Inférence sur graphes et groupes

Jean-Philippe Vert

Ecole des Mines de Paris, France
Jean-Philippe.Vert@mines.org

Séminaire statistiques spatiales, INAPG, 23 janvier 2004

Plan

1. Motivations
2. Analyse harmonique et covariances sur les graphes
3. Covariances sur des groupes
4. Application: analyse de voies métaboliques par puces à ADN

Part 1

Motivations

Contexte

- Problème: estimation d'une fonction $f : \mathcal{X} \to \mathbb{R}$ a partir d'observations ponctuelles $f(x_1), \ldots, f(x_n)$ ou $x_i \in \mathcal{X}$
- Ce dont on a besoin: une fonction de covariance $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ qui definit un processus de Gaussien de moyenne nulle par:

$$Ef(x)f(x') = k(x, x').$$

- On peut alors estimer les lois du processus conditionnellement aux observations, krigeage, SVM, etc...

Cas classique

- $\mathcal{X} = \mathbb{R}^p$, avec $(p = 2, 3)$.
- Hypothese de stationarite:

$$k(x, x') = C(x - x').$$

- Parfois, hypothese d'isotropie:

$$k(x, x') = C(||x - x'||).$$

Conditions sur k et C

- k est une fonction de covariance ssi elle est symetrique:

$$\forall x, x' \in \mathcal{X}, \quad k(x, x') = k(x', x),$$

et definie positive:

$$\forall n \in \mathbb{N}, x_1, \ldots, x_n \in \mathcal{X}, a_1, \ldots, a_n \in \mathbb{R}, \quad \sum_{i,j=1}^{n} a_i a_j k(x_i, x_j) \geq 0.$$

Théorème de Bochner

Théorème 1. *Une fonction stationnaire $k(x, x') = C(x - x')$ sur $\mathbb{R}^d$ est symétrique définie positive si et seulement si c'est la transformée de Fourier d'une mesure réelle positive.*

Exemples de covariances stationnaire isotropiques

- Gaussienne:

$$C(x) = e^{-ax^2}, \quad \hat{C}(\omega) = \sqrt{\frac{\pi}{a}} e^{\pi^2 k^2 / a}.$$

- Exponentielle:

$$C(x) = e^{-2\pi\omega_0|x|}, \quad \hat{C}(\omega) = \frac{1}{\pi} \frac{\omega_0}{\omega^2 + \omega_0^2}$$

Notre probleme

- Comment generaliser ces methodes a des espaces $\mathcal{X}$ differents?
- exemple 1: $\mathcal{X}$ est un graphe
- exemple 2: $\mathcal{X}$ est un groupe

Part 2

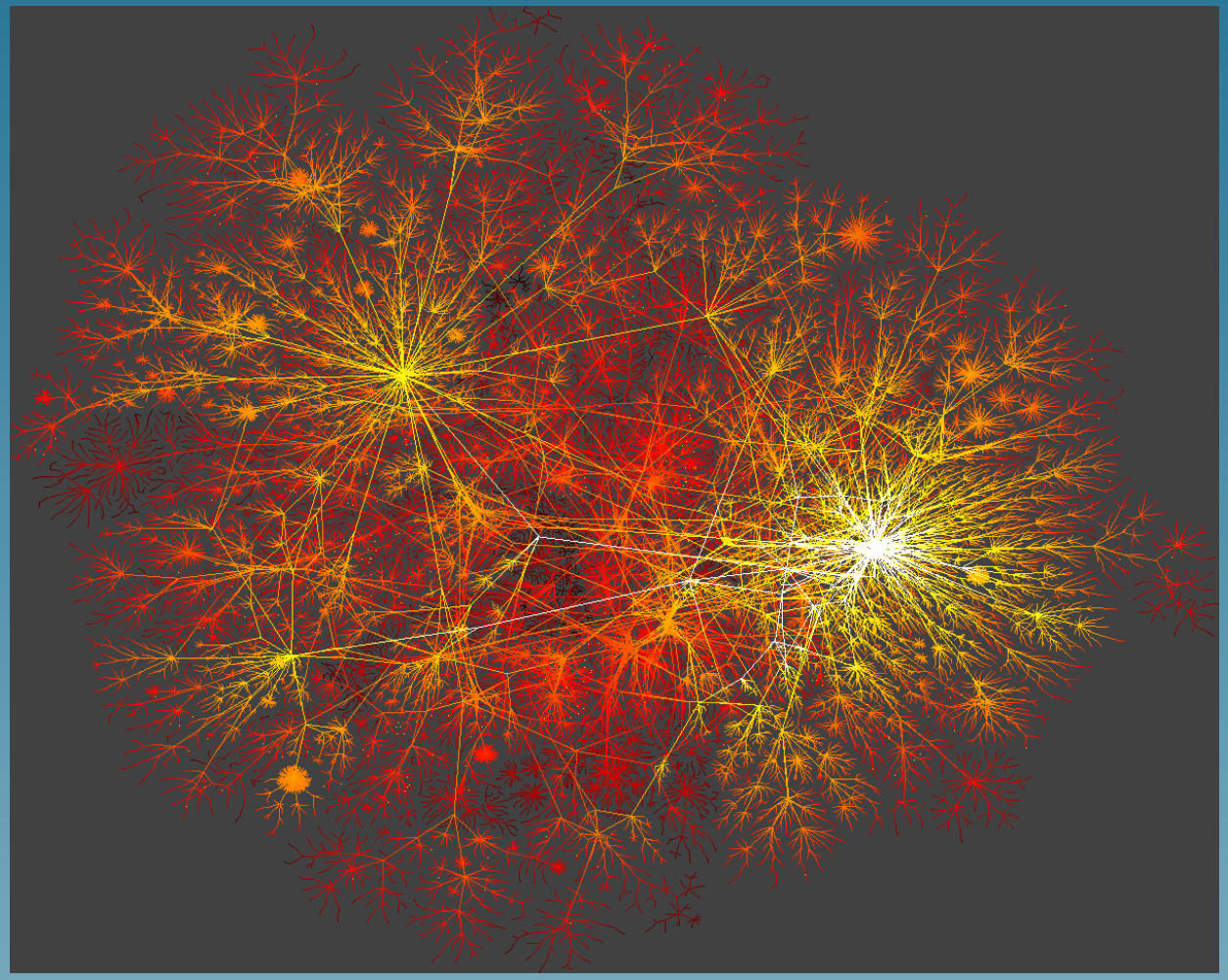
Analyse harmonique et covariances sur les graphes

Motivation

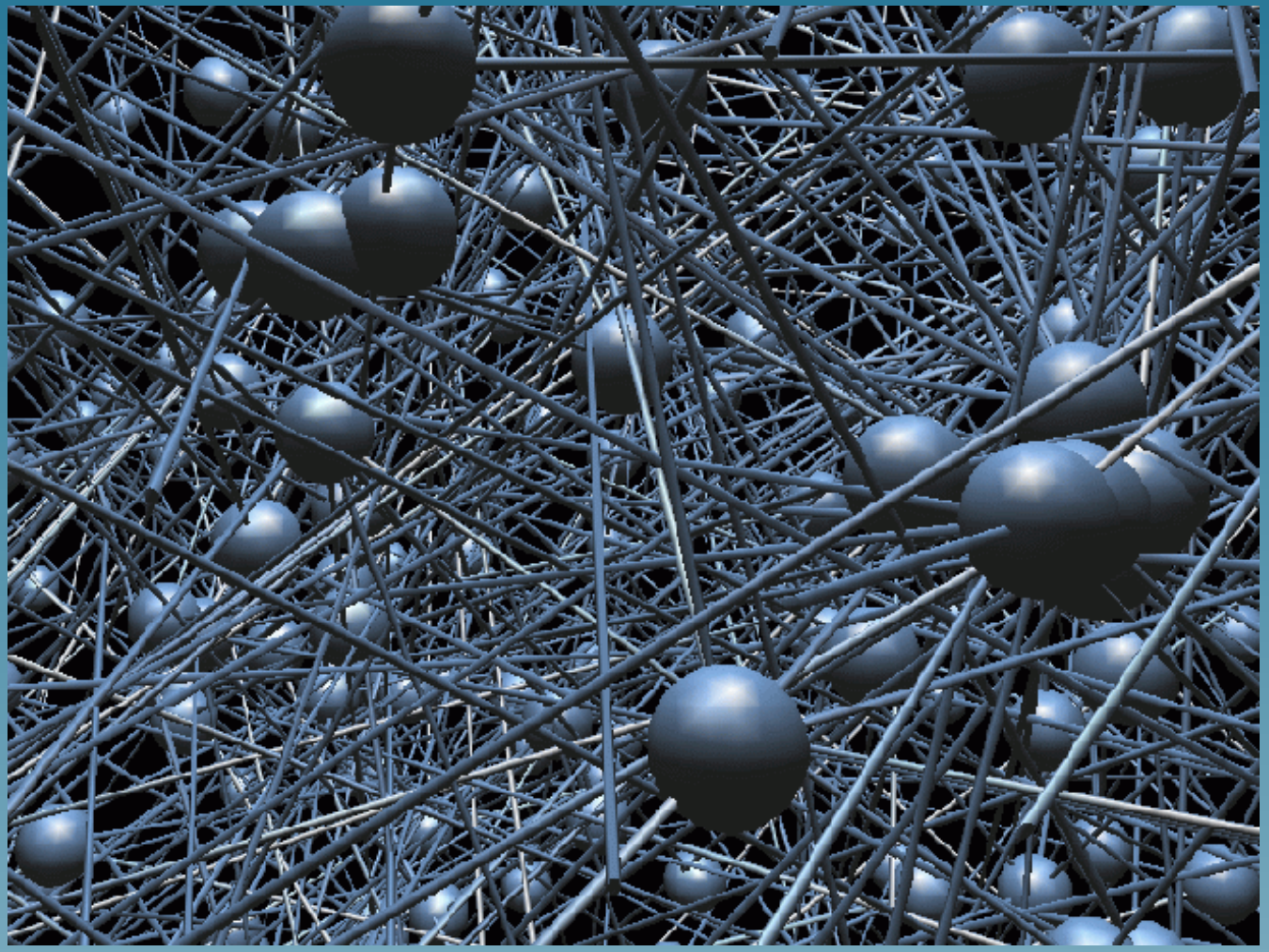
De nombreuses donnees peuvent se representer comme les noeux d'un graphe:

- par nature,
- par discretisation/echantillonage d'un espace continu,
- par necessite

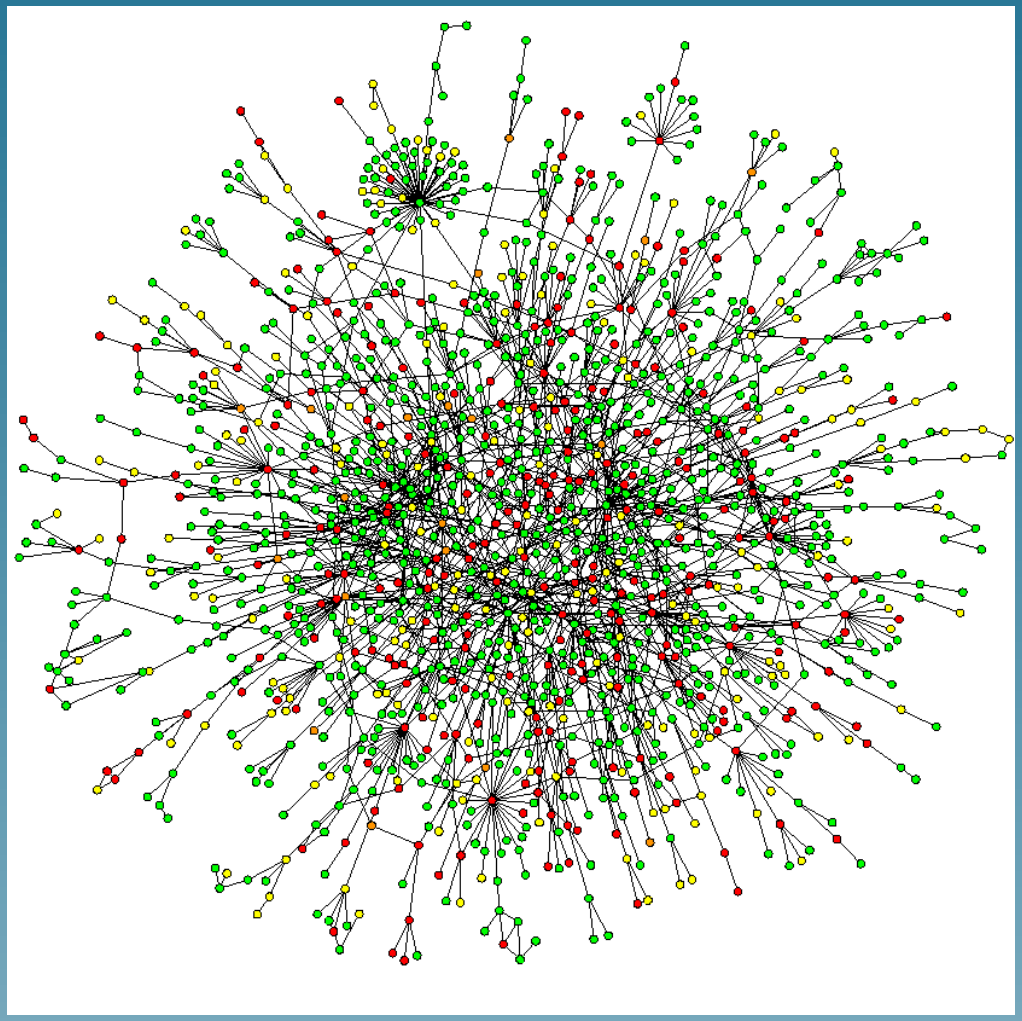
Internet (par nature)



