Machine learning with prior knowledge for genomic data

Jean-Philippe Vert

Jean-Philippe.Vert@mines.org

Mines ParisTech / Curie Institute / Inserm

UCLA, Feb 6, 2012.

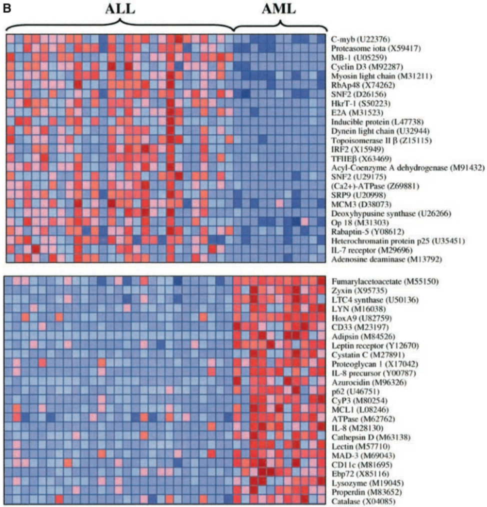
Outline

1. Introduction
2. Inference of gene regulatory networks
3. Cancer prognosis from DNA copy number variations
4. Diagnosis and prognosis from gene expression data
5. Conclusion

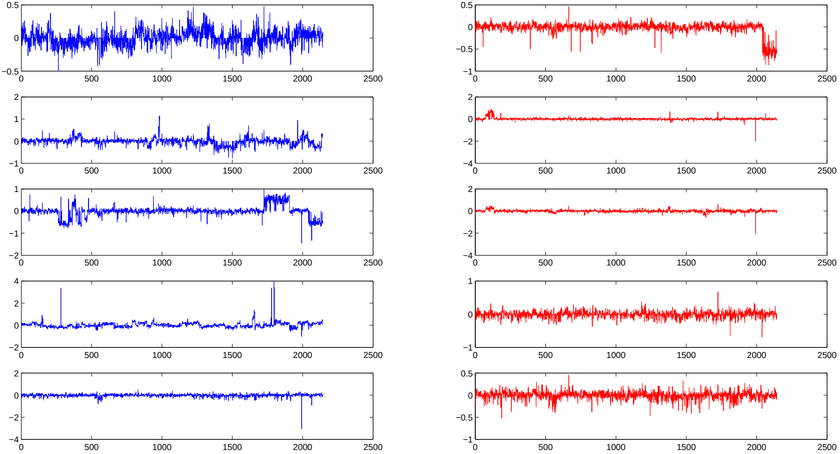
Outline

1. Introduction
2. Inference of gene regulatory networks
3. Cancer prognosis from DNA copy number variations
4. Diagnosis and prognosis from gene expression data
5. Conclusion

