Machine learning in bioinformatics and drug discovery

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

Mines ParisTech / Institut Curie / Inserm

Workshop on Bioinformatics for Medical and Pharmaceutical Research, Institut Curie, Paris, November 16, 2009.

Where we are

institutCurie

MINES PARIS ParisTech

Inserm

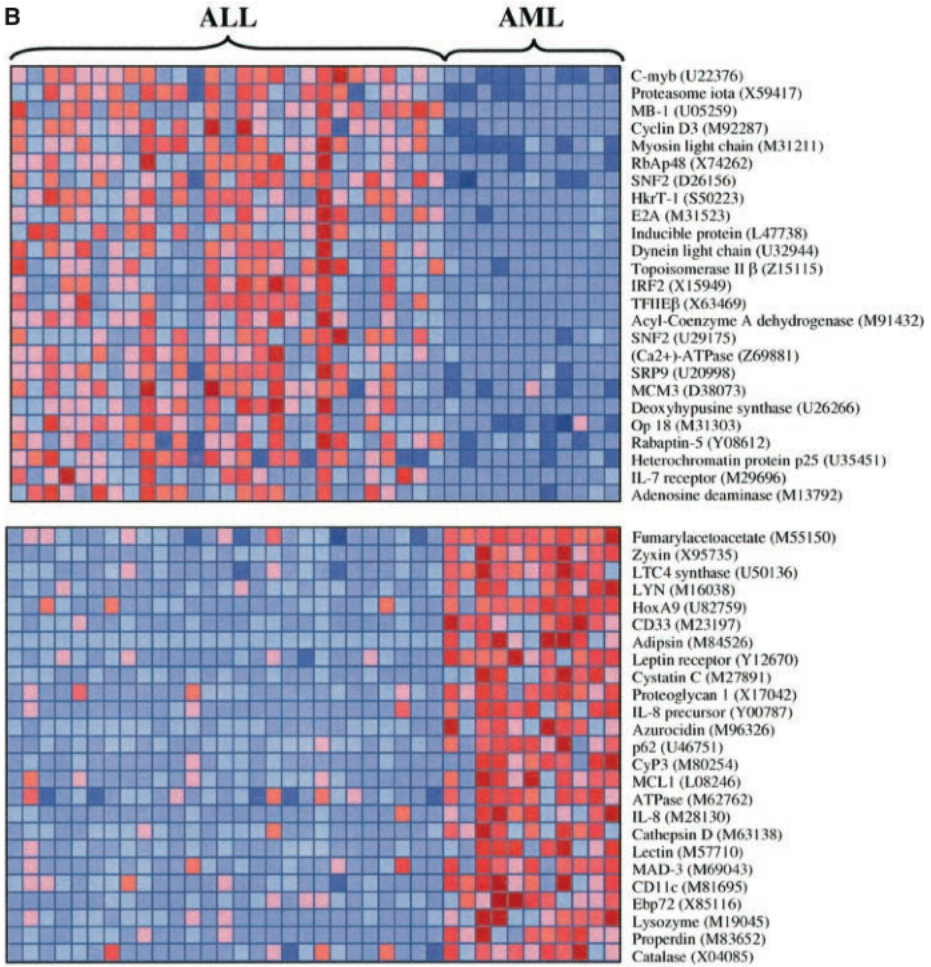
- A joint lab about “Cancer computational genomics, bioinformatics, biostatistics and epidemiology”
- Located in th Institut Curie, a major hospital and cancer research institute in Europe

Statistical machine learning for cancer informatics

Main topics

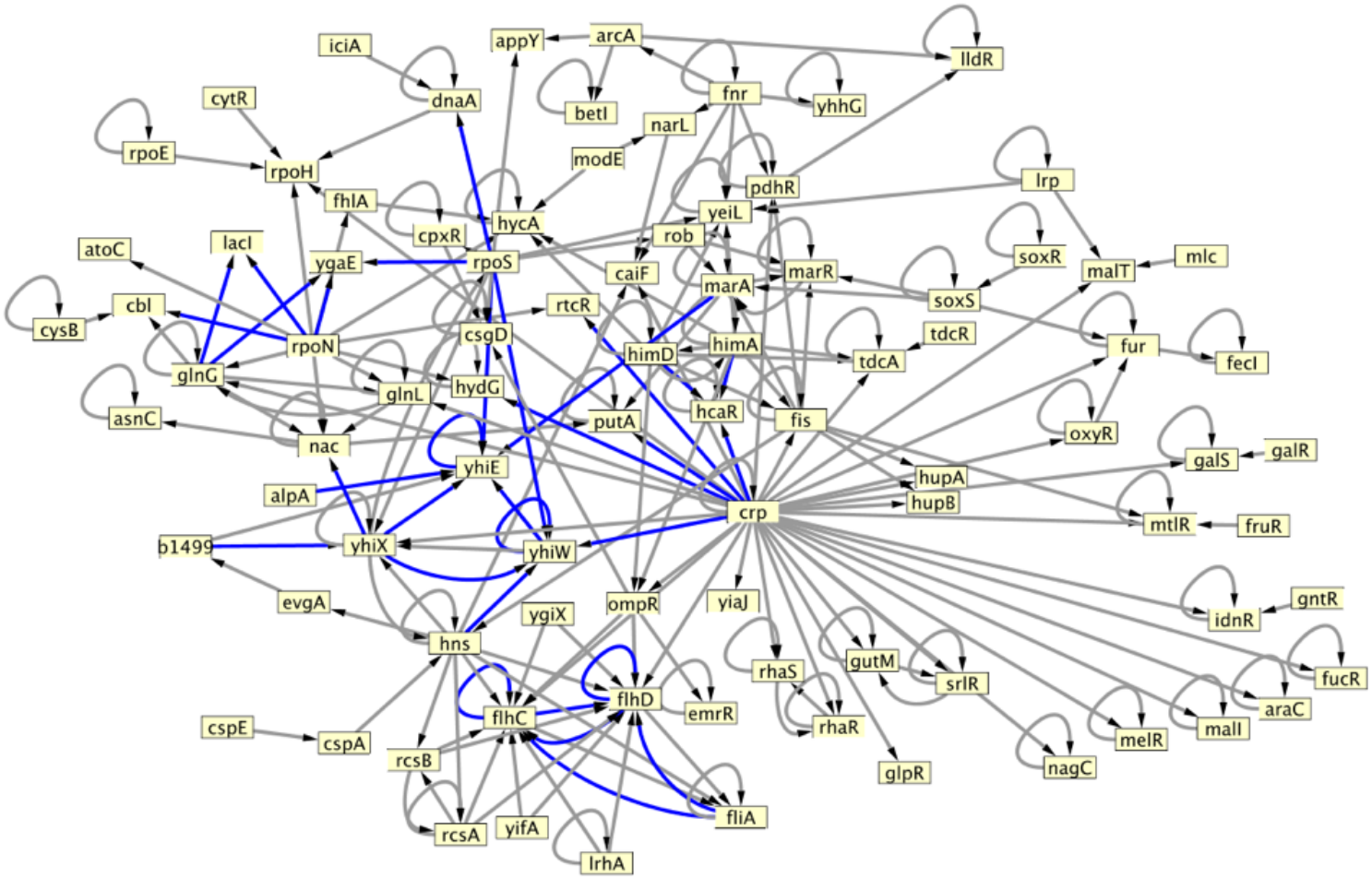
- Towards better diagnosis, prognosis, and personalized medicine
  - Supervised classification of genomic, transcriptomic, proteomic data; heterogeneous data integration
- Towards new drug targets
  - Systems biology, reconstruction of gene networks, pathway enrichment analysis, multidimensional phenotyping of cell populations.
- Towards new drugs
  - Ligand-based virtual screening, *in silico* chemogenomics.

Towards personalized medicine: Diagnosis/prognosis from genome/transcriptome



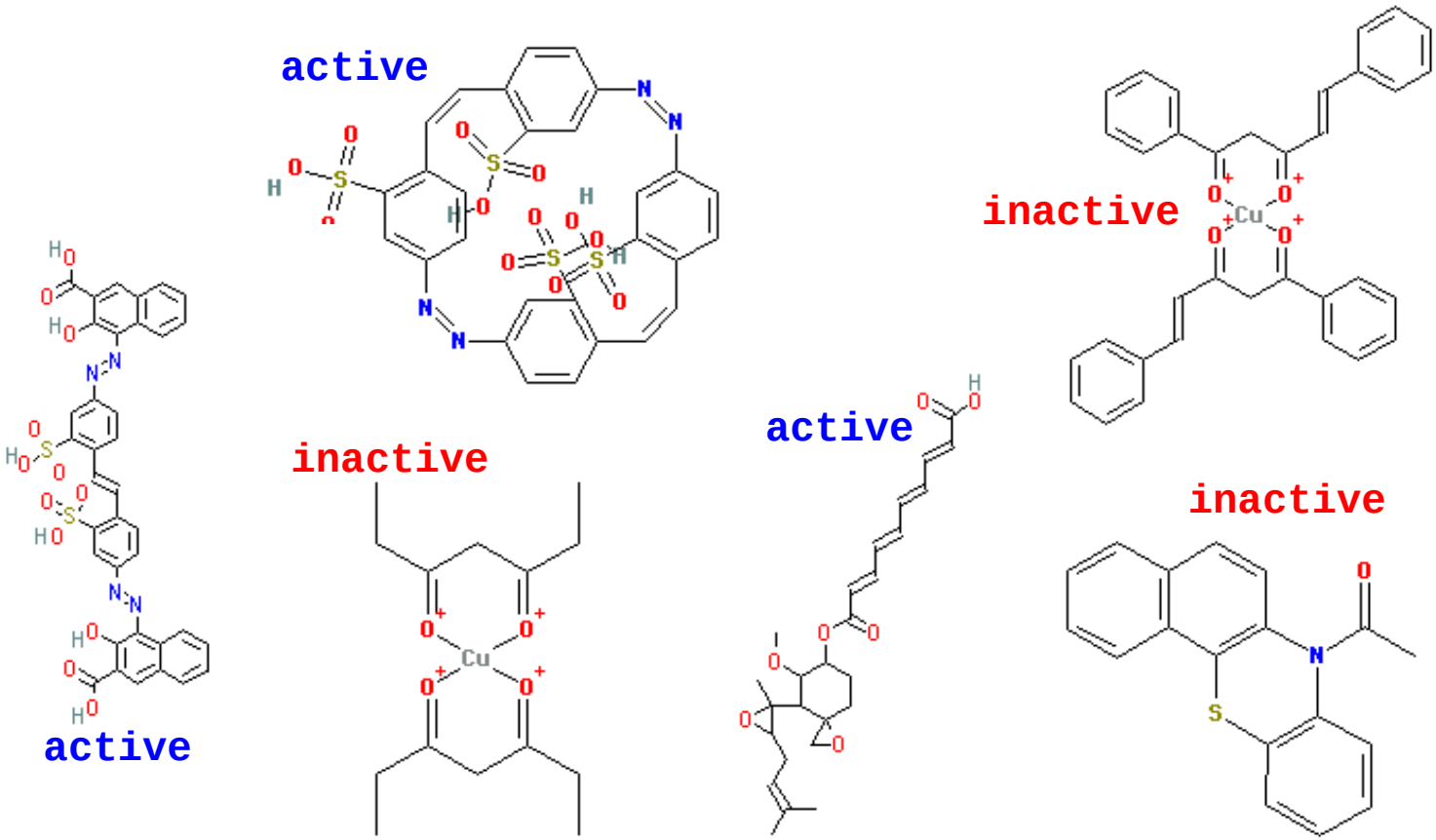
From Golub et al., Science, 1999.

