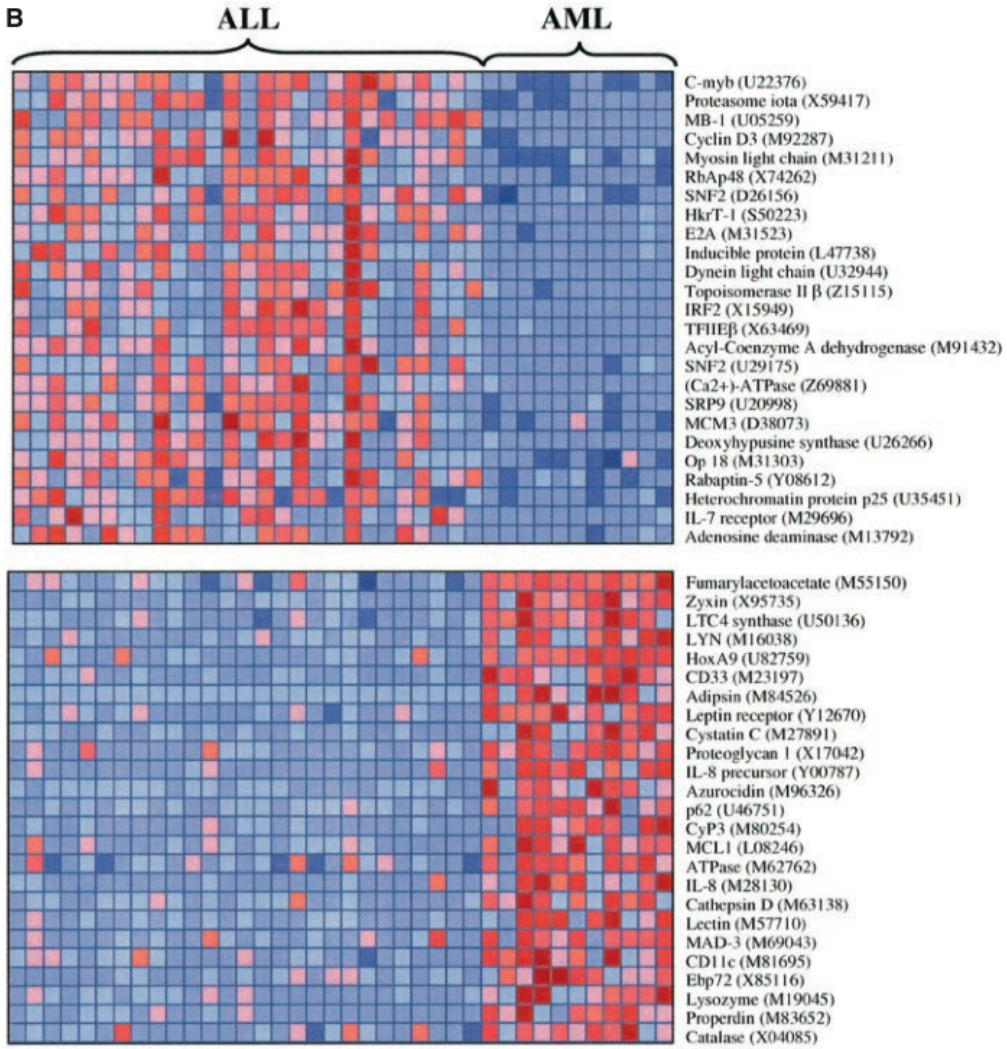
Including prior knowledge in shrinkage classifiers for genomic data

Jean-Philippe Vert
Jean-Philippe.Vert@mines-paristech.fr

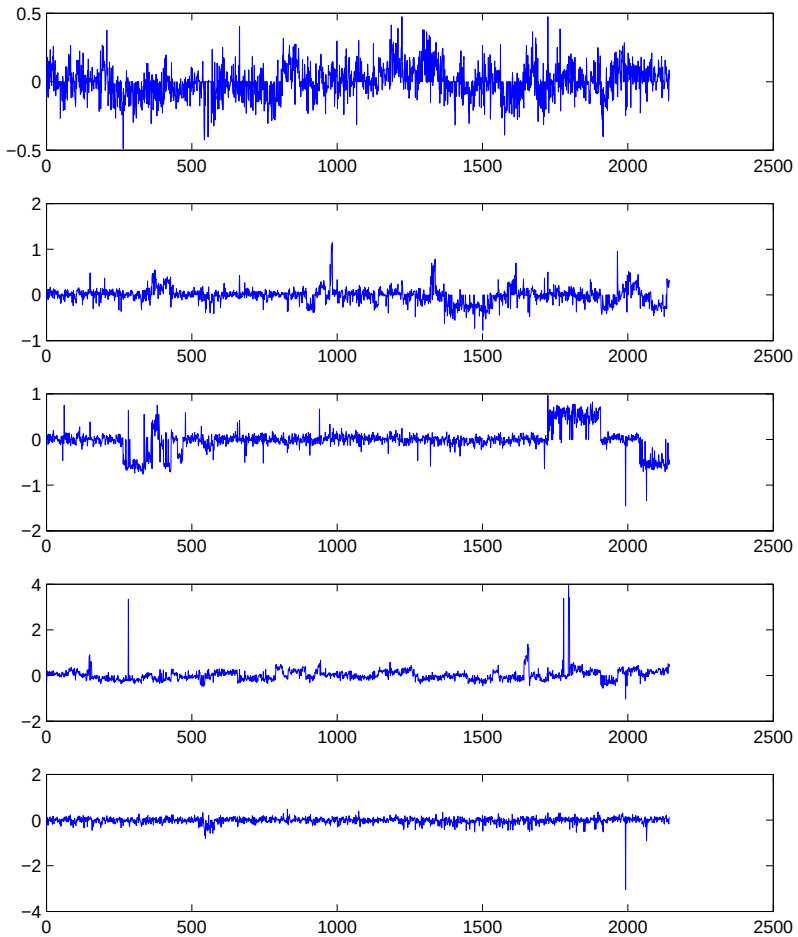
Mines ParisTech / Curie Institute / Inserm

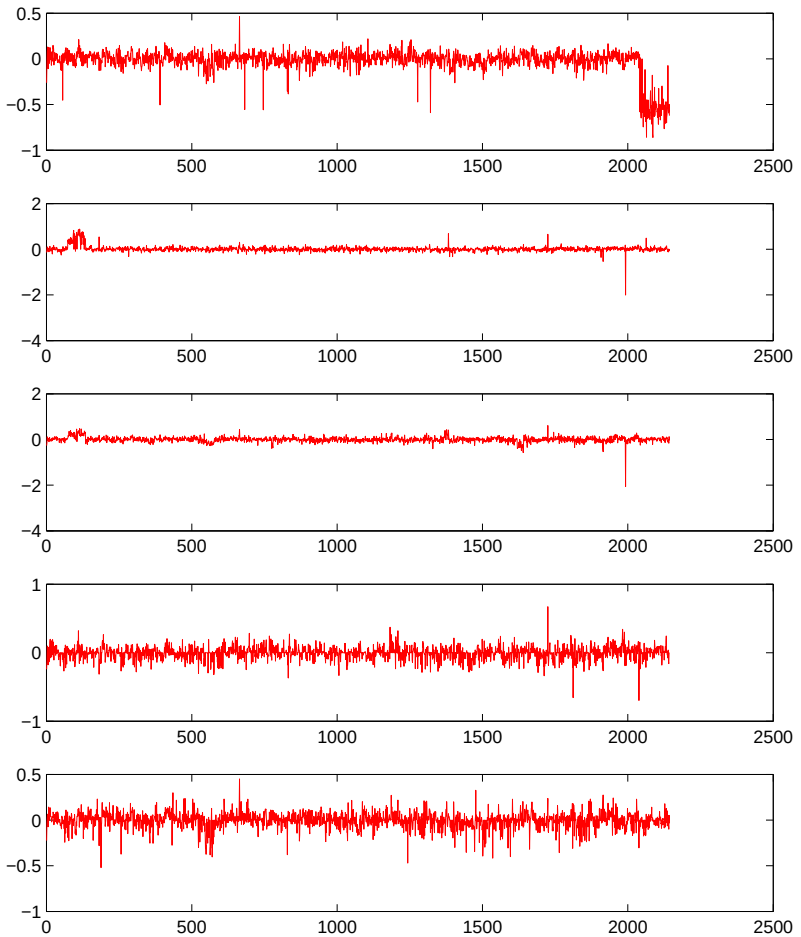
University of Liège, April 30, 2010.

Cancer diagnosis

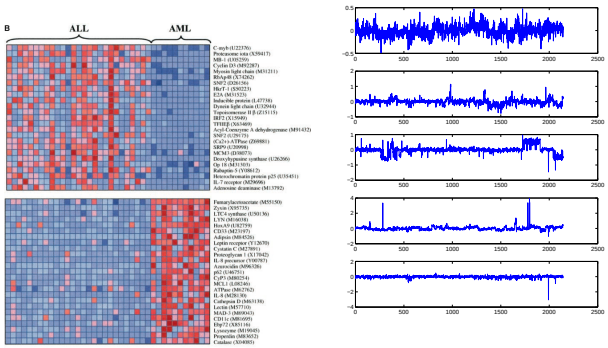


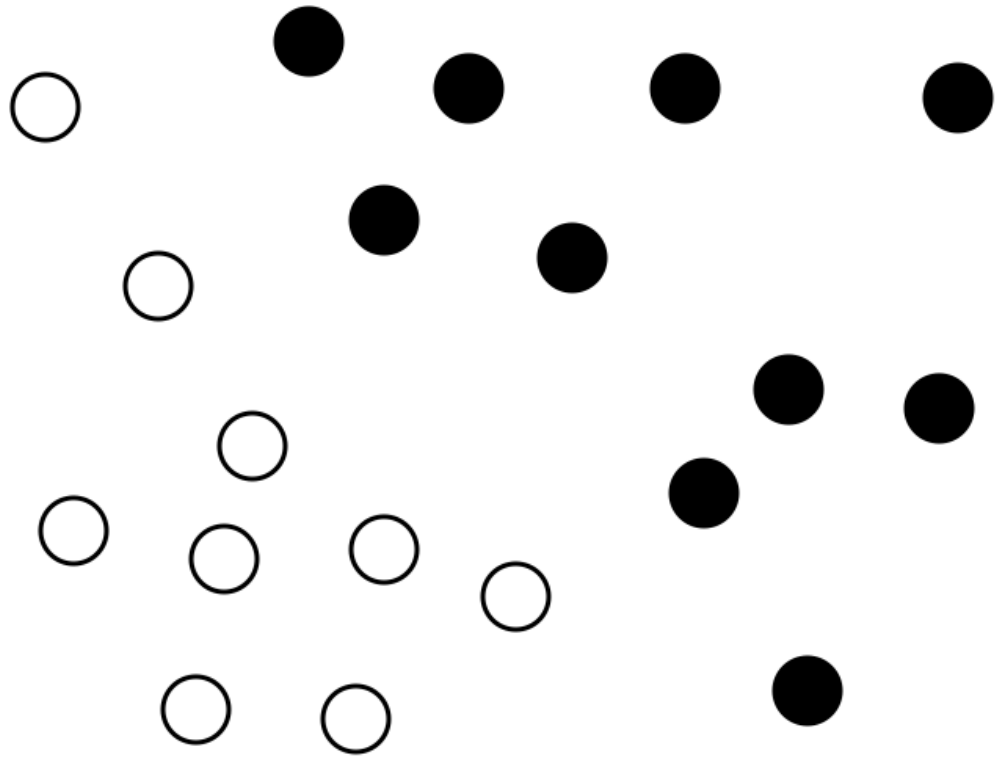
Cancer prognosis



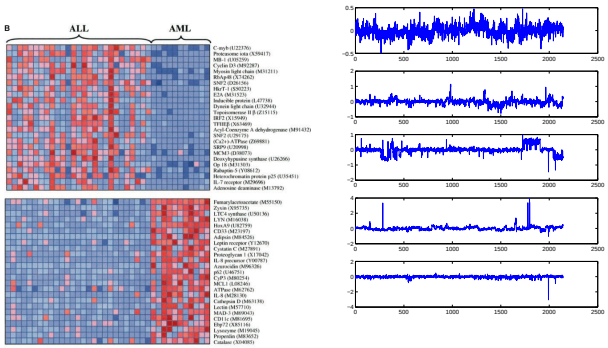


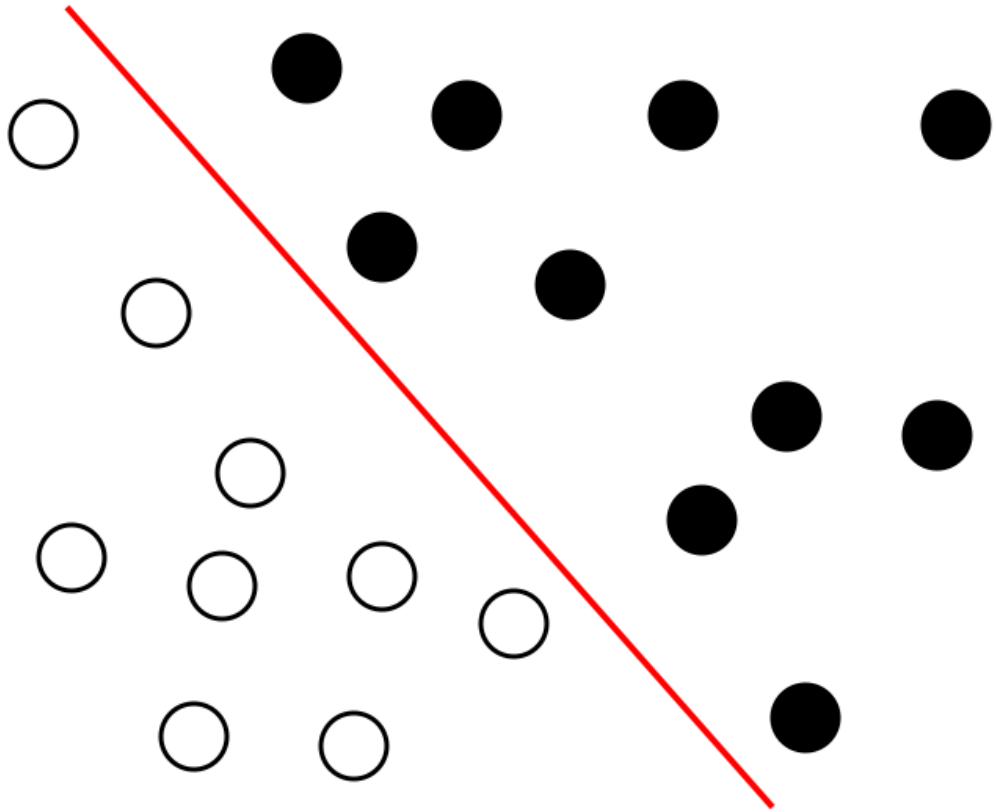
Pattern recognition, *aka* supervised classification