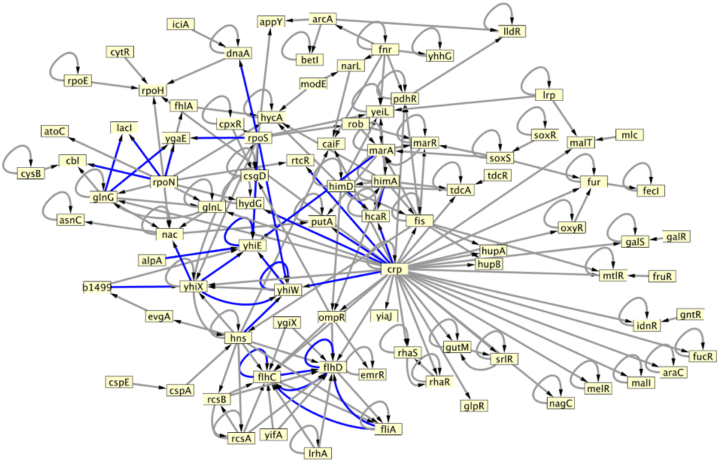
Cancer diagnosis



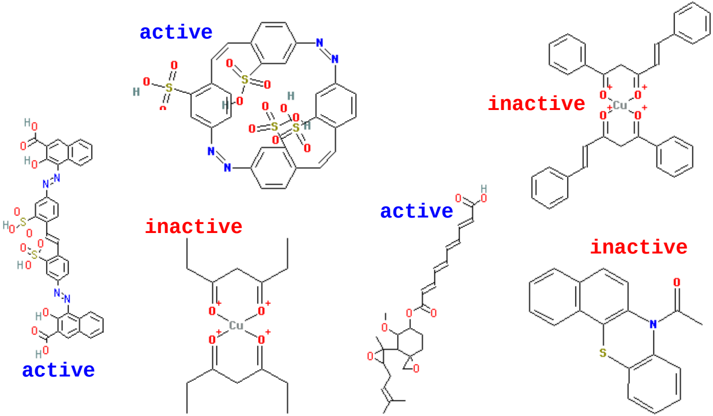
Cancer prognosis



Gene network inference

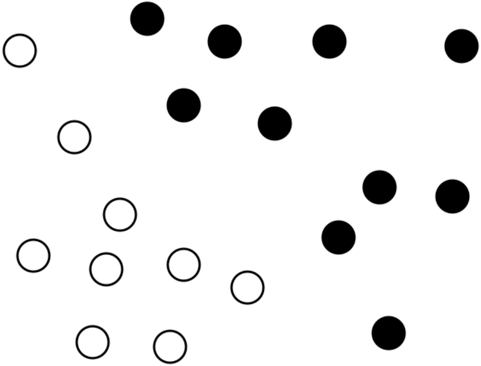


Virtual screening for drug discovery



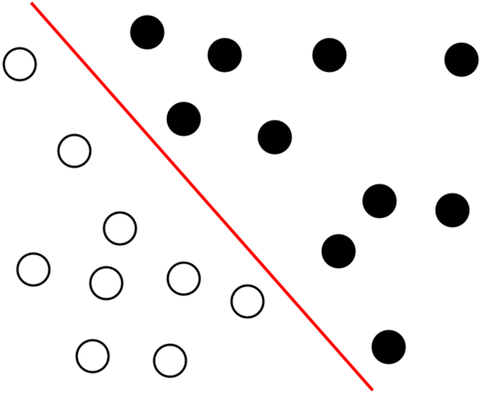
NCI AIDS screen results (from http://cactus.nci.nih.gov).

Machine learning



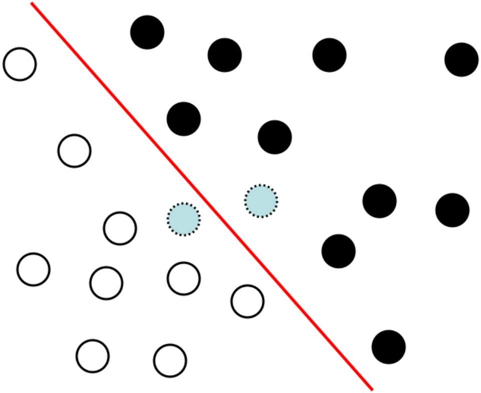
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning



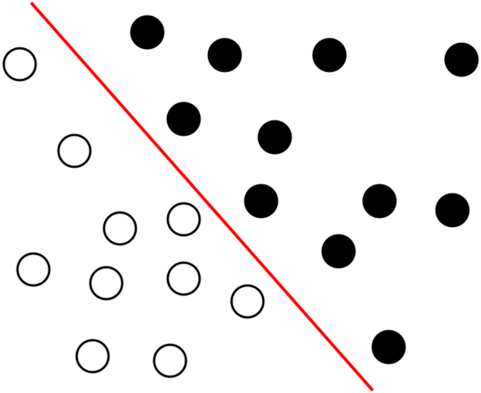
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning



