Spectral approaches to integrate gene expression and gene networks

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

Center for Computational Biology
Ecole des Mines de Paris

ESBIC meeting, Institut Curie, July 6, 2006

ARMINES contribution to ESBIC

- Develop methods for analysis of gene expression data
- Develop methods for integration of heterogeneous data, in particular expression and pathways
- Integrate these tools in the ESBIC standards

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

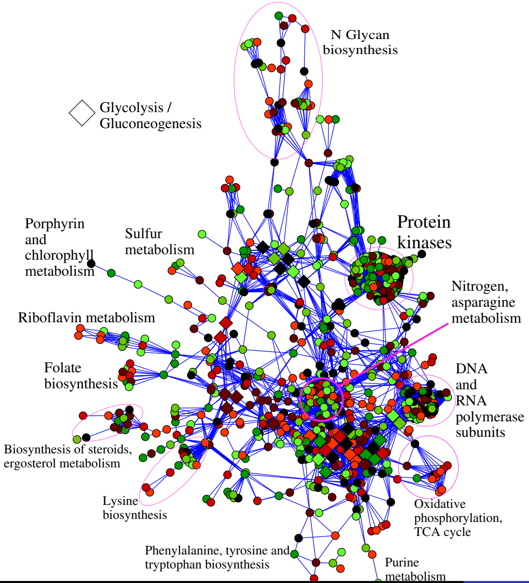
- The gene selection process is usually not robust
- Wrong interpretation is the rule (too much correlation between genes)

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?

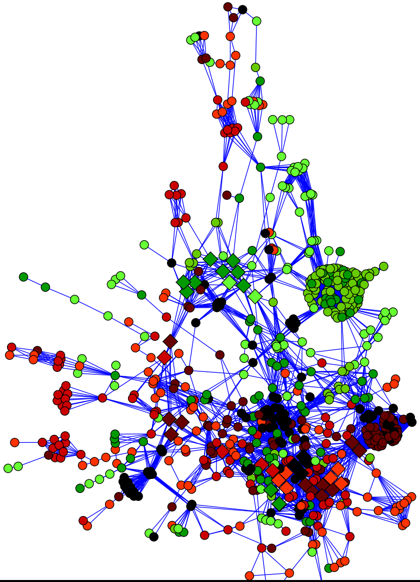
Pathway interpretation


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

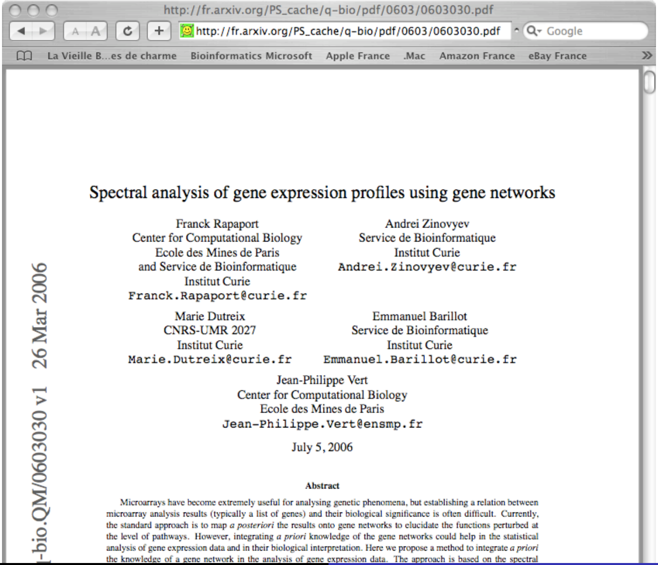
- The graph is the complete known metabolic network of the budding yeast (from KEGG database)
- We project the classifier weight learned by a spectral SVM
- Good classification accuracy, and good interpretation!

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

More details


http://fr.arxiv.org/PS_cache/q-bio/pdf/0603/0603030.pdf
http://fr.arxiv.org/PS_cache/q-bio/pdf/0603/0603030.pdf
Google
La Vieille B...es de charme
Bioinformatics Microsoft
Apple France
.Mac
Amazon France
eBay France

-bio.QM/0603030 v1 26 Mar 2006

Spectral analysis of gene expression profiles using gene networks

Franck Rapaport
Center for Computational Biology
Ecole des Mines de Paris
and Service de Bioinformatique
Institut Curie
Franck.Rapaport@curie.fr

Andrei Zinovyev
Service de Bioinformatique
Institut Curie
Andrei.Zinovyev@curie.fr

Marie Dutreix
CNRS-UMR 2027
Institut Curie
Marie.Dutreix@curie.fr

Emmanuel Barillot
Service de Bioinformatique
Institut Curie
Emmanuel.Barillot@curie.fr

Jean-Philippe Vert
Center for Computational Biology
Ecole des Mines de Paris
Jean-Philippe.Vert@ensmp.fr

July 5, 2006

Abstract

Microarrays have become extremely useful for analysing genetic phenomena, but establishing a relation between microarray analysis results (typically a list of genes) and their biological significance is often difficult. Currently, the standard approach is to map *a posteriori* the results onto gene networks to elucidate the functions perturbed at the level of pathways. However, integrating *a priori* knowledge of the gene networks could help in the statistical analysis of gene expression data and in their biological interpretation. Here we propose a method to integrate *a priori* the knowledge of a gene network in the analysis of gene expression data. The approach is based on the spectral

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.

Ongoing and future work

- Validation on tumour data
- Extension to non-smooth assumption (inhibition...)
- Integration with other softwares

Acknowledgements

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