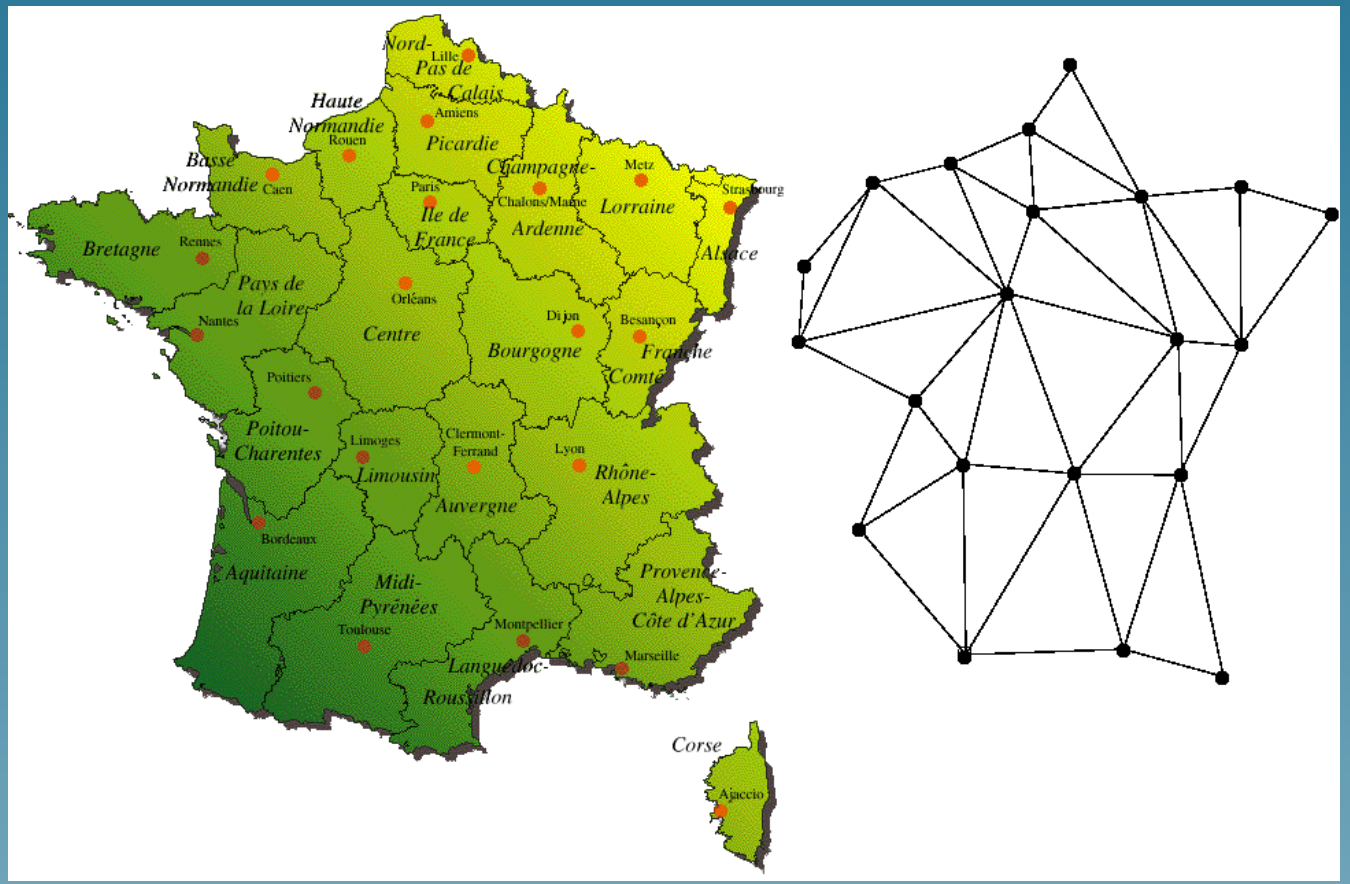
Reseau social (par nature)



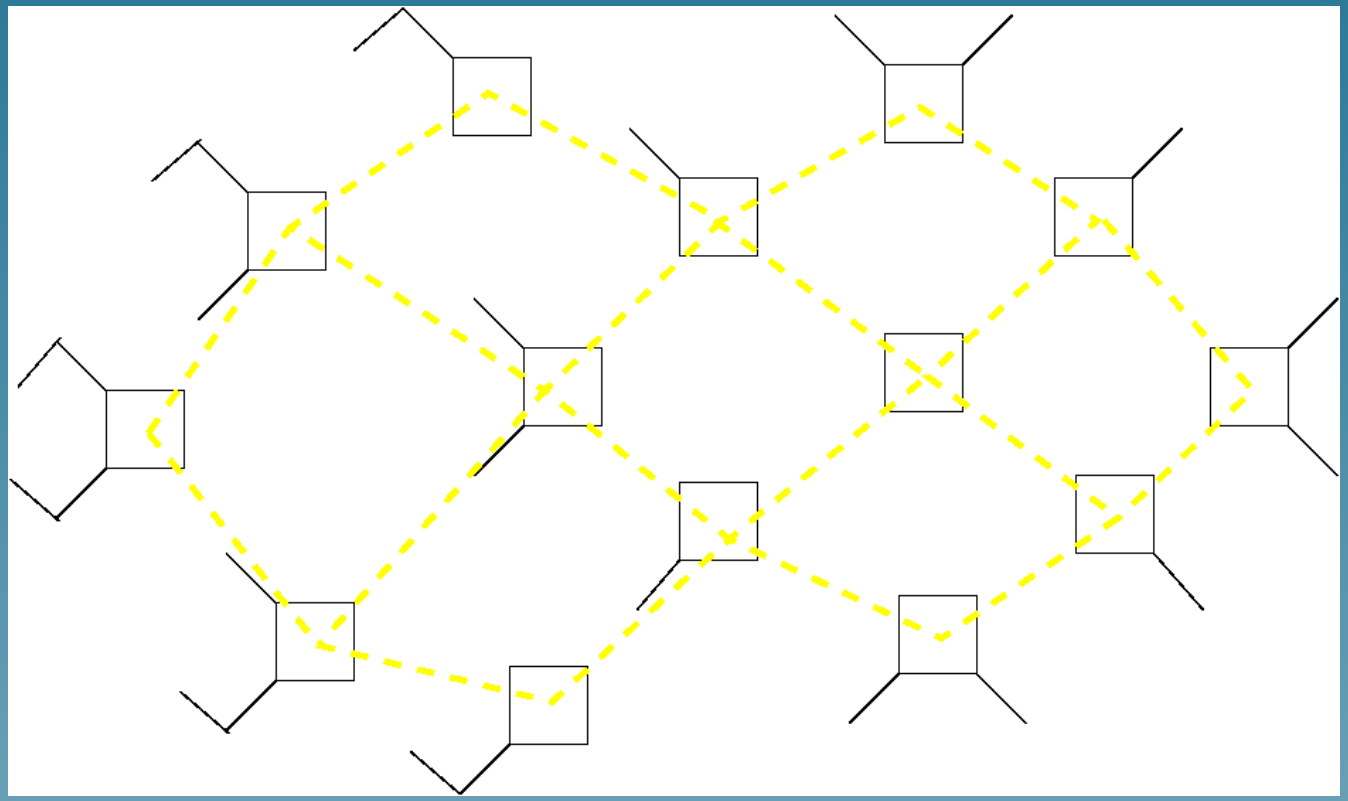
Interaction des proteines (par nature)



Regions (par discretisation)



Molecules (par necessite)



Covariance sur un graphe

- Graphe $G = (\mathcal{X}, E)$ avec $\mathcal{X}$ un ensemble fini de noeuds, $E \subset \mathcal{X} \times \mathcal{X}$ des liens entre les noeuds.
- $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ est une covariance ssi la matrice $K = [k(x, x')]_{x,x' \in \mathcal{X}}$ est symmetrique definie positive.
- Comment trouver une telle matrice qui contienne de l'information sur la structure du graphe?

Premiere approche: distance

- Soit $d(x, x')$ une distance sur le graphe, par exemple la longueur du plus court chemin entre x et x'.
- Soit $k(x, x') = C(d(x, x'))$ ("stationnaire isotropique")
- Probleme: pas de condition generale sur C pour que k soit definie positive...

Deuxieme approche: analyse harmonique

- Dans le cas $\mathcal{X} = \mathbb{R}^p$, la condition sur C est:

$$C(h) = \int_{\omega \in \mathbb{R}^p} e^{2\pi i \omega h} g(\omega) d\omega$$

  avec $g \geq 0$.

- C se decompose sur la base de Fourier $\phi_\omega(h) = e^{2\pi i \omega h}$

- Faisons de l'analyse harmonique sur les graphe...

Laplacien sur les graphe

- Sur $\mathbb{R}^d$, les ϕ_ω sont les fonctions propres de l'operateur Laplacien:

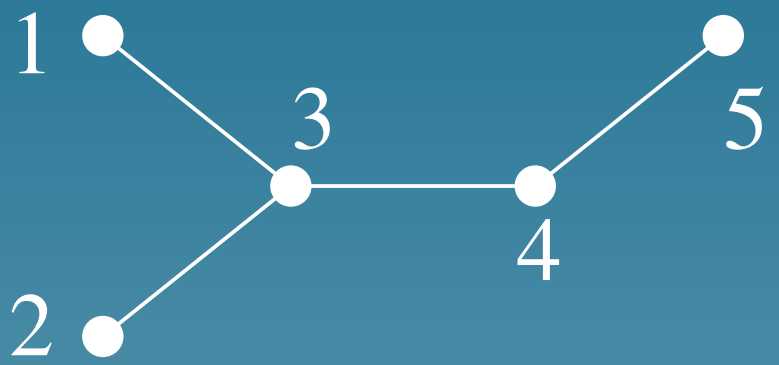
$$\Delta = \sum_{i=1}^{p} \frac{\partial^2}{\partial x_i^2}.$$