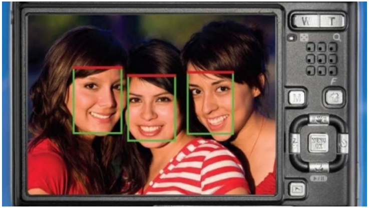
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning



1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning : tools and applications



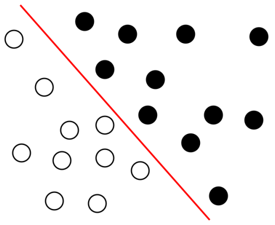
Many applications

Multimedia, image, video, speech recognition, web, social network, online advertising, finance, biology, chemistry

Many tools

Linear discriminant analysis, logistic regression, decision trees, neural networks, support vector machines...

Machine learning in bioinformatics



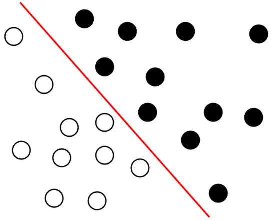
Challenges

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

A strategy: penalized empirical risk minimization

$$\min_f R[f] + \lambda \Omega[f]$$

Machine learning in bioinformatics



Challenges

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

A strategy: penalized empirical risk minimization

$$\min_f R[f] + \lambda\Omega[f]$$

Outline

1. Introduction
2. Inference of gene regulatory networks
3. Cancer prognosis from DNA copy number variations
4. Diagnosis and prognosis from gene expression data
5. Conclusion

Gene expression

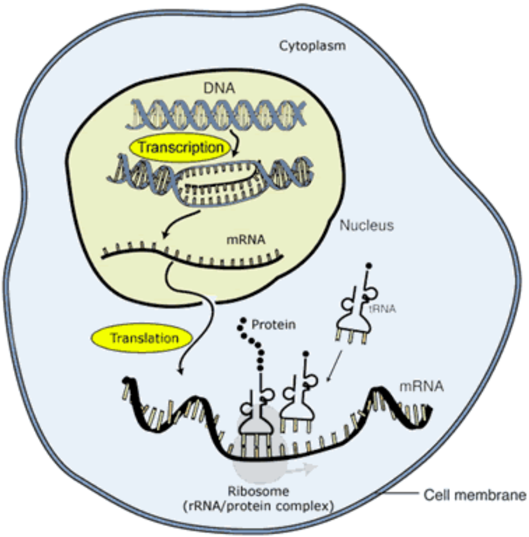
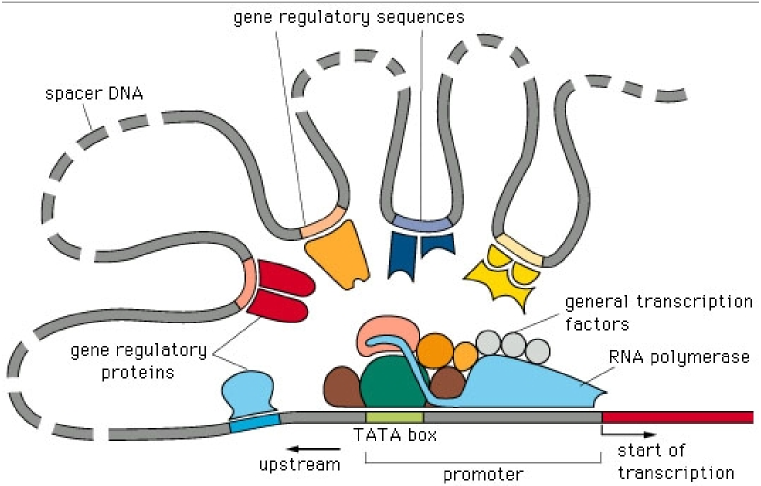
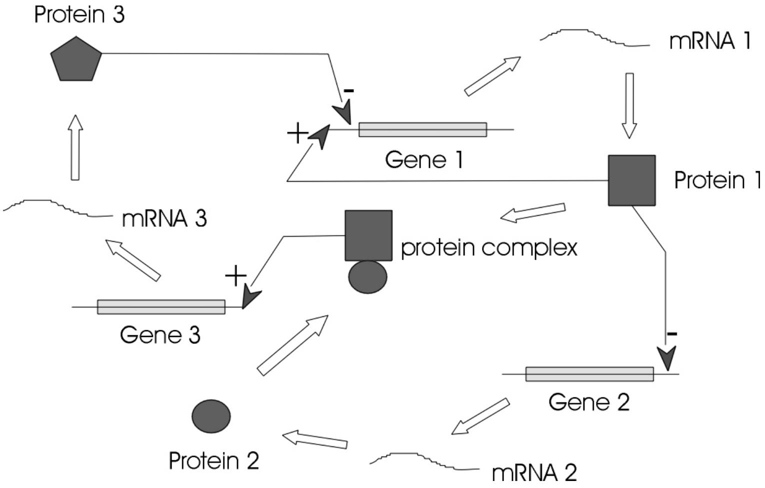