Towards new drug targets: Inference of biological networks



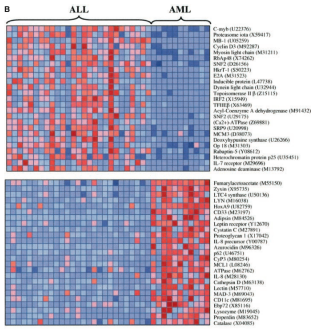
From Mordelet and Vert, Bioinformatics, 2008.

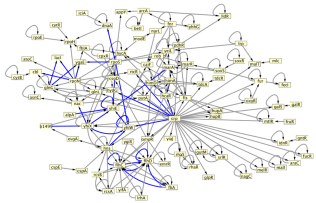
Towards new drugs: Ligand-Based Virtual Screening and QSAR

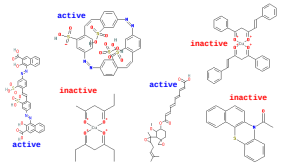


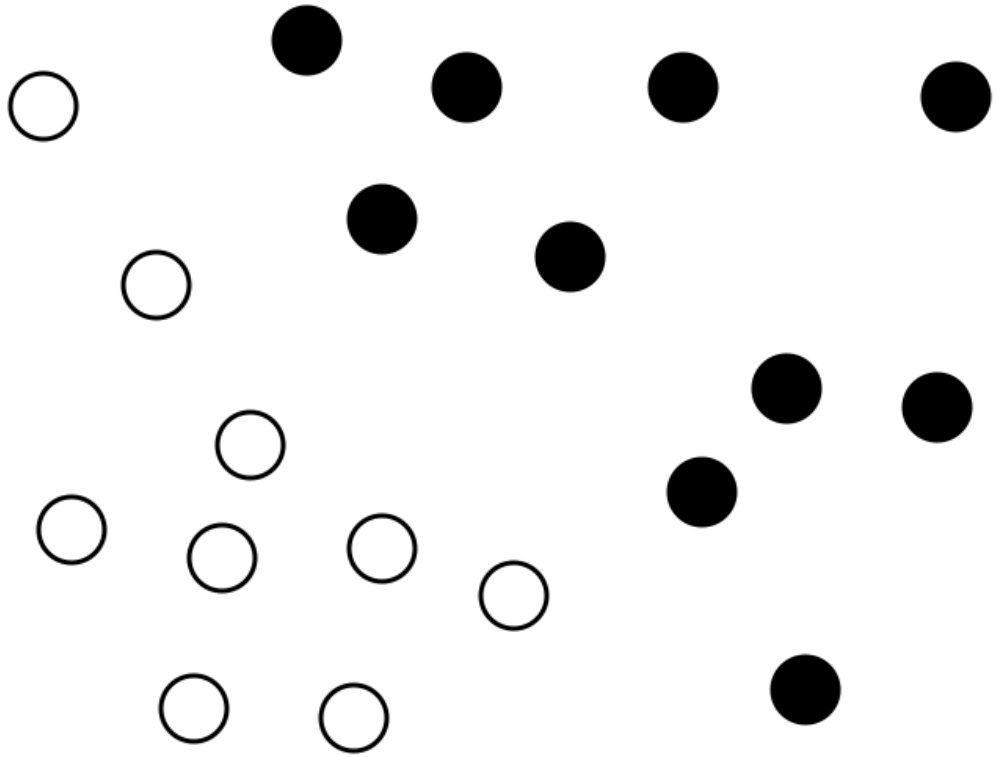
NCI AIDS screen results (from http://cactus.nci.nih.gov).

