Data mining the proteome in reproducible kernel Hilbert spaces

Jean-Philippe Vert

Bioinformatics Center, Kyoto University, Japan
Jean-Philippe.Vert@mines.org

Research Institute for Mathematical Sciences, Kyoto University, Japan, July 12, 2002.

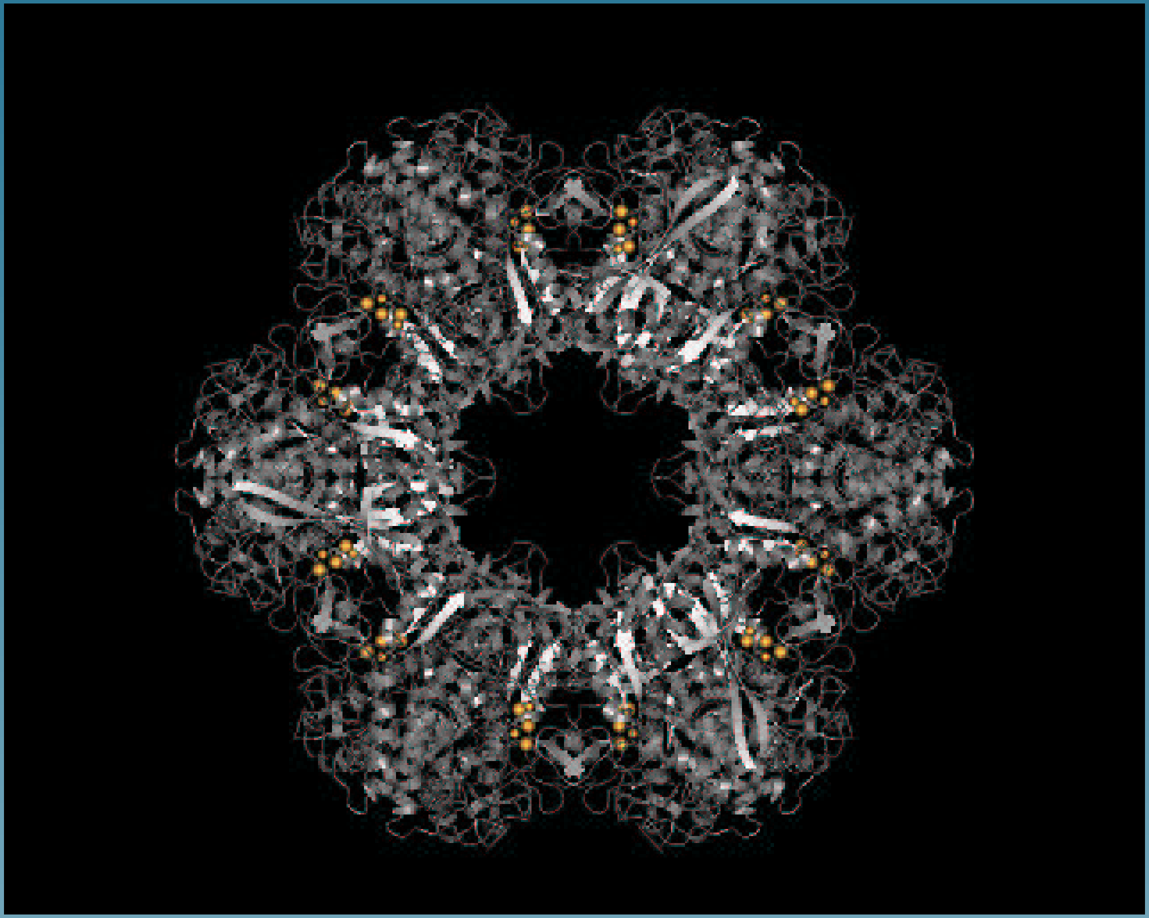
Outline

1. The proteome
2. DNA chips, pathway databases...
3. Kernels and RKHS
4. Example: correlation between microarray data and gene network

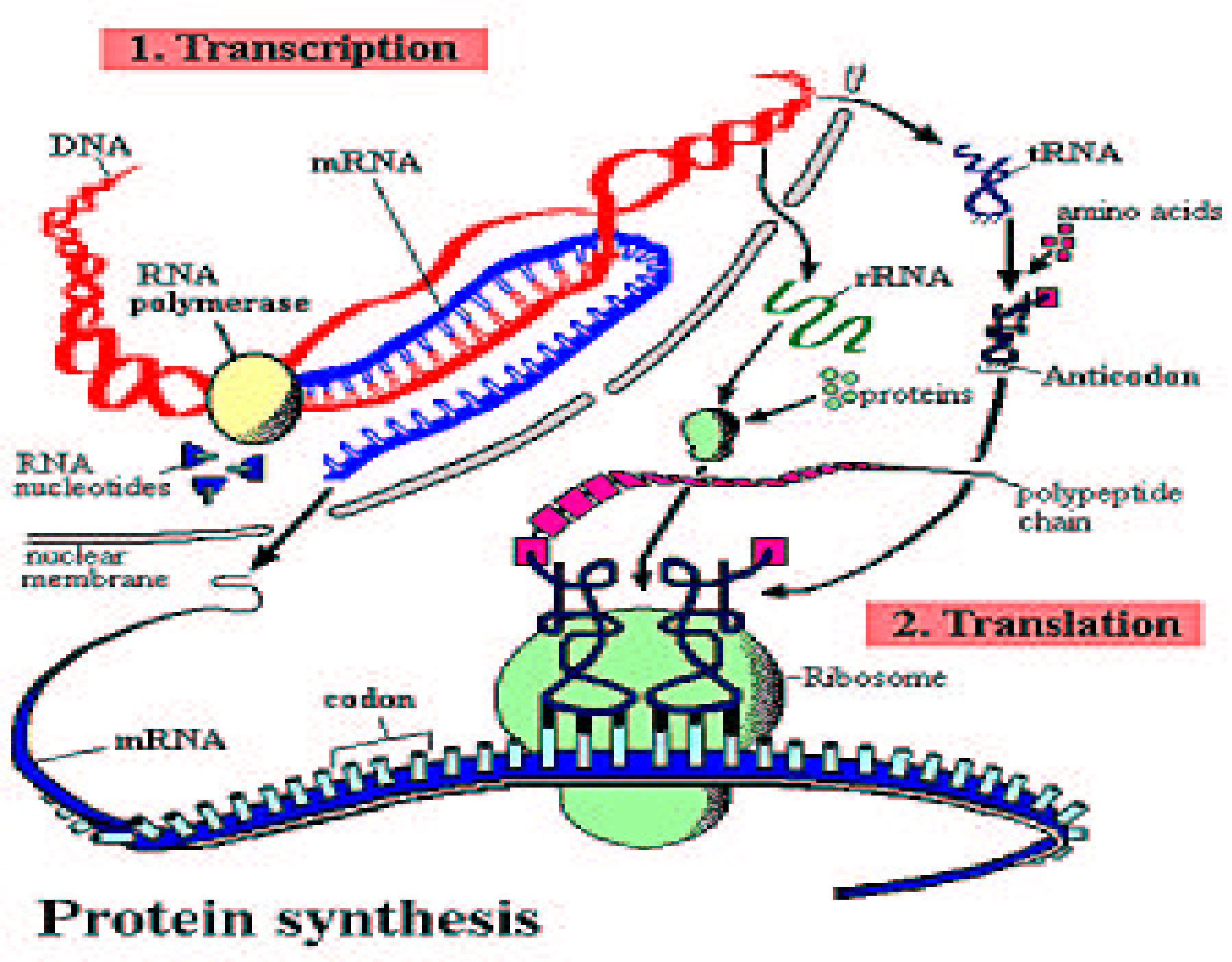
Part 1

Proteomics: a primer

A protein (glutamine synthetase)



