Extracting active metabolic pathways from gene expression data using kernel CCA

Extraction de voies métaboliques actives à partir de données d'expression à l'aide de "kernel CCA"

Jean-Philippe Vert
Ecole des Mines de Paris, Centre de géostatistique,
35 boulevard Saint-Honoré, 77305 Fontainebleau cedex, France
Jean-Philippe.Vert@mines.org

Résumé: Nous présentons un algorithme pour extraire des tendances d'activité de voies métaboliques à partir de données d'expression de gènes. L'algorithme repose sur l'idée que de telles tendances sont susceptibles d'être corrélées avec les profiles d'expression des gènes participants aux réactions concernées, qui forment un sous-graphe connexe du graphe représentant l'ensemble des voies métaboliques connues. L'algorithme consiste à encoder les profiles d'expression et les réseaux métaboliques connus dans deux fonctions noyaux, et à effectuer une forme généralisée d'analyse de corrélation canonique dans les expaces à noyau auto-reproduisants associés.
Keywords: microarray, diffusion kernel, kernel CCA, gene network.

1 Introduction

Almost every chemical reaction taking place in a living organism is catalyzed by proteins, usually synthesized by the organism itself. Evidences suggest that the biochemical activity in a cell is precisely controlled by the quantity of proteins available, which are synthesized or eliminated to ensure the correct activation or inhibition of reactions. Proteins are created from RNA, which are copies of the blueprints of the proteins on the DNA of the organism, and it is believed that the quantity of a given RNA in a cell is strongly correlated with the quantity of the corresponding proteins.

Microarray technology enables the monitoring of the quantity of RNA for virtually all proteins of an organism simultaneously. Independently, many biochemical process (known as metabolic or signalling pathways) have been characterized during decades of biochemical experiments, and recently integrated into databases. The KEGG database (Kanehisa *et al.*, 2002) contains for instance the list of all known reactions arranged into pathways (series of reactions taking place one after another), together with the genes which catalyze each reaction. The algorithm we present in the sequel aims at comparing series of microarray expression data with such a pathway database, in order to extract typical patterns of expression which are likely to correspond to actual biochemical events.

2 The problem

The set of genes of a given organism is represented by a discrete set $\mathcal{X}$ of cardinality $|\mathcal{X}| = n$. The set of expression profiles is a mapping $e : \mathcal{X} \rightarrow \mathbb{R}^p$, where p is the number of measurements and $e(x)$ is the expression profile of gene x. In the sequel we assume that the set of profiles has been centered, i.e., $\sum_{x \in \mathcal{X}} e(x) = 0$, and scaled to unit norm ($\forall x \in \mathcal{X}, ||e(x)|| = 1$).

The pathway database defines a network of genes represented by a graph $\Gamma = (\mathcal{X}, \mathcal{E})$, where two genes are linked whenever then catalyze two successive reactions in the pathway database.

A pattern of expression is a profile $v \in \mathbb{R}^p$. Our goal is to find patterns of expression likely to correspond to biochemical events, more precisely to represent the activity level of particular pathways. To this end we use the biological intuition that for such a pattern v, the quantities of proteins (and of the corresponding RNA) catalyzing the reactions involved are likely to be correlated or anticorrelated with v. For a candidate pattern v let us therefore call $f_v(x) \triangleq v.e(x)$ the correlation between v and $e(x)$. As the genes involved in a given pathway typically form a small connected subgraph of the graph Γ (because they catalyze successive reactions), a pattern v corresponding to the activity level of a pathway could be recognized by the fact that the values of the function $f_v(.)$ are likely be particularly positive or negative in particular connected subparts of the graph. As biochemical events such as response to changes in the environnement or cell growth usually involve several pathways, one way to recognize a relevant pattern v is therefore by the fact that $f_v(.)$ might be particularly *smooth* with respect to the graph topology, at least be smoother than a function $f_{v'}(.)$ where v' is not related to any biochemical process.

On the other hand, the smoothness of $f_v(.)$ on Γ is not a sufficient criterion to ensure that v is relevant. Indeed, one should also ensure that v be reasonably correlated with the directions of large variations between profiles, otherwise smooth functionals can be obtained artificially (consider for example the case $p > n$).

3 Approach using reproducible kernel Hilbert spaces (RKHS)

Any pattern of interest can be written as a linear combination of profiles, $v = \sum_{x \in \mathcal{X}} \alpha_x e(x)$. If we denote by K_1 the linear kernel matrix $K_1(x, y) = e(x).e(y)$, then a simple computation shows that the quantity of variation among profiles captured by the function f_v is:

$$V(f_v) \triangleq \frac{\sum_{x \in \mathcal{X}} f_v(x)^2}{||v||^2} = \frac{\alpha' K_1^2 \alpha}{\alpha' K_1 \alpha} = \frac{||f_v||^2_{L^2(\mathcal{X})}}{||f_v||^2_{\mathcal{H}_1}},$$

where $||f_v||_{\mathcal{H}_1}$ denotes the norm of $f_v(.)$ in the RKHS $\mathcal{H}_1$ defined by K_1. To ensure that v is as correlated as possible with the first principal components of the set of profiles, it is therefore enough to impose that $||f_v||_{\mathcal{H}_1}/||f_v||_{L^2(\mathcal{X})}$ be as small as possible.