Pattern recognition, *aka* supervised classification

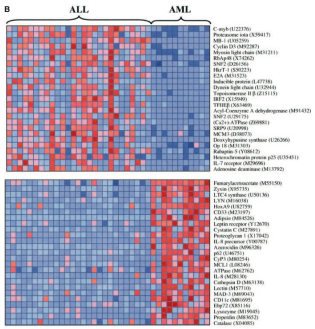


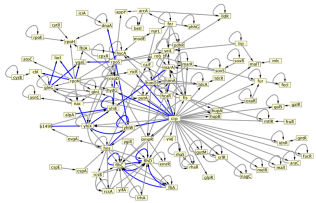


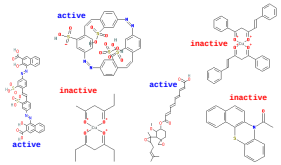


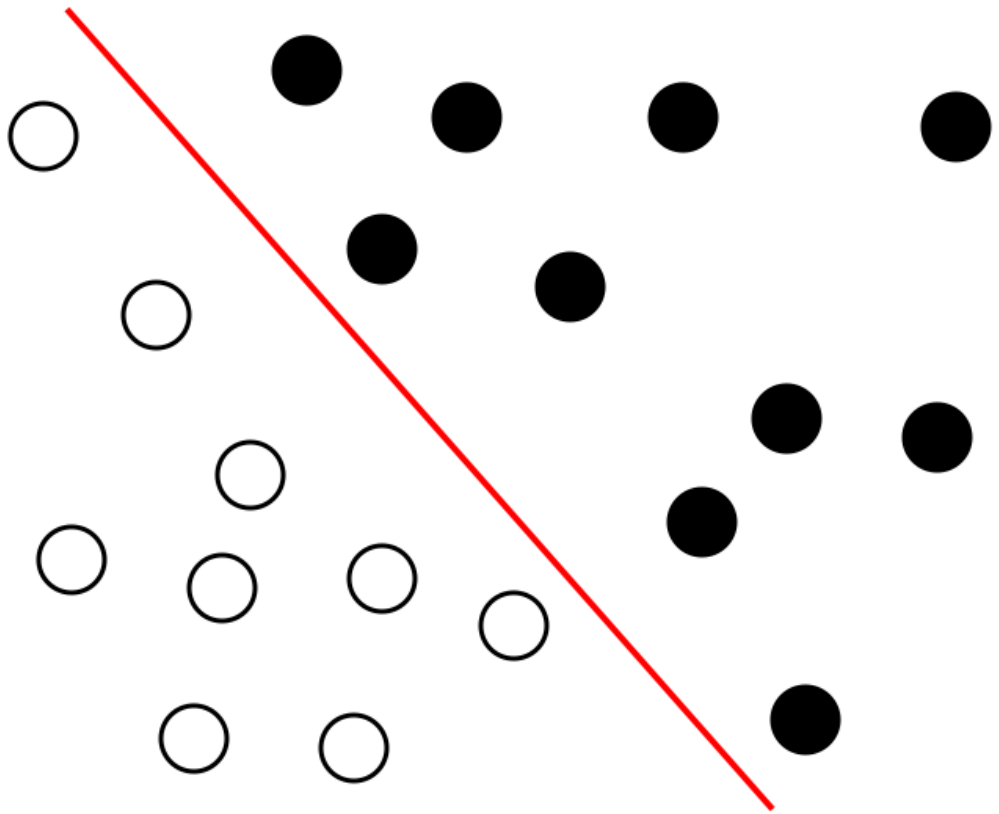


Pattern recognition, *aka* supervised classification

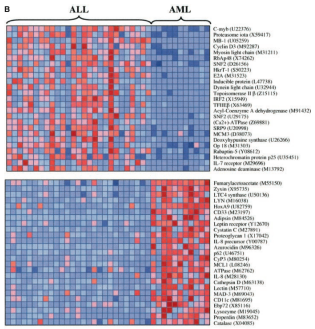


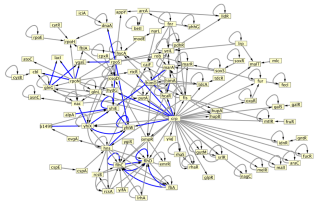


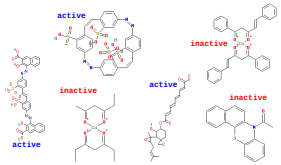


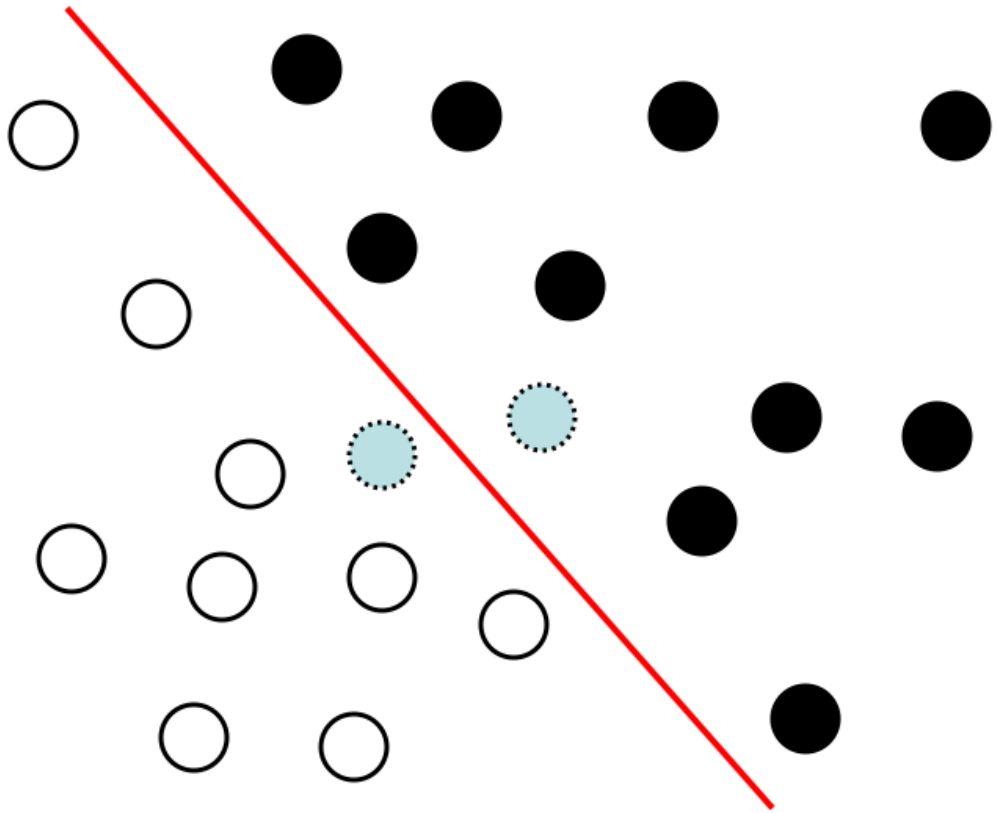


Pattern recognition, *aka* supervised classification

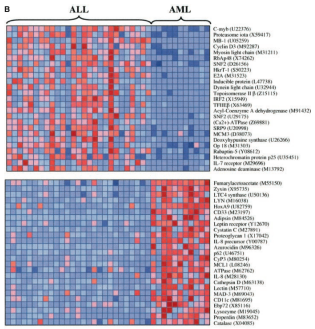


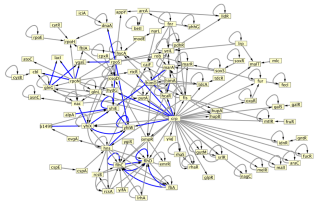


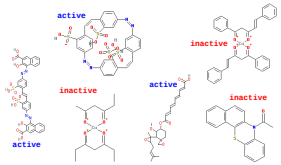


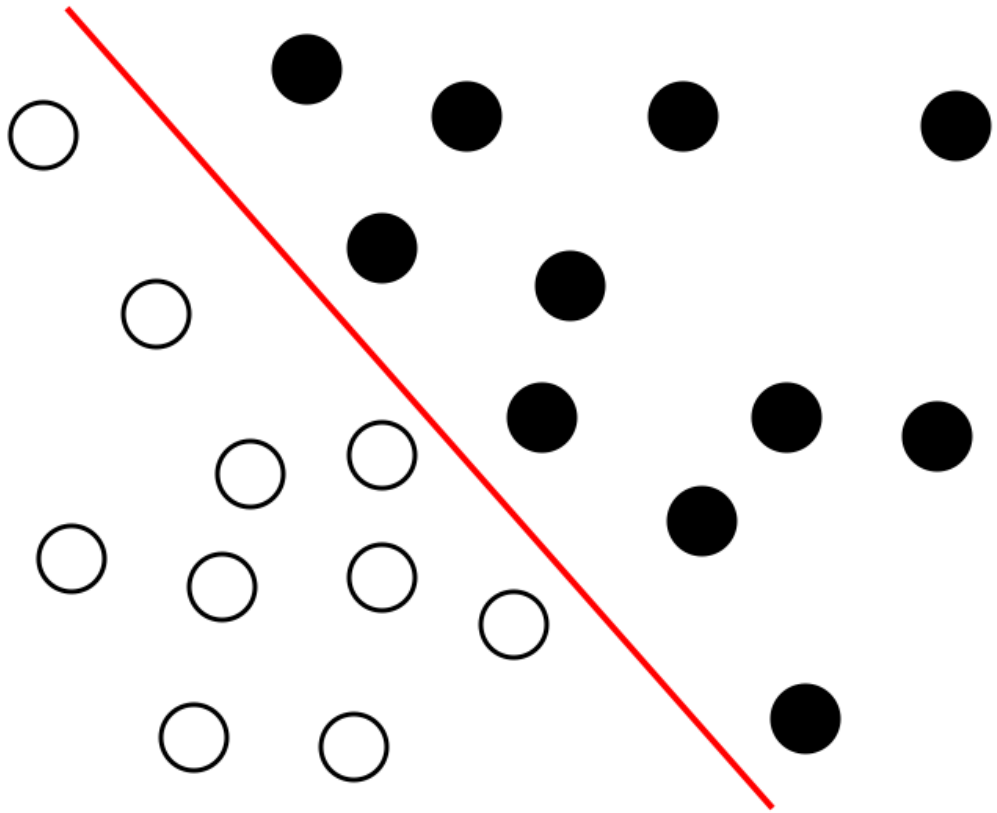


Pattern recognition, *aka* supervised classification

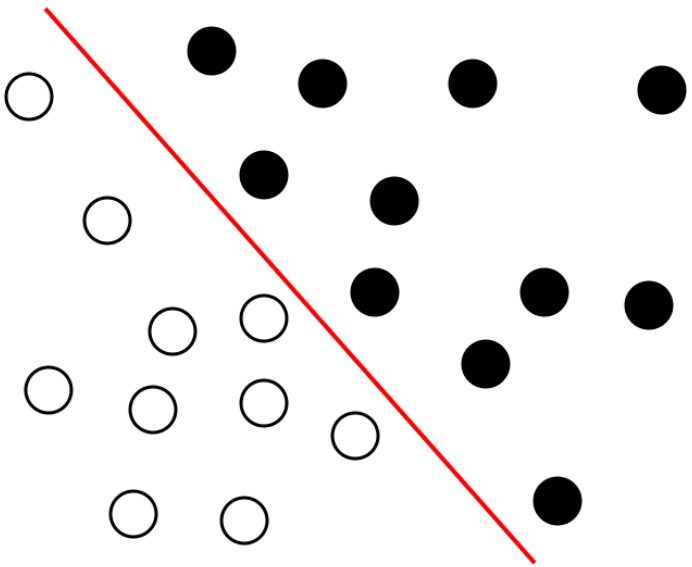








Pattern recognition, *aka* supervised classification



Challenges

- High dimension
- Few samples
- Structured data
- Prior knowledge
- Fast and scalable implementations
- Interpretable models

Linear classifiers

The model

- Each sample is represented by a vector $x = (x_1, \ldots, x_p)$
- Goal: from a training set of samples with known labels, estimate a linear function:

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

  whose sign is a good predictor.
- Interpretability: the weight β_i quantifies the influence of feature i (but...)

Estimating a linear classifiers

- We have a training set of samples $(x^{(1)}, \ldots, x^{(n)})$ with known class $(y^{(1)}, \ldots, y^{(n)})$.
- For any candidate set of weights $\beta = (\beta_1, \ldots, \beta^p)$ we quantify how "good" the linear function f_β is on the training set with some average loss, e.g.,

$$R(\beta) = \frac{1}{n} \sum_{i=1}^{n} l(f_\beta(x^{(i)}), y^{(i)}) \,,$$

- We choose the β that achieves the minimium risk, subject to some constraint on β, e.g.:

$$\Omega(\beta) \leq C \,.$$

Importance of the constraint $\Omega(\beta) < C$

Why it is necessary

- Prevents overfitting (especially when n is small)
- Helps to overcome numerical issues (regularization)

Why it is useful

- Can lead to efficient implementations (convexification)
- Good place to put prior knowledge!