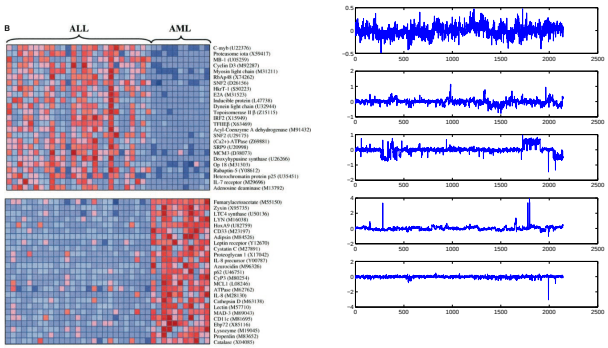


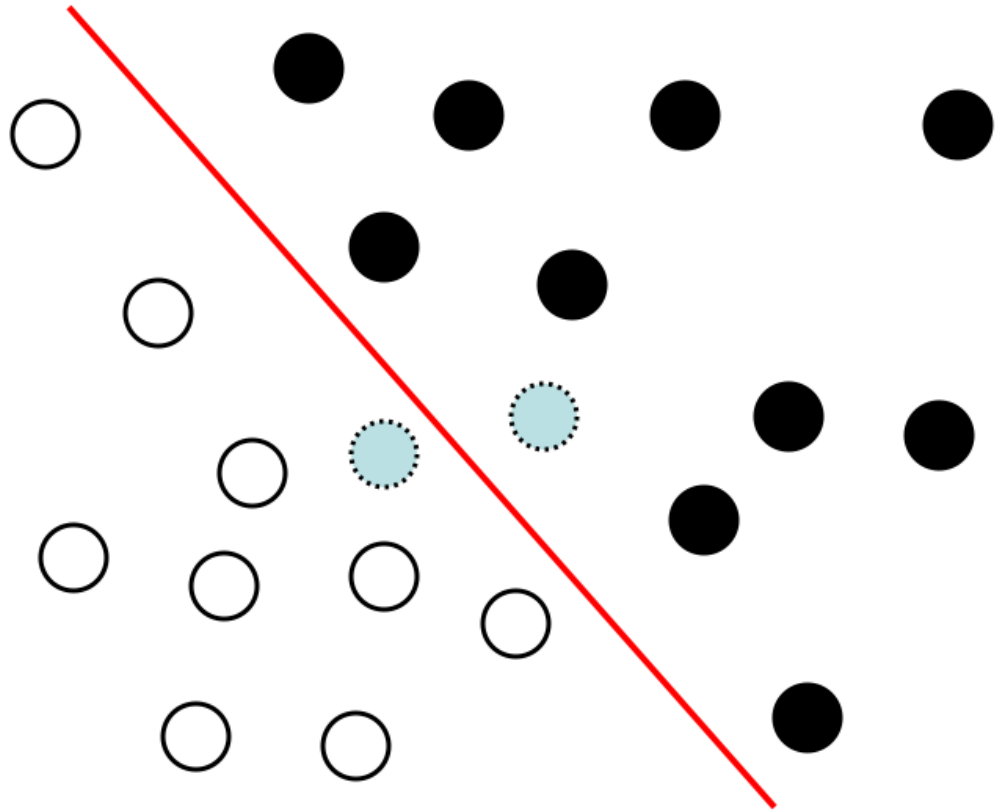
Pattern recognition, *aka* supervised classification



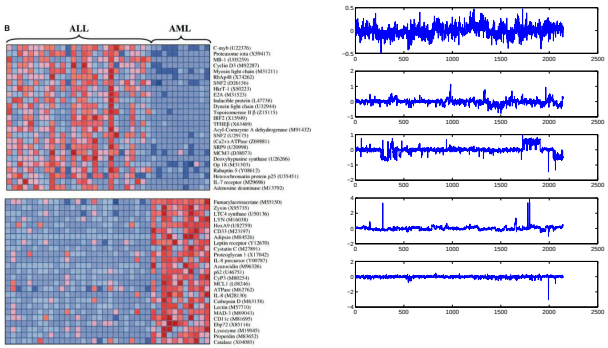


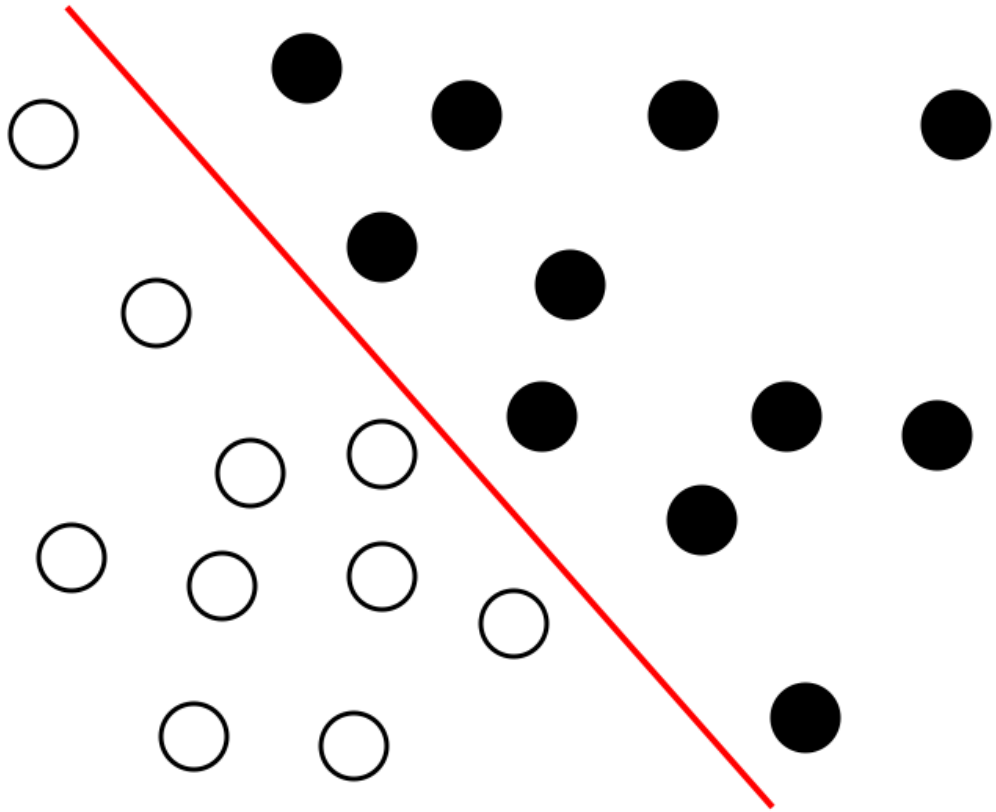
Pattern recognition, *aka* supervised classification



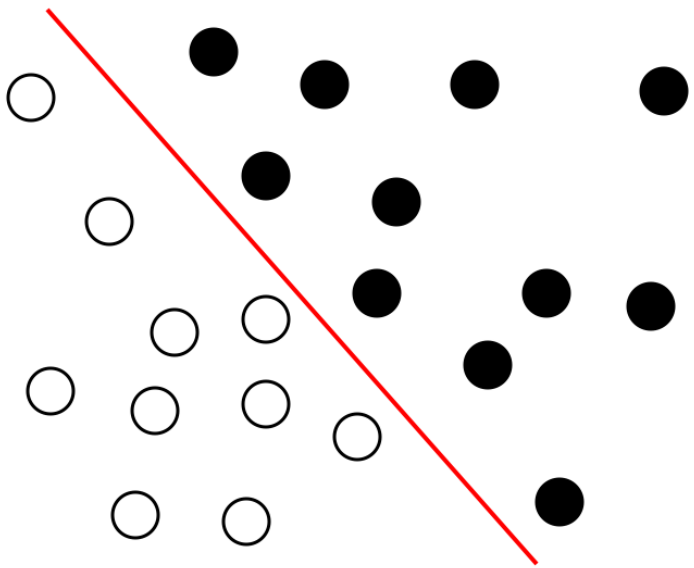


Pattern recognition, *aka* supervised classification





Pattern recognition, *aka* supervised classification



Challenges

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

Formalization

The problem

- Given a set of training instances $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i \in \mathcal{X}$ are data and $y_i \in \mathcal{Y}$ are continuous or discrete variables of interest,
- Estimate a function

$$y = f(x)$$

  where x is any new data to be labeled.
- f should be accurate and intepretable.

Linear classifiers

The model

- Each sample $x \in \mathcal{X}$ is represented by a vector of features (or descriptors, or patterns):

$$\Phi(x) = (\Phi_1(x), \ldots, \Phi_p(x)) \in \mathbb{R}^p \,.$$

- Based on the training set we estimate a linear function:

$$f_\beta(x) = \sum_{i=1}^{p} \beta_i \Phi_i(x) = \beta^\top \Phi(x) \,.$$

Estimating linear classifiers

- 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 empirical risk:

$$R(\beta) = \frac{1}{n} \sum_{i=1}^{n} l(f_\beta(x_i), y_i) \,.$$

- We choose the β that achieves the minimium empirical risk, subject to some constraint:

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

- Equivalently we solve

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

Two important questions

$$f_\beta(x) = \sum_{i=1}^{p} \beta_i \Phi_i(x)$$

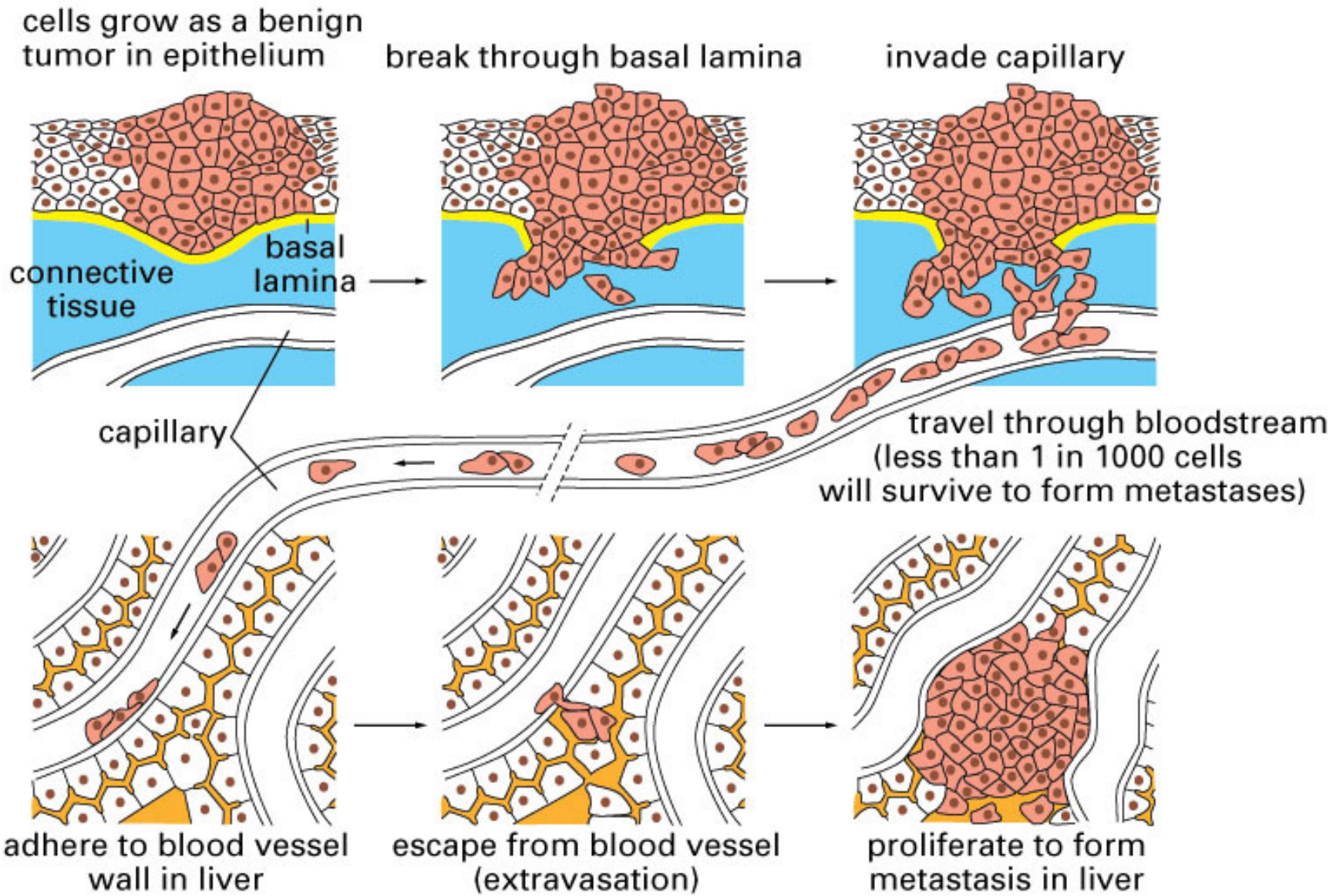
$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^{n} l(f_\beta(x_i), y_i) + \lambda \Omega(\beta)$$

- How to design the features $\Phi(x)$?
- How to choose the penalty $\Omega(\beta)$?