The central dogma : DNA → RNA → protein

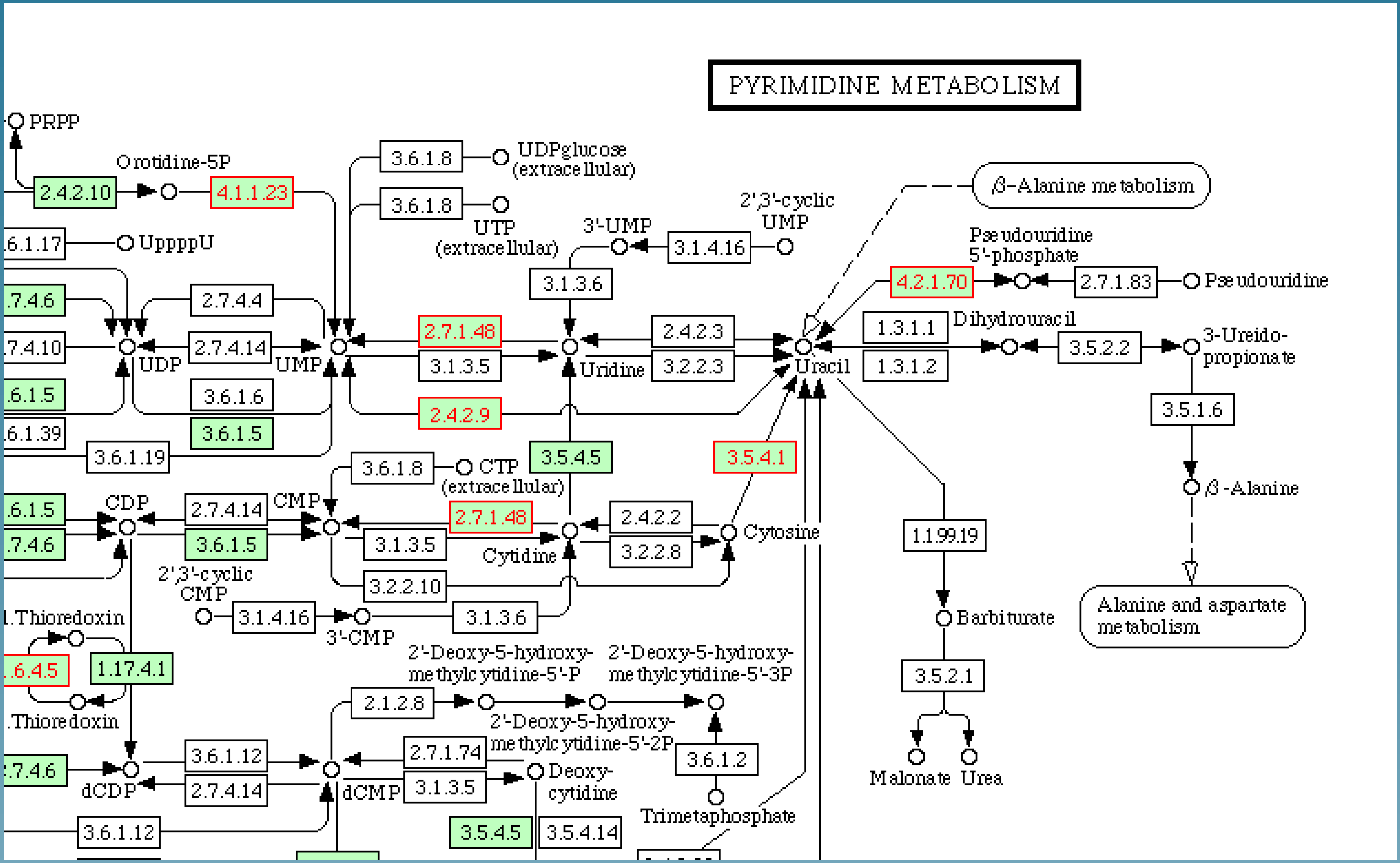


Protein synthesis

The proteome

- 6,000 genes in the budding yeast, 30-100,000 genes in humans
- complex interactions
- complex regulation
- proteins have many functions: structural, functional, ...

Proteins can catalyze chemical reactions



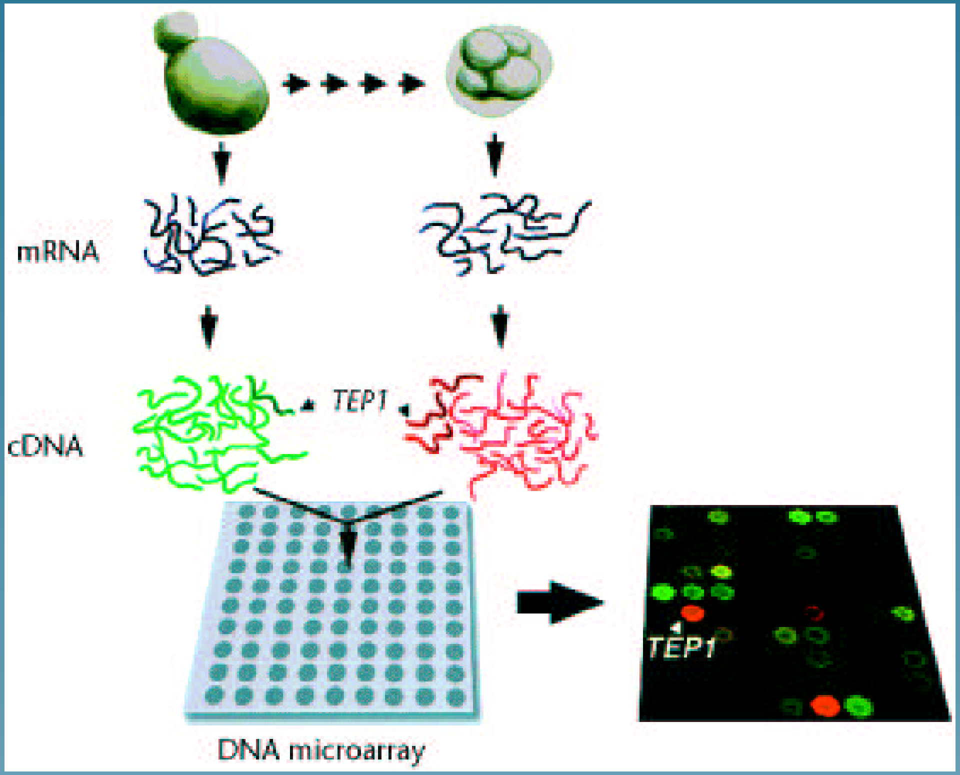
Challenges in proteomics

- Structure, functions of each gene?
- Genetic regulation? System bahaviour?
- Biology is becoming quantitative : need of mathematical frameworks to manipulate biological concepts.

Part 2

Characterizing the proteome: DNA chips, pathways etc...

Microarrays (DNA chips)



(from Brown and Botstein, Nature Genetics, 1999)

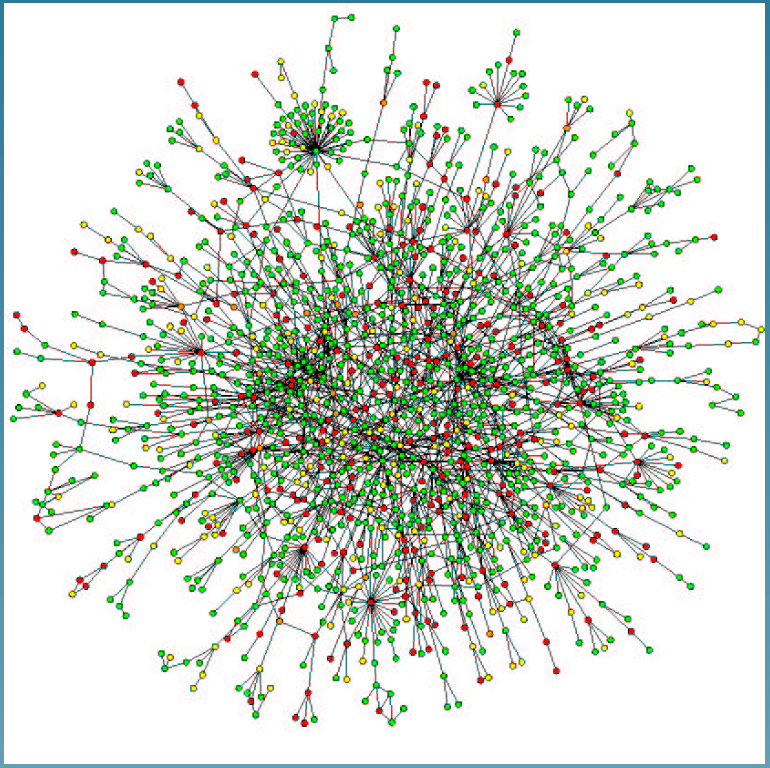
Microarrays (ctd.)

