Extracting active metabolic pathways from gene expression data

Jean-Philippe Vert

Bioinformatics group
Ecole des Mines de Paris

Jean-Philippe.Vert@mines.org

NIPS Bioinformatics workshop, December 12, 2002, Whistler, Canada.
Overview

1. Problem Formulation
2. An approach using RKHS
3. Experimental results
Part 1

Problem formulation
Genes encode proteins which can catalyse chemical reactions

Nicotinamide Mononucleotide Adenylyltransferase With Bound Nad+
Chemical reactions are often parts of pathways

From http://www.genome.ad.jp/kegg/pathway
Microarray technology monitors RNA quantity

(From Spellman et al., 1998)
Comparing gene expression and protein network

Gene network

Expression profiles

Are there "correlations"?
In yellow: a candidate pattern, and the correlation coefficient with each gene profile.
The correlation function with interesting patterns should vary smoothly on the graph.
Pattern relevance

- Interesting patterns involve many genes
- The projection of profiles onto an interesting pattern should capture a lot of variations among profiles
- Relevant patterns can be found by PCA
Problem

Find patterns of expression which are simultaneously

- smooth
- relevant
Part 3

An approach using RKHS
Kernels and RKHS

- A kernel $K(x, y)$ on the set of genes $\mathcal{X}$ is a symmetric positive definite function.

- A kernel defines a Hilbert space on the set of functions:

  $$H = \text{span}\{K(x, .), x \in \mathcal{X}\} \subset \{f : \mathcal{X} \to \mathbb{R}\}$$

  called reproducible kernel Hilbert space (RKHS).

- The norm $\|f\|_H$ can have useful interpretation for particular kernels.
Pattern relevance

- Let $e(x)$ the profile of gene $x$

- Let $K_1(x, y) = e(x) . e(y)$ be the linear kernel, with RKHS $H_1$.

- The norm $\| \cdot \|_{H_1}$ is a relevance functional: the relevance of $f \in H_1$ increases when the following decreases:

$$\frac{\| f \|_{H_1}}{\| f \|_{L_2}}$$
Pattern smoothness

- Let $K_2(x, y)$ be the diffusion kernel obtained from the gene network, with RKHS $H_2$.
- It can be considered as a discretized version of a Gaussian kernel (solving the heat equation with the graph Laplacian).
- The norm $\| \cdot \|_{H_2}$ is a smoothness functional: the smoother a function $f : \mathcal{X} \rightarrow \mathbb{R}$, the larger the function:

$$\frac{\| f \|_{H_1}}{\| f \|_{L_2}}$$
Diffusion kernel (Kondor and Lafferty, 2002)

\[-L = \left( \begin{array}{cccccc} -1 & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 1 & 0 & 0 & 0 \\ 1 & 1 & -3 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 \end{array} \right)\]
Diffusion kernel (Kondor and Lafferty, 2002)

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

Find a linear function $f_1$ and a function $f_2$ such that:

- $f_1$ be relevant: $\|f_1\|_{L^2}/\|f_1\|_{H_1}$ be large

- $f_2$ be smooth: $\|f_2\|_{L^2}/\|f_2\|_{H_2}$ be large

- $f_1$ and $f_2$ be correlated:

$$\frac{f_1 \cdot f_2}{\|f_1\|_{L^2} \|f_2\|_{L^2}}$$

be large
Problem reformulation (2)

The three goals can be combined in the following problem:

$$\max_{f_1, f_2} \frac{f_1 \cdot f_2}{\left(\|f_1\|_{L^2}^2 + \delta \|f_1\|_{H_1}^2\right)^{\frac{1}{2}} \left(\|f_2\|_{L^2}^2 + \delta \|f_2\|_{H_2}^2\right)^{\frac{1}{2}}}$$

where the parameter $\delta$ controls the trade-off between relevance/smoothness on the one hand, correlation on the other hand.
Solving the problem

This formulation is equivalent to a generalized form of CCA (Kernel-CCA, Bach and Jordan, 2002), which is equivalent to the following generalized eigenvector problem

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

Kernel CCA

Diffusion kernel

Linear kernel

Kernel CCA
Part 4

Experimental results
Data

- **Gene network**: two genes are linked if they catalyze successive reactions in the KEGG database.

- **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

![Diagram of Selenoamino Acid Metabolism](image-url)
Opposite pattern
Related genes

- 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...
Related genes
Related genes
Related genes
Conclusion
Conclusion

- Heterogeneous data can be integrated with kernels
- The approach can be generalized (non-linear kernel for gene expression, string kernels...)

Kernel Methods in Bioinformatics
Harnack-Haus, Berlin, April 14, 2003
http://cg.ensmp.fr/vert/kmb03