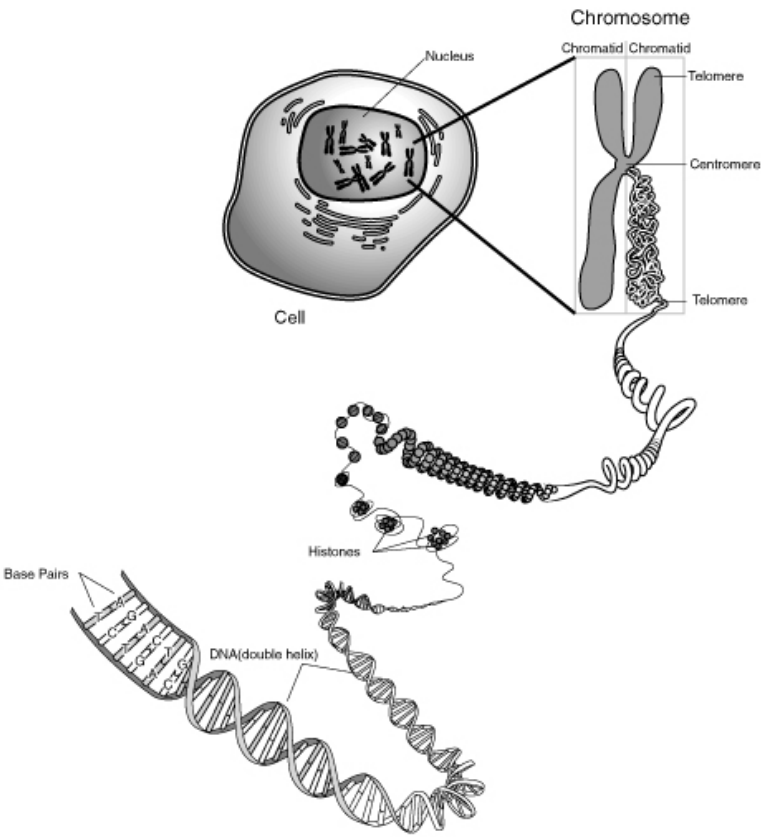
Outline

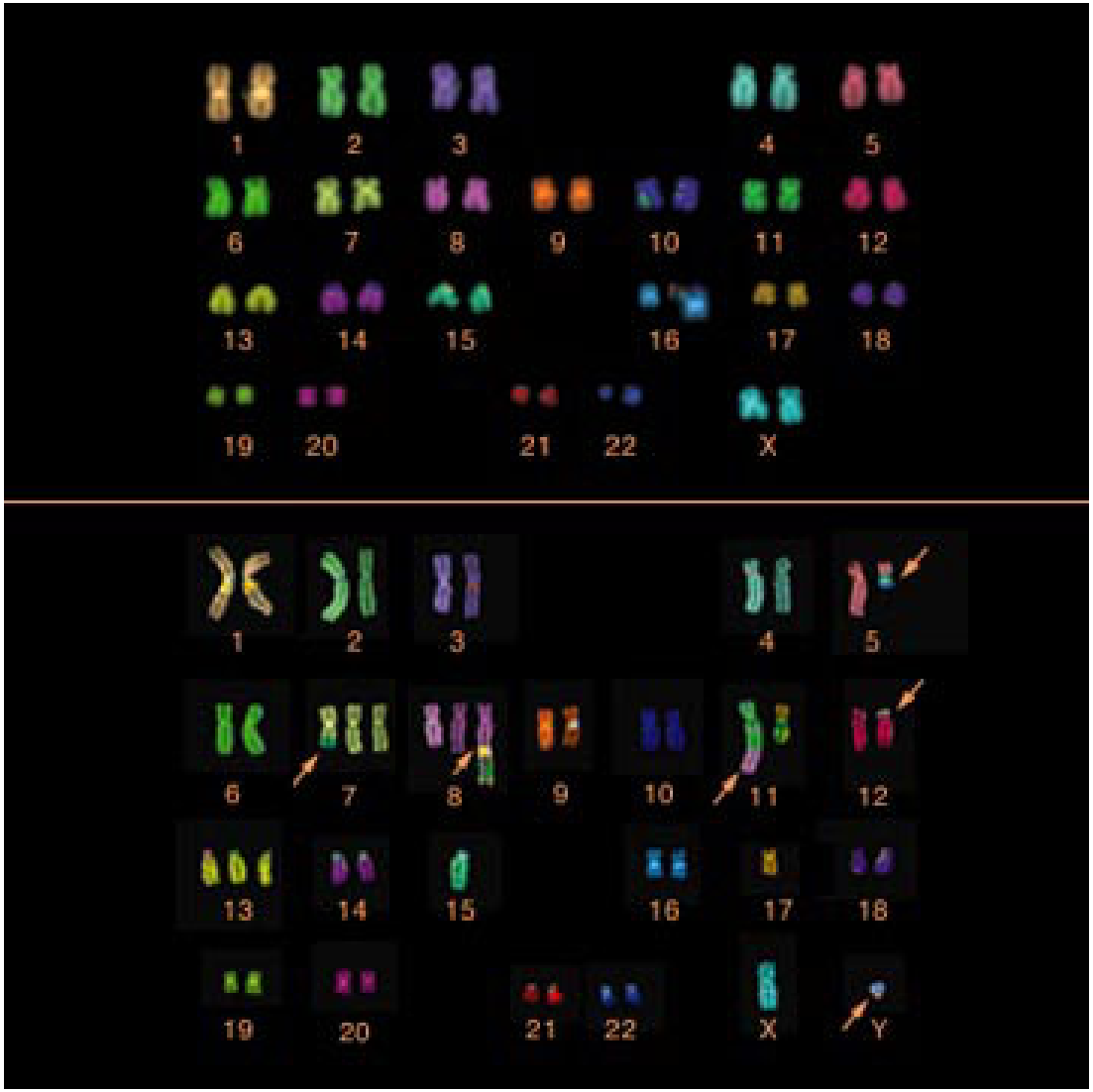
Outline

A simple view of cancer progression



Chromosomic aberrations in cancer

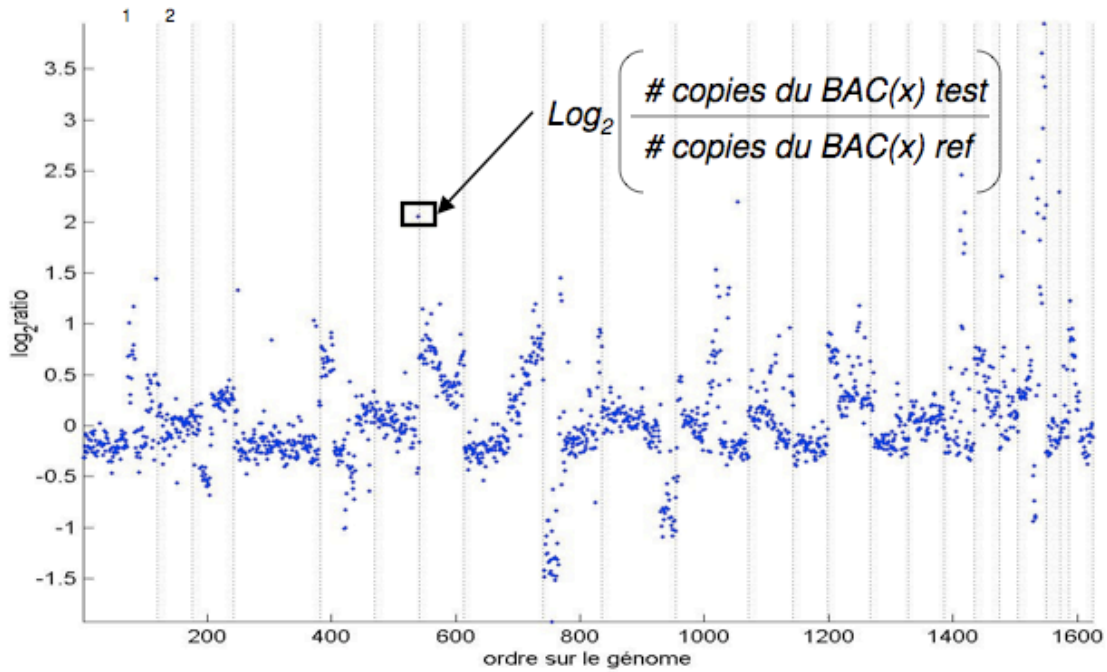




Comparative Genomic Hybridization (CGH)

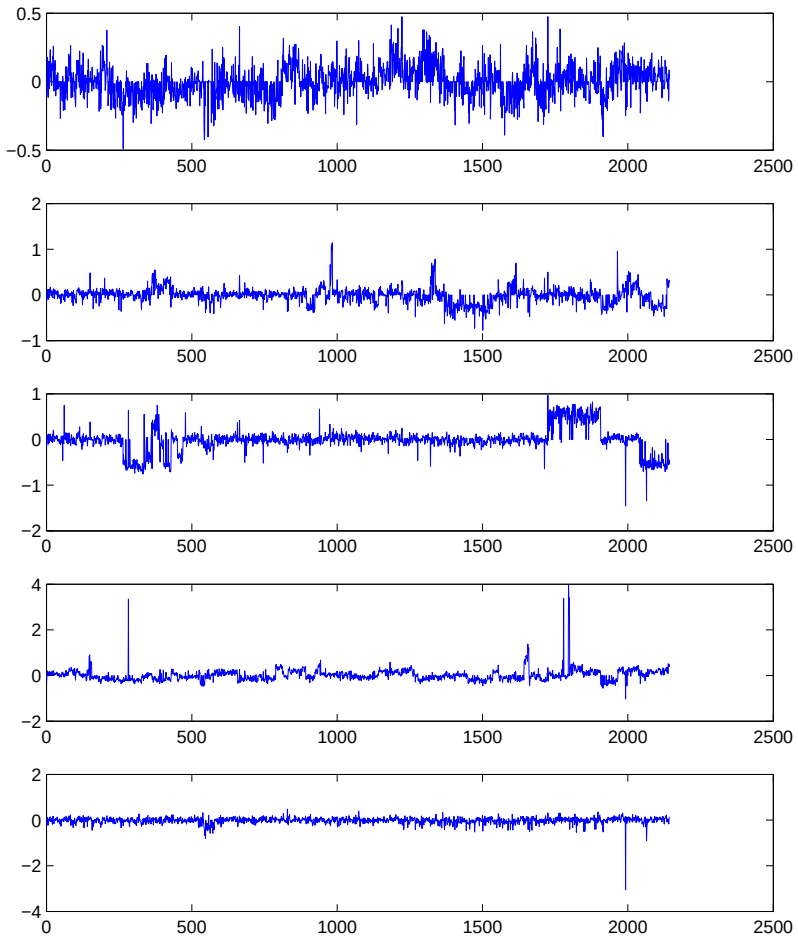
Motivation

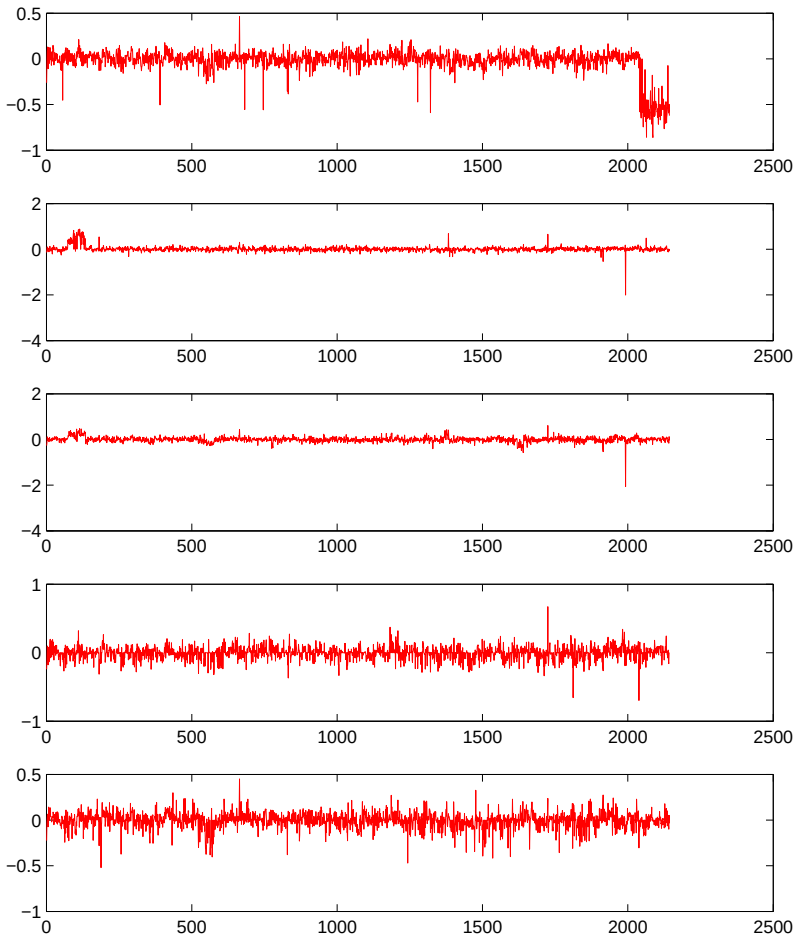
- Comparative genomic hybridization (CGH) data measure the DNA copy number along the genome
- Very useful, in particular in cancer research
- Can we classify CGH arrays for diagnosis or prognosis purpose?