- can monitor the quantity of RNA for several thousands genes simultaneously
- quantity of data increases very fast
- each gene is characterized by an expression profile

Networks of genes

- genes are vertices of a graph
- protein interaction network (recent technology: yeast two-hybrid system...)
- pathway network: two genes are linked when they catalyse two successive reactions

Protein interaction network



(from Jeong et al., Nature 2001)

What is a gene?

- a sequence of letters: nucleotides (4 letters) or amino-acids (20 letters)
- a 3D structure
- a node in a network (protein interactions network, metabolic pathway...)
- an expression profile...

Question

How to represent the various informations about genes in a coherent and useful mathematical framework?

Part 3

Kernels and RKHS (Reproducible Kernel Hilbert Space)

Kernels on finite space

Let $\mathcal{X}$ a finite space (set of genes).

A kernel is a mapping $K : \mathcal{X}^2 \to \mathbb{R}$ such that the Gram matrix:

$$K_{x,x'} = K(x, x')$$

is positive semidefinite (all eigenvalues are ≥ 0).

(Intuition: $K(.,.)$ measures the similarity between two genes).

Mercer kernel map

A kernel K can be expressed as an inner product in a feature space:

$$K = \sum_{i=1}^{n} \lambda_i \phi_i \phi_i',$$

where $\phi_i = (\phi_i(x_1), \ldots, \phi_i(x_n))$ are eigenvectors.

Let

$$\phi(x) = \left(\sqrt{\lambda_1}\phi_1(x), \ldots, \sqrt{\lambda_n}\phi_n(x)\right)'.$$

Then $K(x_i, x_j) = \phi(x_i)'\phi(x_j)$.

RKHS

An other useful way to express a kernel as an inner product. Consider the mapping $\psi : \mathcal{X} \to \mathbb{R}^{\mathcal{X}}$ defined by:

$$\psi(x) = K(x, .).$$

and let $\mathcal{H} \subset \mathbb{R}^{\mathcal{X}}$ be the linear span of $\{K(x, .), x \in \mathcal{X}\}$.

RKHS (ctd.)

Any function $f \in \mathcal{H}$ can be expanded in the eigenvector basis of K as:

$$f = \sum_{i=r+1}^{n} a_i \phi_i .$$

where r is the multiplicity of 0 as eigenvalue.

Define an inner product in $\mathcal{H}$ as:

$$\left\langle \sum_{i=r+1}^{n} a_i \phi_i , \sum_{i=r+1}^{n} b_i \phi_i \right\rangle_{\mathcal{H}} \triangleq \sum_{i=r+1}^{n} \frac{a_i b_i}{\lambda_i} .$$

RKHS (ctd.)

Then the space $\mathcal{H}$ endowed with the inner product $< ., . >_{\mathcal{H}}$ is a Euclidean space, called Reproducible kernel Hilbert space.

Reproducing property:

$$\langle K(x_i, .), K(x_j, .)\rangle = K(x_i, x_j),$$

hence the map $x \mapsto K(x, .)$ is a valid feature space representation.

(Proof: write $K(x, .) = \sum_{i=1}^{n} \lambda_i \phi_i(x)\phi(.)$, and use the definition of the inner product with $a_i = \lambda_i \phi_i(x)$ and $b_i = \lambda_i \phi_i(x'))$

Dual representation in RKHS

Any function $f \in \mathcal{H}$ can be expressed in a dual form:

$$f(.) = \sum_{i=1}^{n} \alpha_i K(x_i, .).$$

α is the dual coordinate of $f = K\alpha$. The inner product in $\mathcal{H}$ can be easily expressed with the dual coordinates:

$$< f, g >_{\mathcal{H}} = \sum_{i,j=1}^{n} \alpha_i \beta_j K(x_i, x_j) = \alpha' K \beta.$$

What is the link between RKHS and the proteome?

- A kernel $K(x, x')$ acts as a similarity measure
- Different representation of the genes (sequences, nodes of a graph, microarray expression) lead to different notions of similarity
- These similarity can be encoded as different kernel functions
- Linear algorithms can be performed implicitly in the feature space.
- The metrics of the RKHS can correspond to useful properties

Metrics in RKHS

Let $f \in \mathcal{H}$ be decomposed in the basis of eigenvectors of K:

$$f = \sum_{i=r+1}^{n} a_i \phi_i .$$

The norm is given by:

$$||f||_{\mathcal{H}}^2 = \sum_{i=r+1}^{n} \frac{a_i^2}{\lambda_i} .$$

A large norm means that f has large components with respect to the eigenvectors with small eigenvalues.

Metrics in RKHS (ctd.)

Example: in the continuous case $(\mathcal{X} = \mathbb{R}^d)$ the eigenvectors of the Gaussian radial basis kernel:

$$K(x, x') = \exp\left(-\frac{||x - x'||^2}{2\sigma^2}\right)$$

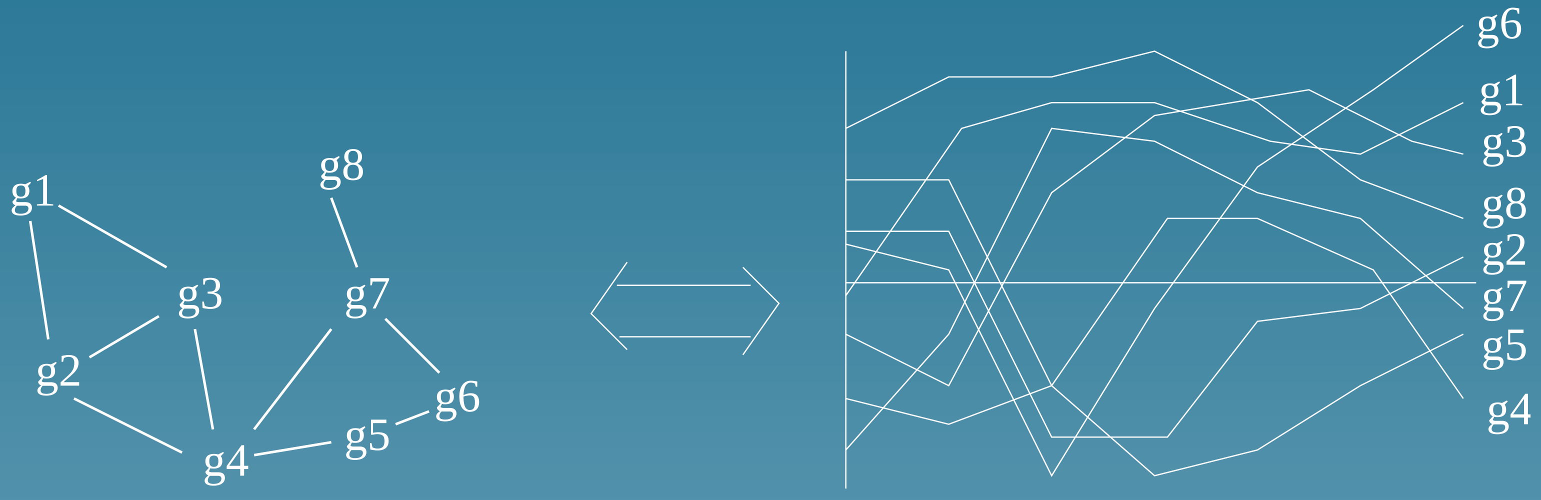
are the Fourier basis function, and the norm in $\mathcal{H}$ is a smoothing functional:

$$||f||_{\mathcal{H}} = \int_{\mathbb{R}^d} e^{\frac{\sigma^2}{2}||\omega||^2} |\hat{f}(\omega)|^2 d\omega.$$

Part 3

Example: correlation between microarray data and gene network

The problem



Gene network

Expression profiles

Are there "correlations"?

