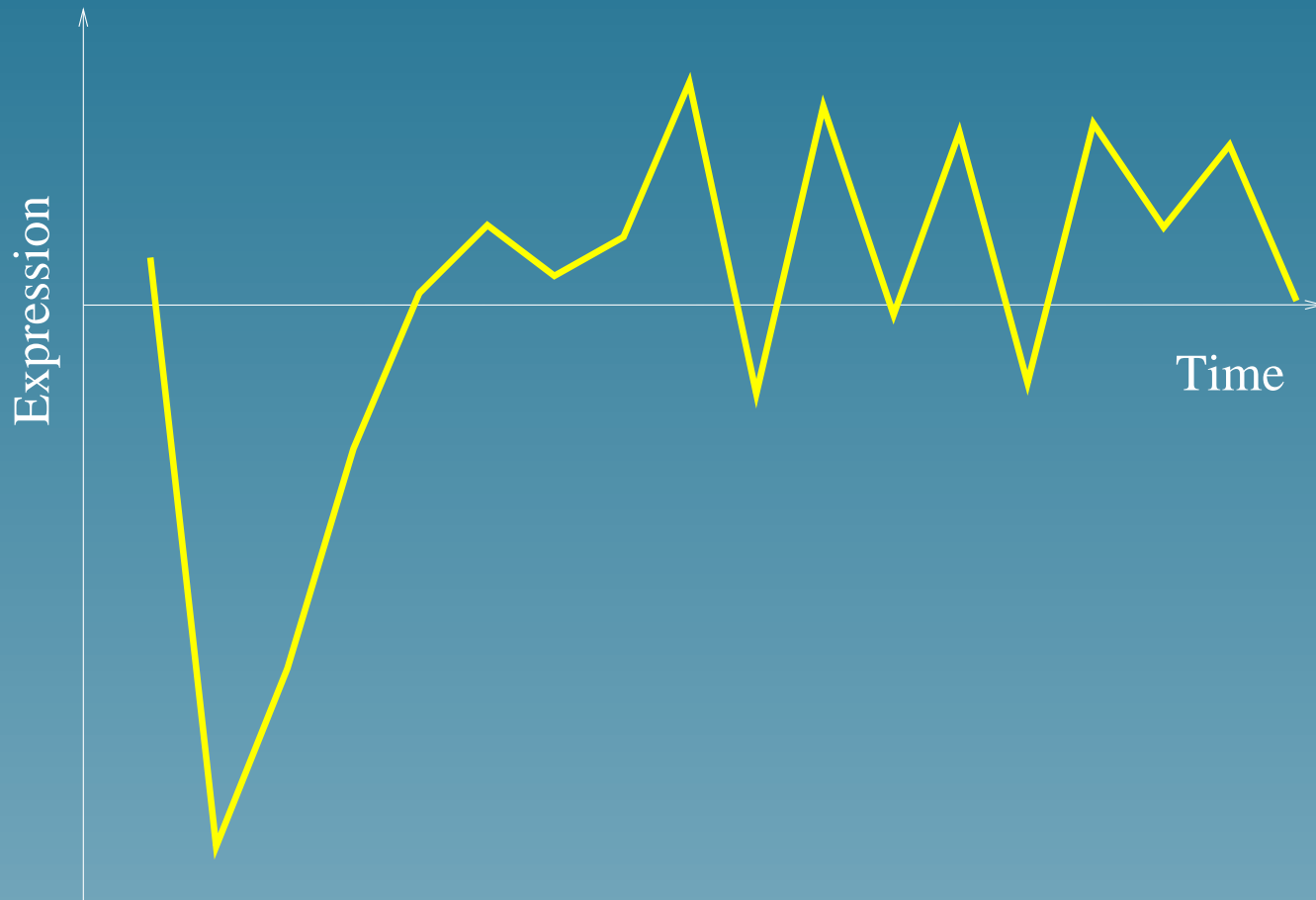
# Gènes associés



# Gènes associés



# Tendance inverse



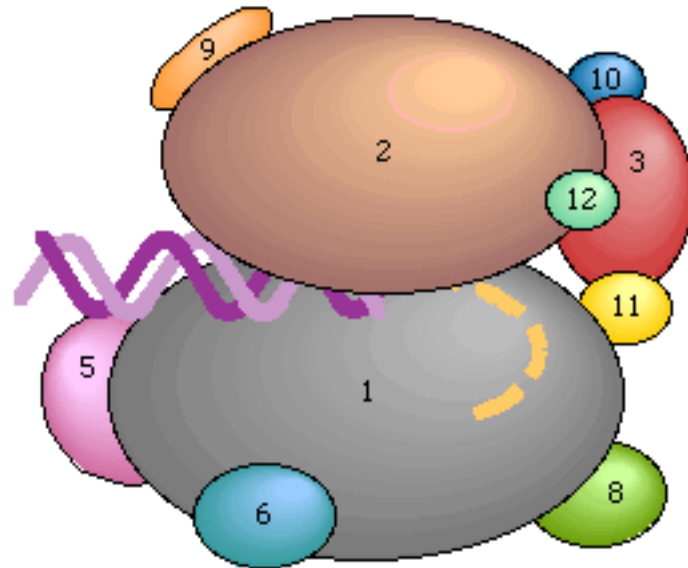


## Gènes associés

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

# Gènes associats

## RNA POLYMERASE



RNA polymerase II (*Saccharomyces cerevisiae*)