Image adapted from: National Human Genome Research Institute.

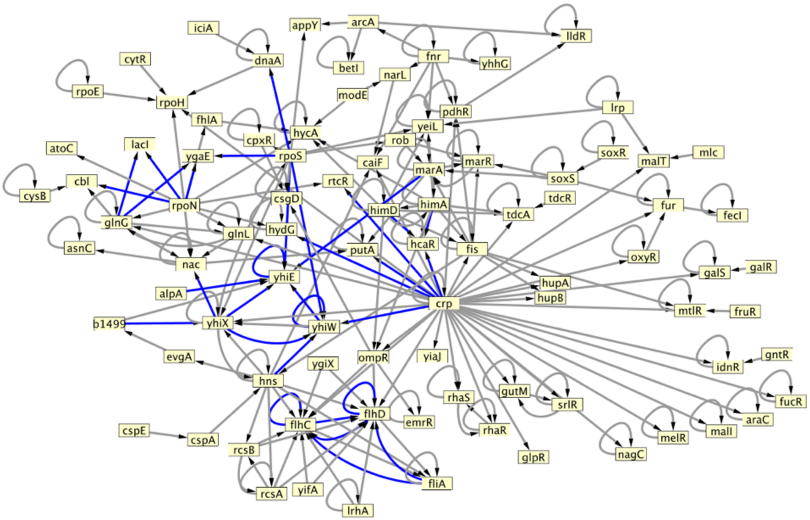
Gene expression regulation



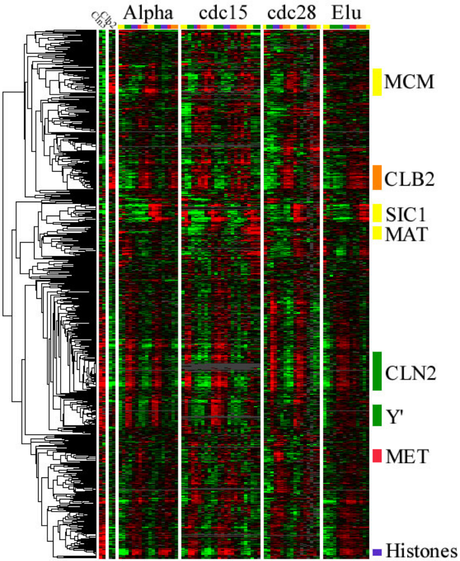
Gene regulatory network



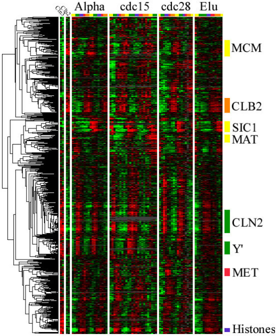
Gene regulatory network of E. coli

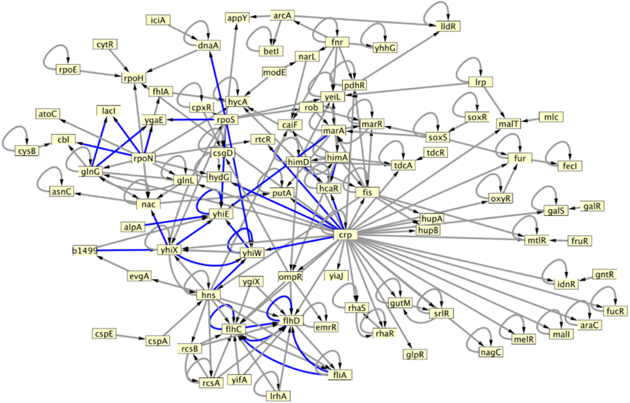


Gene expression data



Reconstruction of gene regulatory network from expression data

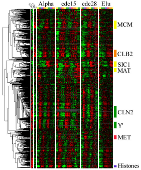


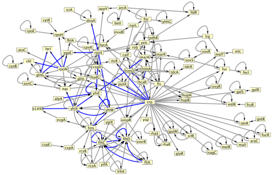


De novo inference

The problem

Given a set of gene expressions, infer the regulations.





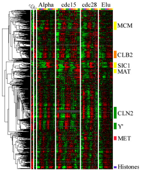
How?

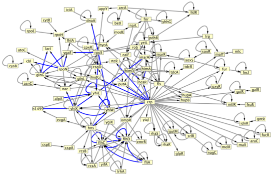
- Connect "similar genes": correlation, mutual-information...
- Model-based approaches: dynamic systems, boolean networks, state-space models, Bayesian networks
- Sparse regression: regulators as the smallest set of TF necessary to predict the expression of the target (GENIE, TIGRESS...)

De novo inference

The problem

Given a set of gene expressions, infer the regulations.





How?

- Connect "similar genes": correlation, mutual-information...