The approach

An interesting feature $f : \mathcal{X} \rightarrow \mathbb{R}$ should be:

- smooth with respect to the graph topology
- capture a lot of variations in the profiles (i.e., be strongly correlated with some the furst principal components)

This can be translated as a canonical correlation analysis (CCA) problem between two RKHS associated with two kernels.

Graph kernel

For a graph let:

- A be the adjacency matrix ($A_{i,j} = 1$ is $x_i \sim x_j$, 0 otherwise)
- D be the diagonal matrix of vertex degrees
- $L = D - A$ be the Laplacian matrix

L can be thought as a discretized version of the continuous Laplacian $\Delta = \sum \frac{\partial}{\partial x_i}$.

Graph kernel (ctd.)

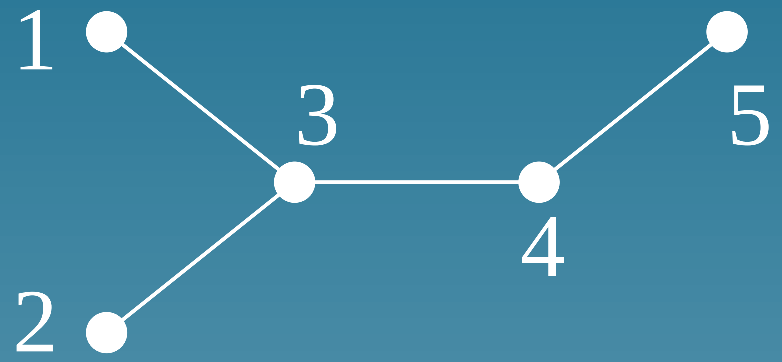
Eigenvectors of L form a Fourier basis of the functions on the vertices of the graph. Frequency increases with the eigenvalue.

By similarity with the continuous case, let

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

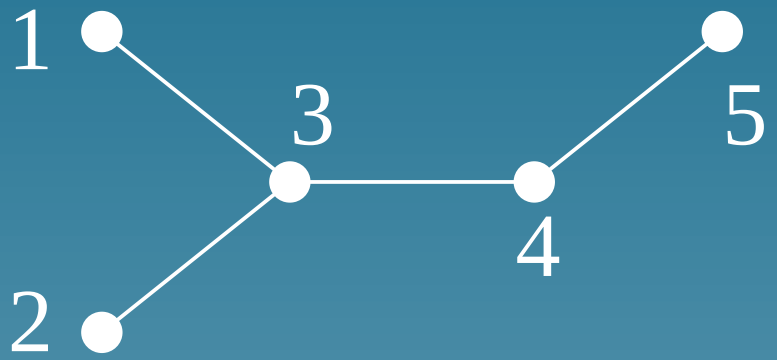
be the diffusion kernel. Its eigenvectors are the Fourier basis, the eigenvalues quickly decrease when the frequency increases. The corresponding norm $||f||_{\mathcal{H}}$ is a smoothing functional.

Example of a graph kernel (1)



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

Example of a graph kernel (2)



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

Microarray kernel

Consider the linear kernel $K(x,x') = e(x).e(x')$, where $e(x) \in \mathbb{R}^p$ is the expression profile (centered).

The corresponding RKHS is the set of linear features:

$$f_v(x) = e(x)'v,$$

for some $v \in span(e(x), x \in \mathcal{X})$. The norm in the RKHS is $||f||_{\mathcal{H}} = ||v||$, and the variance captured by f is

$$V(f_v) = \frac{\sum_{x \in \mathcal{X}} f_v(x)^2}{||v||^2} = \frac{||f_v||_{L^2(\mathcal{X})}}{||f_v||_{\mathcal{H}}}.$$

Combining both kernels

Let K_1 be the graph kernel, and K_2 be the linear kernel, with RKHS $\mathcal{H}_1$ and $\mathcal{H}_2$

The problem can be stated as: find a pair of features $(f_1, f_2) \in \mathcal{H}_1 \times \mathcal{H}_2$ such that:

- $||f_1||_{\mathcal{H}_1} / ||f_1||_{L^2(\mathcal{X})}$ be small (f_1 be smooth)
- $||f_2||_{\mathcal{H}_2} / ||f_2||_{L^2(\mathcal{X})}$ be small (f_2 capture a lot of variation in the profiles)
- f_1 and f_2 be as correlated as possible.

Problem formulation

This can be translated as follows:

$$\max_{(f_1,f_2)\in\mathcal{H}_1\times\mathcal{H}_2} \frac{f_1' f_2}{\sqrt{f_1' f_1 + \delta||f_1||_{\mathcal{H}_1}}\sqrt{f_2' f_2 + \delta||f_2||_{\mathcal{H}_2}}}$$

where δ is a regularization parameter (trade-off correlation vs. smoothness / variation captured).

Dual formulation

Working with the dual coordinates in each feature space, this is equivalent to:

$$\max_{(\alpha,\beta)\in(\mathbb{R}^{\mathcal{X}})^2} \frac{\alpha' K_1 K_2 \beta}{\left(\alpha'(K_1^2+\delta K_1)\alpha\right)^{\frac{1}{2}}\left(\beta'(K_2^2+\delta K_2)\beta\right)^{\frac{1}{2}}}$$

which is equivalent to the generalized eigenvectors problem:

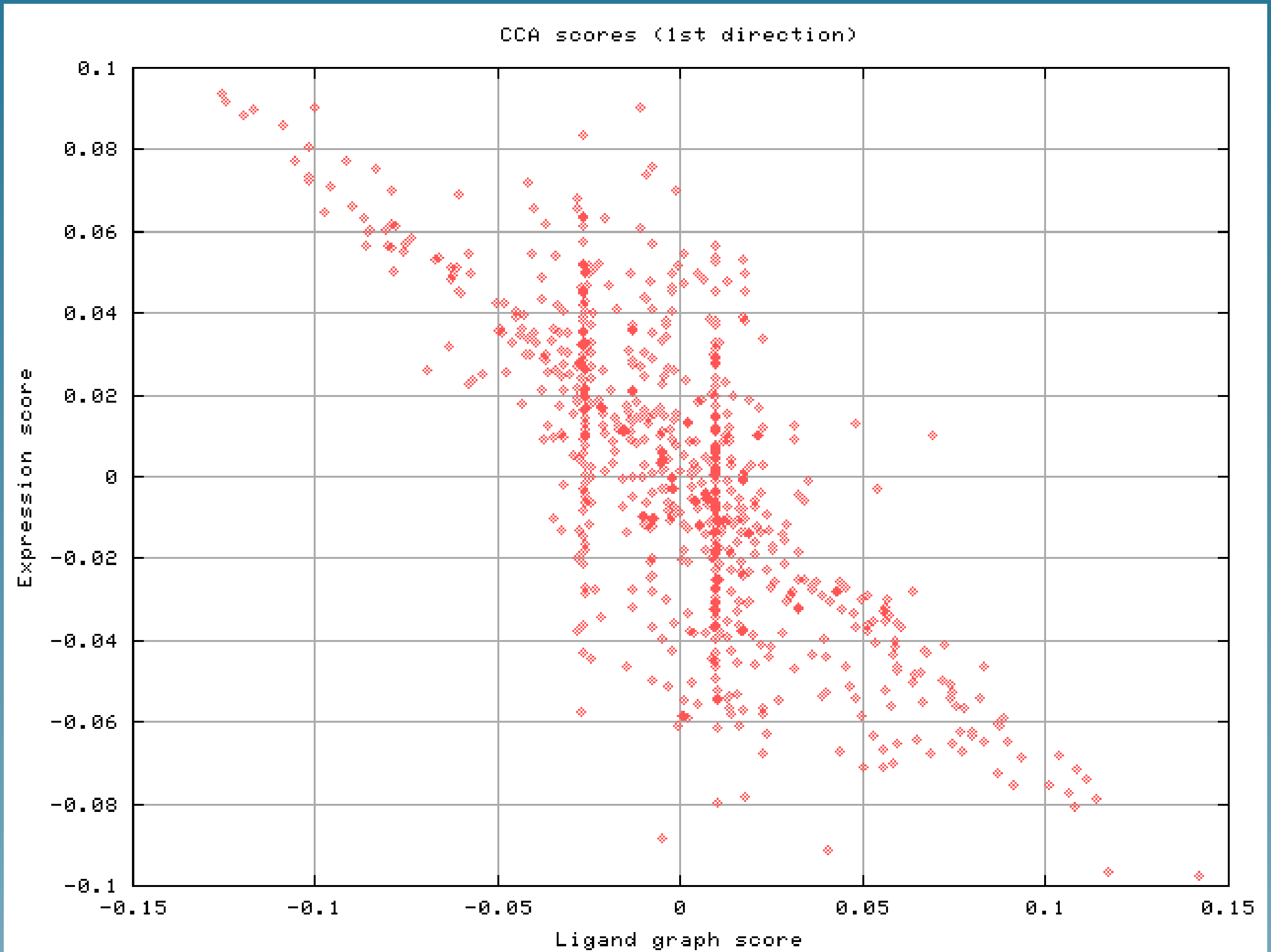
$$\begin{pmatrix} 0 & K_1K_2 \\ K_2K_1 & 0 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho \begin{pmatrix} K_1^2+\delta K_1 & 0 \\ 0 & K_2^2+\delta K_2 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}$$