### Eukaryotic Pol II

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

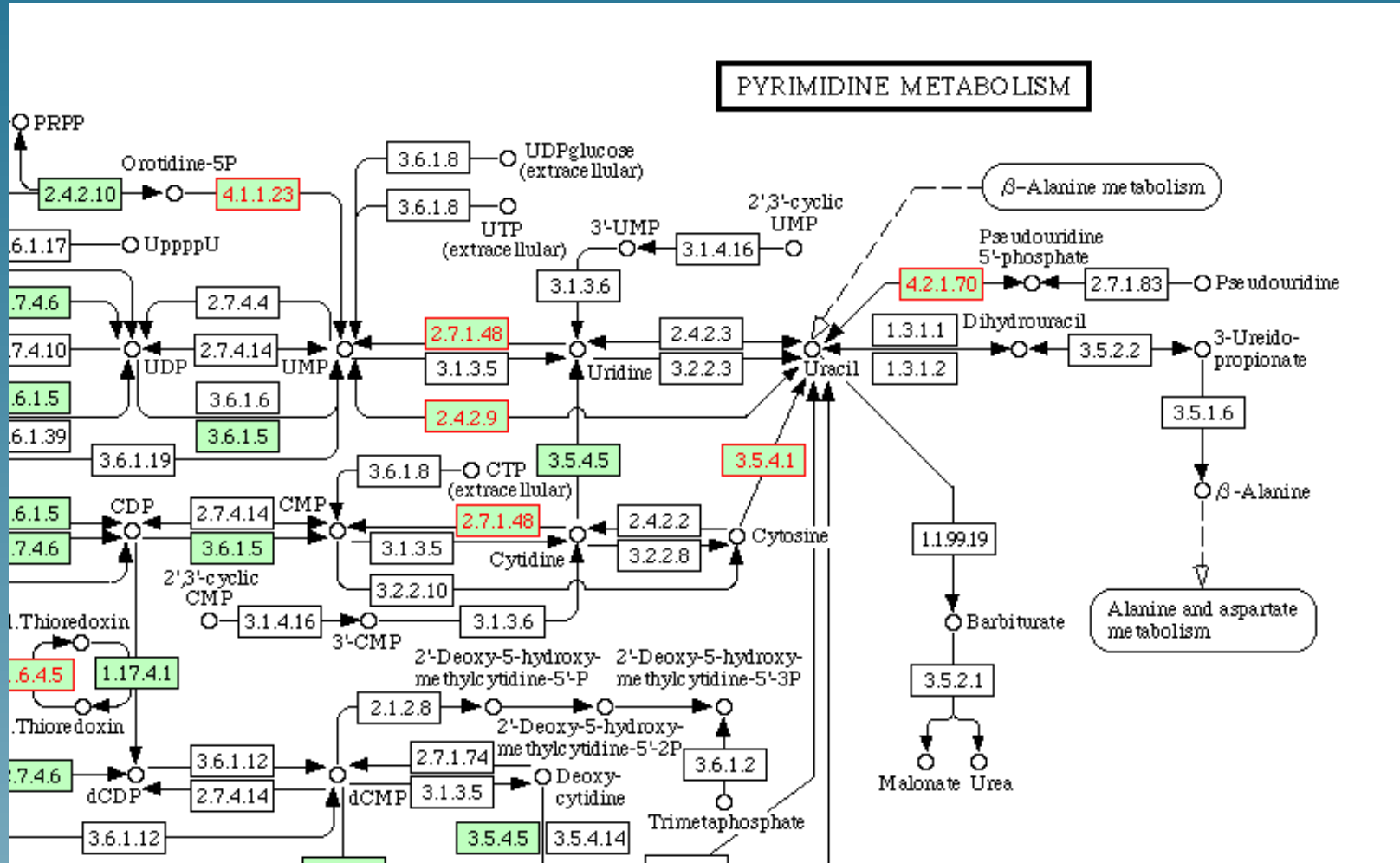
### Eukaryotic Pol III

C2	C3	C4	C5	C11
C1	C19	C25	C31	C34

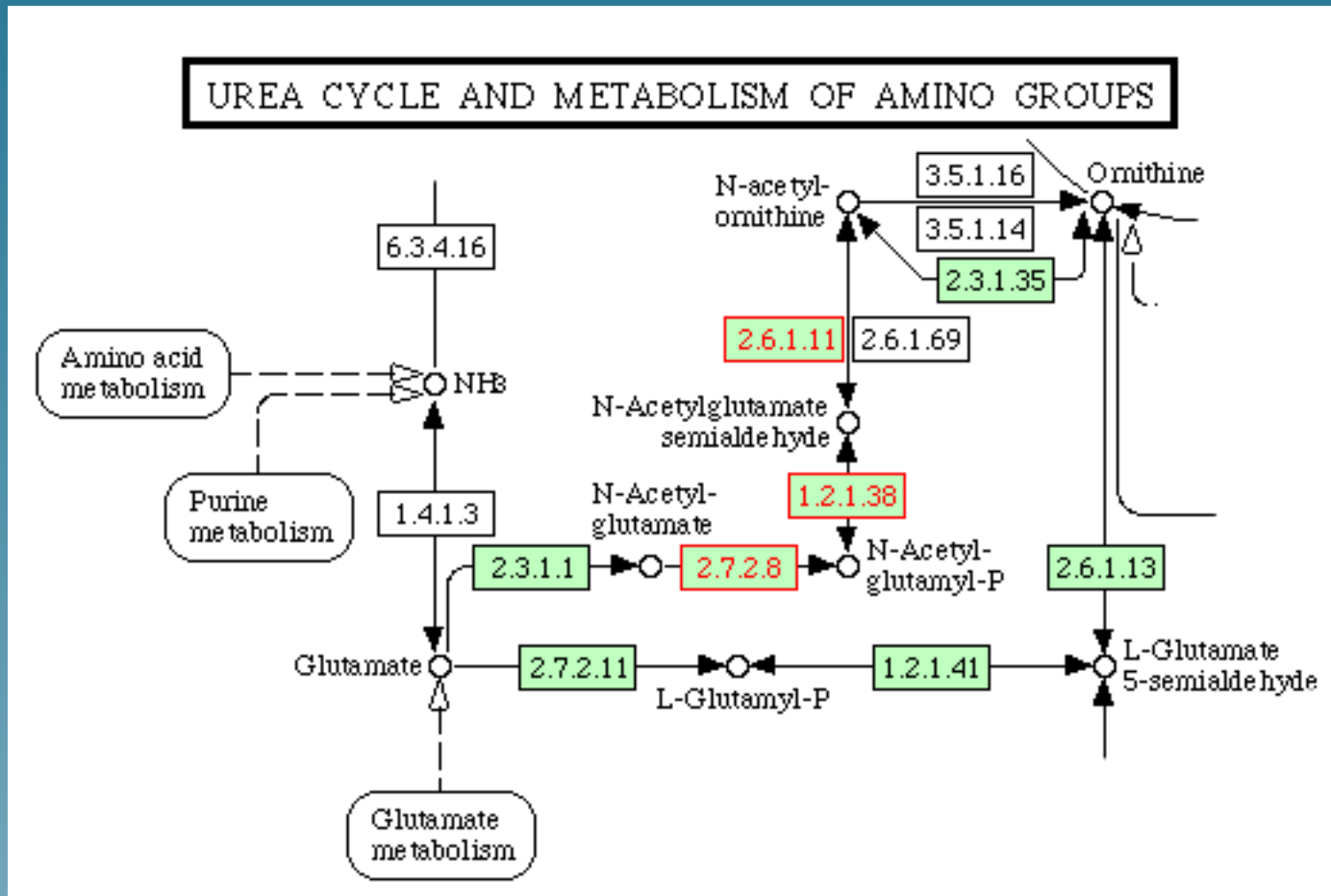
### Eukaryotic Pol I

A2	A12	A14	A34	A43	A49
A1					

# Gènes associés



# Gènes associés



# Conclusion

# Conclusion

- L'approche par noyaux est **flexible**, et permet d'**intégrer des données de nature hétérogènes**.
- Elle permet d'intégrer des **connaissances biologiques** dans la construction de noyaux.
- Elle offre une grande **modularité** : choix du noyau, choix de la méthode.
- Donne de **bons résultats** et ouvre de **nouvelles possibilités d'analyse**.