Classification of gene expression data with gene networks

Jean-Philippe Vert
Jean-Philippe.Vert@ensmp.fr

Centre for Computational Biology
Ecole des Mines de Paris, ParisTech

Systems Biology Group seminar, Insitut Pasteur, May 11th, 2007
Outline

1. Motivation
2. Using gene networks as prior knowledge
3. Application
4. Conclusion
Outline

1. Motivation
2. Using gene networks as prior knowledge
3. Application
4. Conclusion
Outline

1. Motivation
2. Using gene networks as prior knowledge
3. Application
4. Conclusion
Outline

1 Motivation

2 Using gene networks as prior knowledge

3 Application

4 Conclusion
Outline

1. Motivation
2. Using gene networks as prior knowledge
3. Application
4. Conclusion
Tissue profiling with DNA chips

- Gene expression measures for more than 10k genes
- Measured typically on less than 100 samples of two (or more) different classes (e.g., different tumors)
Tissue classification from microarray data

**Goal**
- Design a classifier to automatically assign a class to future samples from their expression profile
- Interpret biologically the differences between the classes
The approach

- Each sample is represented by a vector $x = (x_1, \ldots, x_p)$ where $p > 10^5$ is the number of probes.
- **Classification**: given the set of labeled sample, learn a linear decision function:

$$f_\beta(x) = \sum_{i=1}^{p} \beta_i x_i + \beta_0,$$

that is positive for one class, negative for the other.
- **Interpretation**: the weight $\beta_i$ quantifies the influence of gene $i$ for the classification.
Estimate the weights $\beta_i$ by **minimizing an empirical error** on the training set:

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} l(f_\beta(x_i), y_i),$$

where $l(y, f(x))$ is a loss function.

**Pitfalls**

- Statistics does not apply (?): 100 samples in $10^5$ dimensions!
- It is necessary to reduce the complexity of the problem with prior knowledge.
Empirical risk minimization

Estimate the weights $\beta_i$ by minimizing an empirical error on the training set:

$$\min_{\beta \in \mathbb{R}^{p+1}} \frac{1}{n} \sum_{i=1}^{n} l(f_{\beta}(x_i), y_i),$$

where $l(y, f(x))$ is a loss function.

Pitfalls

- **Statistics does not apply (??):** 100 samples in $10^5$ dimensions!
- It is necessary to reduce the complexity of the problem with prior knowledge.
Example: Norm Constraints

The approach

A common method in statistics to learn with few samples in high dimension is to constrain the Euclidean norm of $\beta$

$$\| \beta \|_2^2 = \sum_{i=1}^{p} \beta_i^2,$$

(ridge regression, support vector machines...)

Pros
- Good performance in classification

Cons
- Limited interpretation (small weights)
- No prior biological knowledge
**Example : Feature Selection**

**The approach**

Constrain most weights to be 0, i.e., select a few genes (< 100) whose expression are enough for classification. Interpretation is then about the selected genes. Examples:

- Greedy feature selection (T-tests, ...)
- Contrain the norm of $\beta$: LASSO penalty ($\| \beta \|_1 = \sum_{i=1}^{p} |\beta_i|$), elastic net penalty ($\| \beta \|_1 + \| \beta \|_2$), ...

<table>
<thead>
<tr>
<th>Pros</th>
<th>Cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good performance in classification</td>
<td>The gene selection process is usually not robust</td>
</tr>
<tr>
<td>Biomarker selection</td>
<td>No use of prior biological knowledge</td>
</tr>
<tr>
<td>Interpretability</td>
<td></td>
</tr>
</tbody>
</table>
Outline

1  Motivation

2  Using gene networks as prior knowledge

3  Application

4  Conclusion
Gene networks

- N-Glycan biosynthesis
- Protein kinases
- DNA and RNA polymerase subunits
- Glycolysis / Gluconeogenesis
- Porphyrin and chlorophyll metabolism
- Sulfur metabolism
- Riboflavin metabolism
- Folate biosynthesis
- Biosynthesis of steroids, ergosterol metabolism
- Lysine biosynthesis
- Phenylalanine, tyrosine and tryptophan biosynthesis
- Nitrogen, asparagine metabolism
- Oxidative phosphorylation, TCA cycle
- Purine metabolism
- Folate biosynthesis
- Biosynthesis of steroids, ergosterol metabolism
- Lysine biosynthesis
- Phenylalanine, tyrosine and tryptophan biosynthesis
- Nitrogen, asparagine metabolism
- Oxidative phosphorylation, TCA cycle
- Purine metabolism
Basic biological functions usually involve the coordinated action of several proteins:
- Formation of protein complexes
- Activation of metabolic, signalling or regulatory pathways

Many pathways and protein-protein interactions are already known

Hypothesis: the weights of the classifier should be “coherent” with respect to this prior knowledge
The idea

1. Use the gene network to extract the “important information” in gene expression profiles by Fourier analysis on the graph.
2. Learn a linear classifier on the smooth components.
Notations

\[ A = \begin{pmatrix}
0 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 \\
0 & 0 & 0 & 1 & 0 \\
\end{pmatrix}, \quad D = \begin{pmatrix}
1 & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & 0 \\
0 & 0 & 3 & 0 & 0 \\
0 & 0 & 0 & 2 & 0 \\
0 & 0 & 0 & 0 & 1 \\
\end{pmatrix} \]
The Laplacian of the graph is the matrix $L = D - A$. 