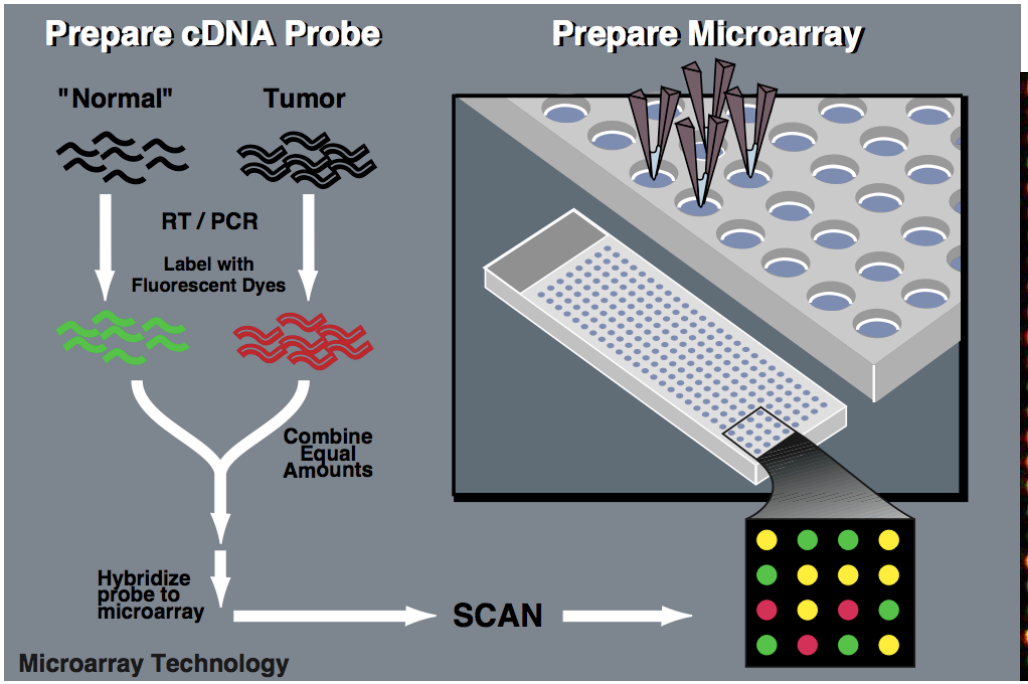
Outline

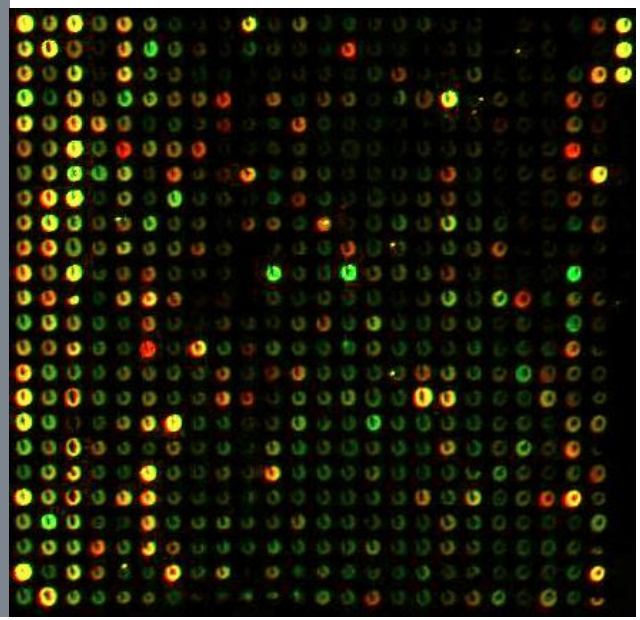
Outline

Outline

Outline

Tissue profiling with DNA chips

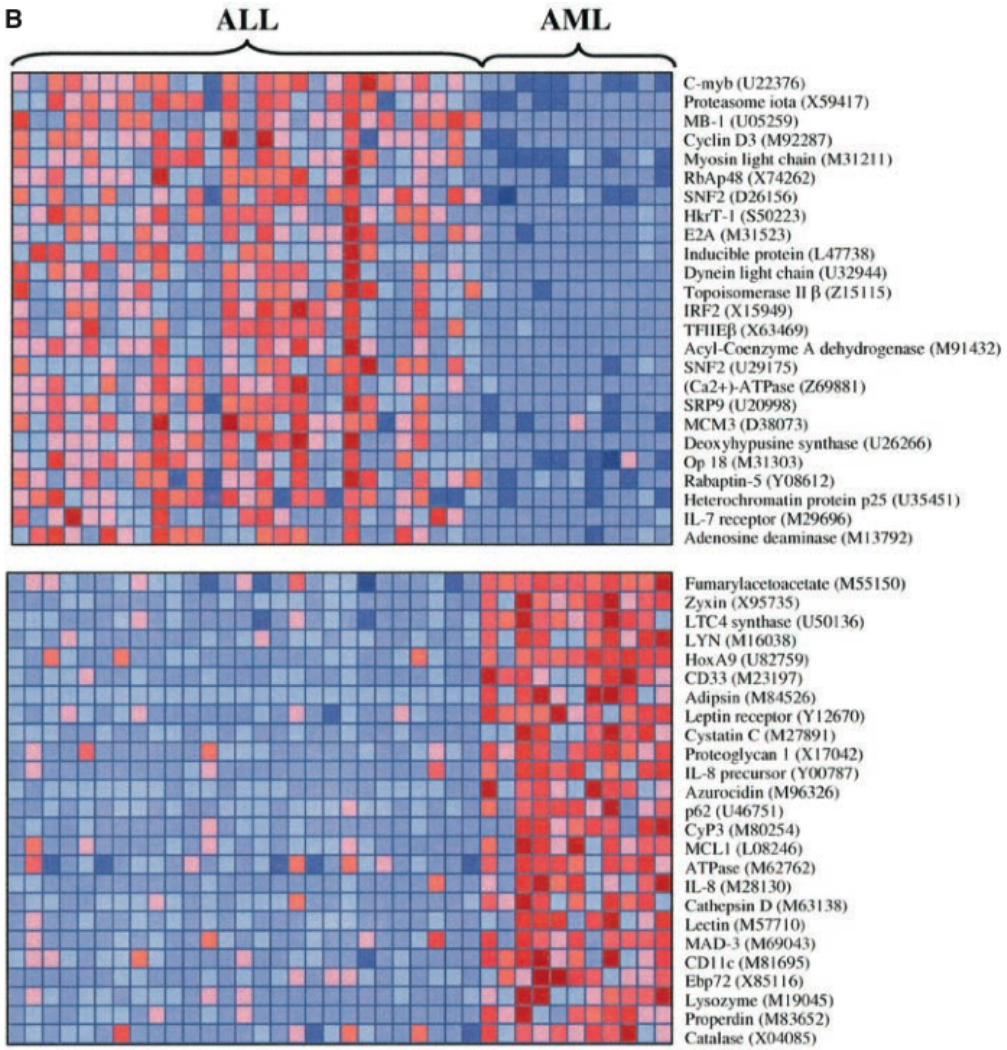




Data

- 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

Difficulty

- Large dimension
- Few samples

Gene signature

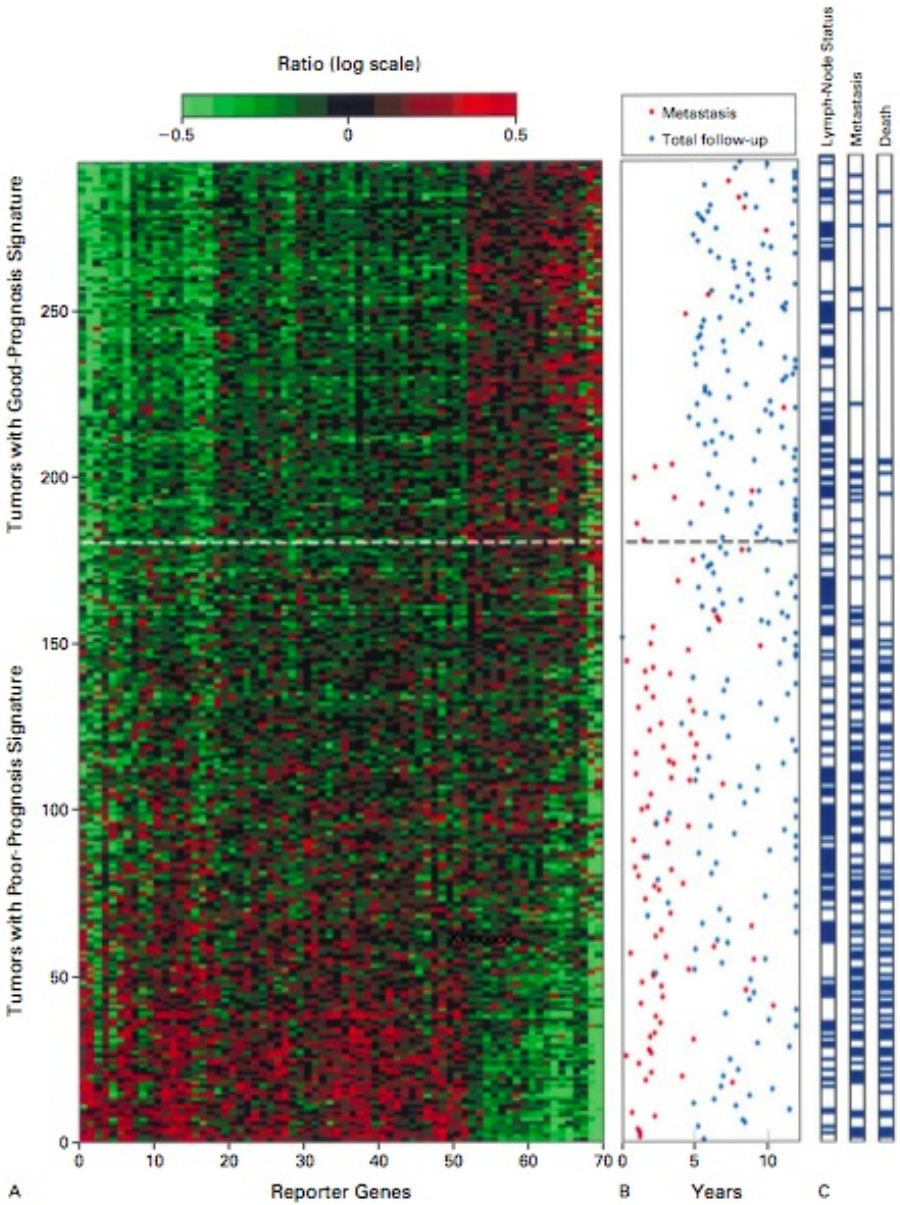
The idea

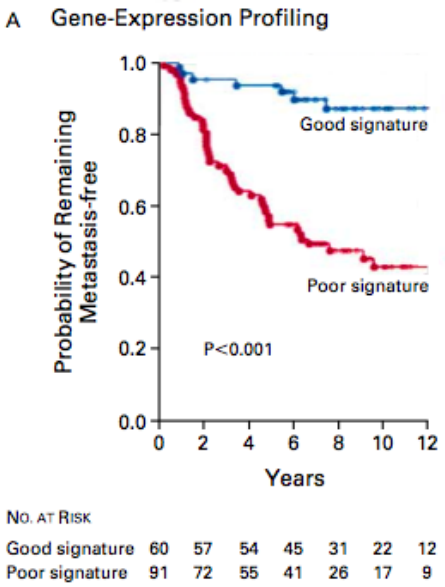
- We look for a limited set of genes that are sufficient for prediction.
- Equivalently, the linear classifier will be sparse

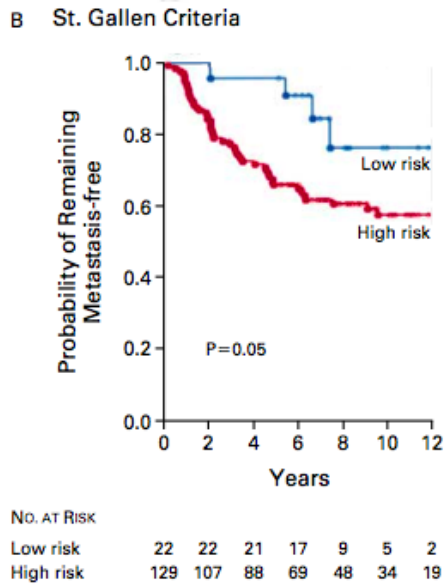
Motivations

- Bet on sparsity: we believe the "true" model is sparse.
- Interpretation: we will get a biological interpretation more easily by looking at the selected genes.
- Accuracy: by restricting the class of classifiers, we "increase the bias" but "decrease the variance". This should be helpful in large dimensions (it is better to estimate well a wrong model than estimate badly a good model).

Example: MAMMAPRINT







How to estimate a sparse linear model?

Best subset selection

- We look for a sparse weight vector β by solving the problem:

$$\min R(f_\beta) \quad \text{s.t.} \quad \| \beta \|_0 \leq k$$

- This is usually a NP-hard problem, feasible for p as large as 30 or 40
- The state-of-the-art is branch-and-bound optimization, known as *leaps and bound* for least squares (Furnival and Wilson, 1974).
- Not useful in practice for us...

Efficient feature selection

To work with more variables, we must use different methods. The state-of-the-art is split among

- Filter methods : the predictors are preprocessed and ranked from the most relevant to the less relevant. The subsets are then obtained from this list, starting from the top.
- Wrapper method: here the feature selection is iterative, and uses the ERM algorithm in the inner loop
- Embedded methods : here the feature selection is part of the ERM algorithm itself (see later the shrinkage estimators).

Filter methods

- Associate a score $S(i)$ to each feature i, then rank the features by decreasing score.
- Many scores / criteria can be used
  - Loss of the ERM trained on a single feature
  - Statistical tests (Fisher, T-test)
  - Other performance criteria of the ERM restricted to a single feature (AUC, ...)
  - Information theoretical criteria (mutual information...)

Pros

Simple, scalable, good empirical success

Cons

- Selection of redundant features
- Some variables useless alone can become useful together

Filter methods

- Associate a score $S(i)$ to each feature i, then rank the features by decreasing score.
- Many scores / criteria can be used
  - Loss of the ERM trained on a single feature
  - Statistical tests (Fisher, T-test)
  - Other performance criteria of the ERM restricted to a single feature (AUC, ...)
  - Information theoretical criteria (mutual information...)

Pros

Simple, scalable, good empirical success

Cons

- Selection of redundant features
- Some variables useless alone can become useful together

Filter methods

- Associate a score $S(i)$ to each feature i, then rank the features by decreasing score.
- Many scores / criteria can be used
  - Loss of the ERM trained on a single feature
  - Statistical tests (Fisher, T-test)
  - Other performance criteria of the ERM restricted to a single feature (AUC, ...)
  - Information theoretical criteria (mutual information...)

Pros

Simple, scalable, good empirical success

Cons

- Selection of redundant features
- Some variables useless alone can become useful together

Wrapper methods

Forward stepwise selection

- Start from no features
- Sequentially add into the model the feature that most improves the fit

Backward stepwise selection (if n>p)

- Start from all features
- Sequentially removes from the model the feature that least degrades the fit

Other variants

Hybrid stepwise selection strategies that consider both forward and backward moves at each stage, and make the "best" move

Wrapper methods

Forward stepwise selection

- Start from no features
- Sequentially add into the model the feature that most improves the fit

Backward stepwise selection (if n>p)

- Start from all features
- Sequentially removes from the model the feature that least degrades the fit

Other variants

Hybrid stepwise selection strategies that consider both forward and backward moves at each stage, and make the "best" move

Wrapper methods

Forward stepwise selection

- Start from no features
- Sequentially add into the model the feature that most improves the fit

Backward stepwise selection (if n>p)

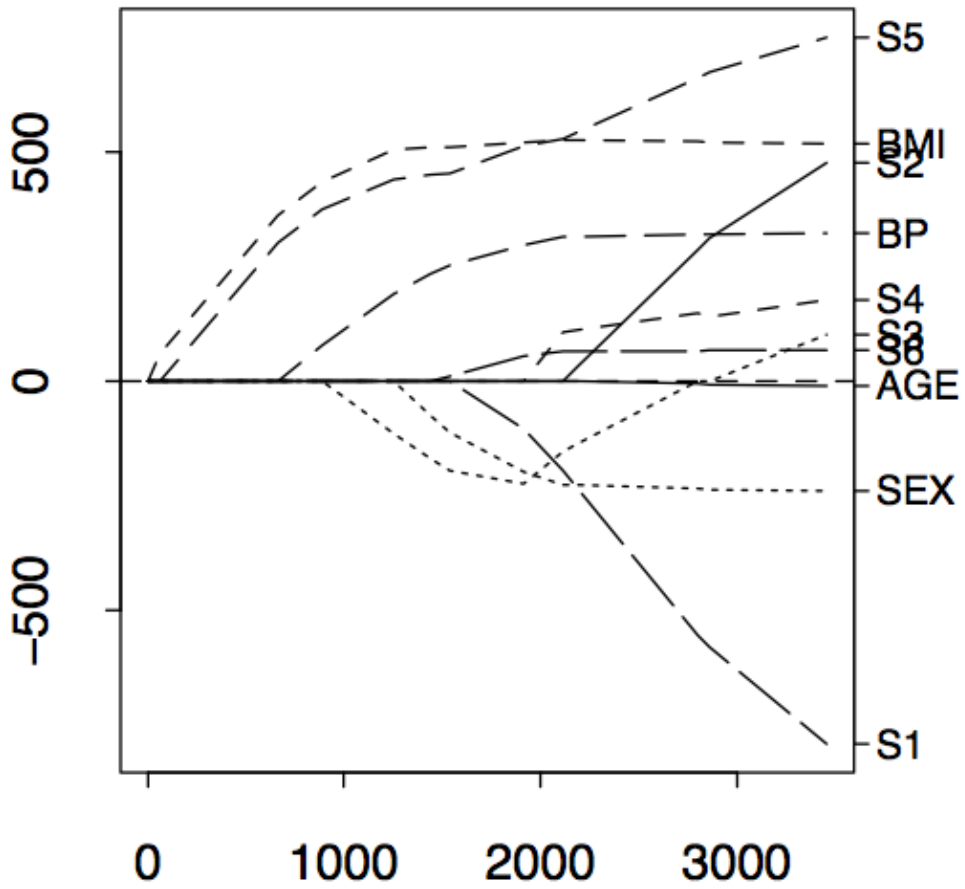
- Start from all features
- Sequentially removes from the model the feature that least degrades the fit

Other variants

Hybrid stepwise selection strategies that consider both forward and backward moves at each stage, and make the "best" move