- Laplacien sur un graphe: si $f : \mathcal{X} \to \mathbb{R}$,

$$\Delta f(x) = \sum_{x' \sim x} [f(x') - f(x)]$$

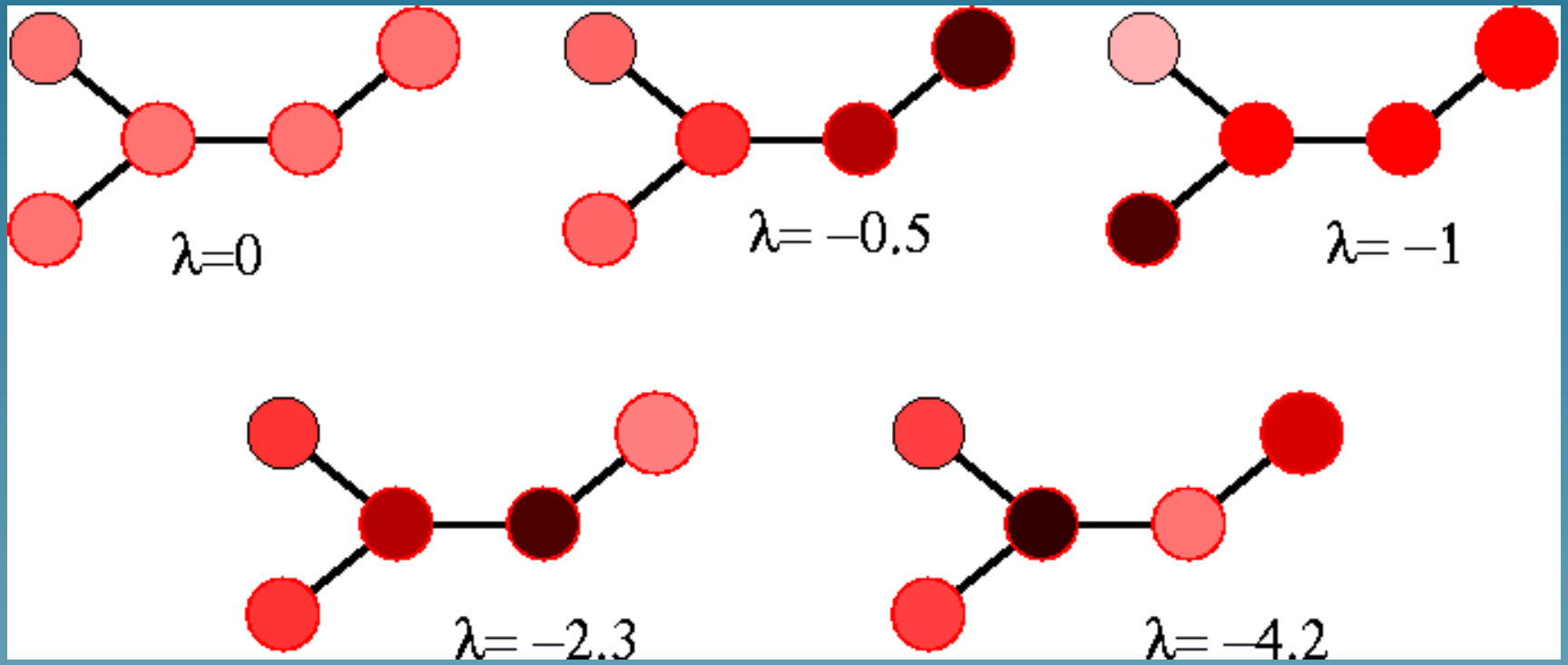
- Forme matricielle: $\Delta = A - D$, avec A la matrice d'ajacence et D la matrice diagonale des degres.

Exemple



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

Fonctions propres du Laplacien



Spectre du Laplacien

- Les fonctions propres $\phi_1, \ldots, \phi_n$ du Laplacien forment une base de Fourier.
- Valeurs propres $\lambda_1 = 0 \geq \ldots \geq \lambda_n$ decroissent quand la frequence augmente.
- Transformee de Fourier d'une fonction $f : \mathcal{X} \to \mathbb{R}$:

$$\forall x \in \mathcal{X}, \quad f(x) = \sum_{i=1}^{n} \hat{f}_i \phi_i(x)$$

Du Laplacien a la covariance

- Si $f(x)$ est une fonction "reguliere", la fonction $k(x, x') = f(x)f(x')$ est une covariance "reguliere".
- Bonne fonction de covariance:

$$K(x, x') = \sum_{i=1}^{n} \gamma(\lambda_i)\phi_i(x)\phi_i(x')$$

avec $\gamma : \mathbb{R}_- \to \mathbb{R}_+$, croissante.

- Matrices: si $\Delta = U^{-1}DU$ alors $K = U^{-1}\gamma(D)U$

Exemple: noyau de la chaleur

- On choisit $\gamma(\lambda) = e^{\beta\lambda}$, et on obtient

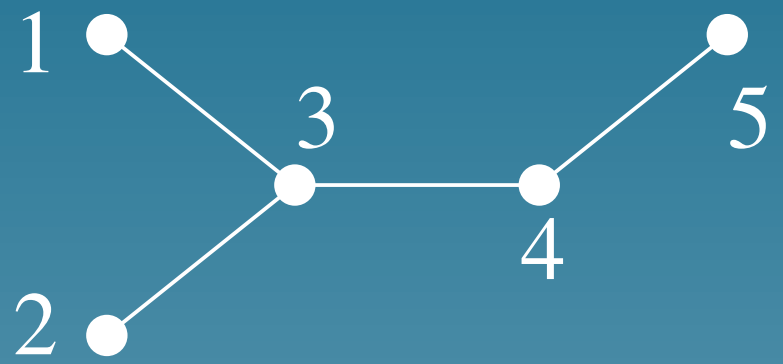
$$K_\beta = U^{-1} e^{\beta D} U = e^{\beta \Delta}.$$

- Ce noyau est la solution de l'equation de la chaleur:

$$\frac{\partial}{\partial \beta} K_\beta = \Delta K_\beta$$

- D'autres covariances sont bien sur possibles!

Exemple de covariance



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

Interpretation: Marche aléatoire

- Soit $\beta < 1/max(d_i)$ et $Z_1, Z_2...$ une marche aléatoire sur le graphe avec

$$P(Z_{i+1} = x | Z_i = x') = \begin{cases} \beta & \text{si } x \sim x', \\ 1 - \beta d_i & \text{si } x = x', \\ 0 & \text{sinon} \end{cases}$$

- Alors $P(Z_N = x | Z_1 = x') = [(I + \beta\Delta)]_{x,x'}$
- Limite diffusion: $N = 1/\delta t$, $\beta = \beta_0 \delta t$:

$$\lim_{\delta t \to 0} P(Z_N = x | Z_1 = x') = [\exp(\beta_0 \Delta)]_{x,x'} = K(x, x').$$

Part 3

Covariance sur les groupes

Motivations

- Un groupe est un ensemble G muni d'une operation $\circ$ associative, avec un element neutre e tel que $e \circ x = x \circ e = x$, et tel que chaque element $x \in G$ a un inverse $x^{-1} \in G$ tel que $x \circ x^{-1} = e$.
- Exemple: $(\mathbb{R}^p, +)$ est un groupe (commutatif)
- L'ensemble des permutations de $(1, \ldots, n)$ muni de la composition est un groupe (non commutatif)
- Exemple: classez les 6 chaines de television par ordre de preference, et je predit vos revenus

Statistique sur les groupes

Persi Diaconis, “Group representation in Probability and Statistics”, IMS lecture notes, 1988.

Covariance sur un groupe

- Sur $(\mathbb{R}^p, +)$, on prend $k(x_1, x_2) = C(x_1 - x_2)$ quand C a une transformee de Fourier reelle positive
- Sur $(g, \circ)$, soit $C : G \to \mathbb{R}$ et

$$k(x_1, x_2) = C(x_1 \circ x_2^{-1})$$

- Quelle condition sur C pour the k soit une covariance?
- Reponse: par l'analyse harmonique sur les groupes...

Cas des groupes finis (resume)

- Theorie classique de la representation lineaire des groupes finies (voir J.-P. Serres, 1966)
- La transformee de Fourier d'une fonction $f : G \to \mathbb{R}$ en un ensemble de matrices carres (on decompose f sur les caracteres des representations irreductibles)
- Theorem (Bochner): k est une covariance ssi la transformee de Fourier de C est composee de matrices semidefinie positive.
- De plus, C est constante sur les classes de conjugaison ssi c'est

une combinaison lineaire de caracteres irreuctibles : peu de degre de liberte!

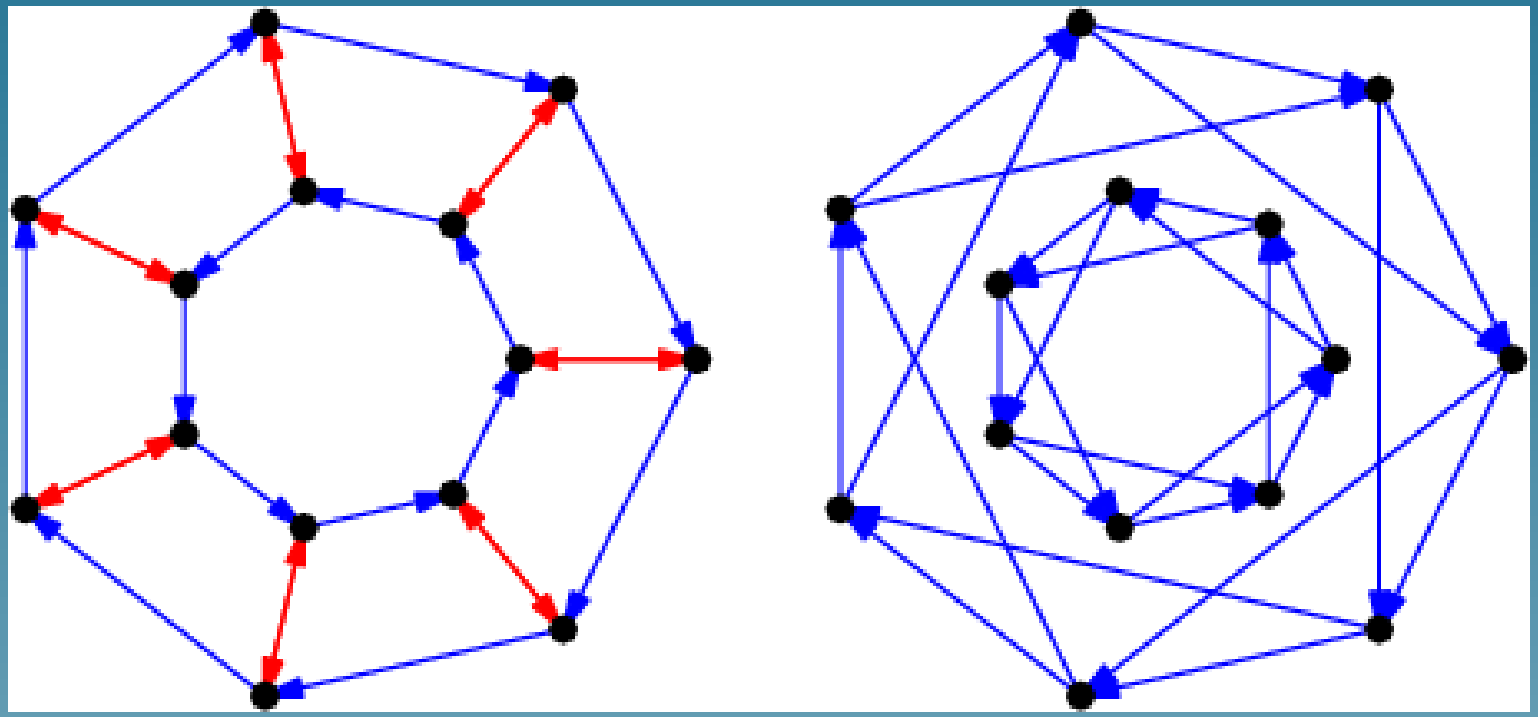
Calcul pratique d'une covariance sur un groupe

- Soit G un groupe
- Soit $S \subset G$ tel que $e \in S$ et si $x \in S$ alors $x^{-1} \in S$
- Le graphe de Cayley a pour neuds G et pour liens:

$$x_1 \sim x_2 \text{ ssi } x_1 \circ x_2^{-1} \in S.$$

- On peut faire un noyau sur le graphe de Cayley!