- Model-based approaches: dynamic systems, boolean networks, state-space models, Bayesian networks
- Sparse regression: regulators as the smallest set of TF necessary to predict the expression of the target (GENIE, TIGRESS...)

Predicting regulation by sparse regression

- Let $Y \in \mathbb{R}^n$ the expression of a gene, and $X_1, \ldots, X_p \in \mathbb{R}^n$ the expression of all TFs. We look for a model

  $$Y = \sum_{i=1}^{p} \beta_i X_i + \text{noise}$$

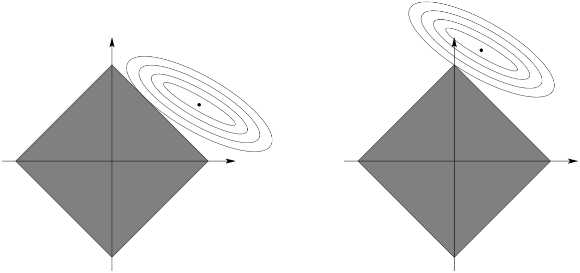
  where β is sparse, i.e., only a few β_i are non-zero.
- We can estimate the sparse regression model from a matrix of expression data.
- Non-zero β_i's correspond to predicted regulators.

Feature selection with the lasso

$$\min_{\beta \in \mathbb{R}^p} \| Y - X\beta \|^2 \quad \text{such that} \quad \sum_{i=1}^{p} |\beta_i| \leq t$$

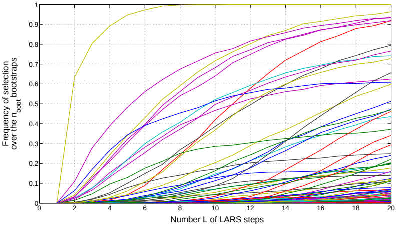
- No explicit solution, but this is just a quadratic program (Tibshirani, 1996; Chen et al., 1998).
- Efficient solution with the LARS (Efron et al., 2004)
- When t is not too large, the solution will usually be sparse

