Extracting correlations between pathways and microarray data

Jean-Philippe Vert

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

4th Biopathways Consortium Meeting, August 1-2, 2002, Edmonton, Canada
Outline

1. Problem formulation
2. An approach using kernel methods
3. Experimental results
Part 1

Problem formulation
The problem

Gene network

Expression profiles

Are there “correlations”? 

What is a correlation?

- “Patterns” of expression shared by genes closed to each others on the network

- Examples:
  - **Activation of a pathway:** enzymes which catalyze successive reactions might share a particular expression pattern
  - **Formation of a protein complex:** the co-expression of several genes closed to each other on a protein interaction network is required.
• An expression pattern is a particular expression profile.

• The correlation between a pattern and a gene expression profile quantifies how each gene shares the profile.
A pattern whose correlation varies smoothly with respect to the graph topology is an interesting pattern.
Part 2

An approach using kernel methods
Overview

- We have developed an algorithm to extract expression patterns smooth with respect to a network topology.

- Based on recent developments in the field of kernel methods (SVM...).

- **Input**: a gene network and a set of expression profiles.

- **Output**: a set of interesting expression patterns, and the groups of genes which share it or not.
The idea
Kernel

For any mapping $\Phi(\cdot)$ from the set of genes to a Euclidean space $\mathbb{R}^n$, the kernel $K(g, g')$ between two genes is the inner product between their images:

$$K(g, g') = \Phi(g) \cdot \Phi(g').$$
**Diffusion kernel (Kondor and Lafferty, 2002)**

- For a given graph, there is a natural mapping $\Phi$ to a (high dimensional) Euclidean space which conserves the topology of the graph.

- The corresponding kernel $K(g, g')$ between any two genes can be computed by:

$$K = \exp(D - A),$$

where $A$ is the adjacency matrix and $D$ the degree diagonal matrix.
Example of a diffusion 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)

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

- Expression profiles are vectors
- The inner product between two profiles is a valid kernel
Kernel CCA (Bach and Jordan, 2002)

- Let $K_1$ be the graph kernel, and $K_2$ be the expression kernel (corresponding to mapping the genes to two Euclidean spaces)

- Finding directions with large correlations is equivalent to solving the generalized eigenvalue problem:

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

Experimental results
Data

- **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 cycle after release of alpha factor.
1st CCA scores
Average expression of the 50 genes with highest $s_2 - s_1$. 

**Upper left expression**
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 metabolism (5)
- Selenoaminoacid metabolism (4), etc…
Upper left genes
Upper left genes
Upper left genes
Average expression of the 50 genes with highest $s_2 - s_1$. 

Lower right expression
Lower right 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...
Lower right genes
Lower right genes
Lower right genes
Conclusion
Conclusion

• A method to extract correlations between microarray data and a gene network

• Accepts noise and errors in the data

• Can be generalized to other types of information by using other kernels (e.g., string kernels to find correlations with sequences)