Graphes de Cayley



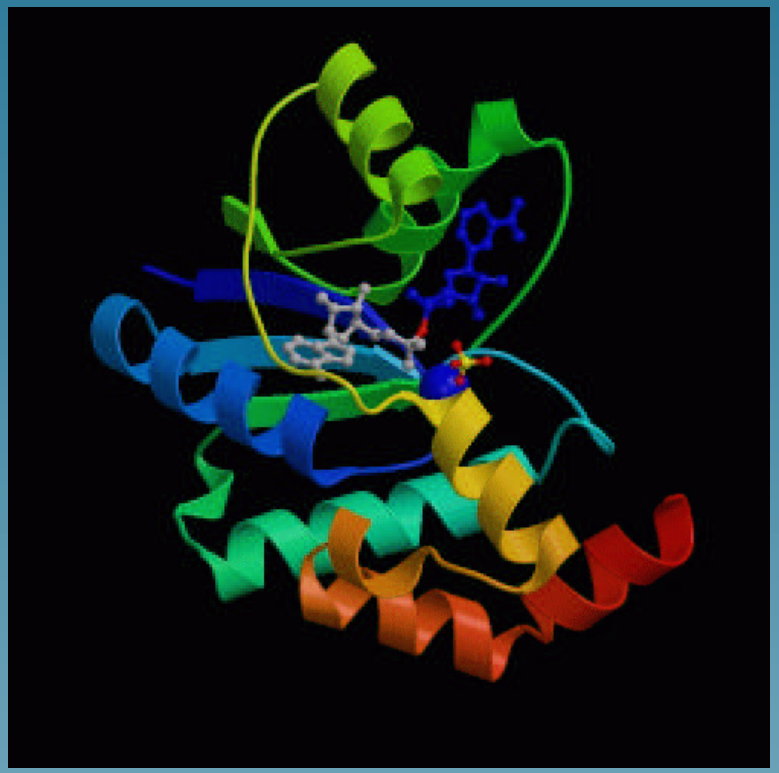
Groupe dihedral D_7, un groupe de permutation de 14 elements, avec differents ensembles S.

Part 4

Detecting pathway activity from microarray data

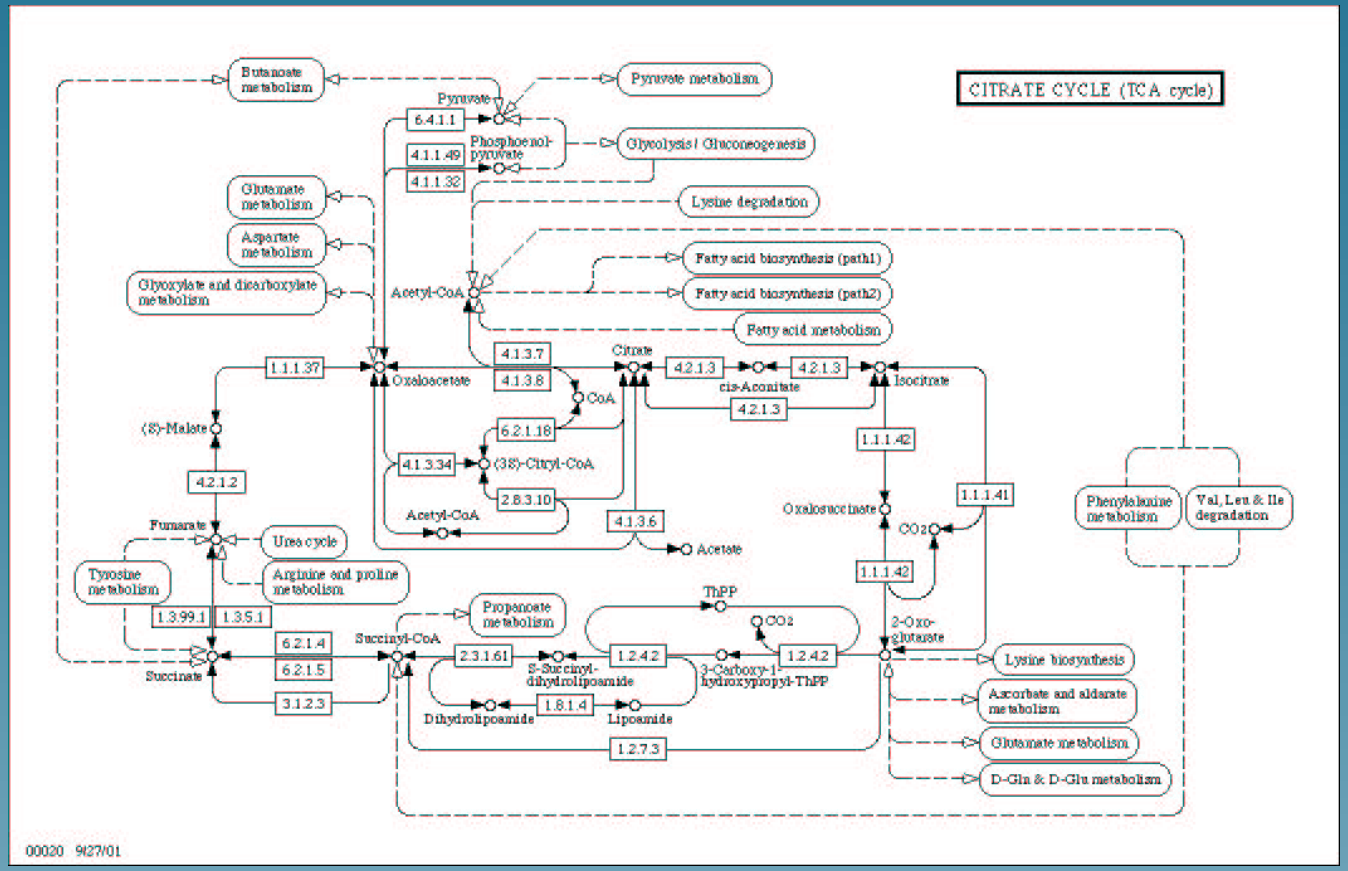
(with M. Kanehisa, *ECCB 2003*)

Genes encode proteins which can catalyse chemical reations



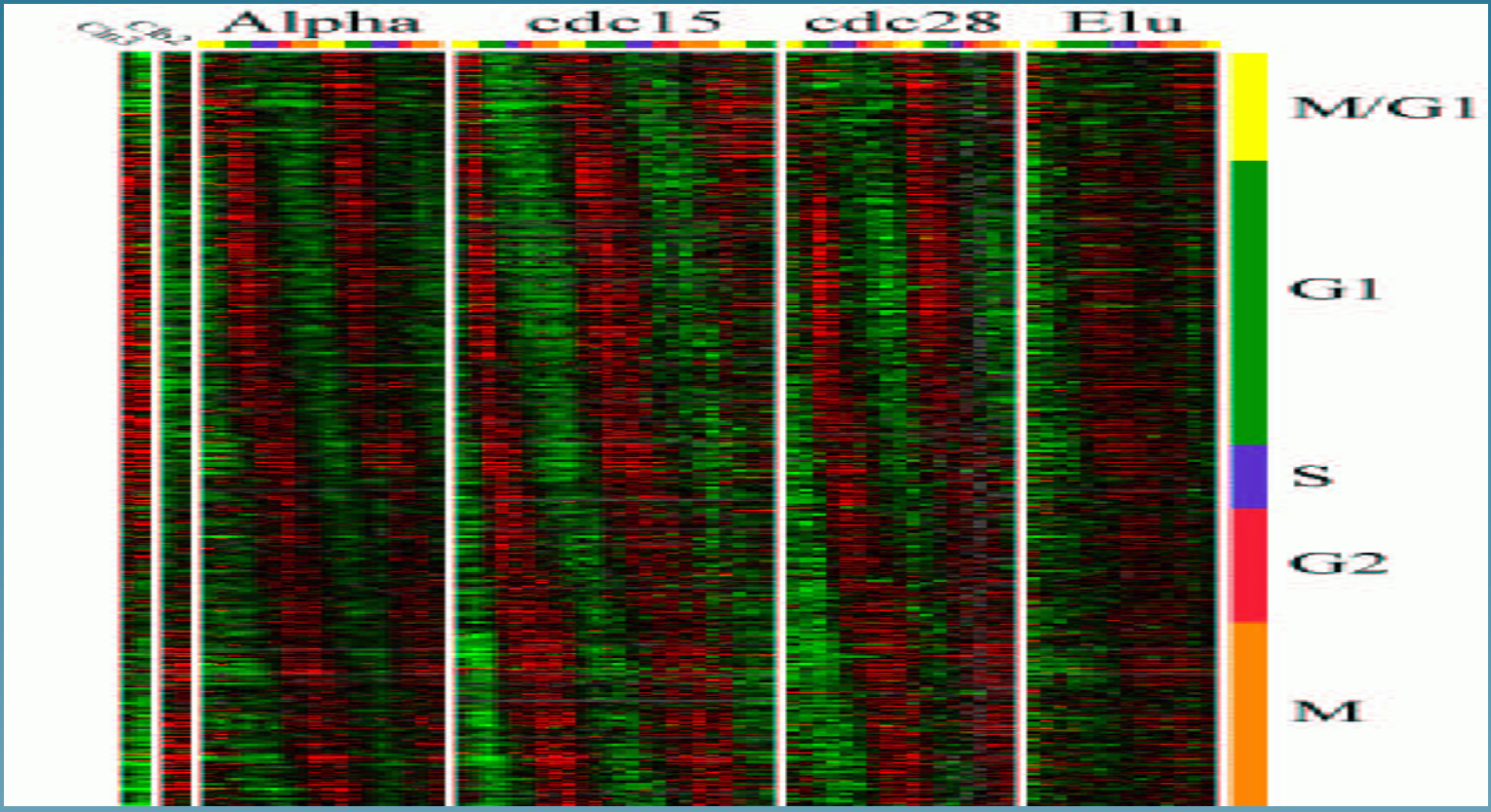
Nicotinamide Mononucleotide Adenylyltransferase With Bound Nad+

Chemical reactions are often parts of pathways



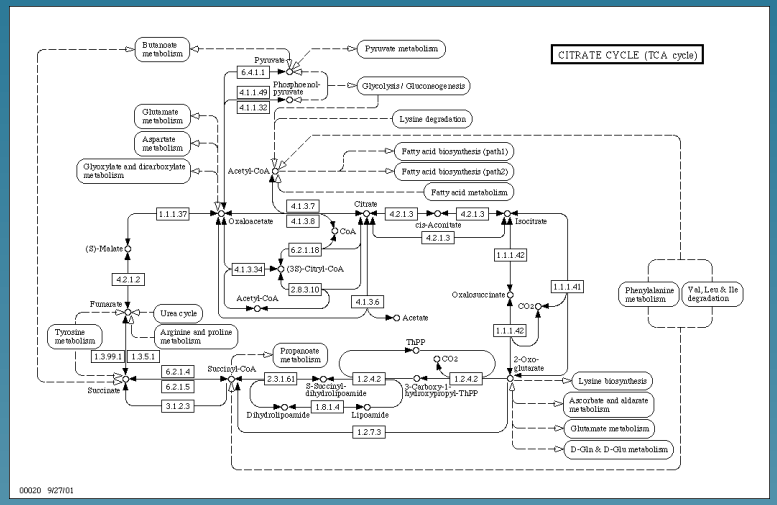
From http://www.genome.ad.jp/kegg/pathway

Microarray technology monitors mRNA quantity

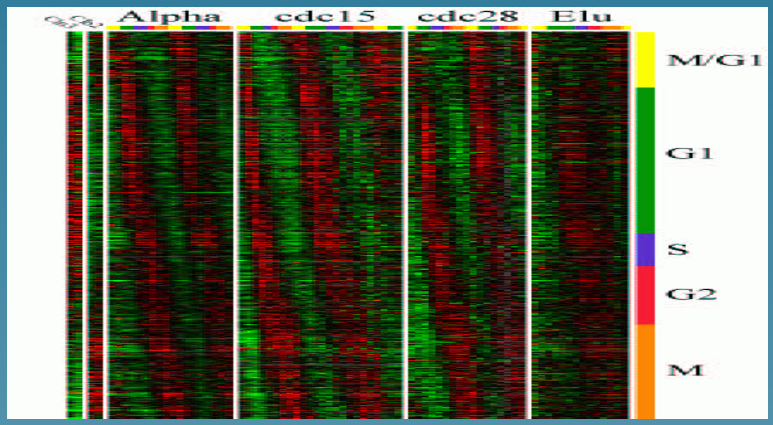


(From Spellman et al., 1998)