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

Aggressive vs non-aggressive melanoma





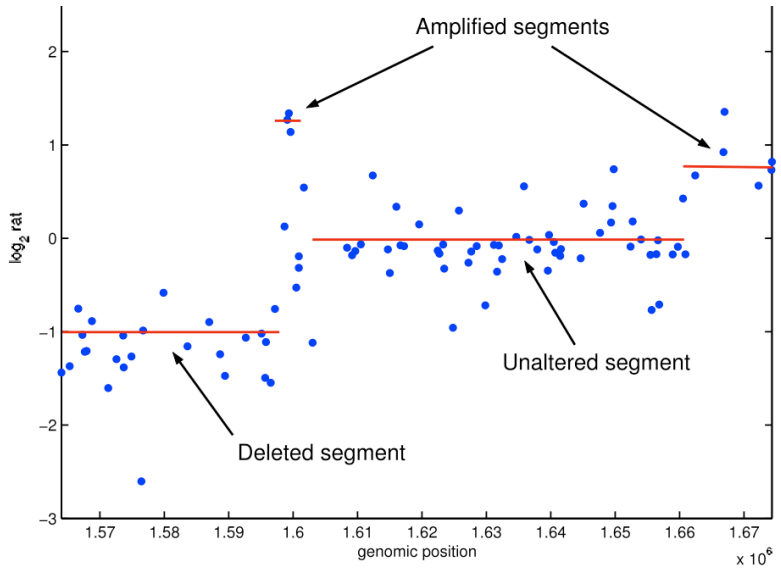
CGH array classification

Prior knowledge

- For a CGH profile $x \in \mathbb{R}^p$, we focus on linear classifiers, i.e., the sign of :

$$f_\beta(x) = \beta^\top x \,.$$

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



Promoting sparsity with the ℓ_1 penalty

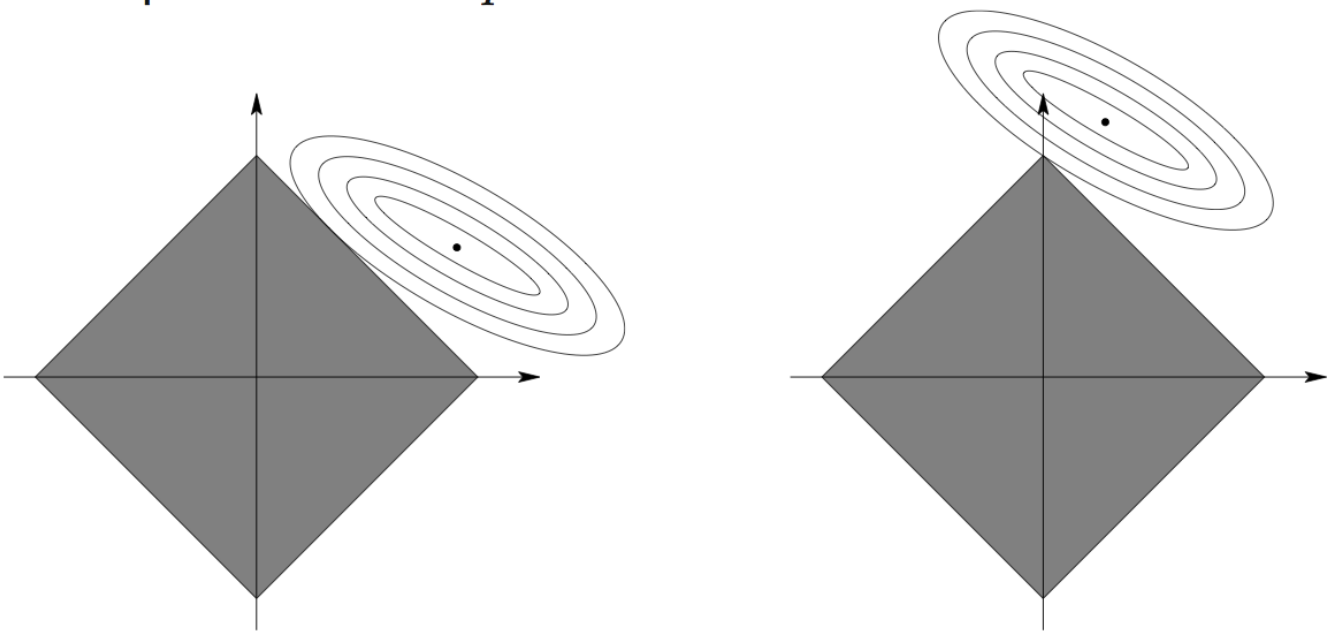
The ℓ_1 penalty (Tibshirani, 1996; Chen et al., 1998)

The solution of

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

is usually sparse.

Geometric interpretation with $p = 2$



Promoting piecewise constant profiles penalty

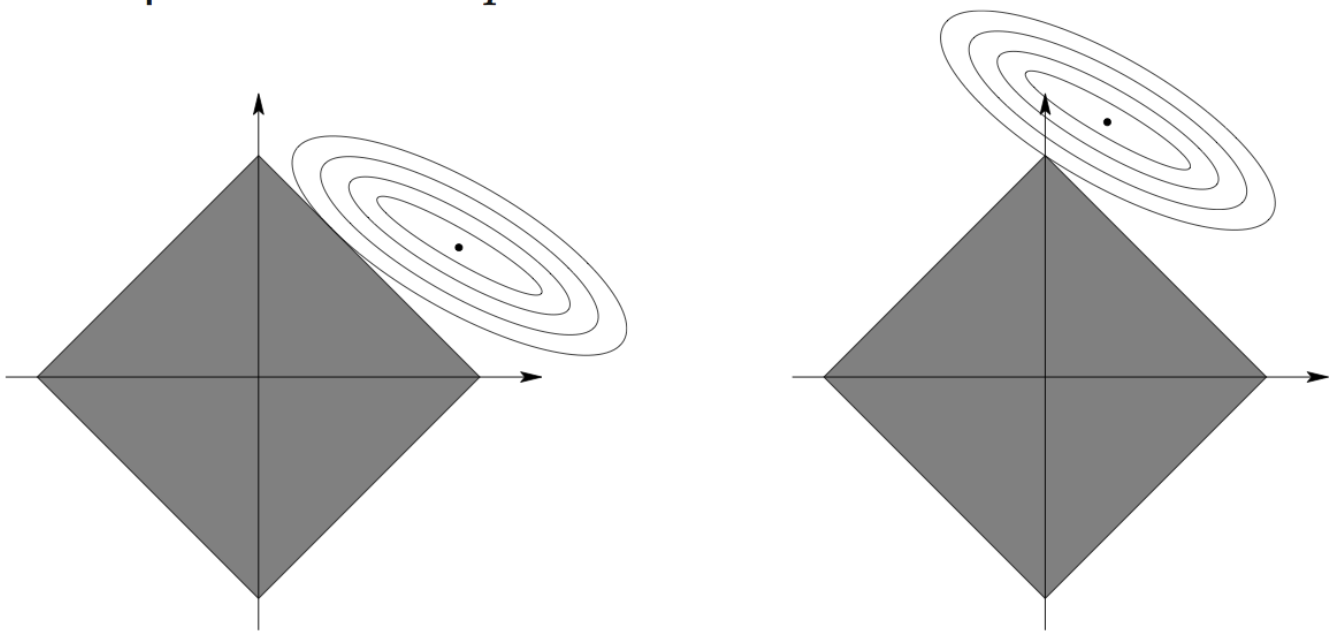
The variable fusion penalty (Land and Friedman, 1996)

The solution of

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|$$

is usually piecewise constant.

Geometric interpretation with $p = 2$



A penalty for CGH array classification

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

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

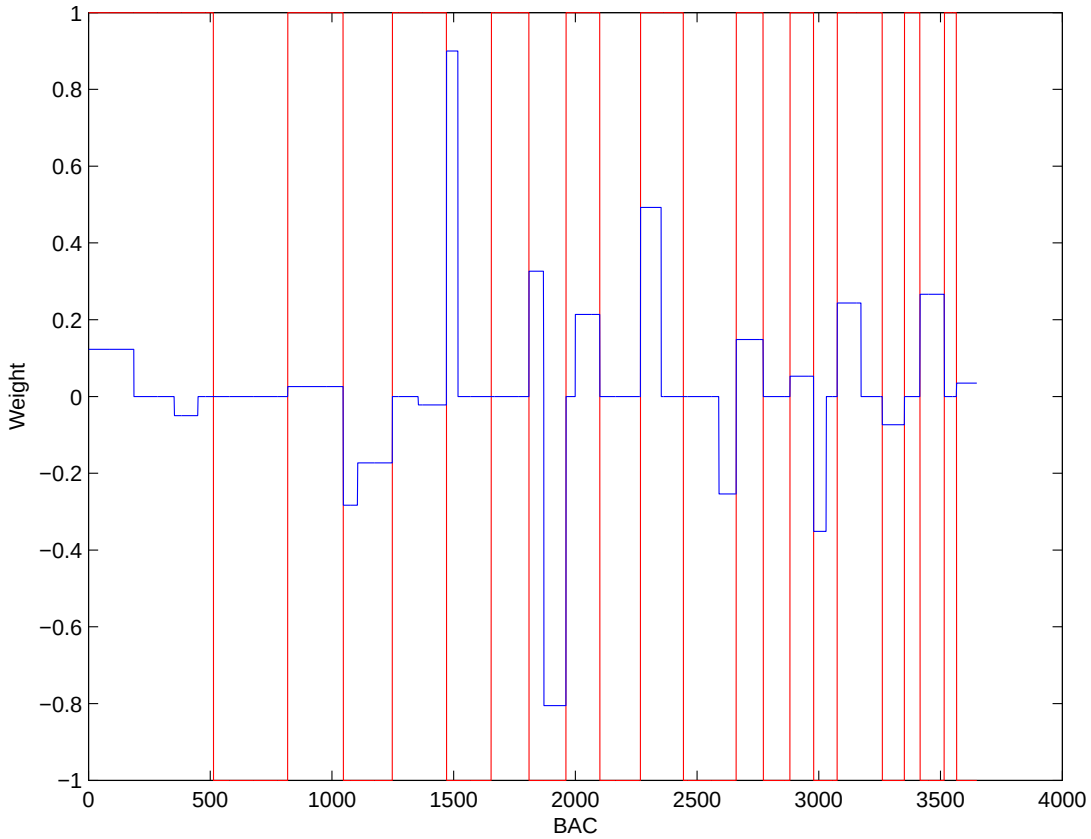
- First term leads to sparse solutions
- Second term leads to piecewise constant solutions

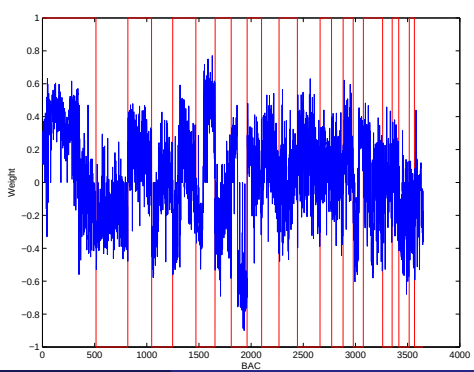
The fused SVM (Rapaport et al., 2008)

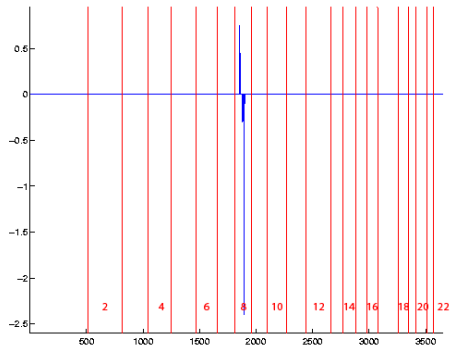
$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \ell\left(y_i, \beta^\top x_i\right) + \lambda \sum_i |\beta_i| + \mu \sum_{i \sim j} |\beta_i - \beta_j| \,.$$

where ℓ is, e.g., the hinge loss $\ell(y,t) = max(1 - yt, 0)$. It is then a LP.