Experiment

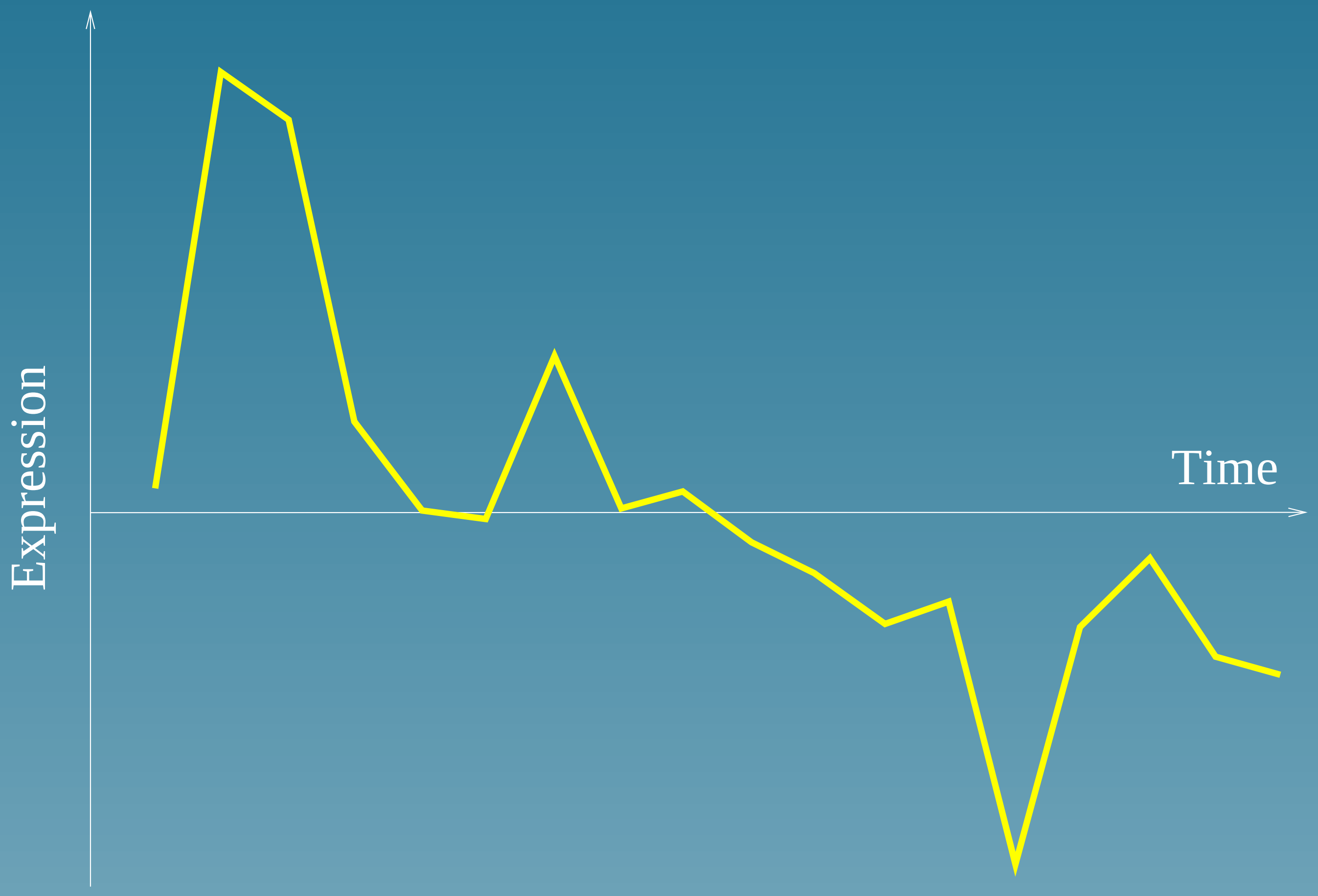
- Gene network: genes are linked if they are known to catalyse two successive reactions (data available in Kyoto University's KEGG database, `www.genome.ad.jp`)

- Microarray data: 18 measures for all genes (6,000) of the budding yeast S. Cerevisiae by Spellman et al. (public data), corresponding to a cell cyle after release of alpha factor.

1st CCA scores



Upper left expression



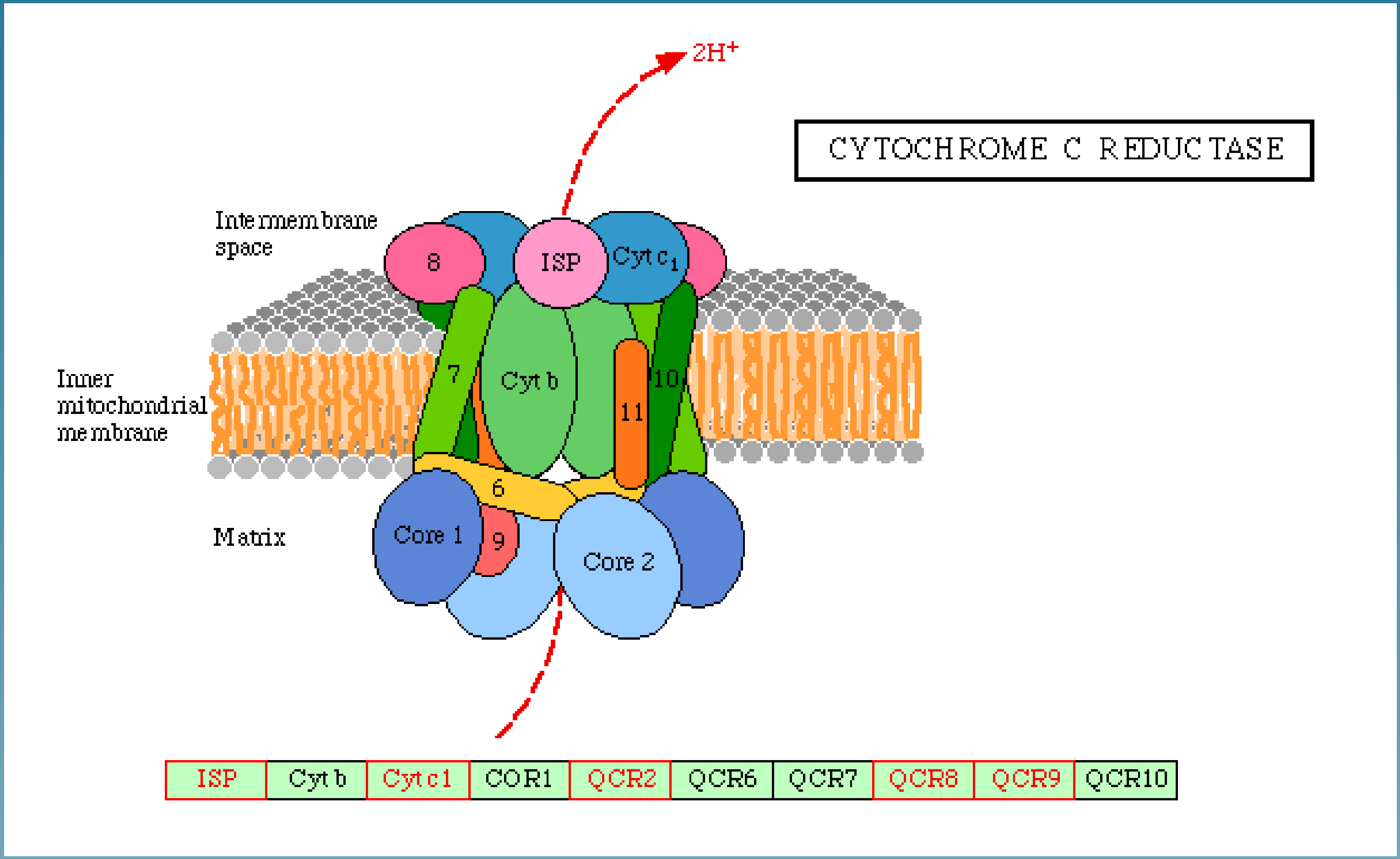
Average expression of the 50 genes with highest $s_2 - s_1$.