Comparing gene expression and pathway databases

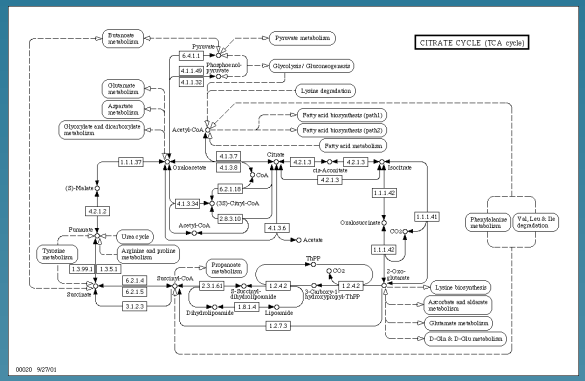


VS

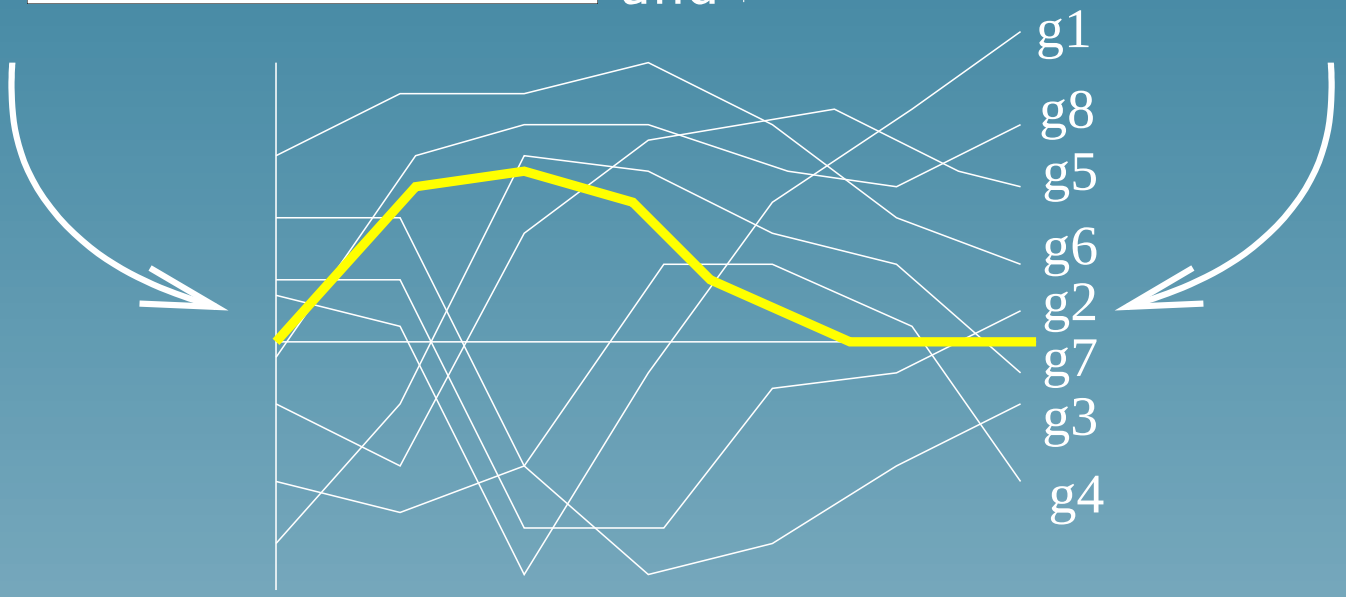


Detect active pathways? Denoise expression data?
Denoise pathway database? Find new pathways?
Are there "correlations"?

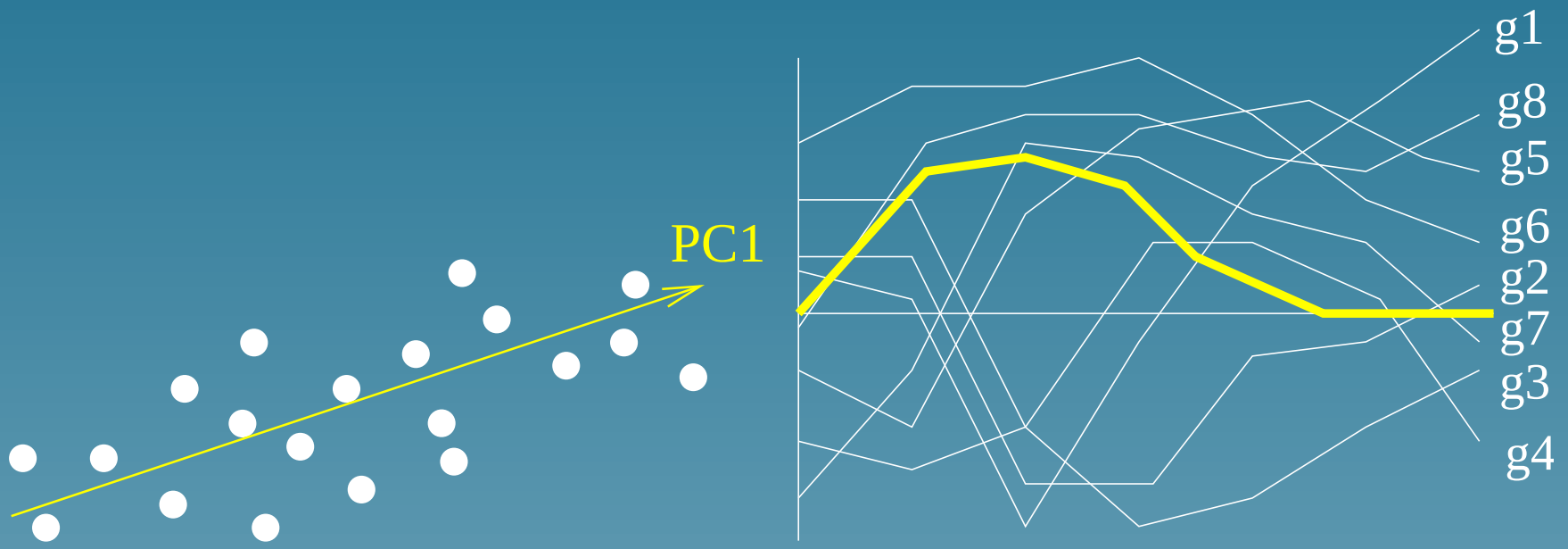
A useful first step



and



Using microarray only



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

PCA formulation

- Let $f_v(i)$ be the projection of the i-th profile onto v.
- The amount of variation captured by f_v is:

$$h_1(v) = \sum_{i=1}^{N} f_v(i)^2$$

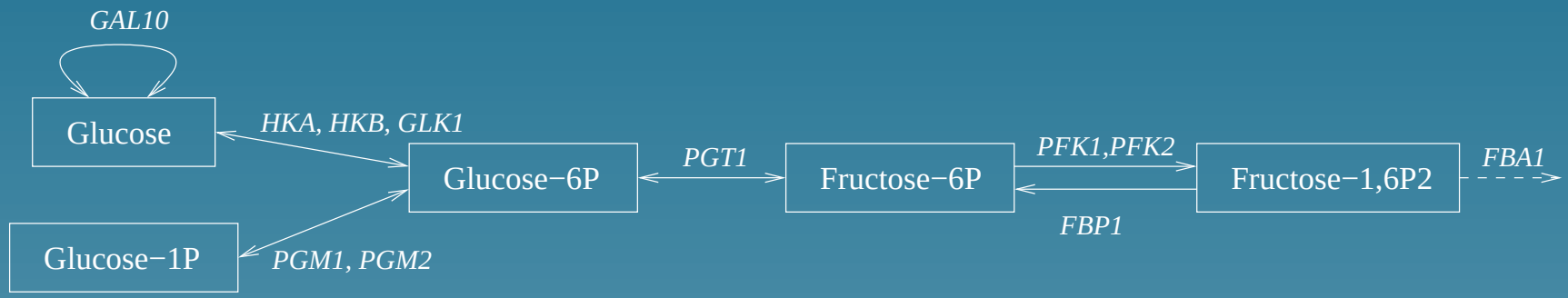
- PCA finds an orthonormal basis by solving successively:

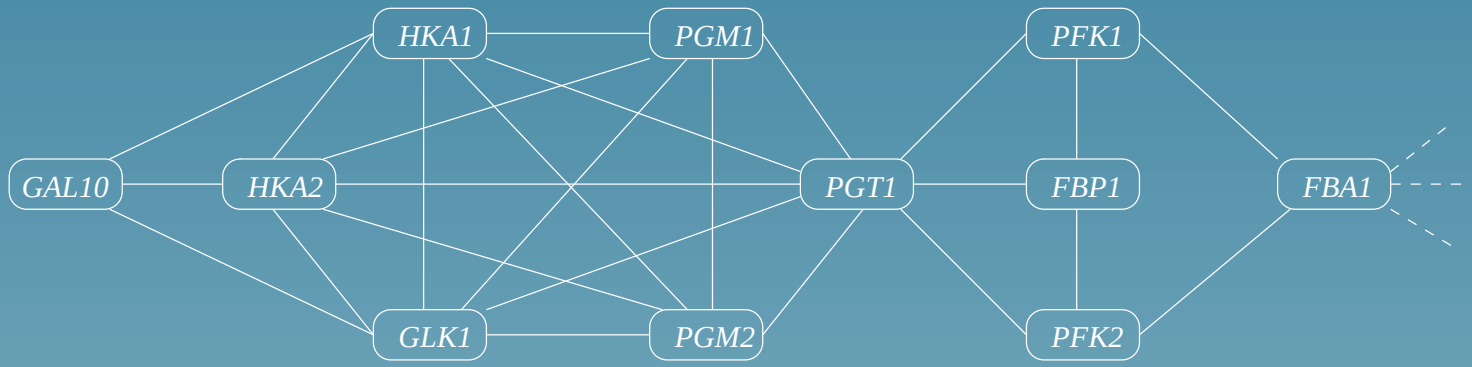
$$\max_v h_1(v)$$

Issues with PCA

- 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

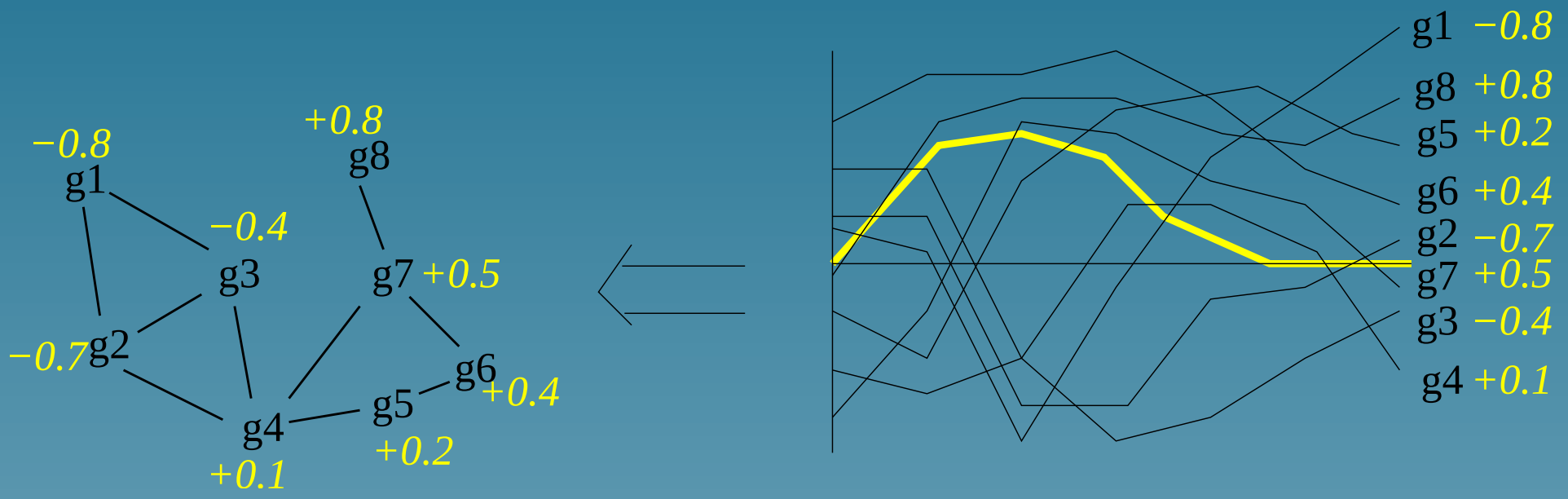
The metabolic gene network





Link two genes when they can catalyze two successive reactions

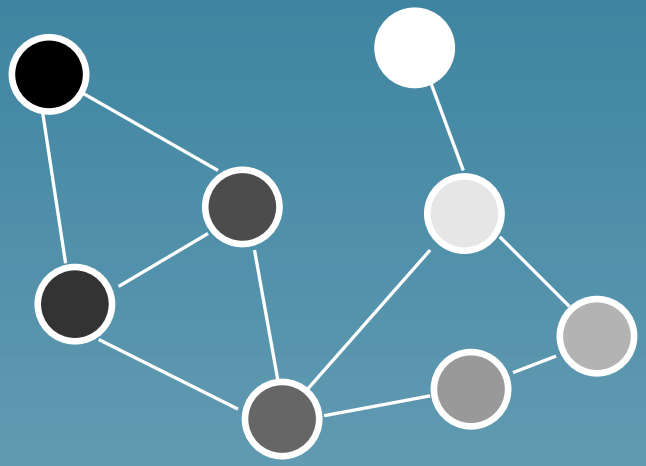
Mapping f_v to the metabolic gene network