Application: predicting metastasis in melanoma

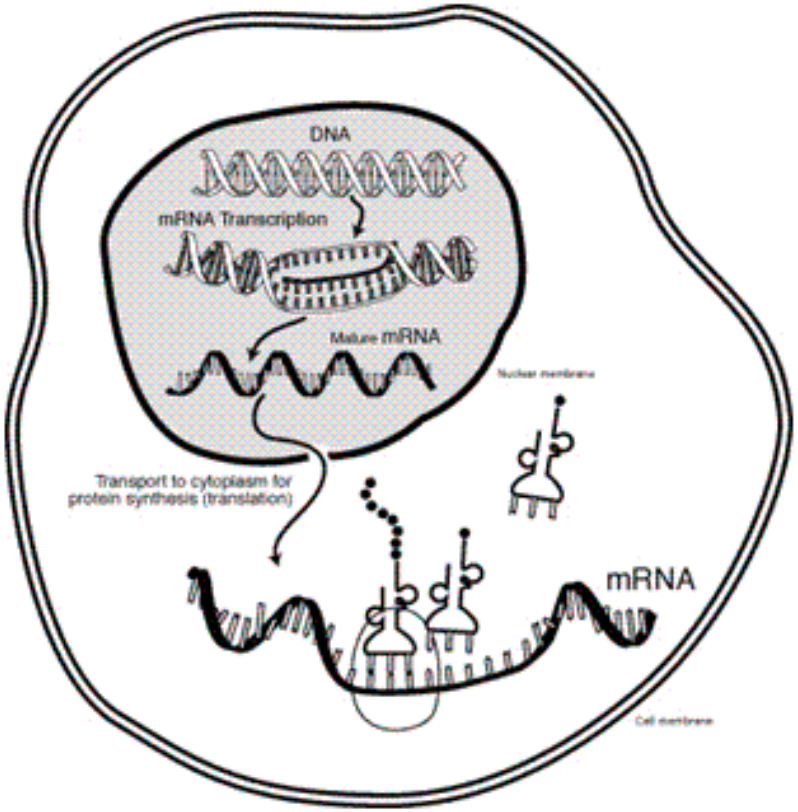


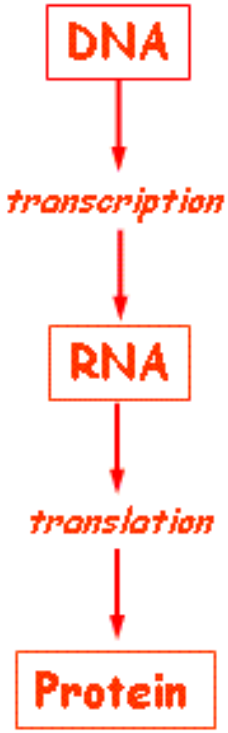




Outline

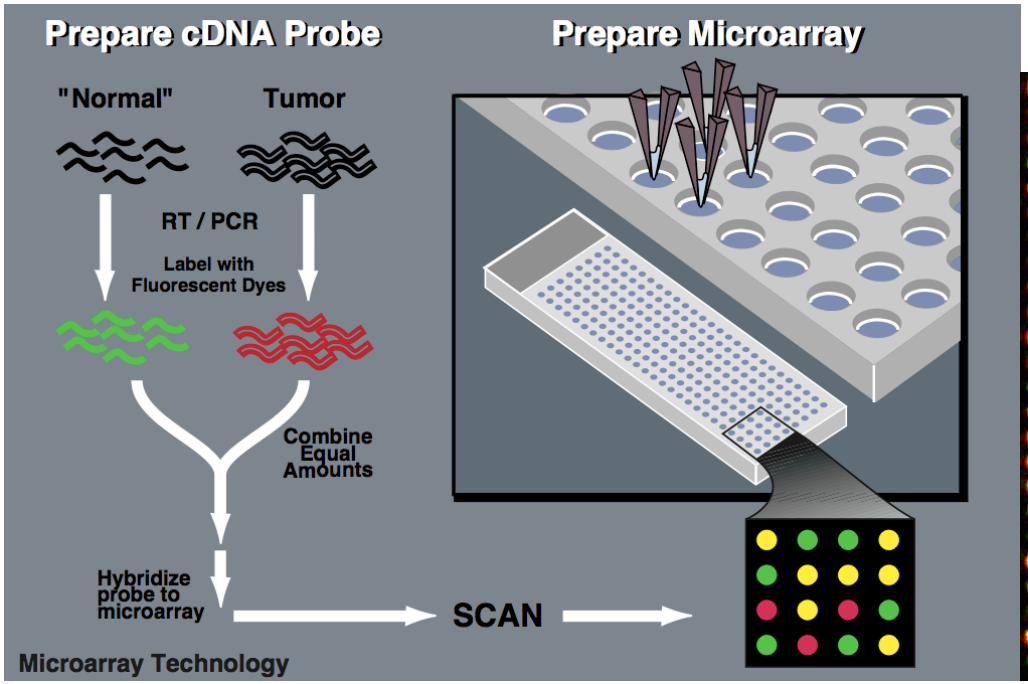
DNA → RNA → protein

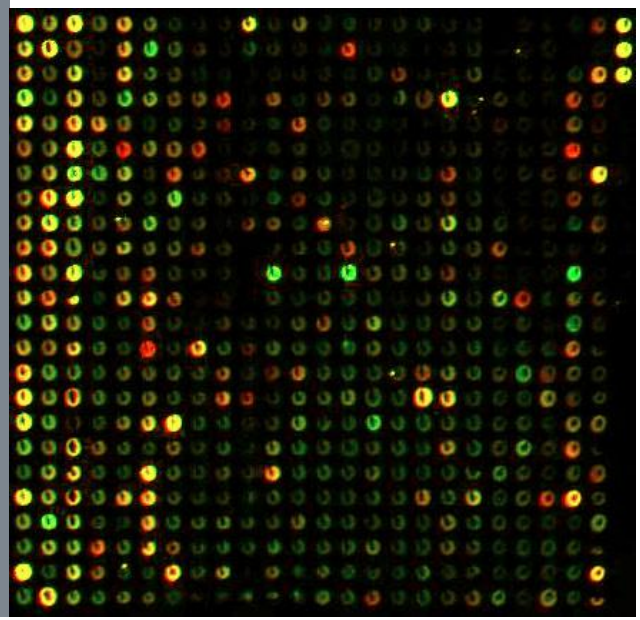




- CGH shows the (static) DNA
- Cancer cells have also abnormal (dynamic) gene expression (= transcription)

Tissue profiling with DNA chips

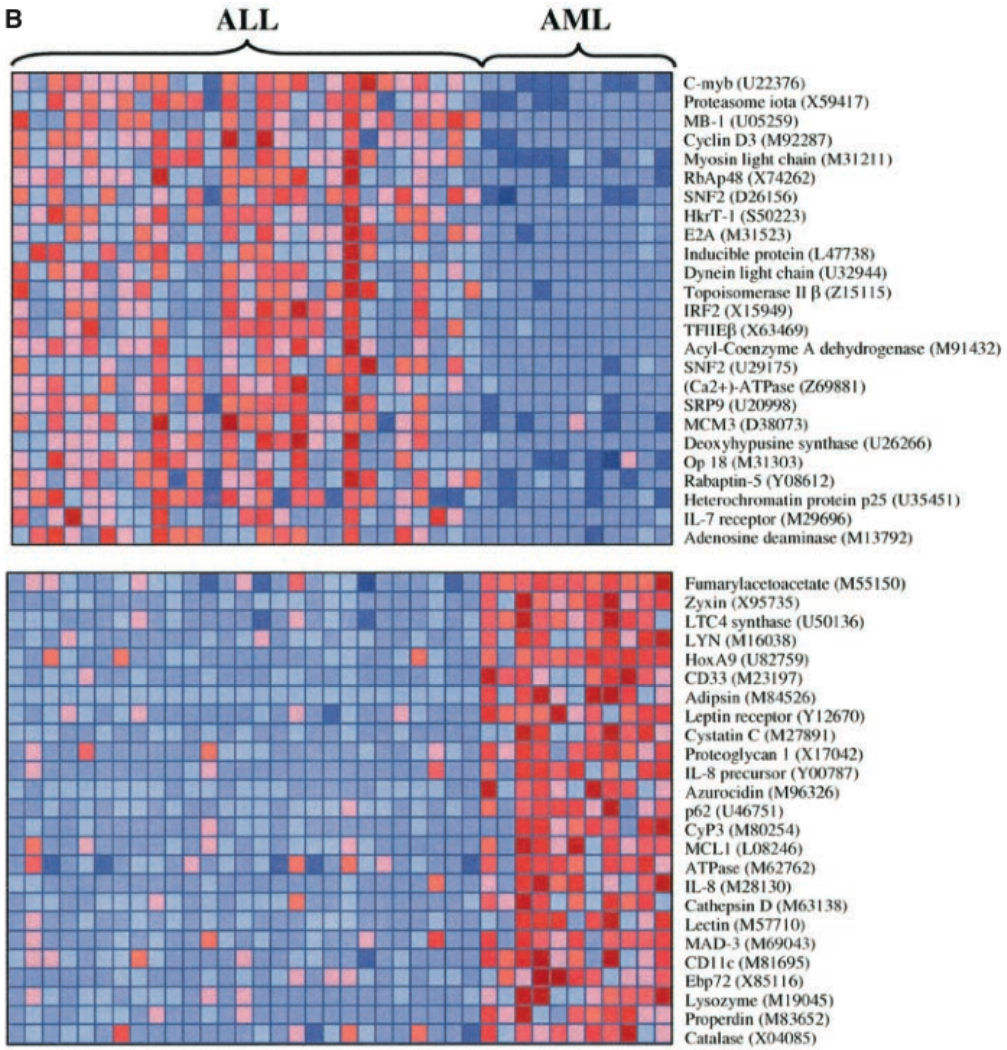




Data

- 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



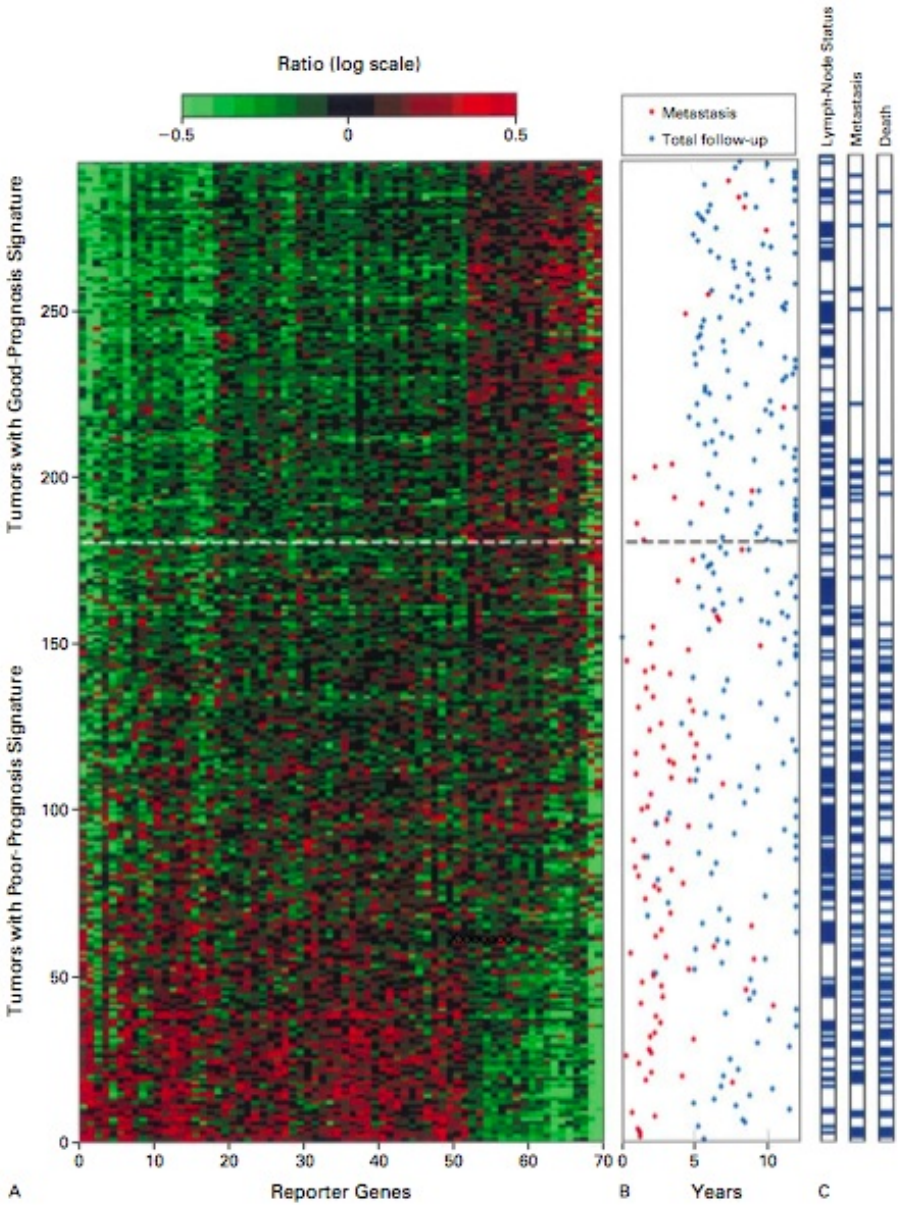
Goal

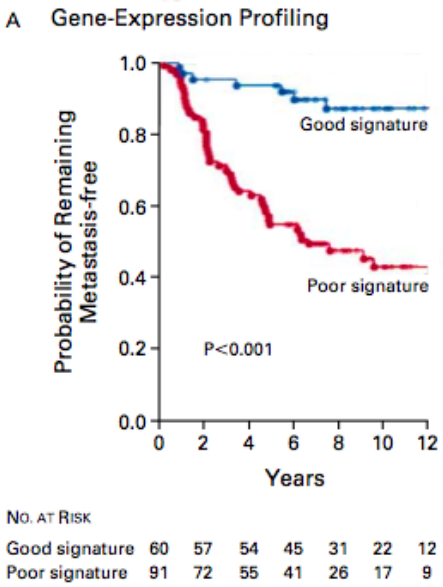
- Design a classifier to automatically assign a class to future samples from their expression profile
- Interpret biologically the differences between the classes

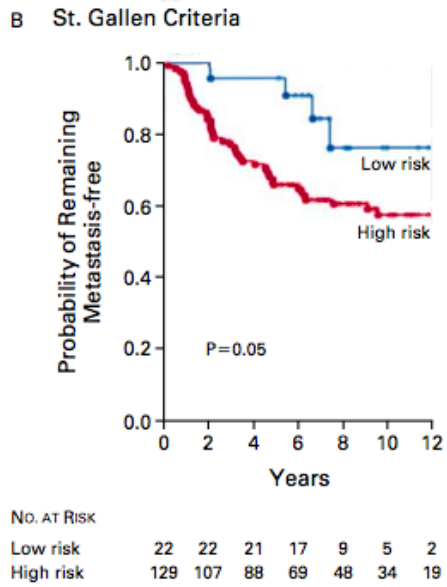
Difficulty

- Large dimension
- Few samples

Prognosis from microarray data (MAMMAPRINT)







Gene signature

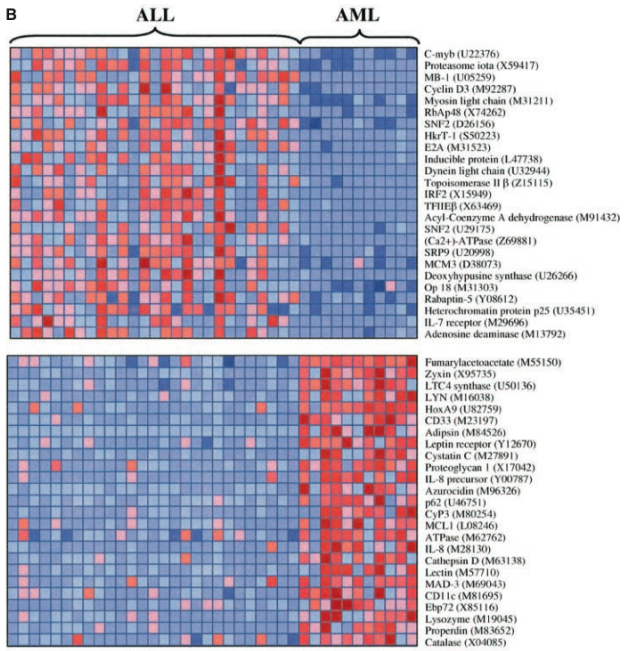
The idea

- 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).