Upper left genes

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

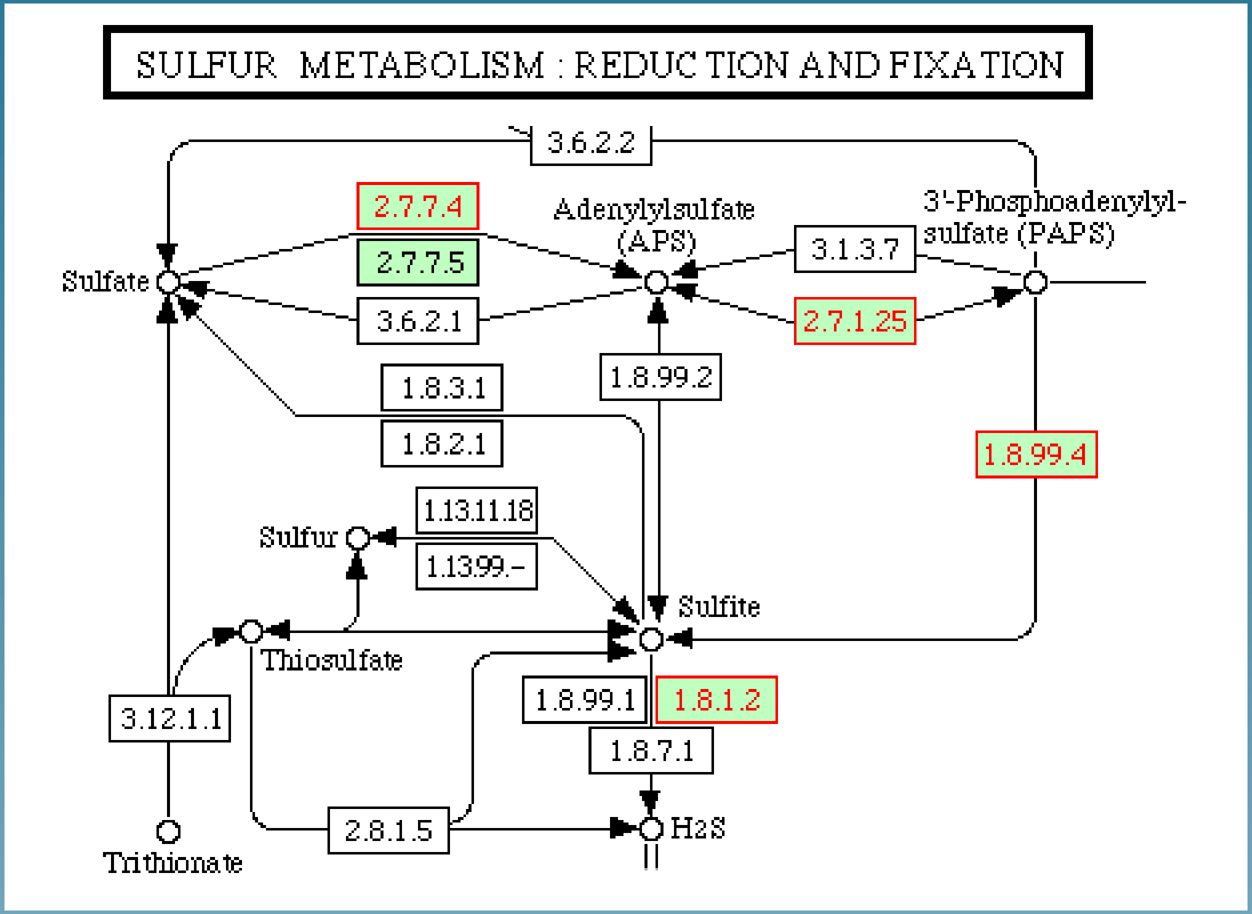
- Oxidative phosphorylation (10 genes)
- Citrate cycle (7)
- Purine metabolism (6)
- Glycerolipid metabolism (6)
- Sulfur metobolism (5)
- Selenoaminoacid metabolism (4) , etc...