$L = D - A = \begin{pmatrix} 
1 & 0 & -1 & 0 & 0 \\
0 & 1 & -1 & 0 & 0 \\
-1 & -1 & 3 & -1 & 0 \\
0 & 0 & -1 & 2 & -1 \\
0 & 0 & 0 & 1 & 1 
\end{pmatrix}$
**Lemma**

Let $L = D - A$ be the Laplacian of the graph:

- For any $f : \mathcal{X} \to \mathbb{R}$,
  \[
  f^\top L f = \sum_{i \sim j} (f(x_i) - f(x_j))^2
  \]

- $L$ is a **symmetric positive semi-definite** matrix
- 0 is an **eigenvalue** with multiplicity equal to the number of connected components.
Proof: link between $\Omega(f)$ and $L$

\[
\sum_{i \sim j} (f(x_i) - f(x_j))^2 = \sum_{i \sim j} \left( (f(x_i))^2 + (f(x_j))^2 - 2f(x_i)f(x_j) \right)
\]

\[
= \sum_{i=1}^{m} D_{i,i} f(x_i)^2 - 2 \sum_{i \sim j} f(x_i)f(x_j)
\]

\[
= f^T D f - f^T A f
\]

\[
= f^T L f
\]
Proof: eigenstructure of $L$

- $L$ is symmetric because $A$ and $D$ are symmetric.
- For any $f \in \mathbb{R}^m$, $f^\top L f \geq 0$, therefore the (real-valued) eigenvalues of $L$ are $\geq 0$ : $L$ is therefore positive semi-definite.
- $f$ is an eigenvector associated to eigenvalue 0
  iff $f^\top L f = 0$
  iff $\sum_{i \sim j} (f(x_i) - f(x_j))^2 = 0$
  iff $f(x_i) = f(x_j)$ when $i \sim j$,
  iff $f$ is constant (because the graph is connected).
Fourier basis

**Definition**

- The eigenvectors $e_1, \ldots, e_n$ of $L$ with eigenvalues $0 = \lambda_1 \leq \ldots \leq \lambda_n$ form a basis called Fourier basis.

- For any $f : V \to \mathbb{R}$, the Fourier transform of $f$ is the vector $\hat{f} \in \mathbb{R}^n$ defined by:

  $$\hat{f}_i = f^\top e_i, \quad i = 1, \ldots, n.$$ 

- Obviously the inverse Fourier formula holds:

  $$f = \sum_{i=1}^{n} \hat{f}_i e_i.$$
Fourier basis

\[
\begin{align*}
\lambda = 0 & : & \text{Figure 1} \\
\lambda = 0.5 & : & \text{Figure 2} \\
\lambda = 1 & : & \text{Figure 3} \\
\lambda = 2.3 & : & \text{Figure 4} \\
\lambda = 4.2 & : & \text{Figure 5}
\end{align*}
\]
Smoothing operator

**Definition**

- Let $\phi : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ be non-increasing.
- A smoothing operator $S_\phi$ transform a function $f : V \rightarrow \mathbb{R}$ into a smoothed version:

$$S_\phi(f) = \sum_{i=1}^{n} \hat{f}_i \phi(\lambda_i) e_i.$$
Smoothing operators

Examples

- Identity operator ($S_\phi(f) = f$):
  \[
  \phi(\lambda) = 1, \quad \forall \lambda
  \]

- Low-pass filter:
  \[
  \phi(\lambda) = \begin{cases} 
  1 & \text{if } \lambda \leq \lambda^*, \\
  0 & \text{otherwise.}
  \end{cases}
  \]

- Attenuation of high frequencies:
  \[
  \phi(\lambda) = \exp(-\beta \lambda).
  \]
Smoothing operators

Examples

- **Identity operator** \((S_\phi(f) = f)\):

  \[
  \phi(\lambda) = 1, \quad \forall \lambda
  \]

- **Low-pass filter**: 

  \[
  \phi(\lambda) = \begin{cases} 
  1 & \text{if } \lambda \leq \lambda^*, \\
  0 & \text{otherwise.}
  \end{cases}
  \]

- **Attenuation of high frequencies**:

  \[
  \phi(\lambda) = \exp(-\beta \lambda).
  \]
Smoothing operators

Examples

- Identity operator \((S_\phi(f) = f)\):

  \[
  \phi(\lambda) = 1, \quad \forall \lambda
  \]

- Low-pass filter:

  \[
  \phi(\lambda) = \begin{cases} 
  1 & \text{if } \lambda \leq \lambda^*, \\
  0 & \text{otherwise}.
  \end{cases}
  \]

- Attenuation of high frequencies:

  \[
  \phi(\lambda) = \exp(-\beta \lambda).
  \]
Supervised classification and regression

Working with smoothed profiles

- Classical methods for linear classification and regression with a ridge penalty solve:

\[
\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l \left( \beta^\top f_i, y_i \right) + \lambda \beta^\top \beta.
\]

- Applying these algorithms on the smooth profiles means solving:

\[
\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l \left( \beta^\top S\phi(f_i), y_i \right) + \lambda \beta^\top \beta.
\]
Smooth solution

Lemma

This is equivalent to:

$$\min_{v \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} I \left( v^\top f_i, y_i \right) + \lambda \sum_{i=1}^{p} \frac{\hat{v}_i^2}{\phi(\lambda_i)} ,$$

hence the linear classifier $v$ is smooth.