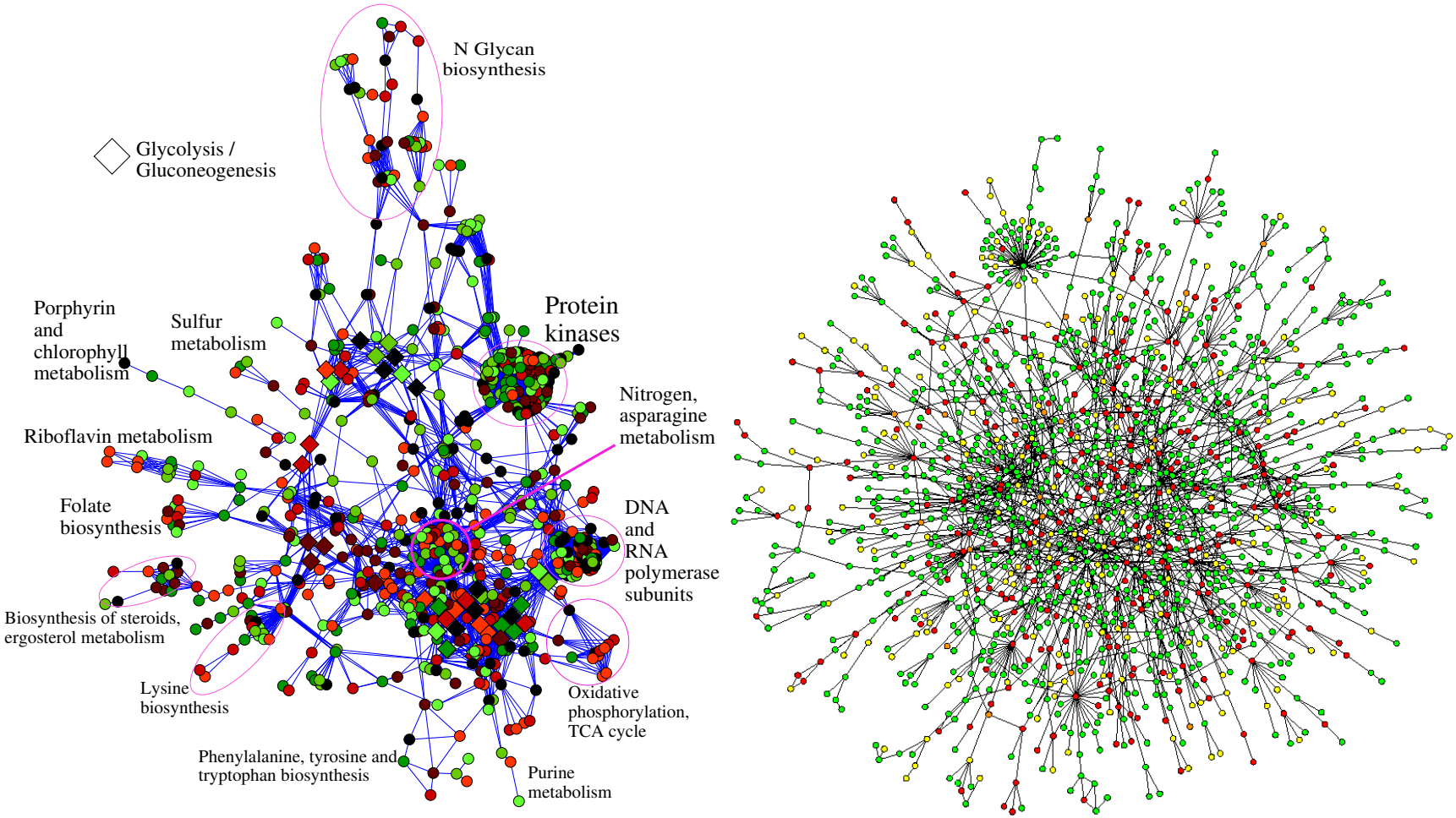
But...



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?

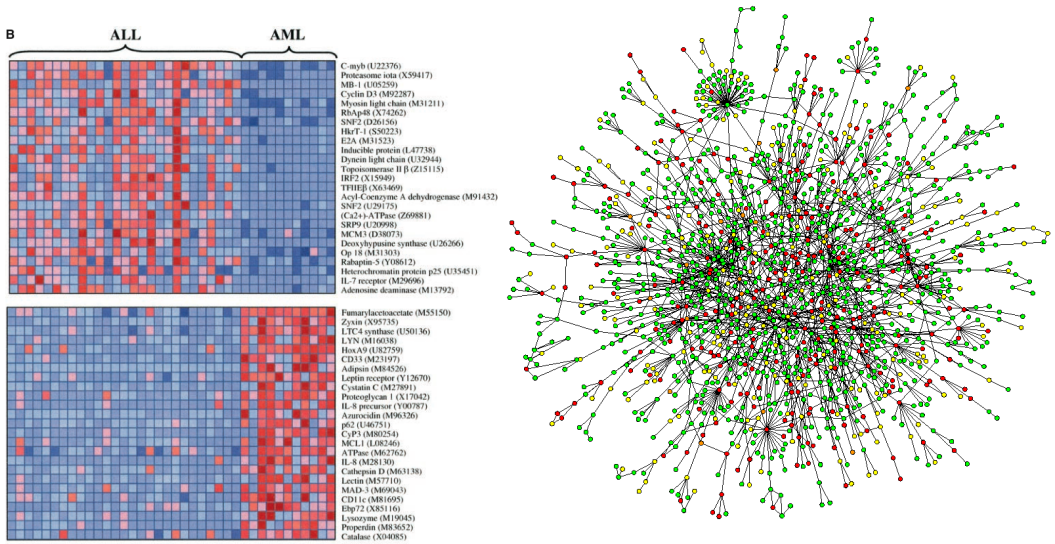
Gene networks



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

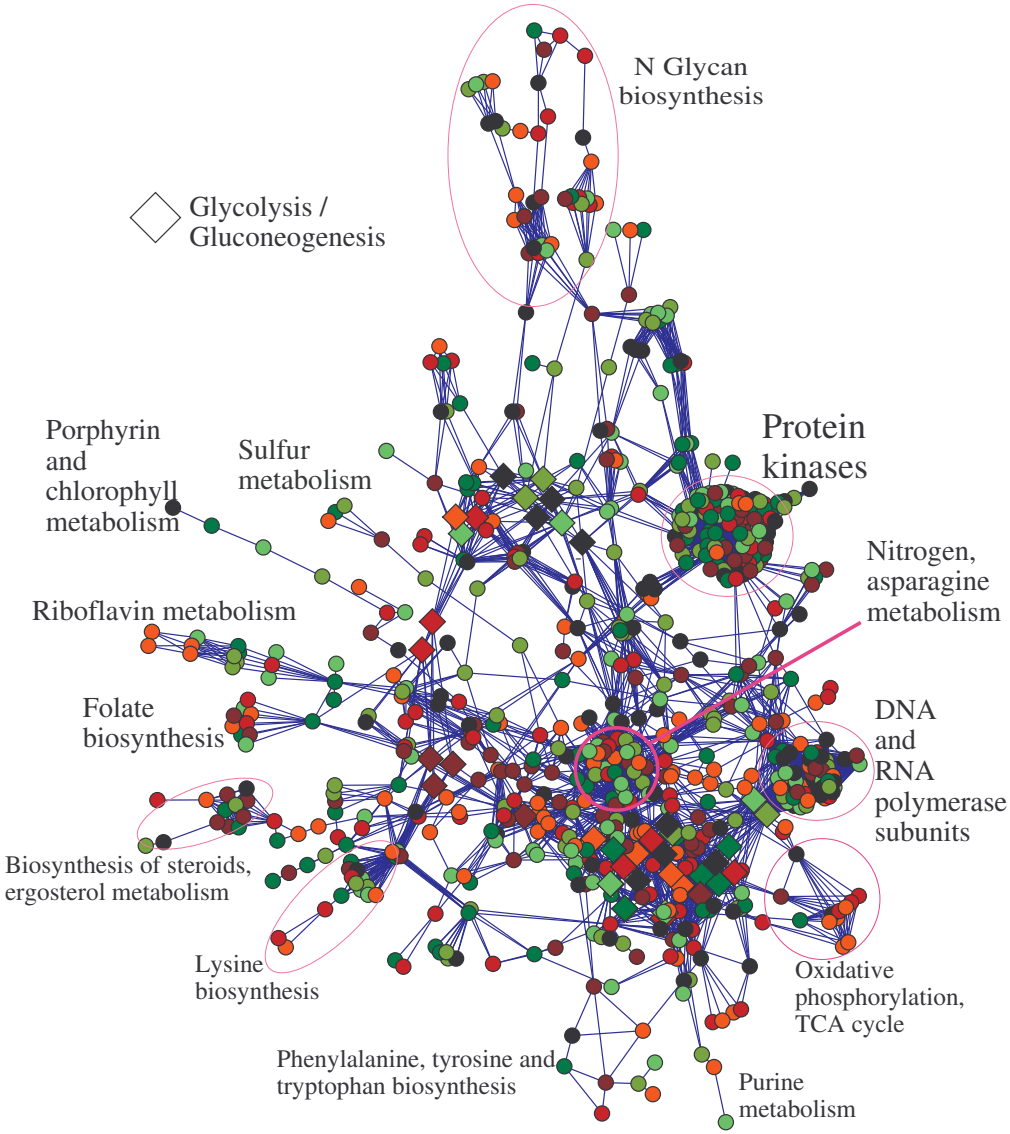
Genes near each other on the graph should have similar weigths.

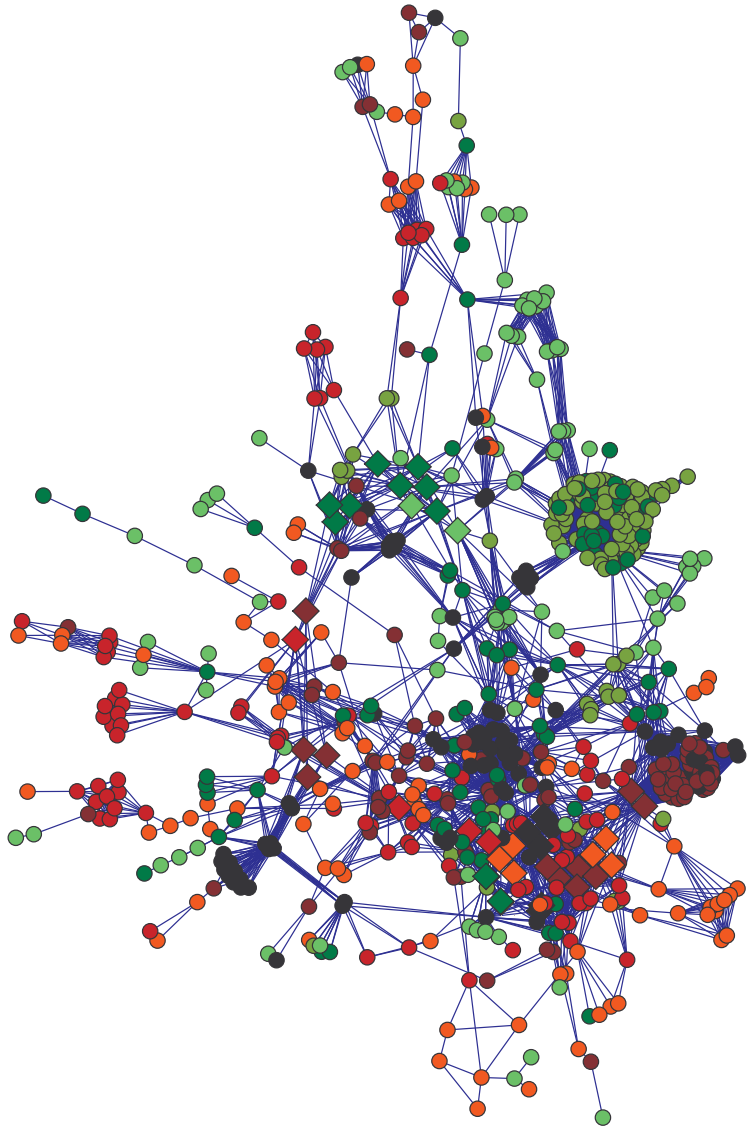
Two solutions (Rapaport et al., 2007, 2008)