Geometric interpretation with $p = 2$

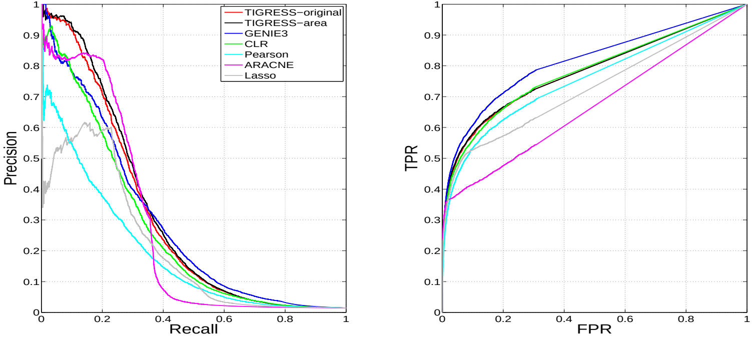


TIGRESS (Haury et al., 2012)

- For $t = 1$ to T do
  - Bootstrap a random sample S_t from the training set
  - Randomly reweight each feature
  - Select L features with the Lasso
- The score of a feature is the number of times it was selected among the T repeats (Bach, 2008; Meinshausen and Bühlmann, 2010).)
- Rank features (TF-TG interactions) by decreasing area under the score curve



Validation

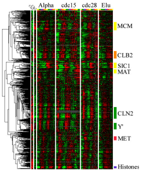


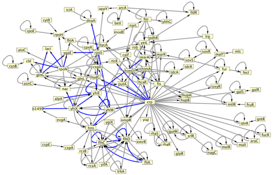
DREAM5: GENIE and TIGRESS ranked 1st and 2nd out or 29 on the *in silico* challenge

Supervised inference

The problem

Given a set of gene expressions AND a set of known regulations, infer missing regulations.





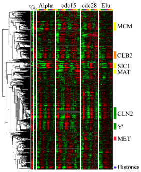
How?

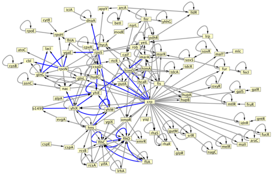
- Local models: for each TF, learn to discriminate the regulated vs non-regulated genes
- Global models: learn to discriminate connected vs non-connected TF-target pairs

Supervised inference

The problem

Given a set of gene expressions AND a set of known regulations, infer missing regulations.



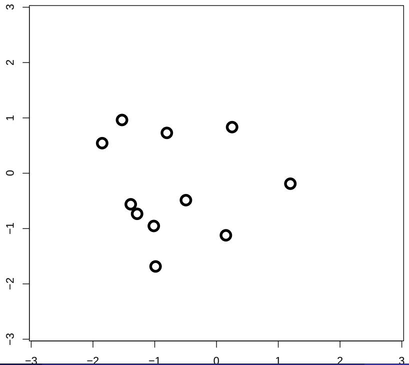


How?

- Local models: for each TF, learn to discriminate the regulated vs non-regulated genes
- Global models: learn to discriminate connected vs non-connected TF-target pairs

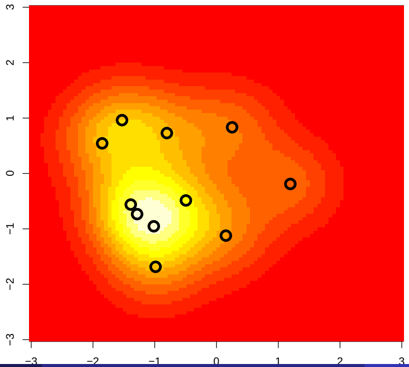
Example: one-class learning approach for local model