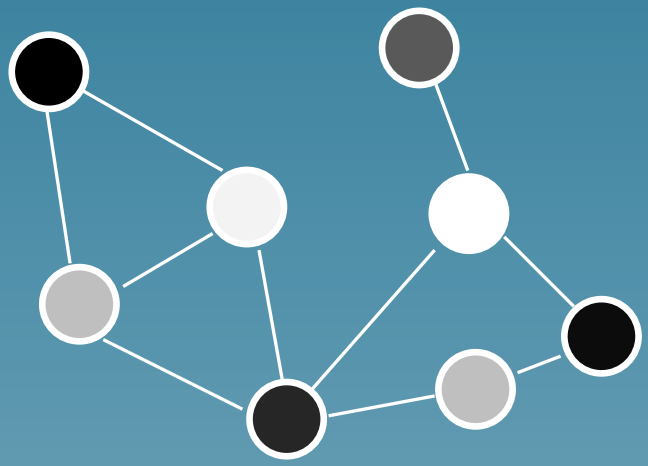
Does it look interesting or not?

Important hypothesis

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

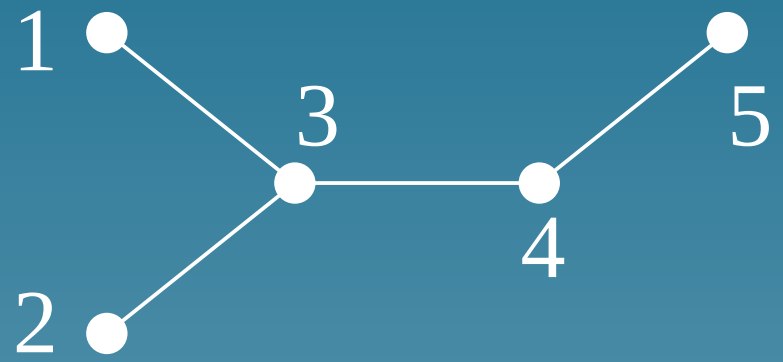


Smooth



Rugged

Graph Laplacian $L = D - A$

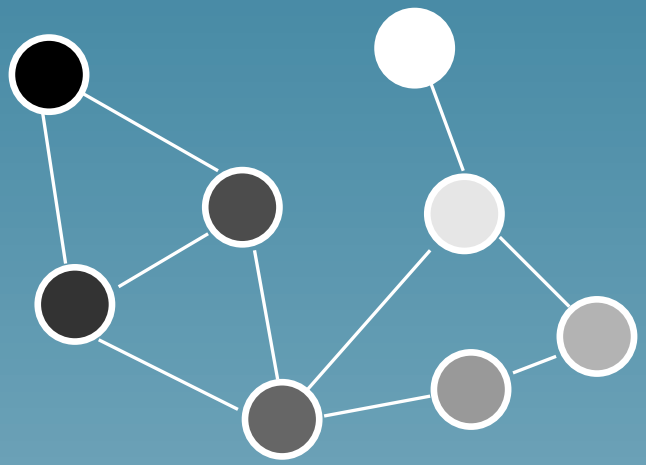


$$L = \begin{pmatrix} -1 & 0 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 \\ 1 & 1 & -3 & 1 & 0 \\ 0 & 0 & 1 & -2 & 1 \\ 0 & 0 & 0 & 1 & -1 \end{pmatrix}$$

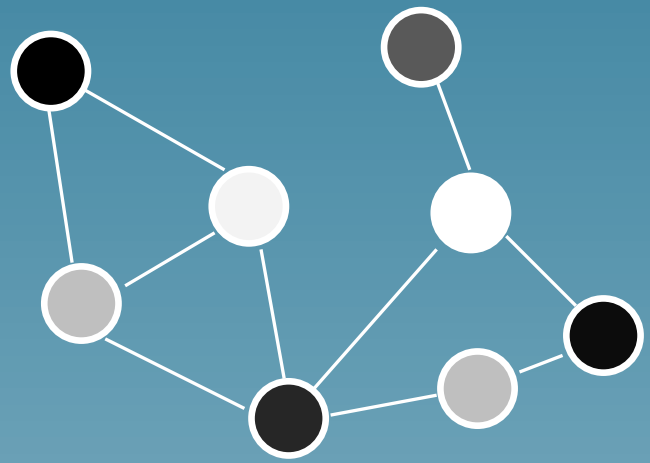
Smoothness quantification

$$h_2(f) = \frac{f^\top \exp(-\beta L) f}{f^\top f}$$

is large when f is smooth



h(f) = 2.5



h(f) = 34.2

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:

- $h_1(f_v)$ be large
- $h_2(f_2)$ be large
- $corr(f_v, f_2)$ be large

by solving:

$$\max_{(f_v, f_2)} corr(f_v, f_2) \times \frac{h_1(f_v)}{h_1(f_v) + \delta} \times \frac{h_2(f_2)}{h_2(f_2) + \delta}$$

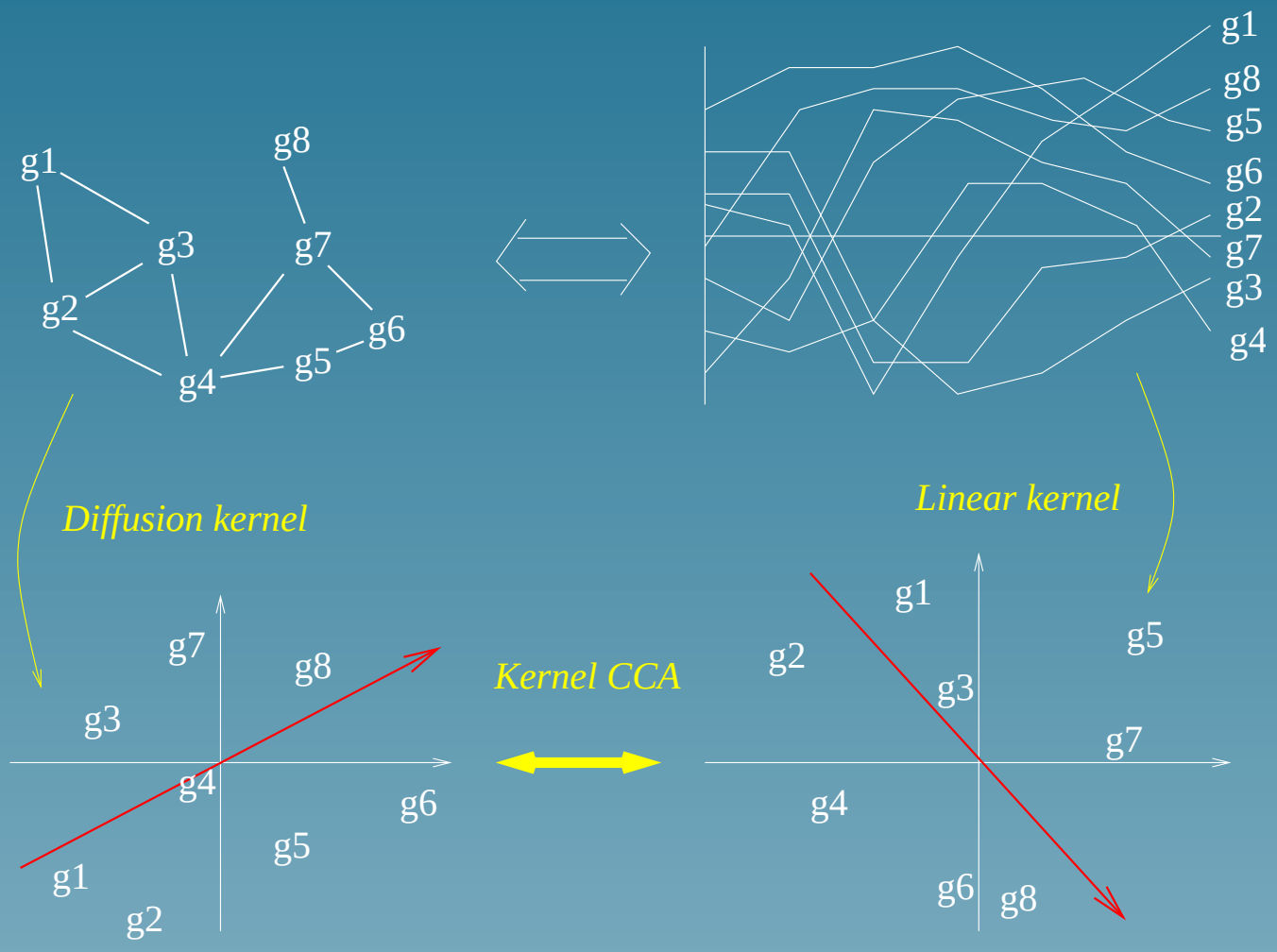
Solving the problem

This formultation is equivalent to a generalized form of CCA (Kernel-CCA, Bach and Jordan, 2002), which is solved by 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 $f_2 = K_2\beta$.

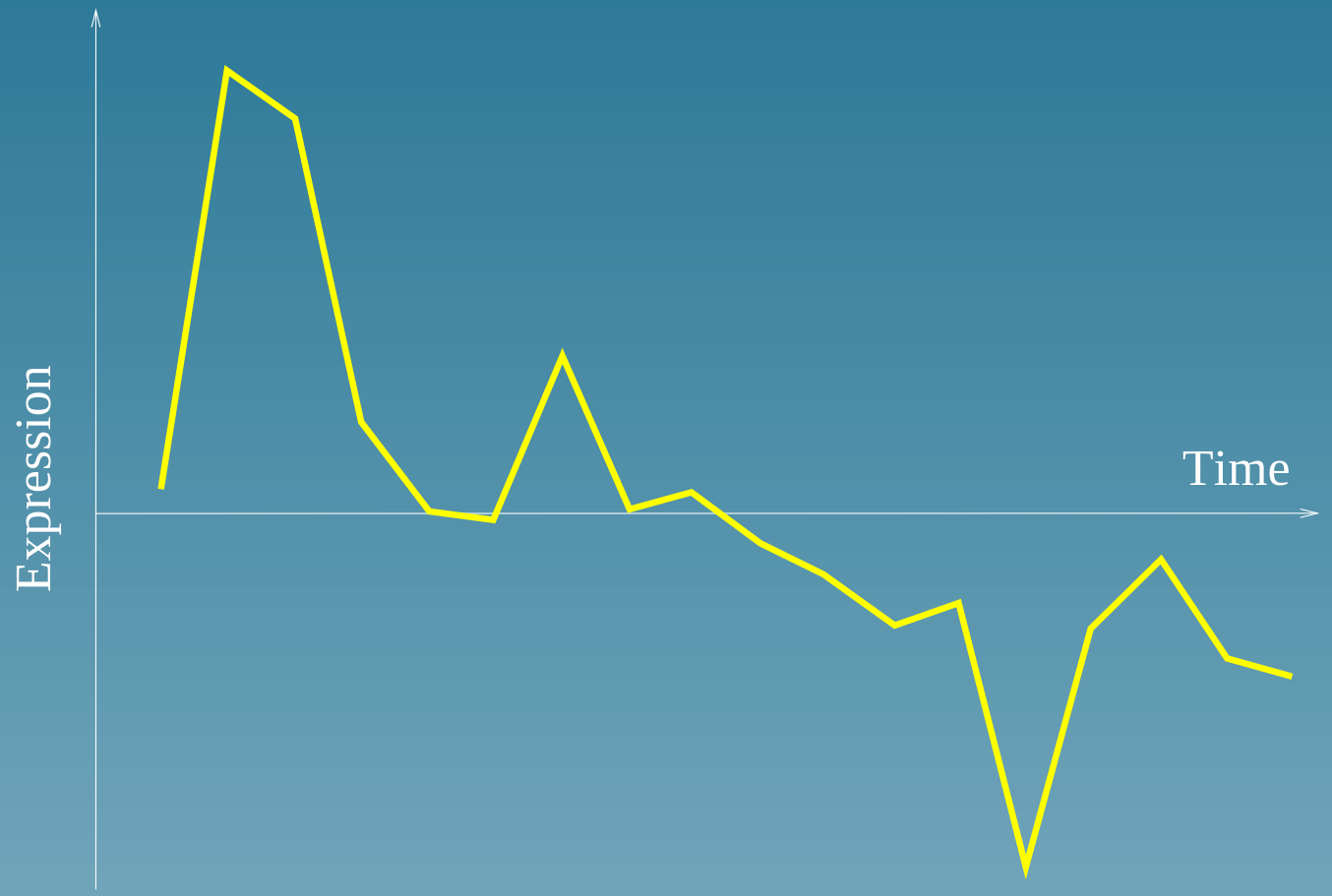
The kernel point of view...



