ECOLE DES MINES
DE PARIS

京都大学化学研究所
バイオインフォマティクスセンター
Bioinformatics Center
Institute for Chemical Research, Kyoto University

Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA

JEAN-PHILIPPE VERT,[a] AND MINORU KANEHISA[b]

[a]Ecole des Mines de Paris, 35 rue Saint-Honoré, 77300 Fontainebleau, France
[b]Bioinformatics Center, Kyoto University, Uji, Kyoto 611-0011, Japan

ABSTRACT

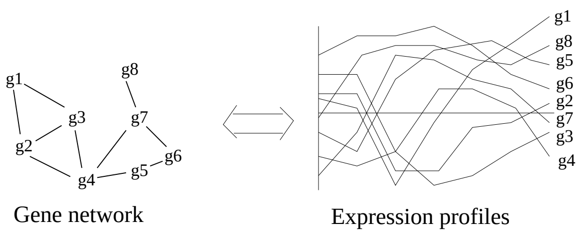
We present an algorithm to *extract features from gene expression profiles*, based on the knowledge of a graph which links together genes known to participate to successive reactions in metabolic pathways. The algorithm is motivated by the intuition that biologically relevant features are likely to exhibit *smoothness with respect to the graph topology*. It involves encoding the graph and the set of expression profiles into *kernel functions*, and performing a generalized form of canonical correlation analysis in the corresponding reproducible kernel Hilbert spaces.

1 Introduction

Microarray technology (DNA chips) is quickly becoming a major data provider in the post-genomics era, enabling the monitoring of the quantity of messenger RNA present in a cell for several thousands genes simultaneously. By submitting cells to various experimental conditions and comparing the expression profiles of different genes, a better understanding of the regulation mechanisms and functions of each gene is expected.

Independently of microarray technology, decades of research in molecular biology have characterized the roles played by many genes as catalyzing chemical reactions in the cell. This information has now been integrated into databases such as KEGG (Kanehisa et al., 2002), where series of successive chemical reactions arranged into pathways are represented, together with the genes catalyzing them.

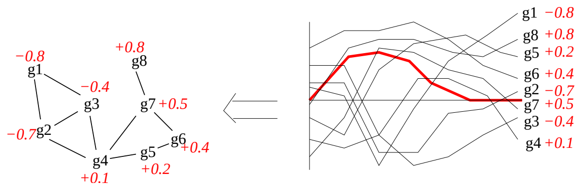
The question motivating this research is whether the knowledge of this graph can help make sense out of the expression profiles, by finding typical patterns of expression which exhibit some form of correlation with the topology of the gene network.



Gene network
Expression profiles

2 Problem formulation

A *pattern of expression* is by definition a profile (one possible pattern is shown in red on the picture below). Our goal is to find patterns of expression which correspond to biologically relevant events, such as activity of a pathway or environnmental changes. For any candidate pattern of expression, the correlation coefficient with gene expression profiles quantifies how each gene is related to the profile. Mapping the correlation coefficients onto the graph, we can check how smooth it is with respect to the graph topology. Biologically relevant patterns are likely to have smooth correlation coefficients; such patterns are called *smooth*



Moreover, a relevant pattern should be characteristic of large variations among gene expression profiles, i.e., should be correlated with the first principal components of the set of expression profiles. Such patterns are called *relevant*

The question is: how to automatically detect smooth and relevant patterns?

3 An approach using kernel methods

- Let f_1 a function defined on the set of genes. By discrete Fouriere analysis, it can be shown that the smoother f_1, the smaller the function:

$$\frac{||f_1||_{\mathcal{H}_1}}{||f||_{L^2}},$$

  where $||.||_{\mathcal{H}_1}$ is the norm in the reproducible kernel Hilbert space (R.K.H.S.) associated with a diffusion kernel (Kondor and Lafferty, 2002).

- Les f_2 be a correlation function on the set of genes, associated with a candidate pattern v. Then the relevance of v increases when the function

$$\frac{||f_2||_{\mathcal{H}_2}}{||f_2||_{L^2}}$$

  decreases, where $||.||_{\mathcal{H}_1}$ is the norm in the R.K.H.S. associated with a linear kernel (when a gene is mapped to its expression profile).

- The correlation between f_1 and f_2:

$$\frac{f_1.f_2}{||f_1||_{L^2}||f_2||_{L^2}}, \tag{1}$$

  increases when f_1 and f_2 get similar.

