Analysis of microarray data with pathway information

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

Center for Computational Biology
Ecole des Mines de Paris

Centre d'Ecologie Cellulaire, Faculté de médecine
Pitié-Salpêtrière, May 19, 2006

CBIO overview

- The newest research center of Ecole des Mines
- Started in 2002, became an autonomous research center in 2006
- Objective: develop mathematical approaches and computational tools to process and analyze biological and chemical data
- http://cbio.ensmp.fr

CBIO research

1. Machine learning and statistics (theory and algorithms)
2. Analysis of post-genomic data and systems biology (focus on cancer and malaria)
3. Data analysis methods for new technologies (DNA chips, cell chips, high-throughput microscopy)
4. Virtual screening (docking, ligand-based)

Outline

1. Classification and interpretation of microarray data

2. Including pathway information

Classical setting

Data available

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

Goal

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

Classical setting

Data available

- Gene expression measures for more than 10*k* genes
- Measured on less than 100 samples of two (or more) different classes (e.g., different tumors)

Goal

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

Linear classifiers

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:

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

  that is positive for one class, negative for the other
- Interpretation: the weight β_i quantifies the influence of gene i for the classification

Linear classifiers

Pitfalls

- No robust estimation procedure exist for 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 norm of β, e.g.:

- Euclidean norm (support vector machines, ridge regression): $\parallel \beta \parallel_2 = \sum_{i=1}^{p} \beta_i^2$
- L_1-norm (lasso regression) : $\parallel \beta \parallel_1 = \sum_{i=1}^{p} \mid \beta_i \mid$

Pros

- Good performance in classification

Cons

- Limited interpretation (small weights)
- No prior biological knowledge

Example 2: Feature Selection

The approach

Constrain most weights to be 0, i.e., select a few genes (< 20) whose expression are enough for classification. Interpretation is then about the selected genes.

Pros

- Good performance in classification
- Useful for biomarker selection
- Apparently easy interpretation

Cons

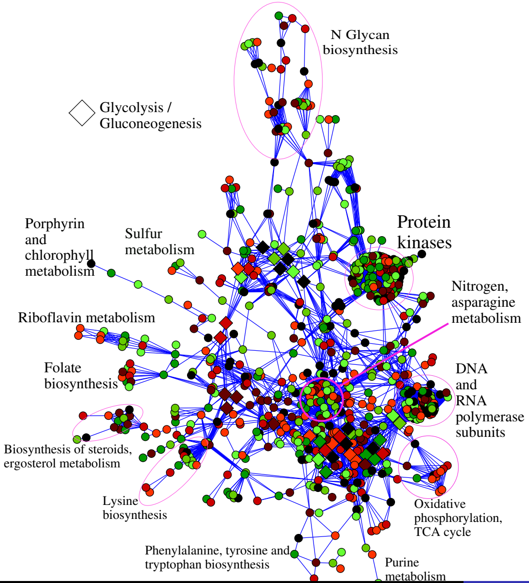
- The gene selection process is usually not robust
- Wrong interpretation is the rule (too much correlation between genes)

Pathway interpretation

Motivation

- Basic biological functions are usually expressed in terms of pathways and not of single genes (metabolic, signaling, regulatory)
- Many pathways are already known
- How to use this prior knowledge to constrain the weights to have an interpretation at the level of pathways?

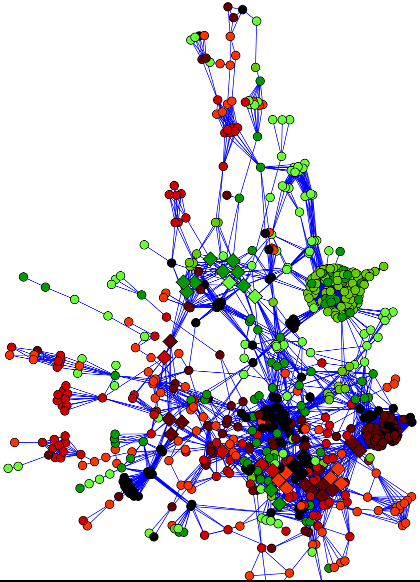
Pathway interpretation


N Glycan biosynthesis
Glycolysis / Gluconeogenesis
Porphyrin and chlorophyll metabolism
Sulfur metabolism
Protein kinases
Nitrogen, asparagine metabolism
Riboflavin metabolism
Folate biosynthesis
DNA and RNA polymerase subunits
Biosynthesis of steroids, ergosterol metabolism
Lysine biosynthesis
Oxidative phosphorylation, TCA cycle
Phenylalanine, tyrosine and tryptophan biosynthesis
Purine metabolism

Bad example

- The graph is the complete known metabolic network of the budding yeast (from KEGG database)
- We project the classifier weight learned by a SVM
- Good classification accuracy, but no possible interpretation!

Pathway interpretation



Good example

- The graph is the complete known metabolic network of the budding yeast (from KEGG database)
- We project the classifier weight learned by a spectral SVM
- Good classification accuracy, and good interpretation!

Spectral SVM

Short description

1. Pre-process each microarray profile to filter out the high frequencies with respect to the known pathways. This involves discrete Fourier transforms + spectral graph theory.
2. Perform classical SVM on the smoothed expression profiles

Discussion

You will always have an interpretable model because you enforce it. Can we trust is?

- Any method must use prior knowledge because of the $n << p$ problem.
- In many cases the “true” classifier is more likely to have a pathway interpretation than to be based on a few genes only.

There are many cases where smoothness is not expected on the pathway (negative regulation...)

- We just enforce a global smoothness, local jumps are possible (although penalized).
- As more data are available, a more precise estimation is possible.

Conclusion

- Manipulating gene expression data is difficult for statistical reasons.
- Inclusion of prior knowledge is required (e.g., feature selection)
- Known pathways form a natural prior knowledge
- This results in classifiers with good accuracy and interpretability.

Acknowledgements

Franck Rapaport, Emmanuel Barillot, Andrei Zynoviev, Marie Dutreix (Institut Curie)