- For a given TF, let $P \subset [1, n]$ be the set of genes known to be regulated by it
- From the expression profiles $(X_i)_{i \in P}$, estimate a score $s(X)$ to assess which expression profiles X are similar
- Then classify the genes not in P by decreasing score



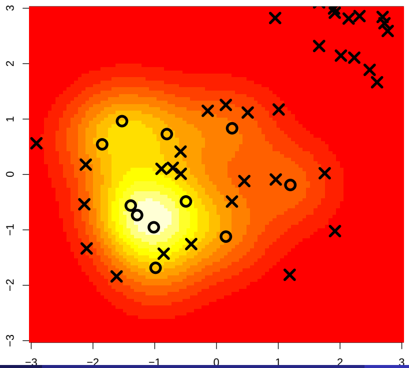
Example: one-class learning approach for local model

- For a given TF, let $P \subset [1, n]$ be the set of genes known to be regulated by it
- From the expression profiles $(X_i)_{i \in P}$, estimate a score $s(X)$ to assess which expression profiles X are similar
- Then classify the genes not in P by decreasing score

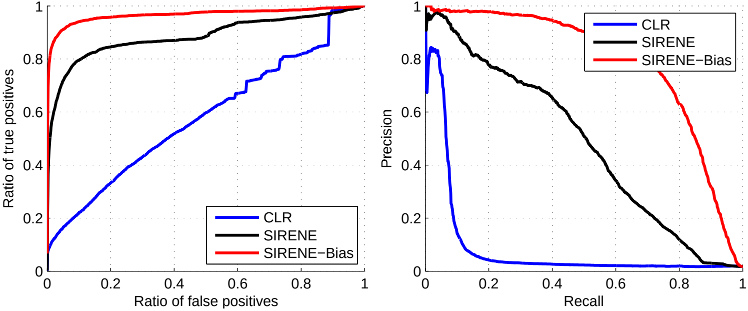


Example: one-class learning approach for local model

- For a given TF, let $P \subset [1, n]$ be the set of genes known to be regulated by it
- From the expression profiles $(X_i)_{i \in P}$, estimate a score $s(X)$ to assess which expression profiles X are similar
- Then classify the genes not in P by decreasing score



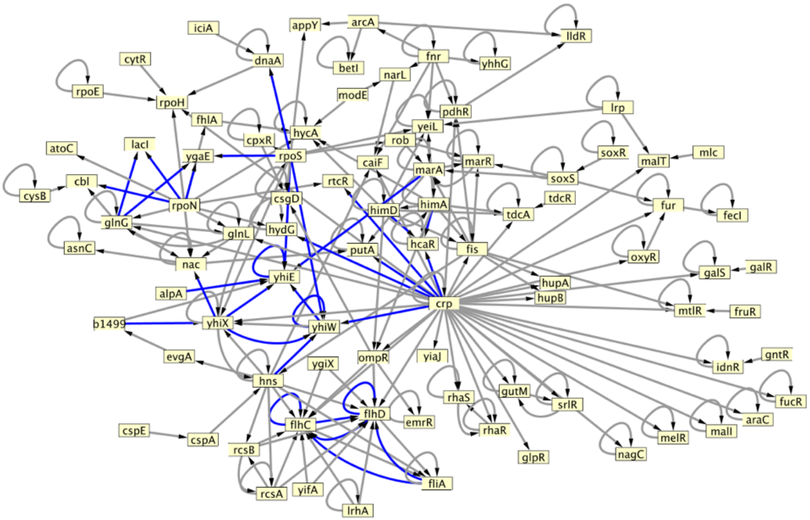
Validation (Mordelet and V., 2008)



Method	Recall at 60%	Recall at 80%
SIRENE	**44.5%**	**17.6%**
CLR	7.5%	5.5%
Relevance networks	4.7%	3.3%
ARACNe	1%	0%
Bayesian network	1%	0%

SIRENE = Supervised Inference of REgulatory NEtworks (Mordelet and V., 2008)