Upper left genes

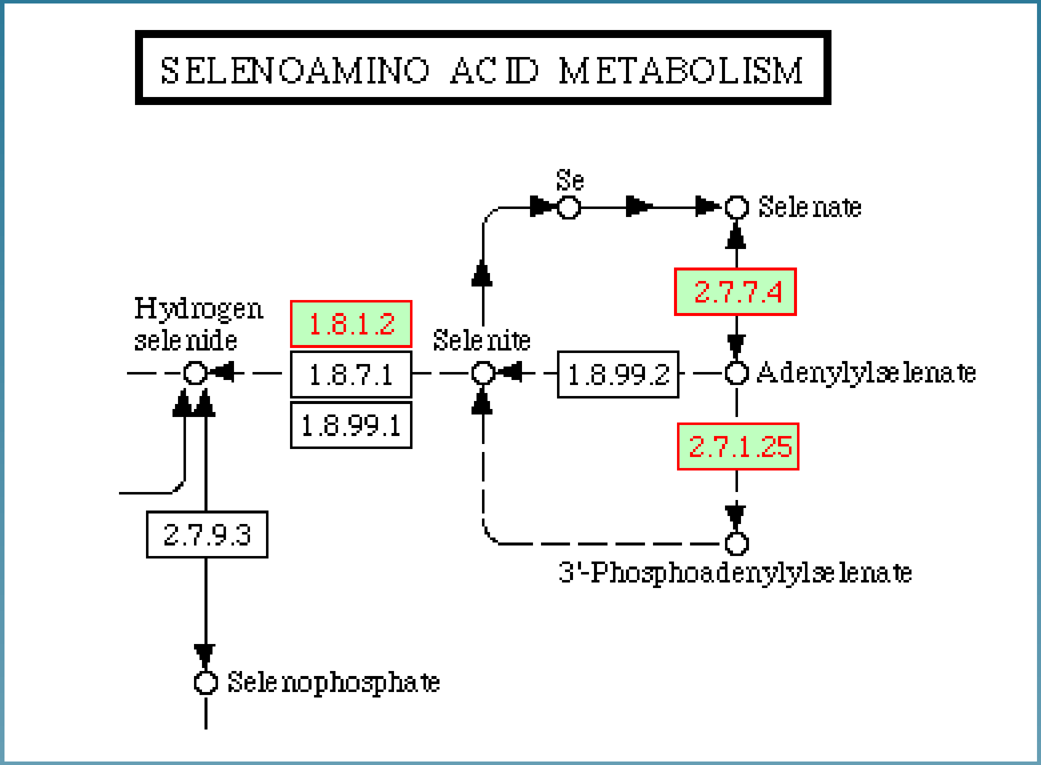


ISP	Cyt b	Cyt c1	COR1	QCR2	QCR6	QCR7	QCR8	QCR9	QCR10

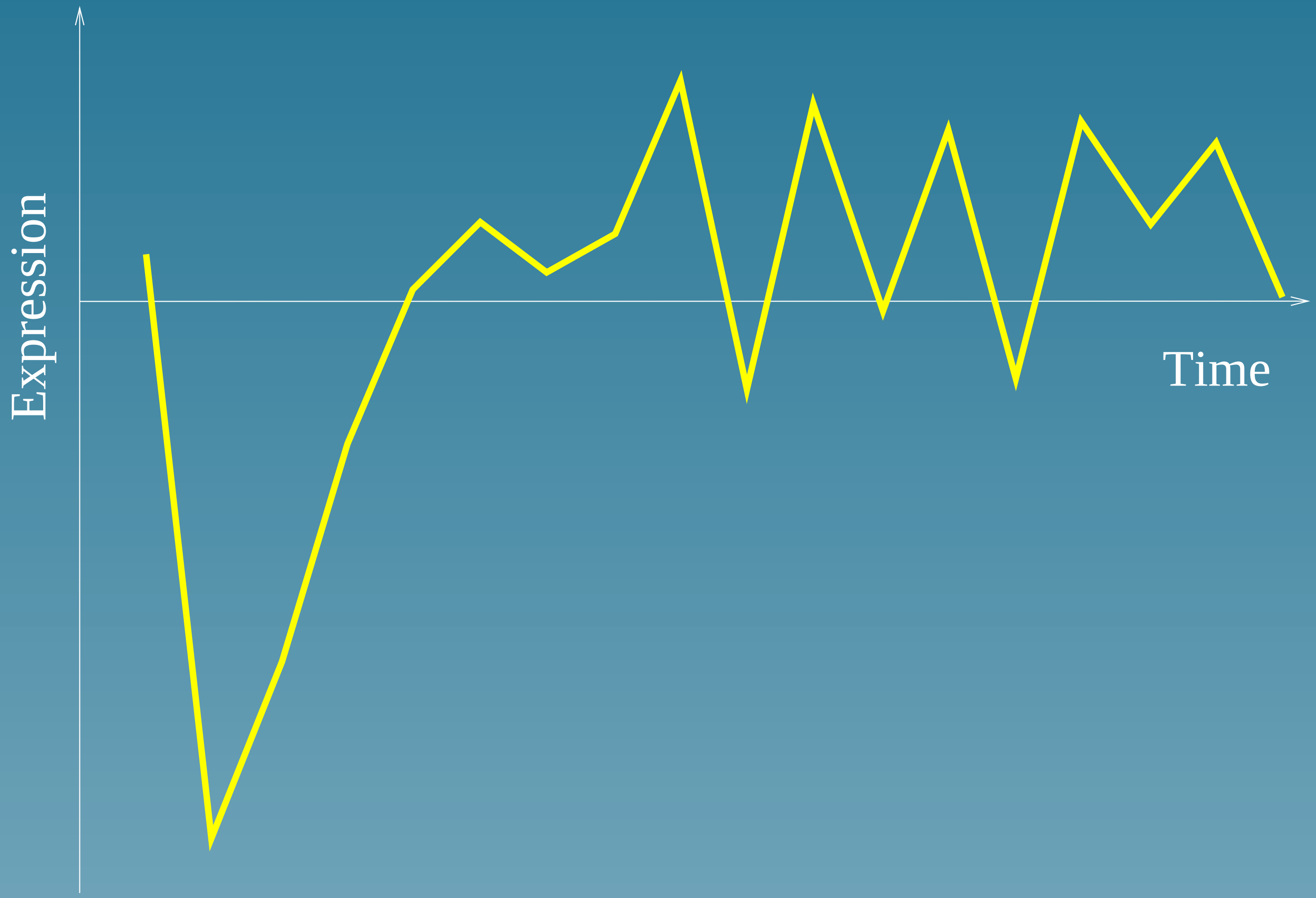
Upper left genes



Upper left genes



Lower right expression

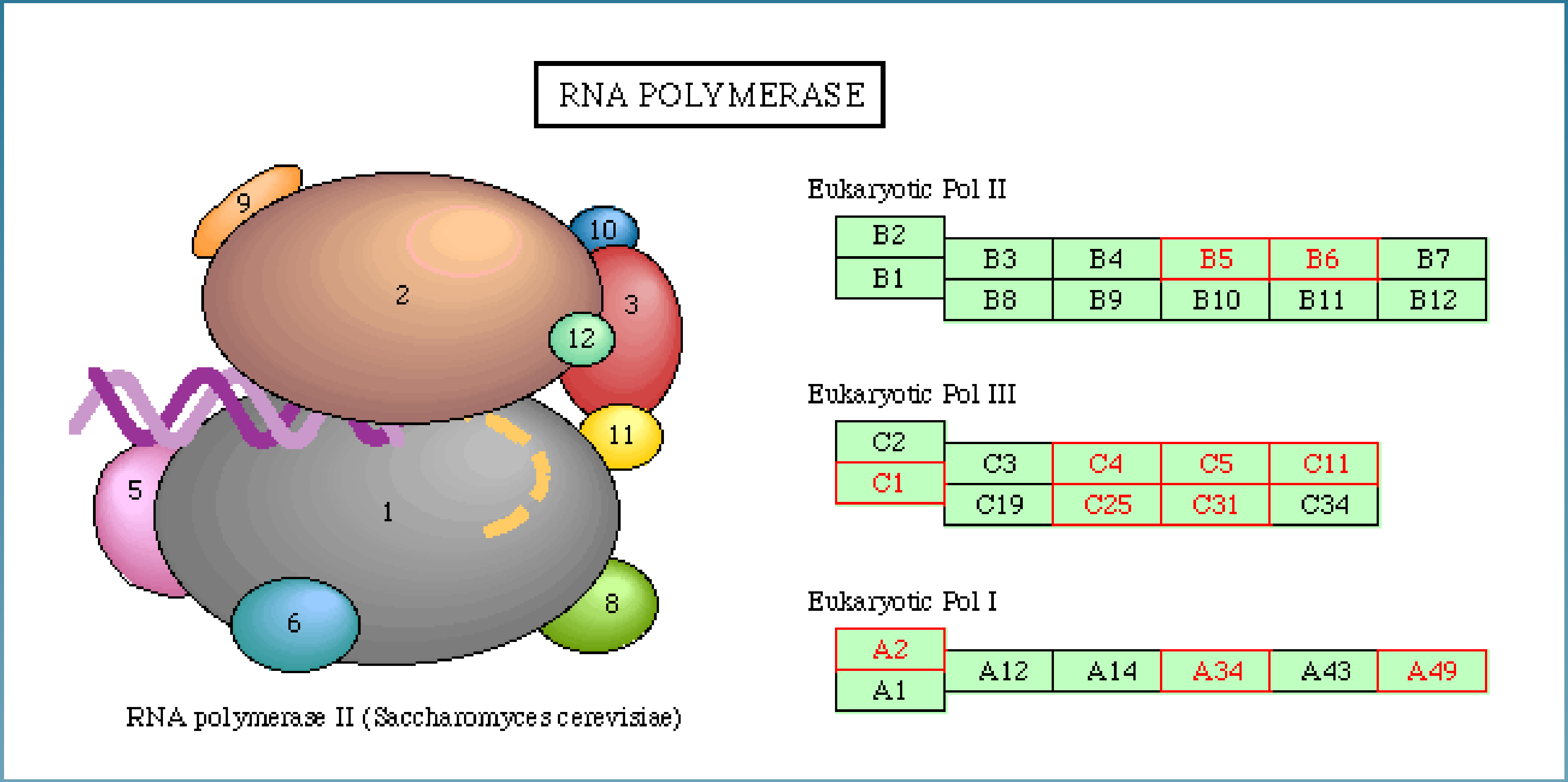


Average expression of the 50 genes with highest $s_2 - s_1$.

Lower right 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...