In order to quantify the smoothness of a function on the nodes of the graph, let us now consider the diffusion kernel K_2 defined by $K_2 = \exp(-\tau L)$, where L is the graph Laplacian of Γ, τ is a parameter and exp(.) denotes the matrix exponential (Kondor and Lafferty, 2002). The Laplacian matrix itself is defined by $L_{x,y} = -1$ if there is an edge between x and y, 0 otherwise (for $x \neq y$), and $L_{x,x}$ is the degree of x in Γ. It is known in spectral graph theory that the discrete Laplacian shares many properties with the continuous Laplacian on a Riemannian manifold. It is symmetric, positive semidefinite, and singular. If $\{\phi_i, i = 1, \ldots, n\}$ denotes an orthonormal set of eigenvectors of L with eigenvalues $0 = \lambda_1 \leq \ldots \leq \lambda_n$, it is known that ϕ_i oscillates more and more on the graph as i increases and is called a Fourier basis (Chung, 1997). The kernel K_2 has the same eigenvectors as L, but eigenvalues $1 = \exp(-\lambda_1) \geq \ldots \geq \exp(-\lambda_n)$. Hence the norm of a function $f \in \mathcal{X}^{\mathbb{R}}$ in the RKHS $\mathcal{H}_2$ defined by the diffusion kernel K_2 is given by:

$$||f||^2_{\mathcal{H}_2} = \sum_{i=1}^{n} \hat{f}_i^2 \exp(\tau \lambda_i), \tag{1}$$

where $\hat{f}$ is the discrete Fourier transform ($\forall i \in [1, n], \hat{f}_i = f.\phi_i$). This shows that the energy at high frequency is strongly penalized by this norm, and that $||.||_{\mathcal{H}_2}$ is a smoothness functional. In other words, scaling f to unit norm, we finally get that the smoother a function f, the smaller $||f||_{\mathcal{H}_2}/||f||_{L^2(\mathcal{X})}$.

Let us now come back to the problem of finding a pattern v such that f_v be smooth on the graph and capture a lot of variation between profiles. For any function $f \in \mathcal{X}^{\mathbb{R}}$ on the graph, the correlation between f and f_v is given by:

$$\frac{f_v.f}{||f_v||_{L^2(\mathcal{X})}||f||_{L^2(\mathcal{X})}}, \tag{2}$$

which is maximized for any pair $f = f_v$. An indirect way to find a function f_v smooth and which captures a lot of variation in the same time is to constraint f to be smooth, f_v to capture a lot of variation, and f and f_v to be as correlated as possible. From the previous paragraphs this can be done by modifying (2) and considering the following problem:

$$\max_{(f,f_v)} \frac{f_v.f}{\left(||f_v||^2_{L^2(\mathcal{X})} + \delta ||f_v||^2_{\mathcal{H}_1}\right)^{1/2} \left(||f||^2_{L^2(\mathcal{X})} + \delta ||f||^2_{\mathcal{H}_2}\right)^{1/2}}, \tag{3}$$

where δ is a regularization parameter which controls the trade-off between correlation on the one hand, smoothness of f and statistical relevance of f_v on the other hand. Formulated as (3) 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:

$$\begin{pmatrix} 0 & K_1 K_2 \\ K_2 K_1 & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho \begin{pmatrix} K_1^2 + \delta K_1 & 0 \\ 0 & K_2^2 + \delta K_2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} \tag{4}$$

Solving (4) 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. Figure 1 shows 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.

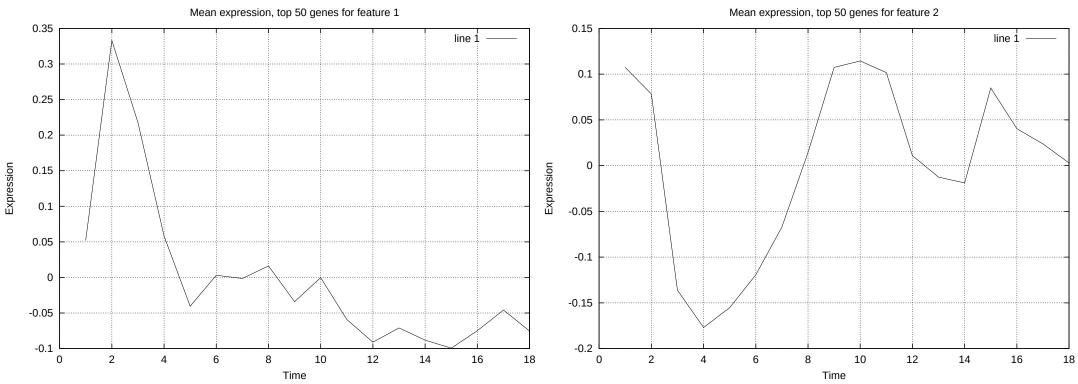

Figure 1: First 2 extracted patterns

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.*