- In order to find an interesting pattern v, we try to find two functions f_1 and f_2 such that f_1 be as relevant as possible, f_2 be as smooth as possible, and f_1 and f_2 be as correlated as possible. This can be done by modifying (1) into the following problem:

$$\max_{(f_1,f_2)} \frac{f_1.f_2}{\left(||f_1||^2_{L^2} + \delta||f_1||^2_{\mathcal{H}_1}\right)^{1/2}\left(||f_2||^2_{L^2} + \delta||f_2||^2_{\mathcal{H}_2}\right)^{1/2}}, \tag{2}$$

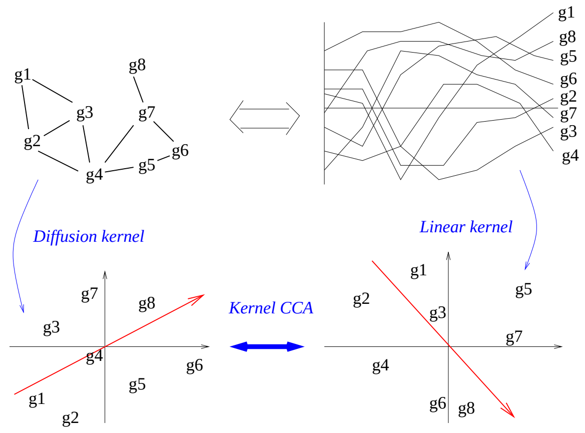
  where δ is a regularization parameter which controls the trade-off between correlation on the one hand, smoothness of f_1 and relevance of f_2 on the other hand.

- Formulated as (2) the problem appears to be a generalization of canonical correlation analysis (CCA) known as kernel-CCA, discussed in (Bach and Jordan, 2002). The authors show in particular that (3) is equivalent to the following generalized eigenvalue problem, which can be solved using classical mathematical softwares:

$$\begin{pmatrix} 0 & K_1K_2 \\ K_2K_1 & 0 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho\begin{pmatrix} K_1^2 + \delta K_1 & 0 \\ 0 & K_2^2 + \delta K_2 \end{pmatrix}\begin{pmatrix} \alpha \\ \beta \end{pmatrix}, \tag{3}$$

  where K_1 and K_2 are the kernel Gram matrices of the diffusion and the linear kernel, respectively.

  Solving (3) provides a series of pairs of functions (f, f_v), equivalent to the extraction of successive canonical directions with decreasing correlation in classical CCA.

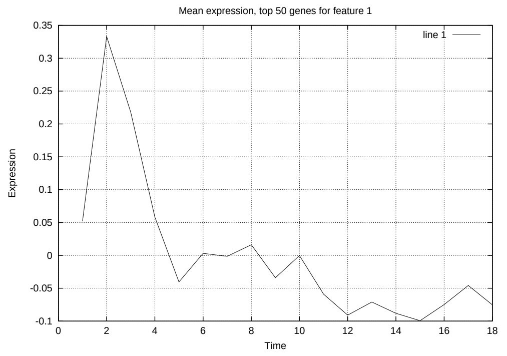


4 Experiments

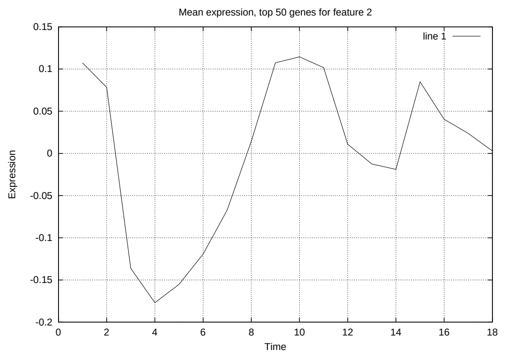
We extracted from the LIGAND database of chemical compounds of reactions in biological pathways (Kanehisa *et al.*, 2002) a graph made of 774 genes of the budding yeast *S. Cerevisiae*, linked through 16,650 edges, where two genes are linked when they catalyze two successive reactions in the LIGAND database (i.e, two reactions such that the main product of the first one be the main substrate of the second one).

We compared this graph with a set of 18 time series expression data points corresponding to two cell cycles of the yeast *S. cereviciae* after release of alpha factor. The following Figures show the first two patterns extracted.

- The first pattern is essentially a strong positive signal immediately after the beginning of the experiment. Several pathways positively correlated with this pattern are involved in energy metabolism (oxidative phosphorylation, TCA cycle, glycerolipid metabolism), while pathways negatively correlated concern mainly pathways involved in protein synthesis (aminoacyl-tRNA biosynthesis, RNA polymerase, pyrimidine metabolism). Hence the first pattern clearly detects the sudden change of environment, and the priority to fuel the start of the cell cycle with fresh energetic molecules rather than to synthesize proteins.



- The second pattern detects the progression in the cell cycle, and is correlated with cyclic pathways such as DNA duplication.



References

Bach, F., Jordan, M. (2002) Kernel independent component analysis, *Journal or Machile Learning Research*, 3, 1-48.

Chung F. (1997) *Spectral Graph Theory*, Regional Conference Series in Mathematics, number 92, AMS.

Kanehisa, M., Goto, S., Kawashima, S, Nakaya, A. (2002) The KEGG databases at GenomeNet, *Nucleic Acid Research*, 30, 42-46.

Kondor, R.I., Lafferty, J. (2002) Diffusion kernels on graphs and other discrete inputs, *Proceedings of ICML 2002*.