Lower right genes



RNA POLYMERASE

RNA polymerase II (Saccharomyces cerevisiae)

Eukaryotic Pol II

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

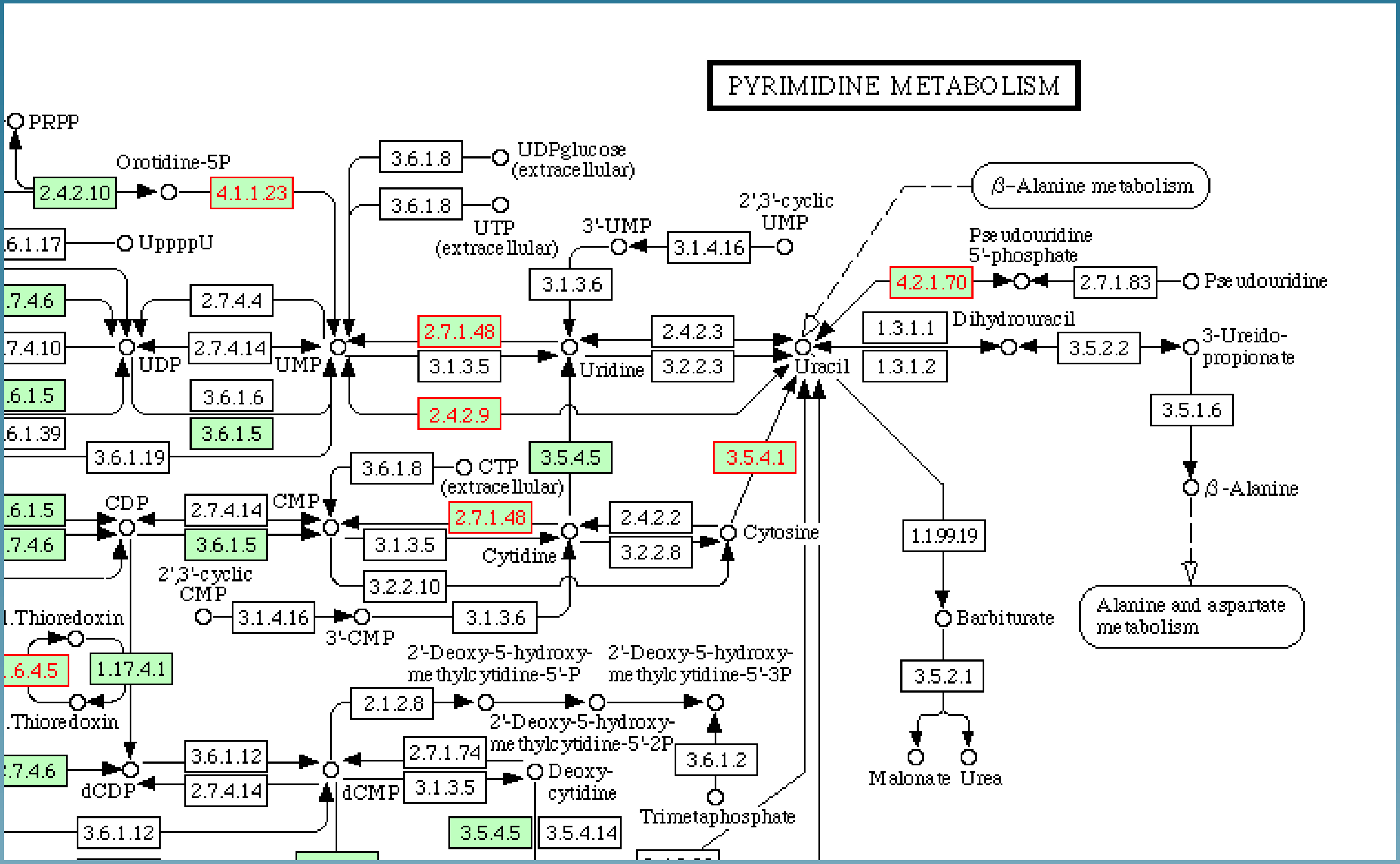
Eukaryotic Pol III

C2	C3	C4	C5	C11
C1	C19	C25	C31	C34

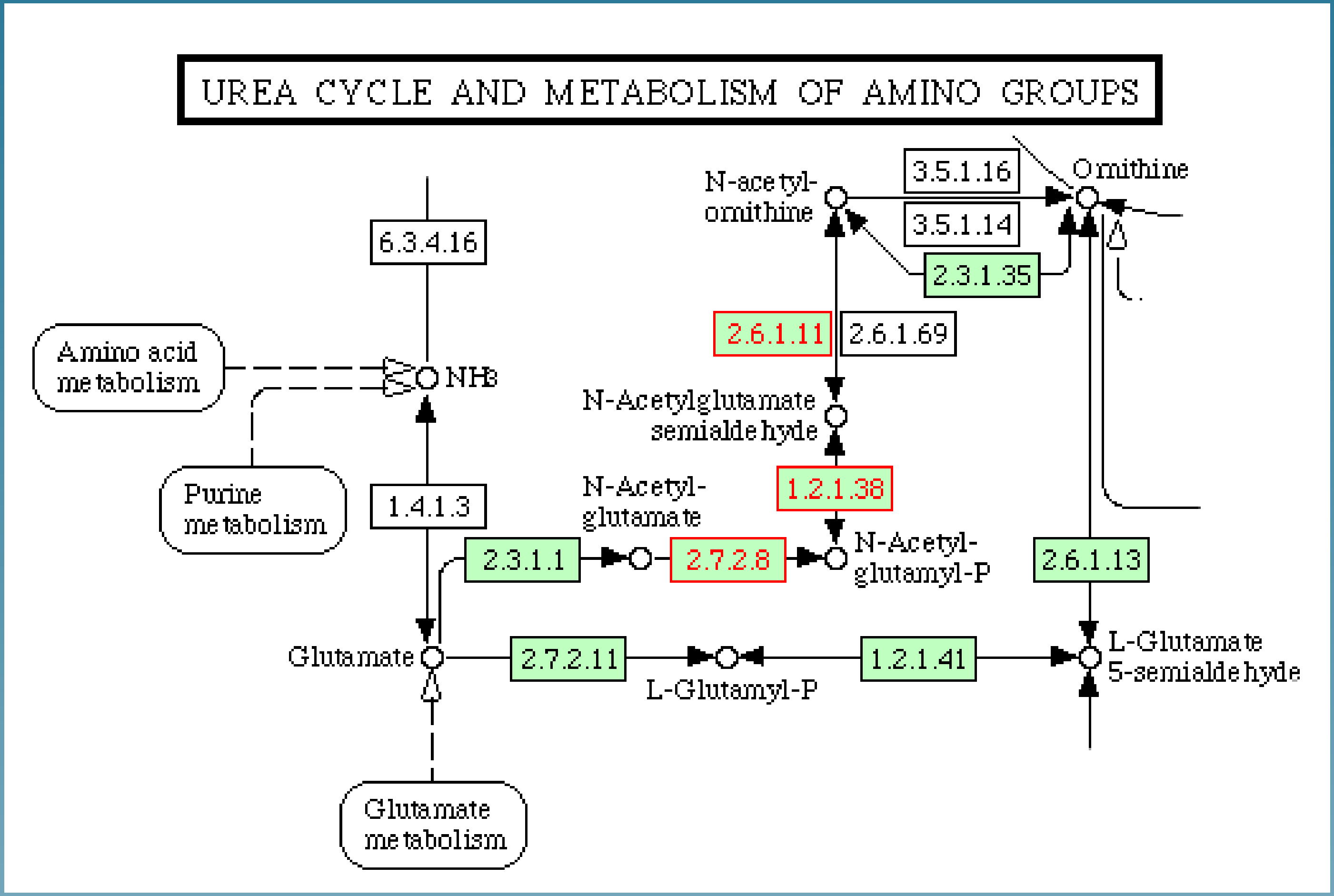
Eukaryotic Pol I

A2	A12	A14	A34	A43	A49
A1					

Lower right genes



Lower right genes



Conclusion

Conclusion

- New technologies, new data: biology is changing quickly, need for new mathematical ideas (not only in statistics)

- We proposed a way to encode different kinds of informations about genes into kernel functions, and to work in the corresponding RKHS

- This is still an over-simplified model of the reality. More interesting structures might be imagined for the proteome (the idea of gene itself is more and more controversial...)

- Thank you!