Proof

- Let $v = \sum_{i=1}^{n} \phi(\lambda_i) e_i e_i^\top \beta$, then

  $$\beta^\top S_{\phi}(f_i) = \beta^\top \sum_{i=1}^{n} \hat{f}_i \phi(\lambda_i) e_i = f^\top v .$$

- Then $\hat{v}_i = \phi(\lambda_i) \hat{\beta}_i$ and $\beta^\top \beta = \sum_{i=1}^{n} \frac{\hat{v}_i^2}{\phi(\lambda_i)^2}$. 
Smooth solution

**Lemma**

This is equivalent to:

\[
\min_{v \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l \left( v^\top f_i, y_i \right) + \lambda \sum_{i=1}^{p} \frac{\hat{v}_i^2}{\phi(\lambda_i)},
\]

**Proof**

- Let \( v = \sum_{i=1}^{n} \phi(\lambda_i) e_i e_i^\top \beta \), then

\[
\beta^\top S_\phi(f_i) = \beta^\top \sum_{i=1}^{n} \hat{f}_i \phi(\lambda_i) e_i = f^\top v.
\]

- Then \( \hat{v}_i = \phi(\lambda_i) \hat{\beta}_i \) and \( \beta^\top \beta = \sum_{i=1}^{n} \frac{\hat{v}_i^2}{\phi(\lambda_i)^2} \).
Kernel methods (SVM, kernel ridge regression..) only need the inner product between smooth profiles:

\[
K(f, g) = S_\phi(f)^\top S_\phi(g) \\
= \sum_{i=1}^{n} \hat{f}_i \hat{g}_i \phi(\lambda_i)^2 \\
= f^\top \left( \sum_{i=1}^{n} \phi(\lambda_i)^2 e_i e_i^\top \right) g \\
= f^\top K_\phi g ,
\]

with

\[
K_\phi = \sum_{i=1}^{n} \phi(\lambda_i)^2 e_i e_i^\top .
\]
Examples

- For $\phi(\lambda) = \exp(-t\lambda)$, we recover the diffusion kernel:

  $$K_{\phi} = \exp_{M}(-2tL).$$

- For $\phi(\lambda) = 1/\sqrt{1 + \lambda}$, we obtain

  $$K_{\phi} = (L + I)^{-1},$$

  and the penalization is:

  $$\sum_{i=1}^{n} \frac{\hat{v}_i^2}{\phi(\lambda_i)} = v^\top (L + I) v = \|v\|_2^2 + \sum_{i \sim j} (v_i - v_j)^2.$$
Examples

- For $\phi(\lambda) = \exp(-t\lambda)$, we recover the diffusion kernel:

  $$K_\phi = \exp_M(-2tL).$$

- For $\phi(\lambda) = 1/\sqrt{1 + \lambda}$, we obtain

  $$K_\phi = (L + I)^{-1},$$

  and the penalization is:

  $$\sum_{i=1}^{n} \frac{\hat{v}_i^2}{\phi(\lambda_i)} = v^\top (L + I) v = \|v\|_2^2 + \sum_{i \sim j} (v_i - v_j)^2.$$
Outline

1 Motivation

2 Using gene networks as prior knowledge

3 Application

4 Conclusion
Data

Expression

- Study the effect of low irradiation doses on the yeast
- 12 non irradiated vs 6 irradiated
- Which pathways are involved in the response at the transcriptomic level?

Graph

- KEGG database of metabolic pathways
- Two genes are connected if they code for enzymes that catalyze successive reactions in a pathway (metabolic gene network).
- 737 genes, 4694 vertices.
Spectral analysis of gene expression profiles using gene networks

PC1

pyruvate metabolism

A large network helps keep the biochemical relationships between genes without the constraints of pathway limits.
Spectral analysis of gene expression profiles using gene networks

Fig. 5... work was supported by the grant ACI-IMPBIO-2004-47 of the French Ministry for Research and New Technologies.
Outline

1. Motivation
2. Using gene networks as prior knowledge
3. Application
4. Conclusion
Use the gene graph to encode prior knowledge about the classifier.

Prior knowledge is always needed to classify few examples in large dimensions (sometimes implicitly)

Future work: validation of the method on more data, other formulations, directed graphs...
Acknowledgements

KernelChip project (ACI IMPBIO)

- Franck Rapaport (Ecole des Mines and Curie Institute)
- Emmanuel Barillot (Curie Institute)
- Andrei Zynoviev (Curie Institute)
- Marie Dutreix (Curie Institute)

Reference