Data

- Gene network: two genes are linked if the catalyze successive reactions in the KEGG database (669 yeast genes)
- Expression profiles: 18 time series measures for the 6,000 genes of yeast, during two cell cycles

First pattern of expression

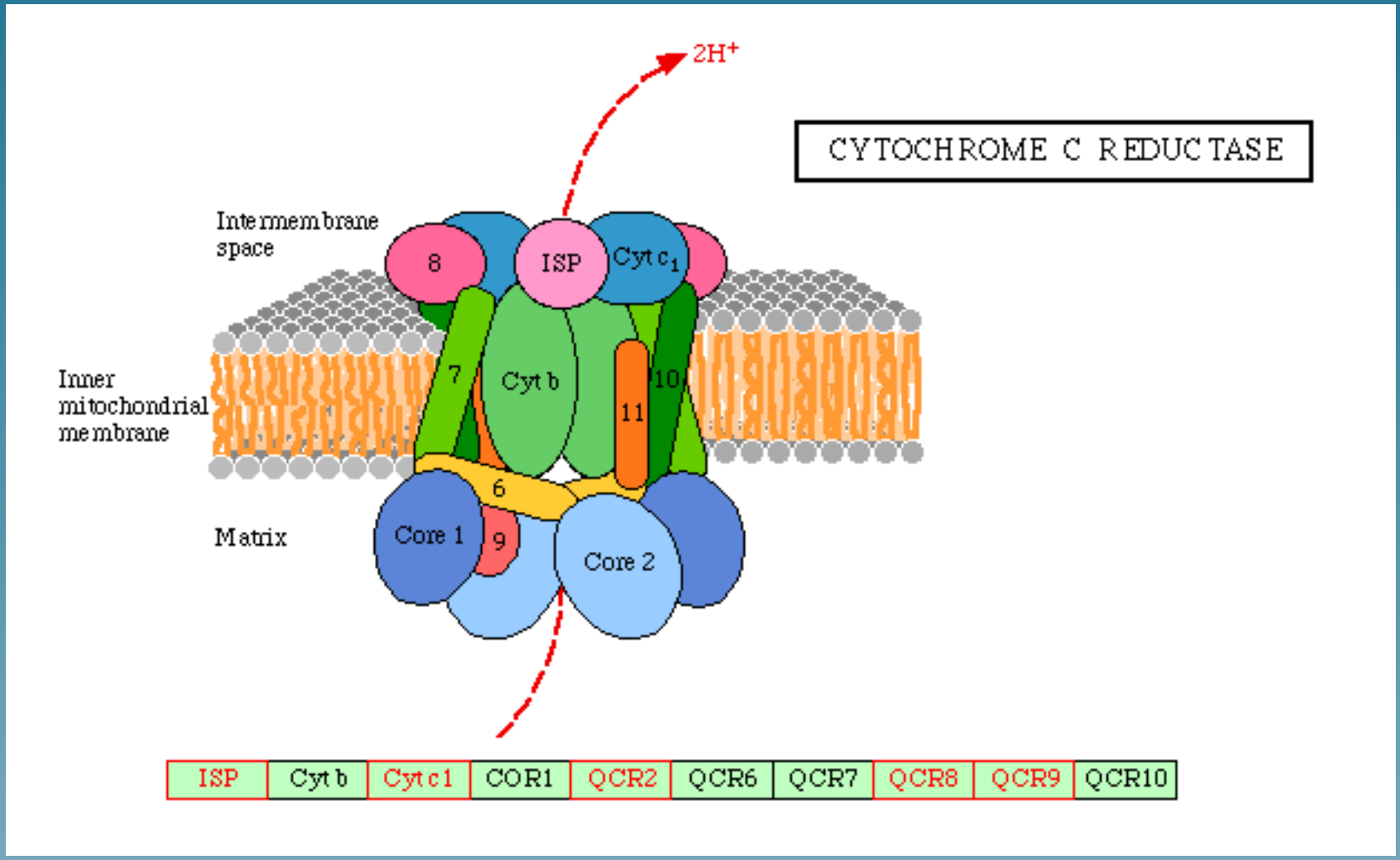


Related metabolic pathways

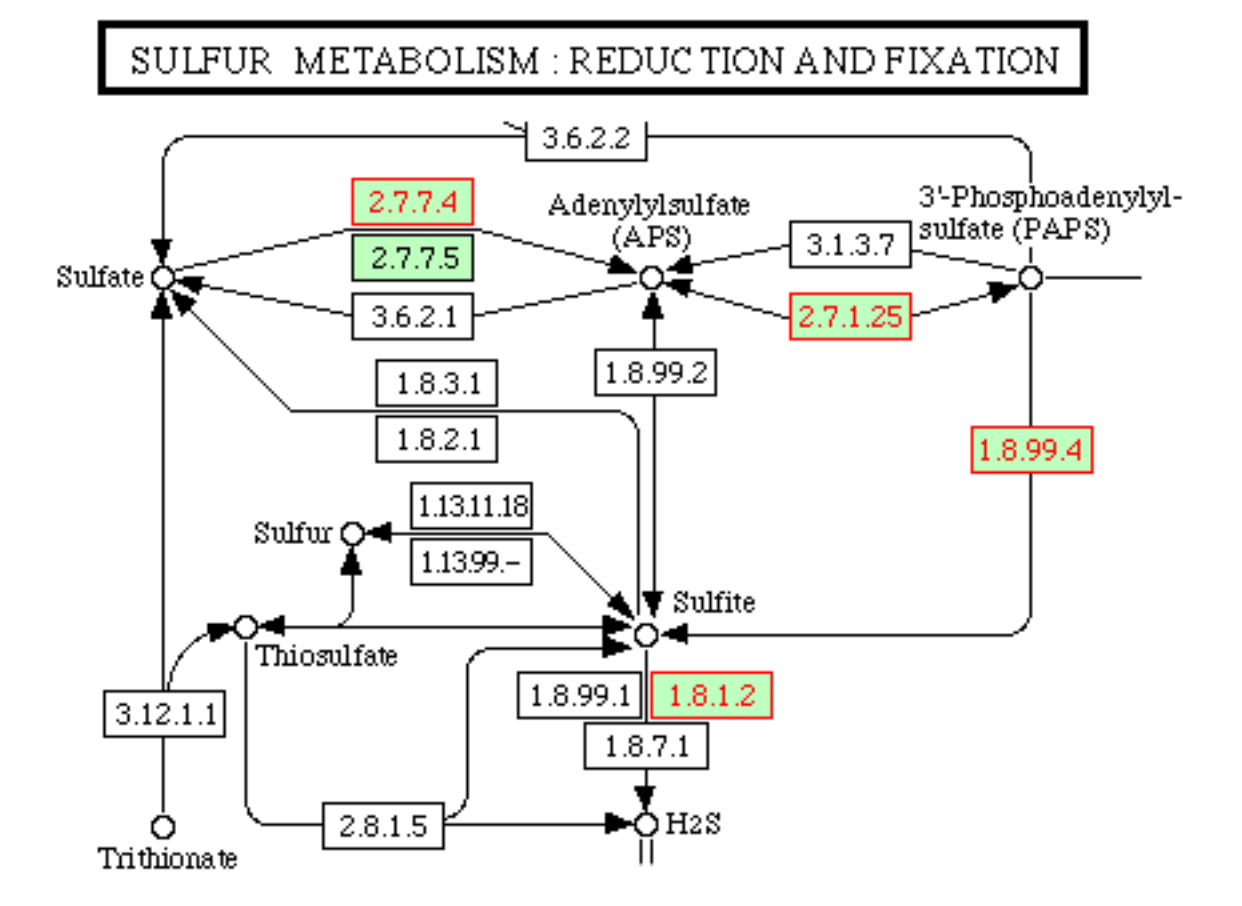
50 genes with highest $s_2 - s_1$ belong to:

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

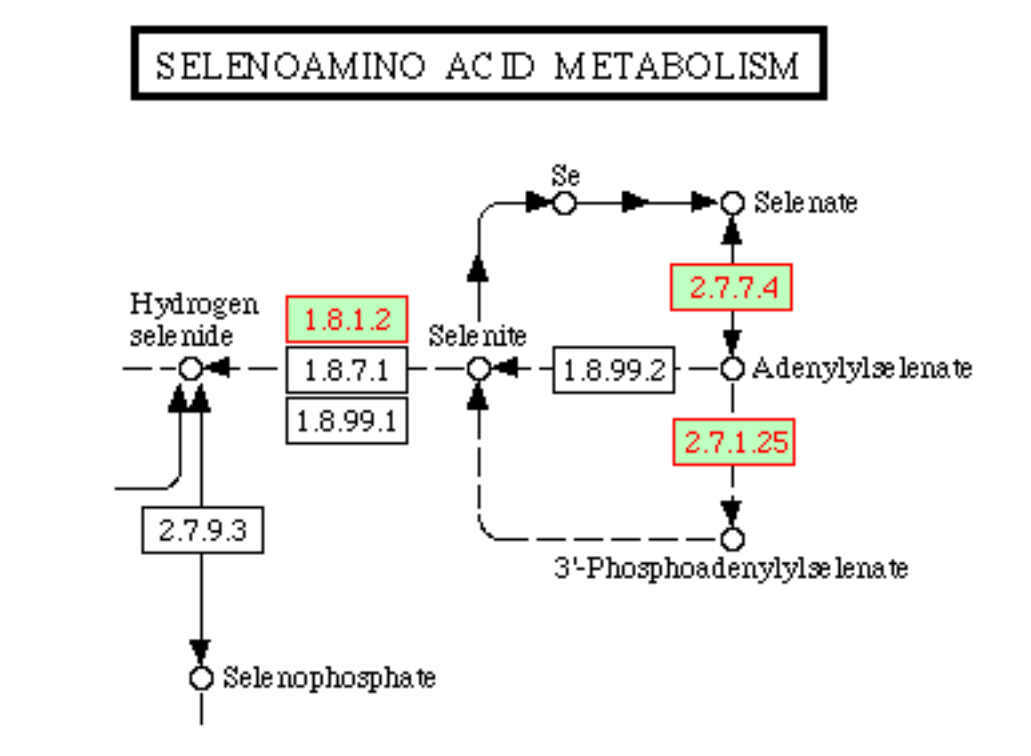
Related genes



Related genes



Related genes



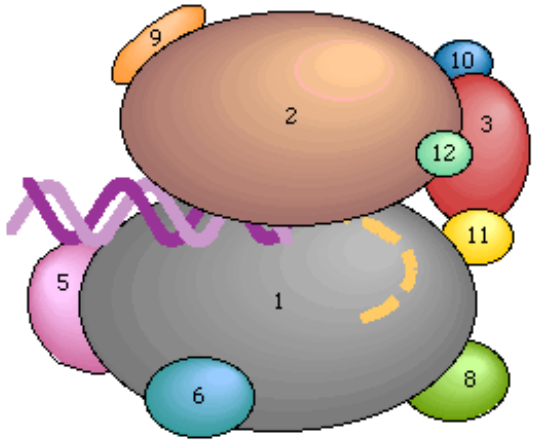
Opposite pattern

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

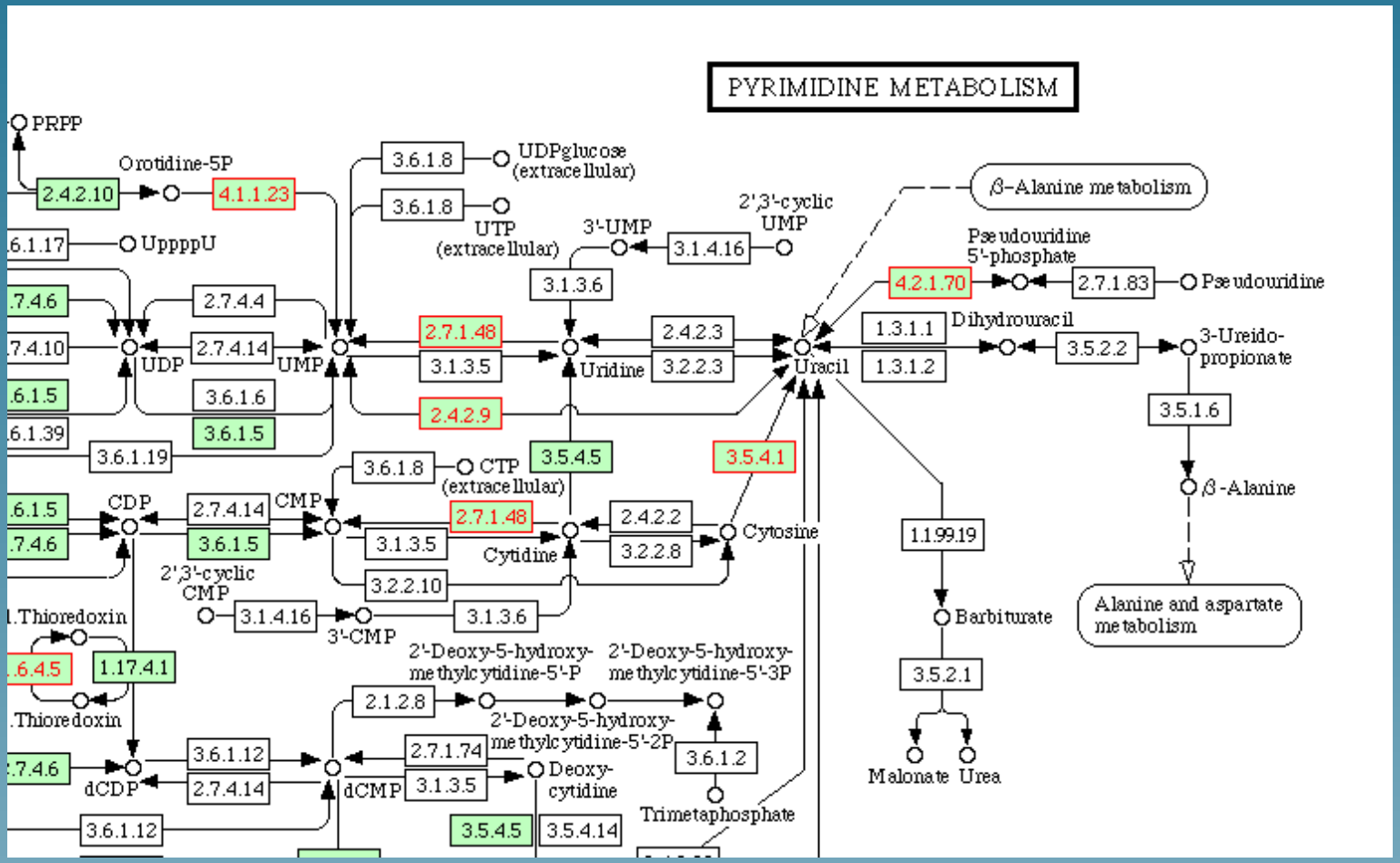
Eukaryotic Pol III

C2	C3	C4	C5	C11
C1	C19	C25	C31	C34

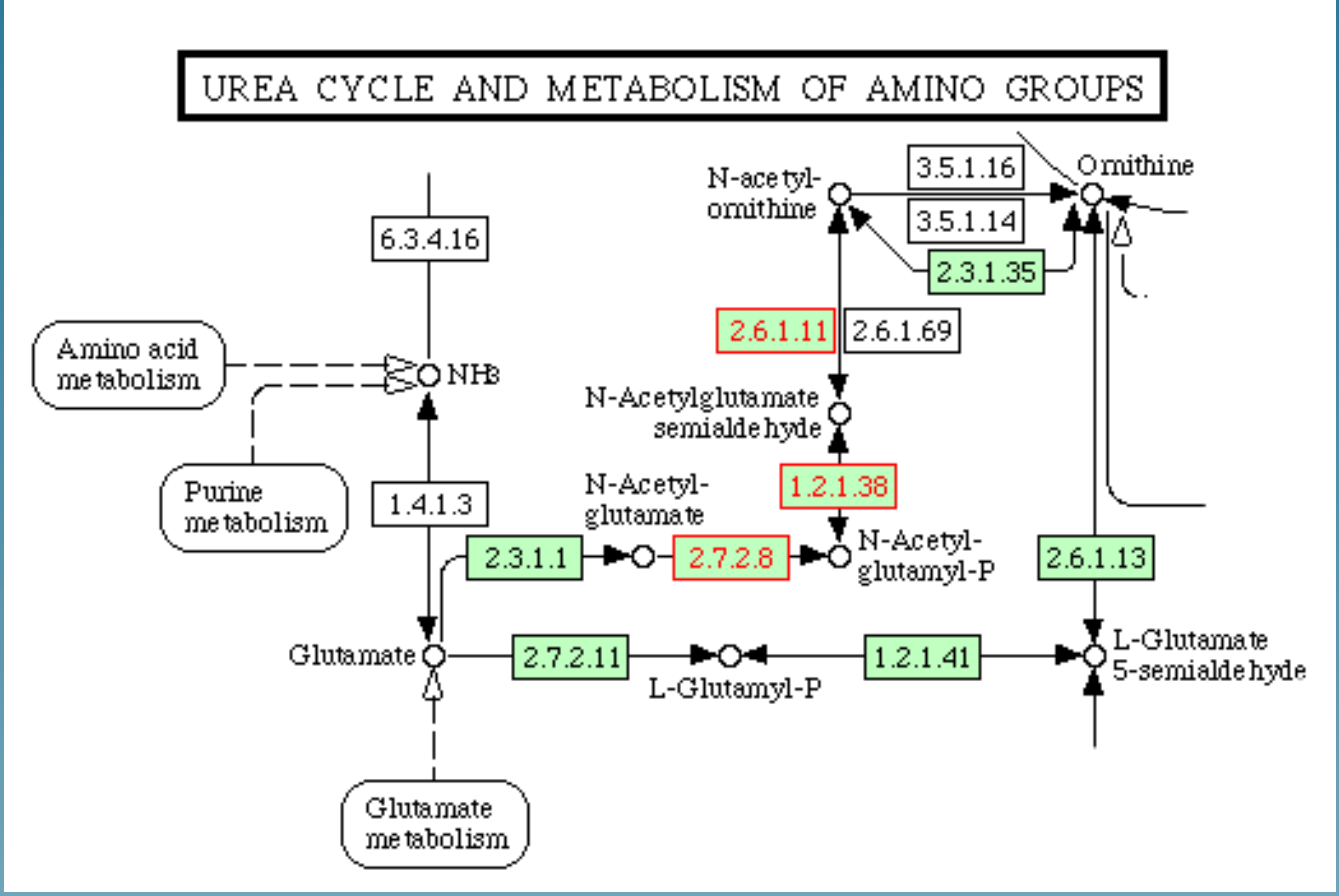
Eukaryotic Pol I

A2	A12	A14	A34	A43	A49
A1					

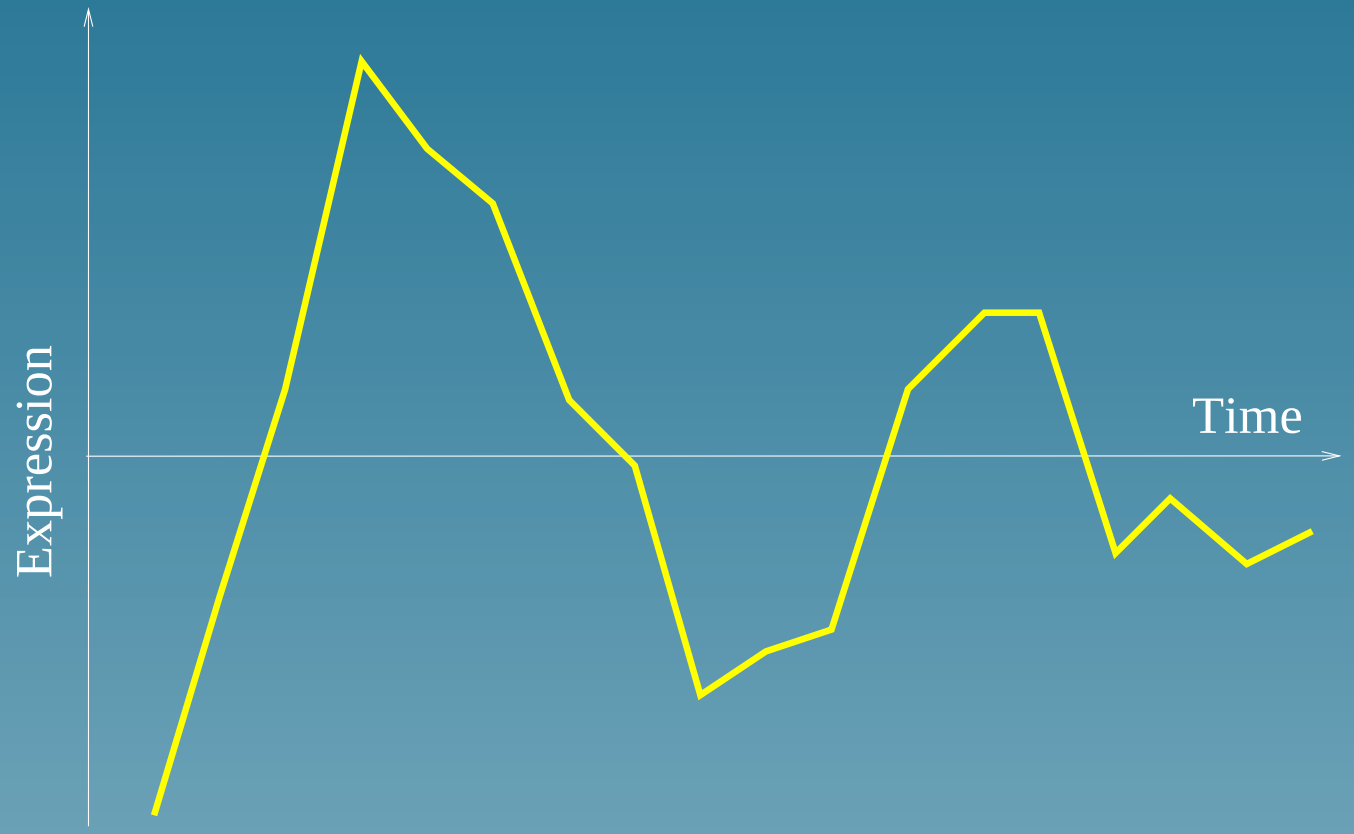
Related genes



Related genes



Second pattern



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

- On peut etendre la methodologie de krigeage (et processus Gaussien, machines a vecteurs de support, methodes a noyau...) a des espaces structures
- L'approche par analyse harmonique est tres generale pour les espaces avec une structure algebrique
- Idem pour la generalisation sur des varietes Riemannienne (analyse harmonique par calcul differentiel.
- Encore tres peu applique