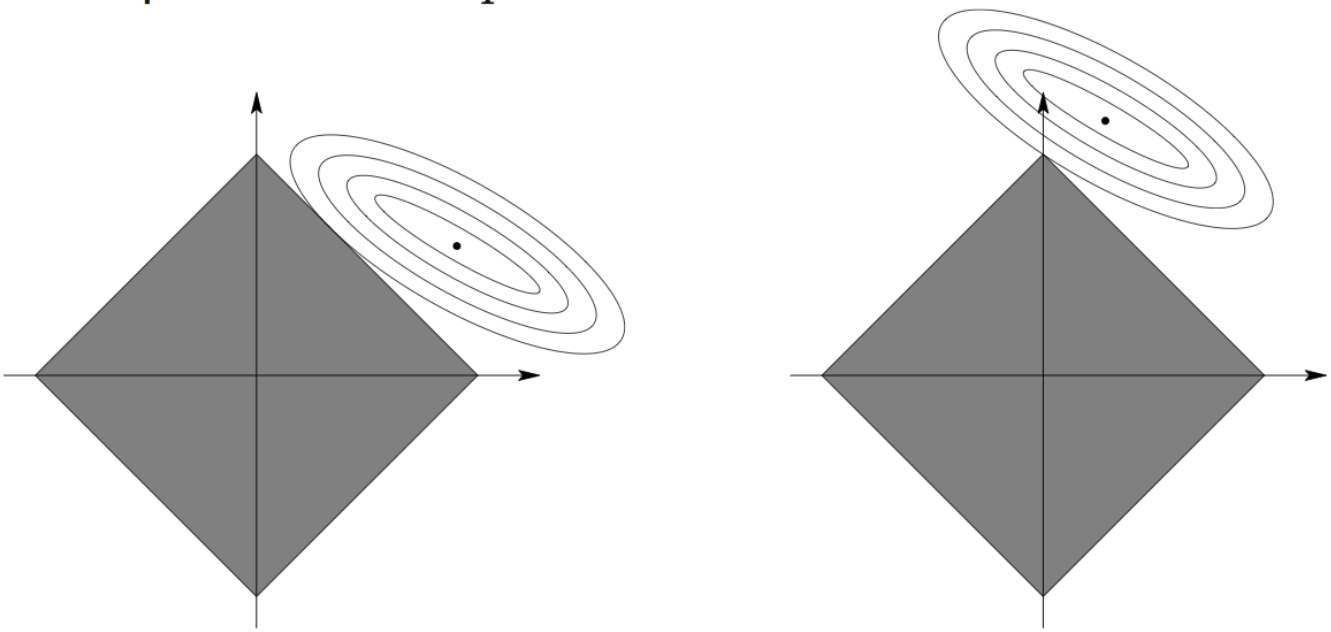
Embedded methods (LASSO)

$$\min_{\beta} R(\beta) + \sum_{i=1}^{p} |\beta_i|$$



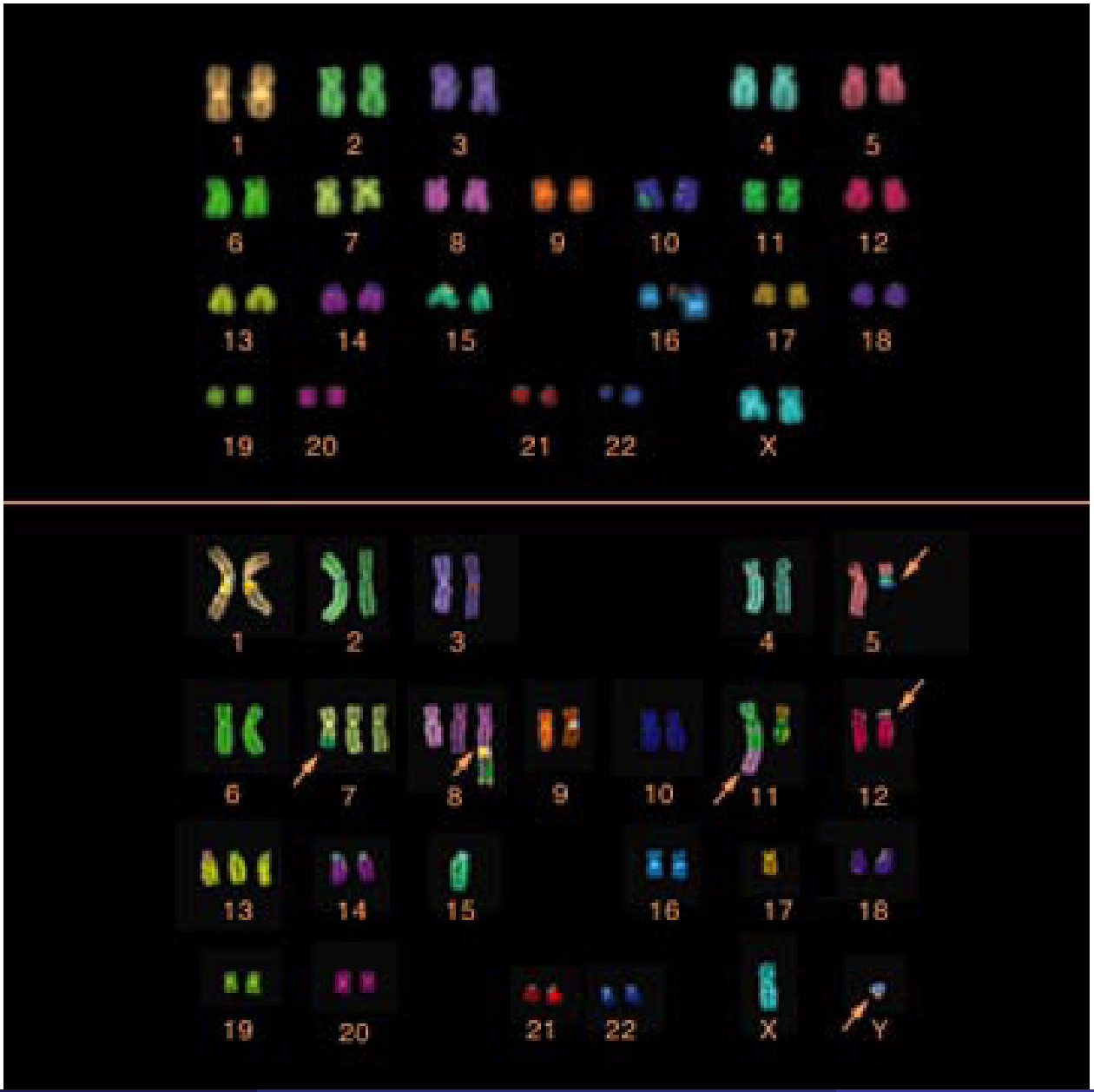
Why LASSO leads to sparse solutions

Geometric interpretation with $p = 2$



Outline

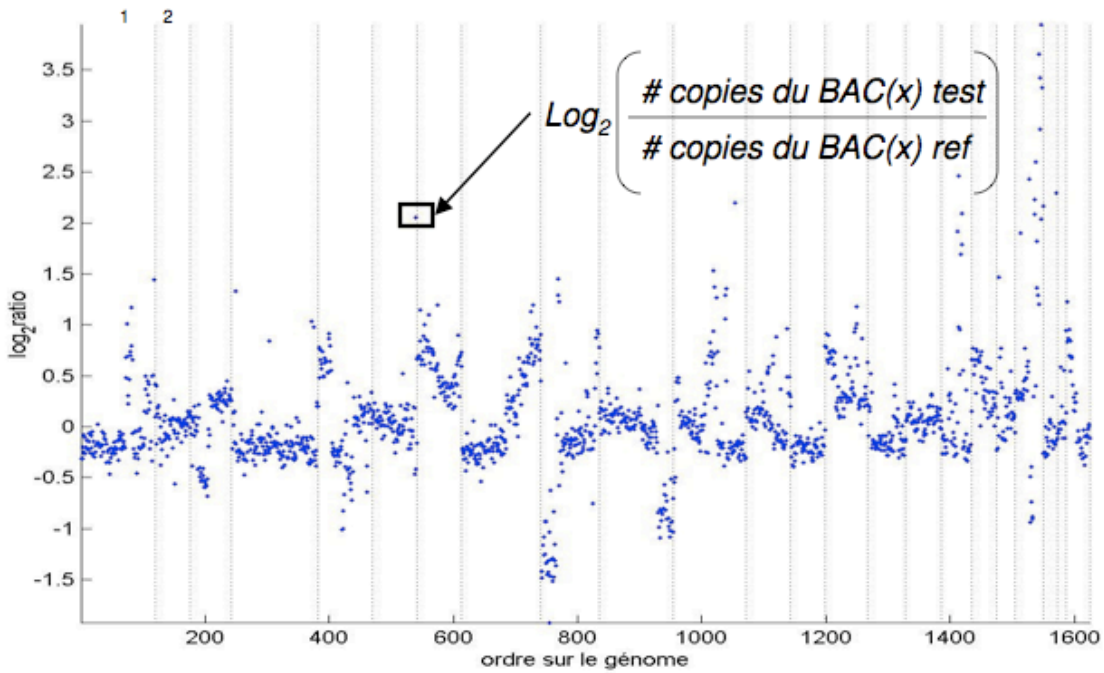
Chromosomic aberrations in cancer



Comparative Genomic Hybridization (CGH)