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

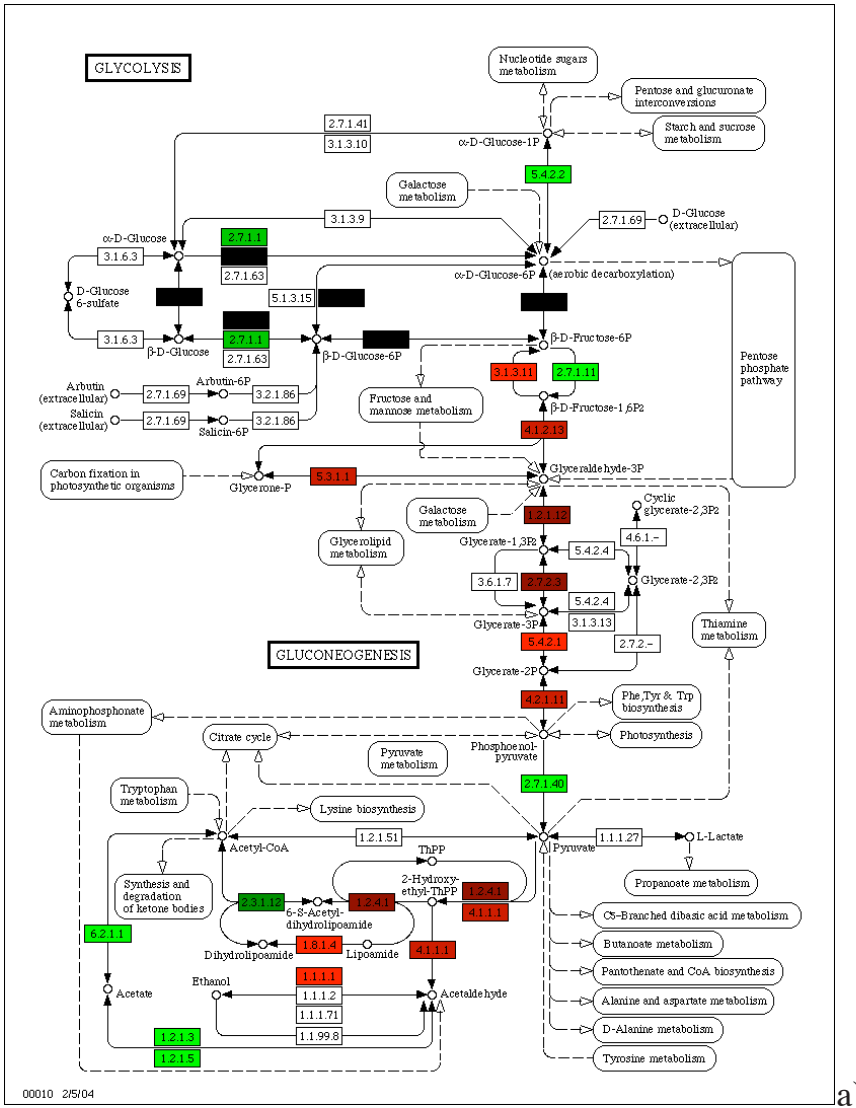
$$\Omega_{graphfusion}(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_i |\beta_i| \,.$$

Classifiers

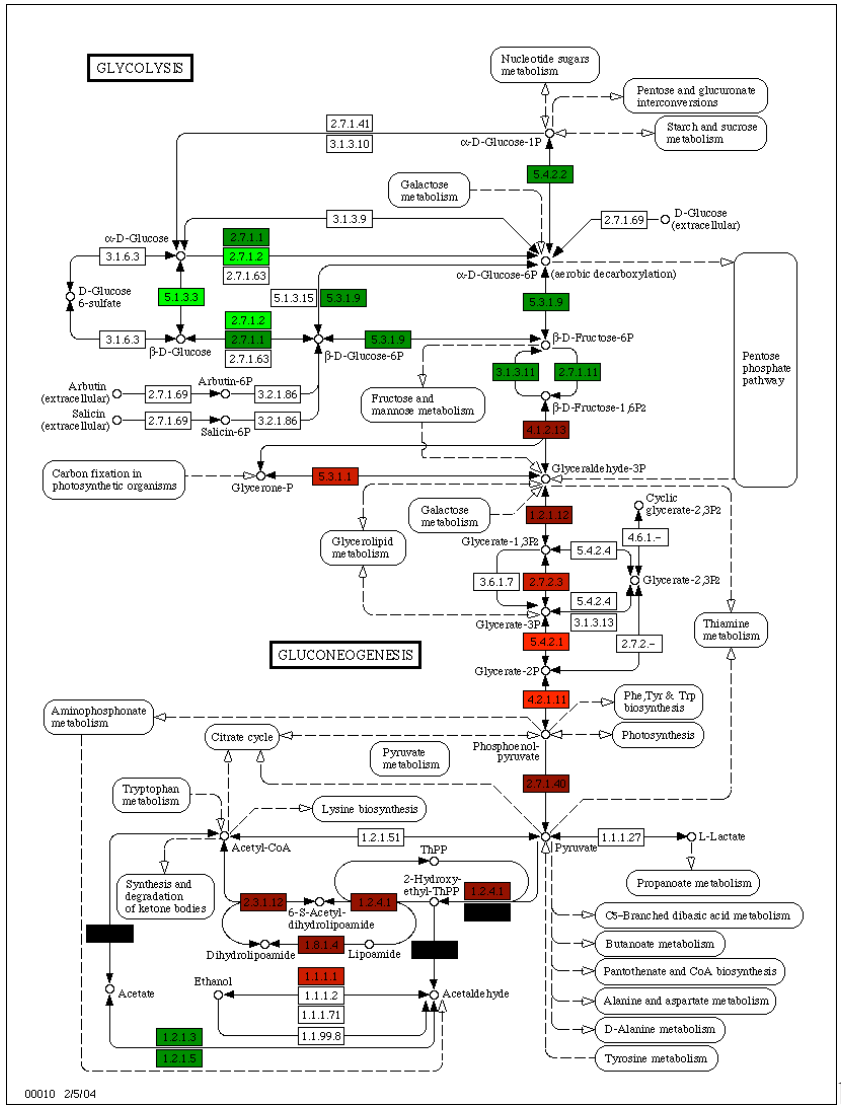




Classifier

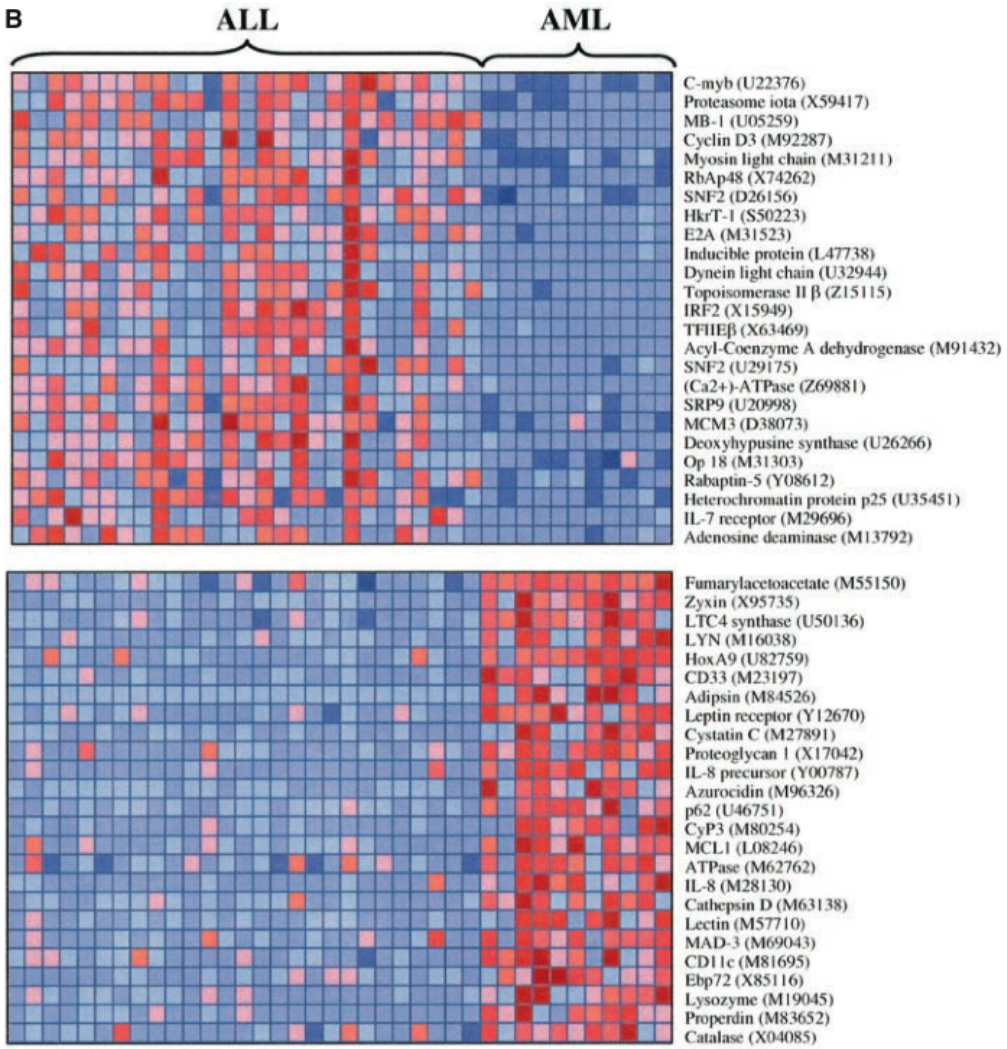


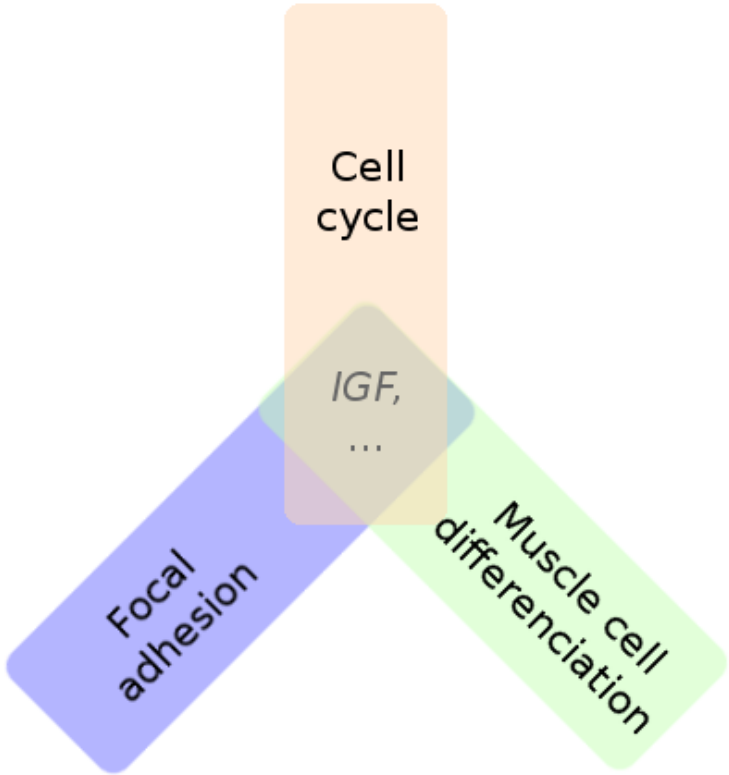
a)



b)

How to select jointly genes belonging to predefined pathways?



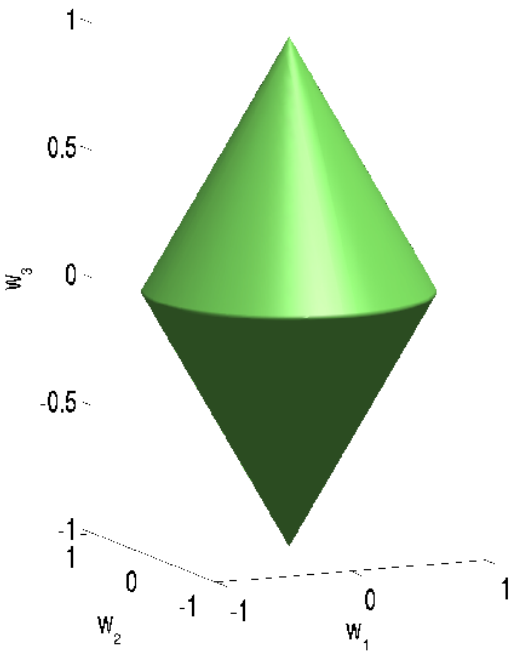


Selecting pre-defined groups of variables

Group lasso (Yuan & Lin, 2006)

If groups of covariates are likely to be selected together, the ℓ_1/ℓ_2-norm induces sparse solutions *at the group level*:

$$\Omega_{group}(w) = \sum_g \|w_g\|_2$$

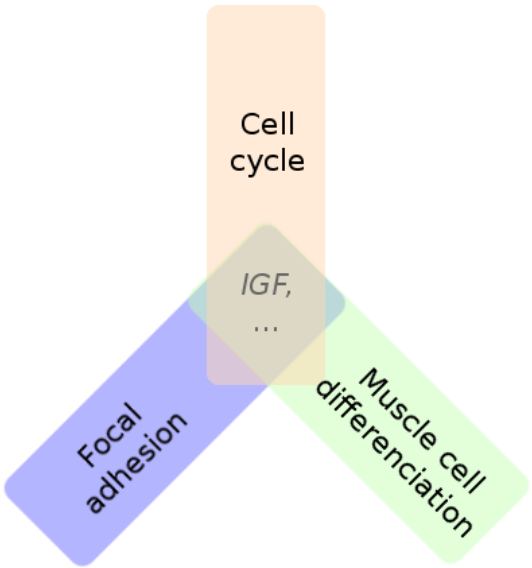


$$\Omega(w_1, w_2, w_3) = \|(w_1, w_2)\|_2 + \|w_3\|_2$$

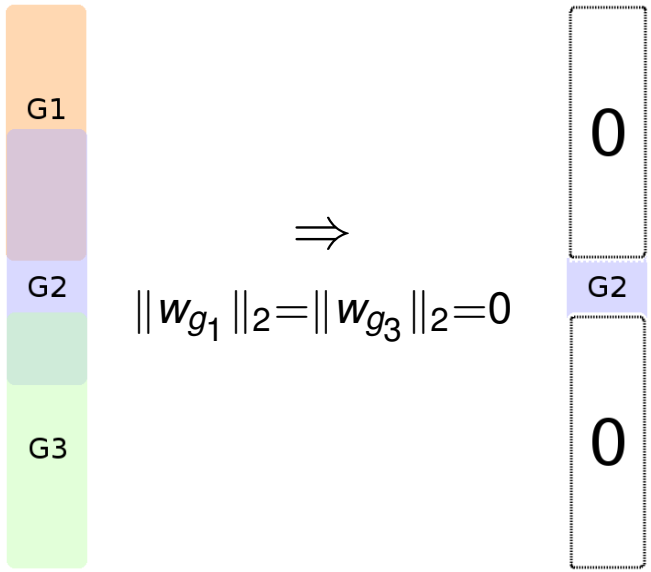
What if a gene belongs to several groups?