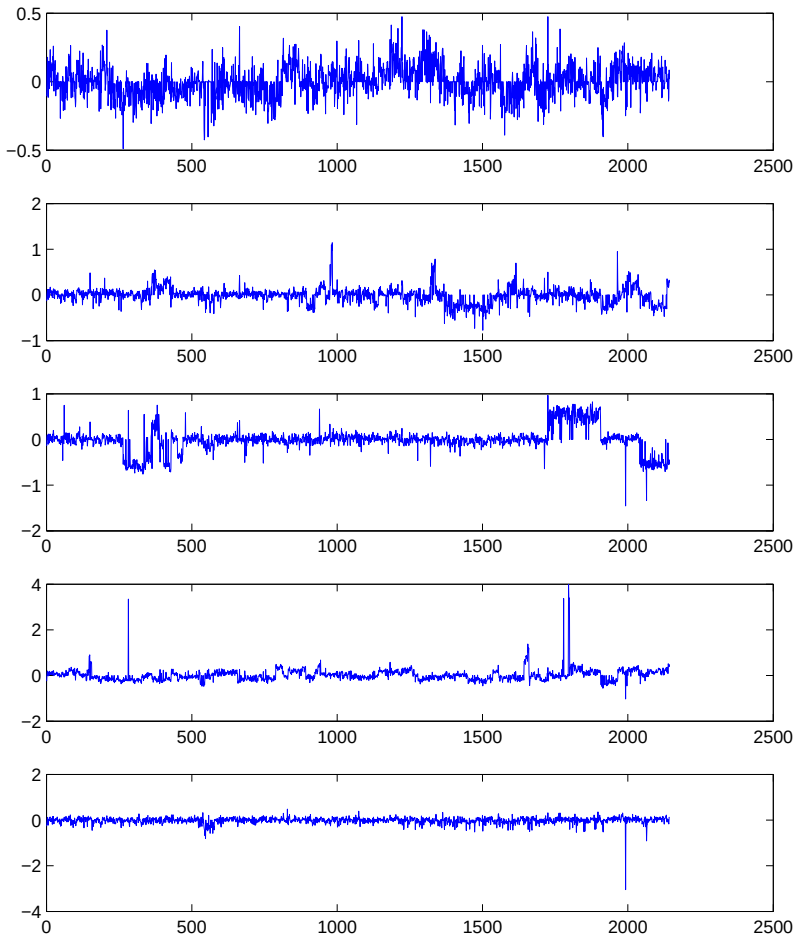
Motivation

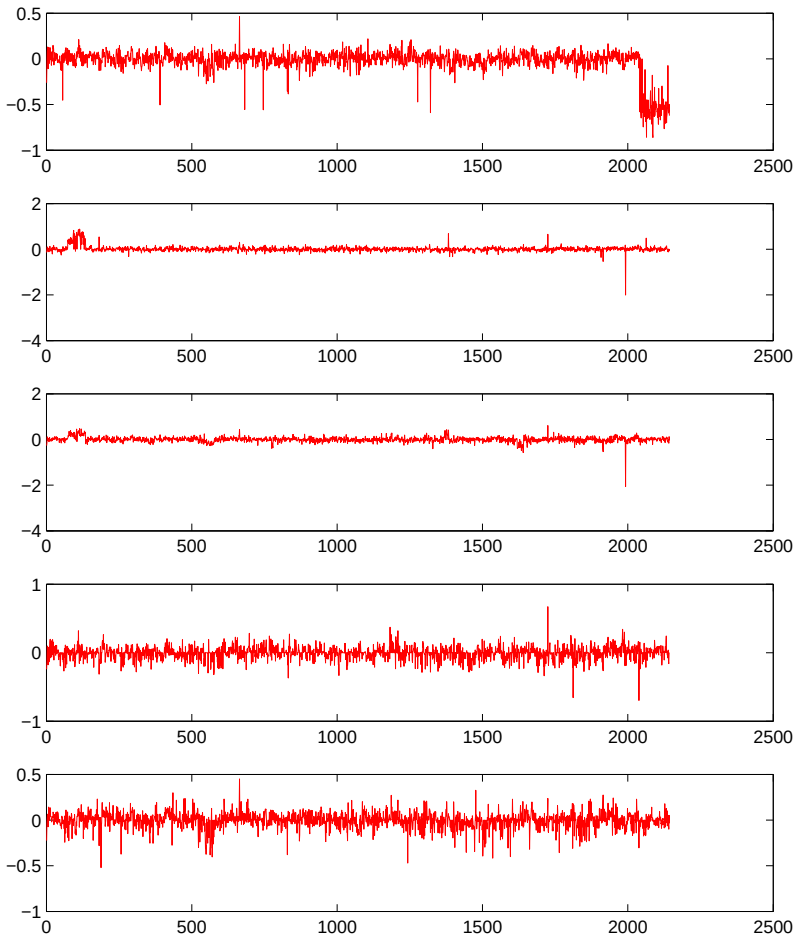
- Comparative genomic hybridization (CGH) data measure the DNA copy number along the genome
- Very useful, in particular in cancer research
- Can we classify CGH arrays for diagnosis or prognosis purpose?



Jain et al. Genome research 2002 12:325-332

Aggressive vs non-aggressive melanoma





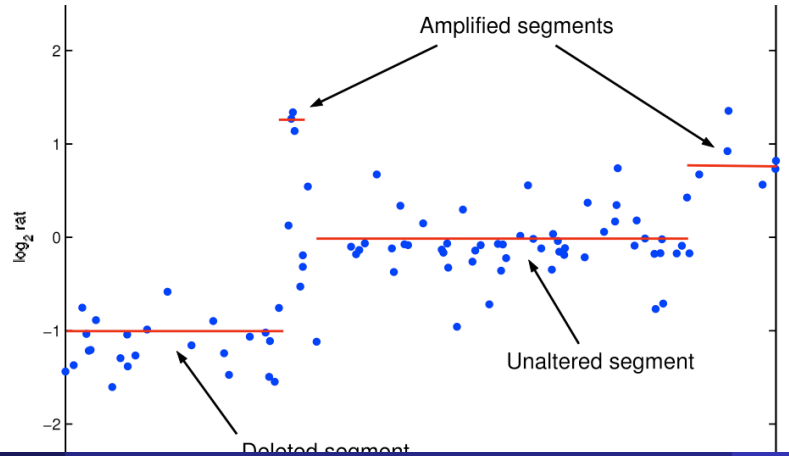
Example: CGH array classification

Prior knowledge

- For a CGH profile $x = (x_1, \ldots, x_p)$, we focus on linear classifiers, i.e., the sign of :

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

- We expect β to be
  - sparse : not all positions should be discriminative
  - piecewise constant : within a selected region, all probes should contribute equally



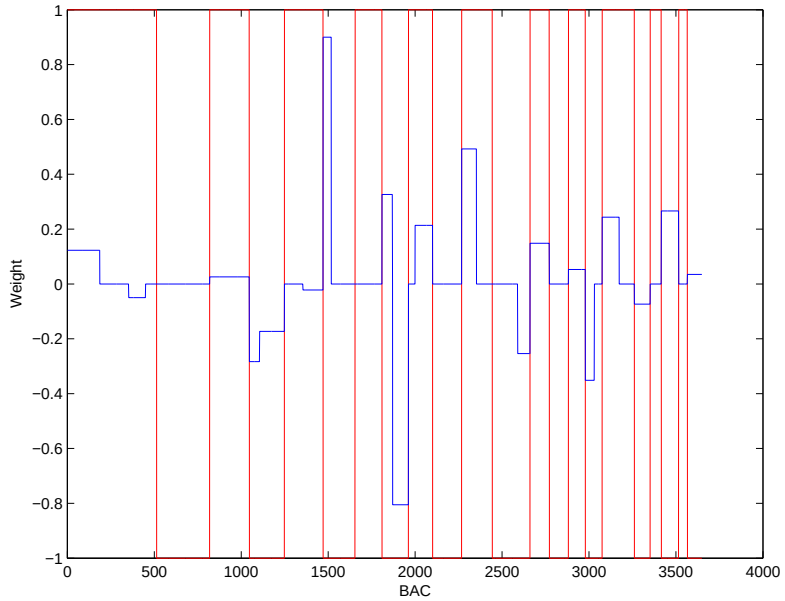
A penalty for CGH array classification

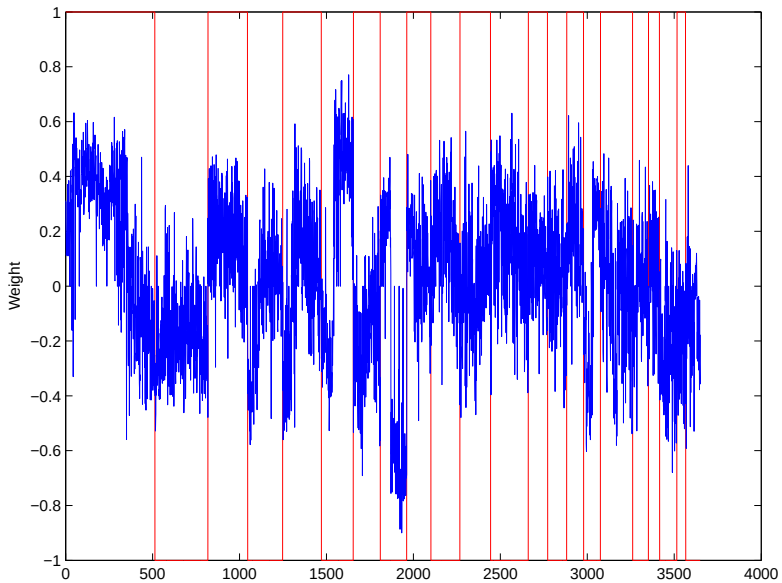
The fused LASSO penalty (Tibshirani et al., 2005)

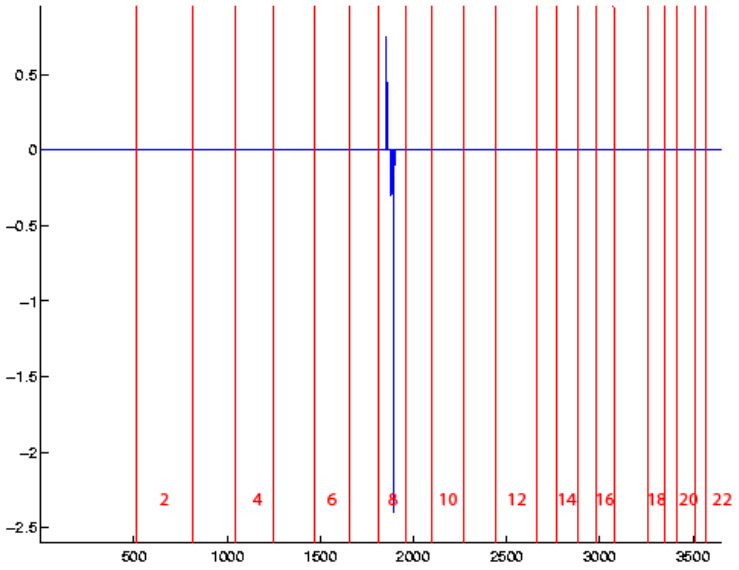
$$\Omega_{fusedlasso}(\beta) = \sum_i |\beta_i| + \sum_{i \sim j} |\beta_i - \beta_j| \,.$$

- First term leads to sparse solutions
- Second term leads to piecewise constant solutions
- Combined with a hinge loss leads to a fused SVM (Rapaport et al., 2008);