Issue of using the group-lasso

- $\Omega_{group}(w) = \sum_g \|w_g\|_2$ sets groups to 0.
- One variable is selected $\Leftrightarrow$ all the groups to which it belongs are selected.



IGF selection $\Rightarrow$ selection of unwanted groups



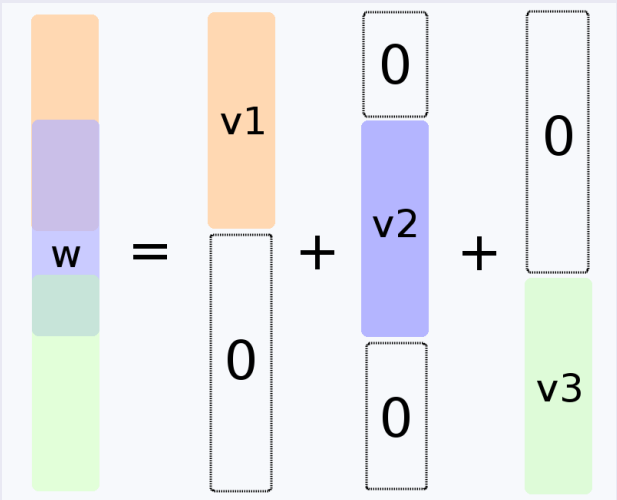
Removal of *any* group containing a gene $\Rightarrow$ the weight of the gene is 0.

Overlap norm (Jacob et al., 2009)

An idea

Introduce latent variables v_g:

$$\begin{cases} \min_{w,v} L(w) + \lambda \sum_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \text{supp}\left(v_g\right) \subseteq g. \end{cases}$$



Properties

- Resulting support is a *union* of groups in $\mathcal{G}$.
- Possible to select one variable without selecting all the groups containing it.
- Equivalent to group lasso when there is no overlap

A new norm

Overlap norm

$$\begin{cases} \min_{w,v} L(w) + \lambda \sum_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \text{supp}\left(v_g\right) \subseteq g. \end{cases} = \min_{w} L(w) + \lambda \Omega_{overlap}(w)$$

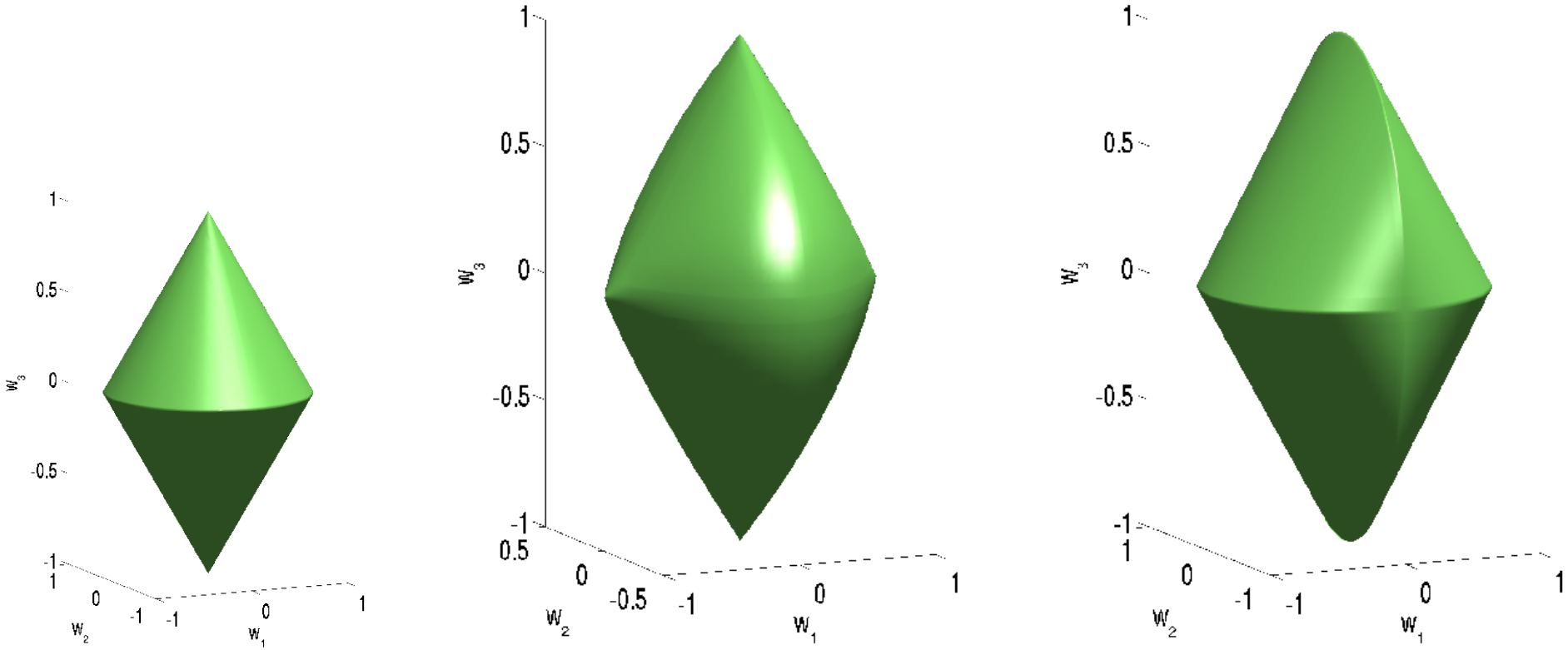
with

$$\Omega_{overlap}(w) \triangleq \begin{cases} \min_{v} \sum_{g \in \mathcal{G}} \|v_g\|_2 \\ w = \sum_{g \in \mathcal{G}} v_g \\ \text{supp}\left(v_g\right) \subseteq g. \end{cases} \quad (*)$$

Property

- $\Omega_{overlap}(w)$ is a norm of w.
- $\Omega_{overlap}(.)$ associates to w a specific (not necessarily unique) decomposition $(v_g)_{g \in \mathcal{G}}$ which is the argmin of $(*)$.

Overlap and group unity balls



Balls for $\Omega^{\mathcal{G}}_{\text{group}}(\cdot)$ (middle) and $\Omega^{\mathcal{G}}_{\text{overlap}}(\cdot)$ (right) for the groups $\mathcal{G} = \{\{1,2\},\{2,3\}\}$ where w_2 is represented as the vertical coordinate. Left: group-lasso ($\mathcal{G} = \{\{1,2\},\{3\}\}$), for comparison.

Theoretical results

Consistency in group support (Jacob et al., 2009)

- Let $\bar{w}$ be the true parameter vector.
- Assume that there exists a unique decomposition $\bar{v}_g$ such that $\bar{w} = \sum_g \bar{v}_g$ and $\Omega^{\mathcal{G}}_{\text{overlap}}(\bar{w}) = \sum \|\bar{v}_g\|_2$.
- Consider the regularized empirical risk minimization problem $L(w) + \lambda \Omega^{\mathcal{G}}_{\text{overlap}}(w)$.

Then

- under appropriate mutual incoherence conditions on X,
- as $n \to \infty$,
- with very high probability,

the optimal solution $\hat{w}$ admits a unique decomposition $(\hat{v}_g)_{g \in \mathcal{G}}$ such that

$$\left\{ g \in \mathcal{G} | \hat{v}_g \neq 0 \right\} = \left\{ g \in \mathcal{G} | \bar{v}_g \neq 0 \right\}.$$

Theoretical results

Consistency in group support (Jacob et al., 2009)

- Let $\bar{w}$ be the true parameter vector.
- Assume that there exists a unique decomposition $\bar{v}_g$ such that $\bar{w} = \sum_g \bar{v}_g$ and $\Omega^{\mathcal{G}}_{\text{overlap}}(\bar{w}) = \sum \|\bar{v}_g\|_2$.
- Consider the regularized empirical risk minimization problem $L(w) + \lambda \Omega^{\mathcal{G}}_{\text{overlap}}(w)$.

Then

- under appropriate mutual incoherence conditions on X,
- as $n \to \infty$,
- with very high probability,

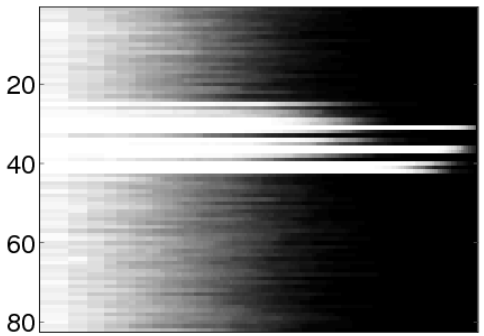
the optimal solution $\hat{w}$ admits a unique decomposition $(\hat{v}_g)_{g \in \mathcal{G}}$ such that

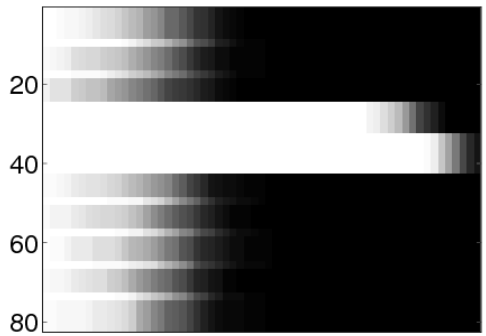
$$\left\{ g \in \mathcal{G} | \hat{v}_g \neq 0 \right\} = \left\{ g \in \mathcal{G} | \bar{v}_g \neq 0 \right\}.$$

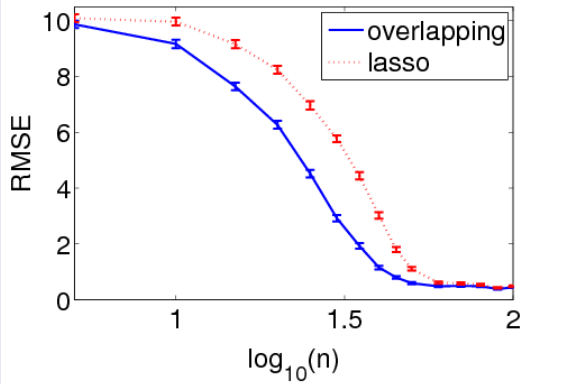
Experiments

Synthetic data: overlapping groups

- 10 groups of 10 variables with 2 variables of overlap between two successive groups :$\{1, \ldots, 10\}, \{9, \ldots, 18\}, \ldots, \{73, \ldots, 82\}$.
- Support: union of 4*th* and 5*th* groups.
- Learn from 100 training points.

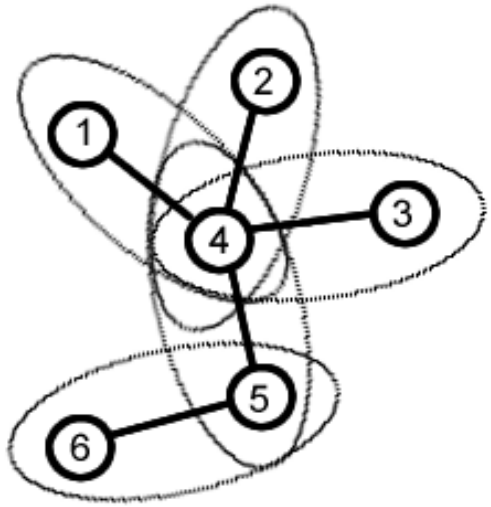






Frequency of selection of each variable with the lasso (left) and $\Omega^{\mathcal{G}}_{\text{overlap}}(.)$ (middle), comparison of the RMSE of both methods (right).

Graph lasso



Two solutions

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

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

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	$\Omega^{\mathcal{G}}_{\text{OVERLAP}}(.)$
ERROR	0.38 ± 0.04	0.36 ± 0.03
MEAN $\sharp$ PATH.	130	30

- Graph on the genes.

METHOD	ℓ_1	$\Omega_{graph}(.)$
ERROR	0.39 ± 0.04	0.36 ± 0.01
AV. SIZE C.C.	1.03	1.30

Outline

Conclusion

- Modern machine learning methods for regression / classification lend themselves well to the integration of prior knowledge in the penalization / regularization function.
- Several computationally efficient approaches (structured LASSO, kernels...)
- Tight collaborations with domain experts can help develop specific learning machines for specific data
- Natural extensions for data integration

People I need to thank

Franck Rapaport (now MSKCC), Emmanuel Barillot, Andrei Zynoviev (Institut Curie), Laurent Jacob (UC Berkeley) Guillaume Obozinski (INRIA)