Application: metastasis prognosis in melanoma

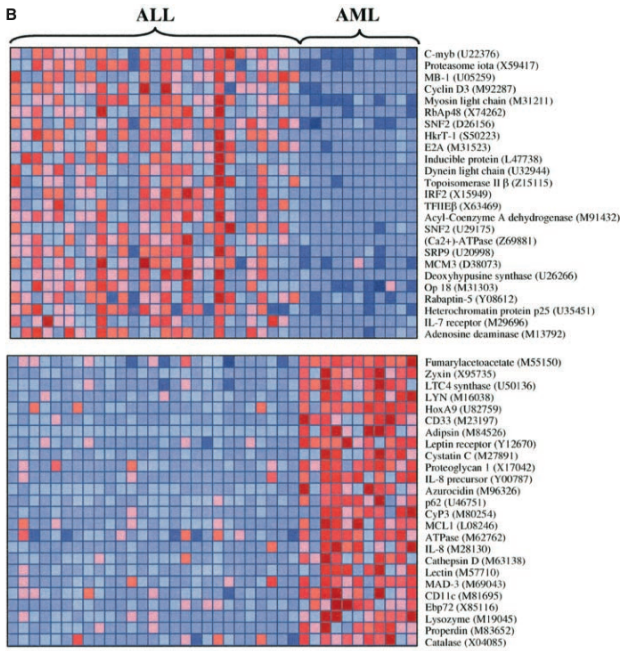






Outline

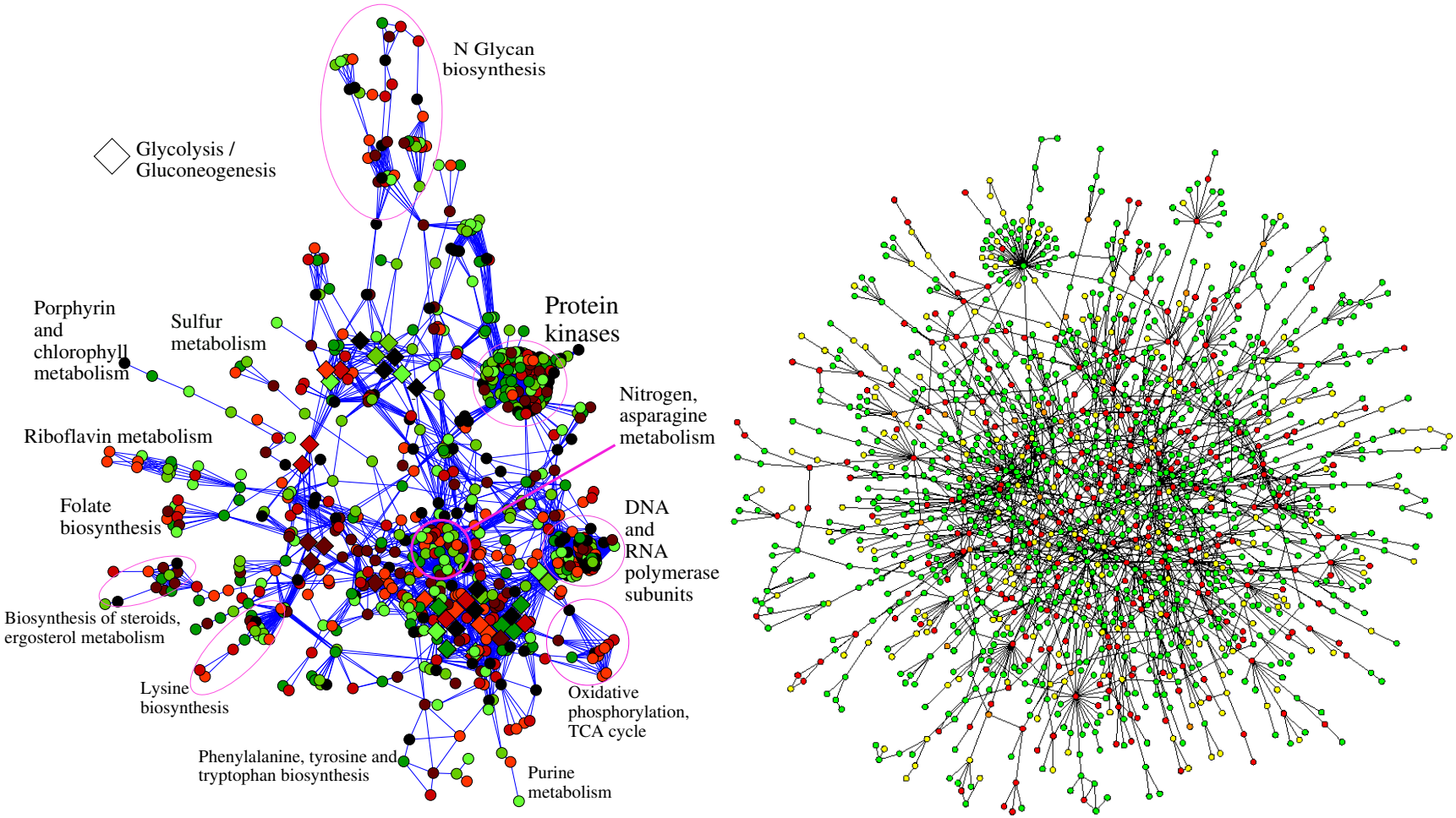
Motivation



Challenging the idea of gene signature

- We often observe little stability in the genes selected...
- Is gene selection the most biologically relevant hypothesis?
- What about thinking instead of "pathways" or "modules" signatures?

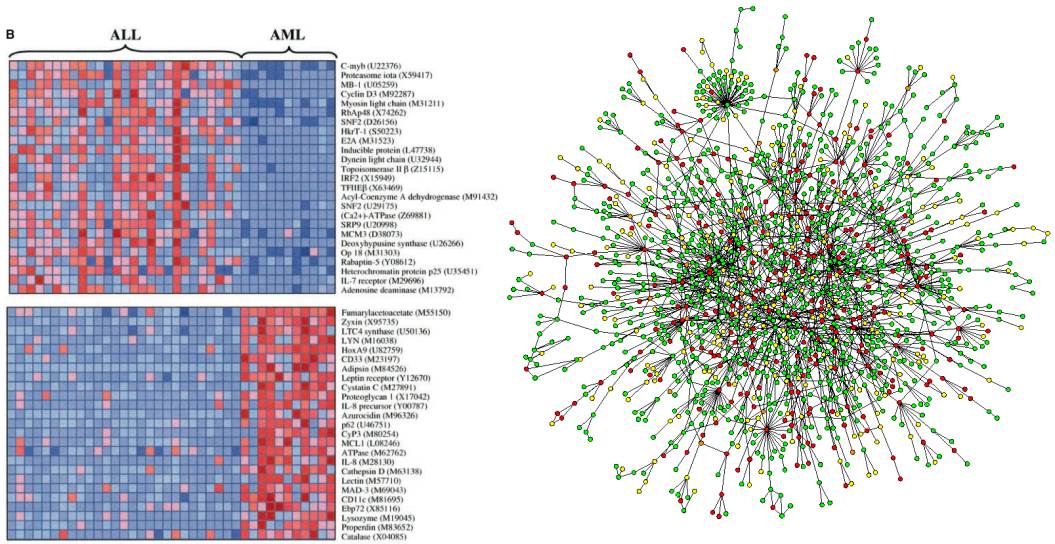
Gene networks



Gene networks and expression data

Motivation

- 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



Graph based penalty

Prior hypothesis

Genes near each other on the graph should have similar weigths.

Two solutions (Rapaport et al., 2007, 2008)

$$\Omega_{spectral}(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 \,,$$

$$\Omega_{graphfusion}(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_i |\beta_i| \,.$$

Graph based penalty

Prior hypothesis

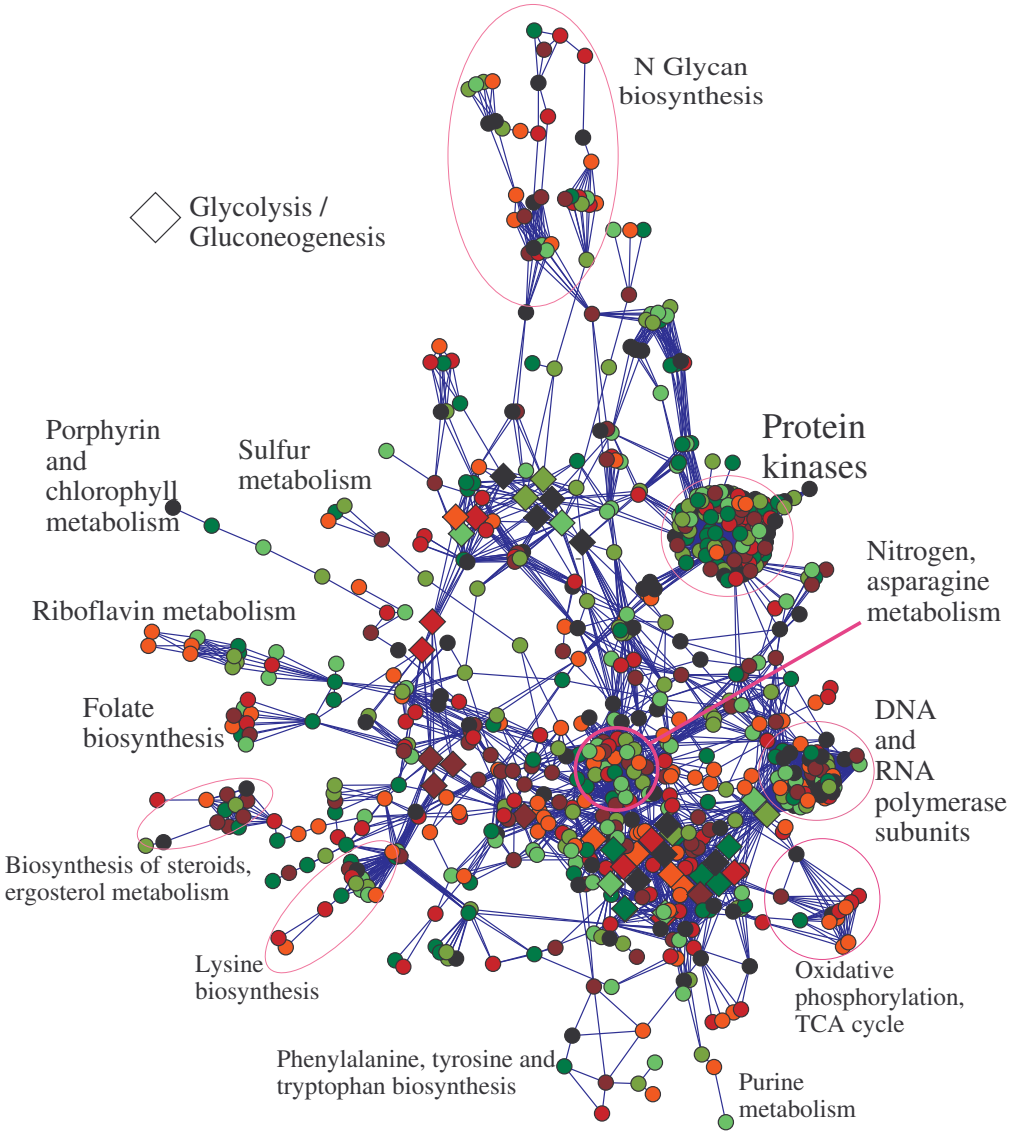
Genes near each other on the graph should have similar weigths.

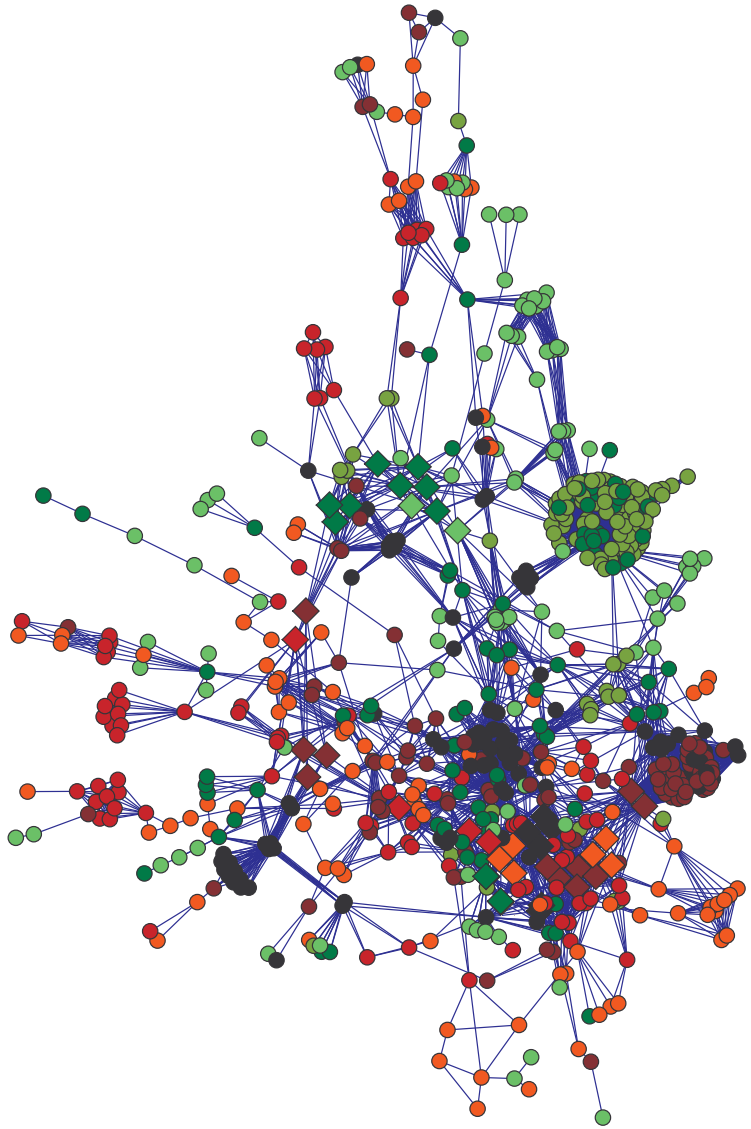
Two solutions (Rapaport et al., 2007, 2008)

$$\Omega_{spectral}(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 \,,$$

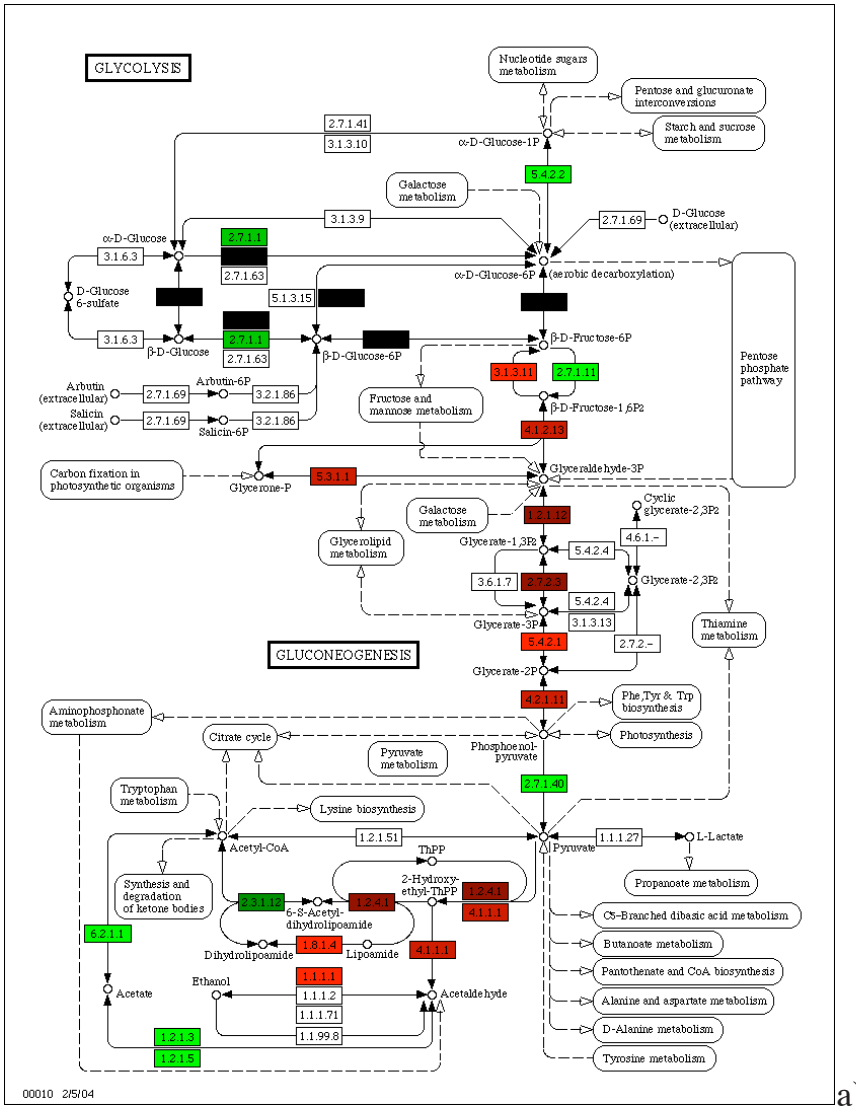
$$\Omega_{graphfusion}(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_i |\beta_i| \,.$$

Classifiers

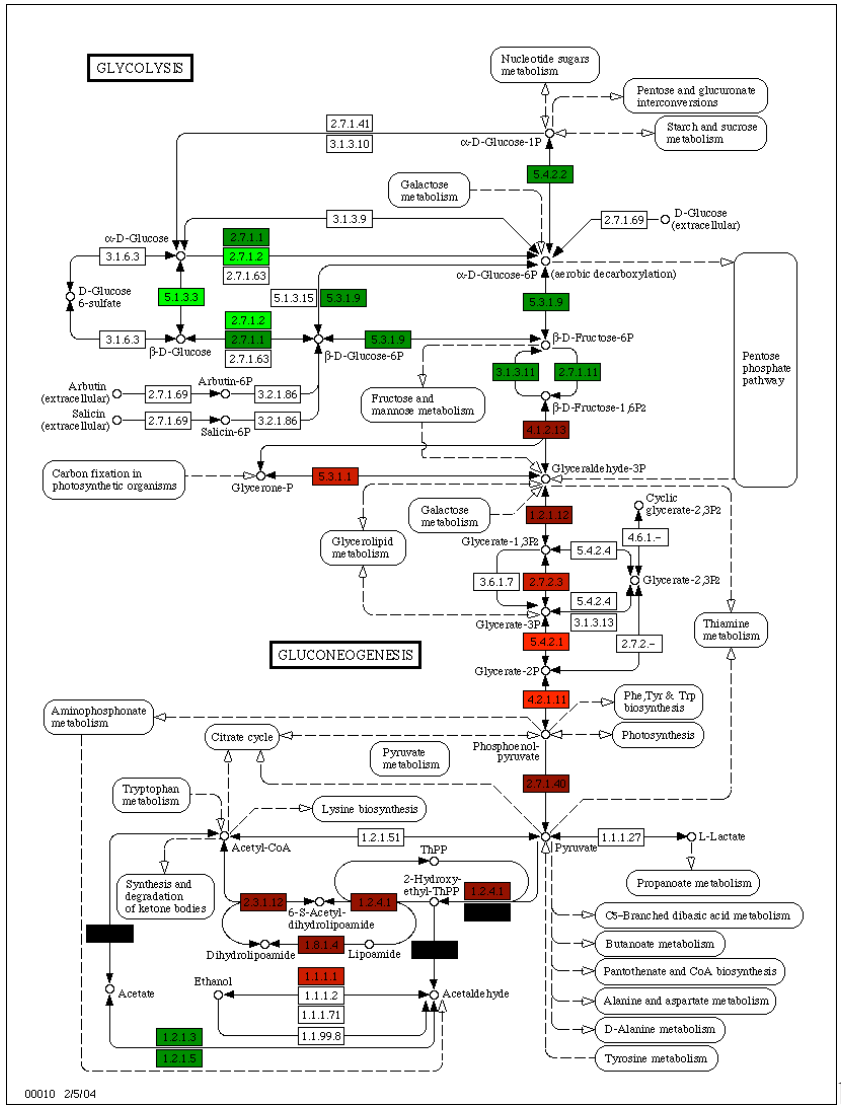




Classifier



a)



b)

Example: finding discriminant modules in gene networks

Prior hypothesis

Genes near each other on the graph should have non-zero weigths (i.e., the support of β should be made of a few connected components).

Two solutions?

$$\Omega_{intersection}(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2}\,,$$

$$\Omega_{union}(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta\,.$$

Example: finding discriminant modules in gene networks

Prior hypothesis

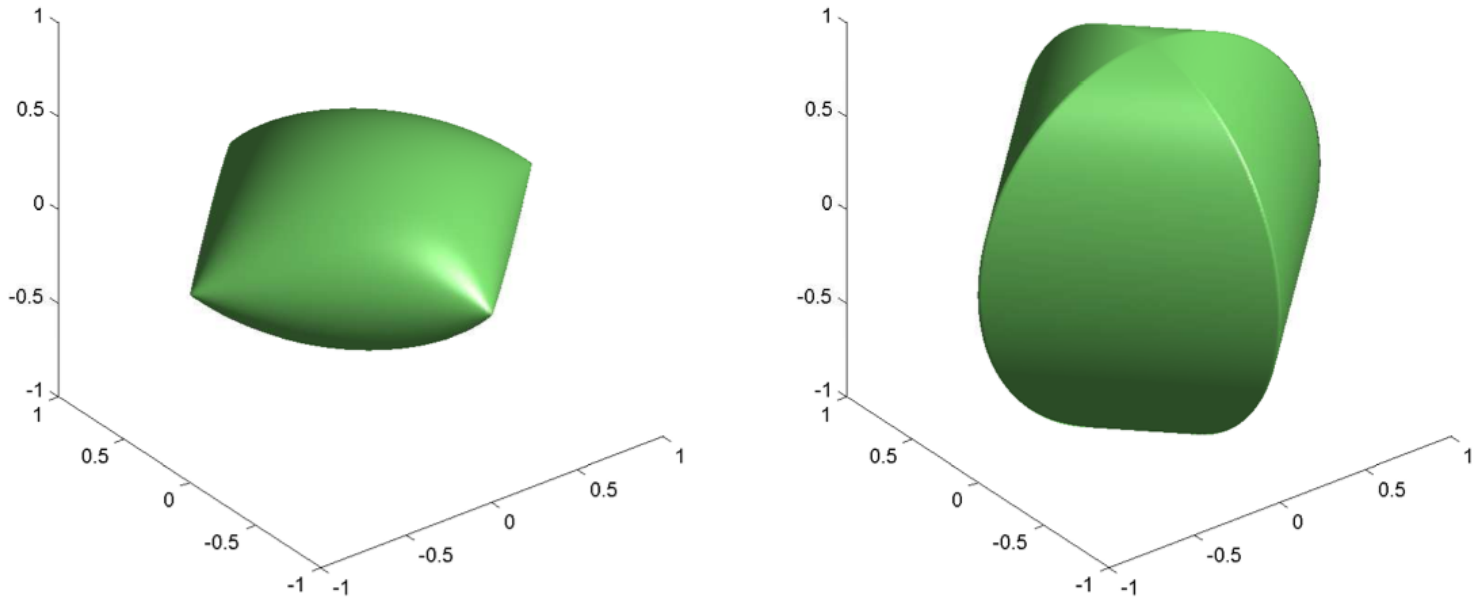
Genes near each other on the graph should have non-zero weigths (i.e., the support of β should be made of a few connected components).

Two solutions?

$$\Omega_{intersection}(\beta) = \sum_{i \sim j} \sqrt{\beta_i^2 + \beta_j^2}\,,$$

$$\Omega_{union}(\beta) = \sup_{\alpha \in \mathbb{R}^p : \forall i \sim j, \|\alpha_i^2 + \alpha_j^2\| \leq 1} \alpha^\top \beta\,.$$

Example: finding discriminant modules in gene networks



Groups $(1,2)$ *and* $(2,3)$. *Left:* $\Omega_{intersection}(\beta)$. *Right:* $\Omega_{union}(\beta)$. *Vertical axis is* β_2.

Graph lasso vs kernel on graph

- Graph lasso:

$$\Omega_{\text{graph lasso}}(w) = \sum_{i \sim j} \sqrt{w_i^2 + w_j^2}\,.$$

  constrains the sparsity, not the values
- Graph kernel

$$\Omega_{\text{graph kernel}}(w) = \sum_{i \sim j} (w_i - w_j)^2\,.$$

  constrains the values (smoothness), not the sparsity

Preliminary results

Breast cancer data

- Gene expression data for 8,141 genes in 295 breast cancer tumors.
- Canonical pathways from MSigDB containing 639 groups of genes, 637 of which involve genes from our study.

METHOD	ℓ_1	$\Omega_{group}(.)$
ERROR	0.38 ± 0.04	0.36 ± 0.03
♯ PATH.	148, 58, 183	6, 5, 78
PROP. PATH.	0.32, 0.14, 0.41	0.01, 0.01, 0.17

- Graph on the genes.

METHOD	ℓ_1	$\Omega_{graph}(.)$
ERROR	0.39 ± 0.04	0.36 ± 0.01
AV. SIZE C.C.	1.1, 1, 1.0	1.3, 1.4, 1.2

Conclusion

- Machine learning provides many solutions for the analysis of high-throughput data (more examples later..)
- The development of dedicated method is increasingly important to overcome the challenges (few samples, high-dimension, structures..)
- This increasingly requires tight collaboration with domain experts

Thanks!

- Franck Rapaport, Emmanuel Barillot, Andrei Zynoviev, Laurent Jacob, Anne-Claire Haury (Institut Curie / Mines ParisTech)
- Guillaume Obozinski (UC Berkeley / INRIA)

... and the INSERM-JSPS grant for collaborative research which support this workshop

Inserm

Institut national
de la santé et de la recherche médicale