Application: predicted regulatory network (E. coli)



Outline

1 Introduction

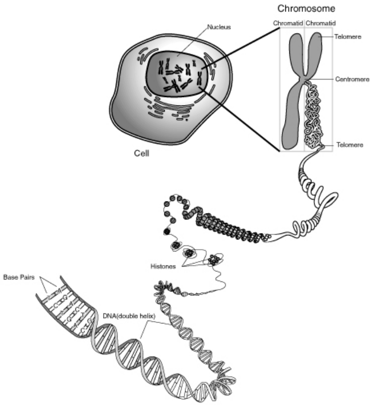
2 Inference of gene regulatory networks

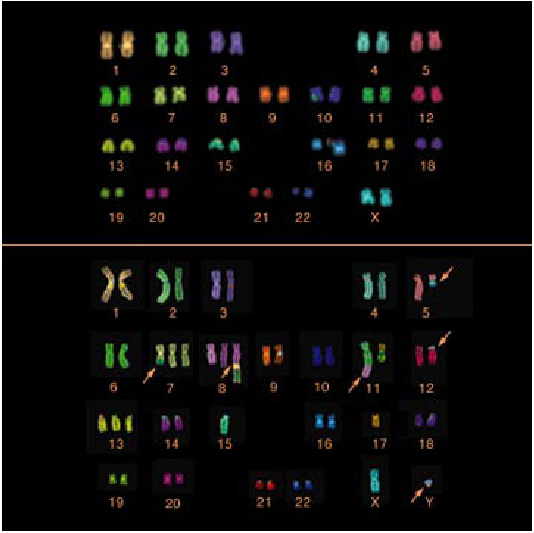
3 Cancer prognosis from DNA copy number variations

4 Diagnosis and prognosis from gene expression data

5 Conclusion

Chromosomic aberrations in cancer

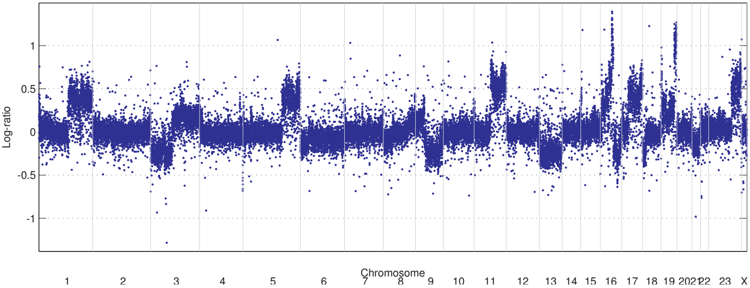




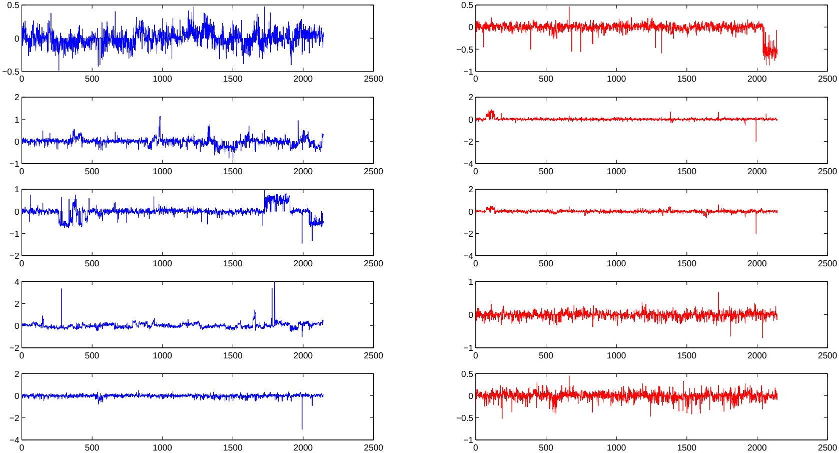
Comparative Genomic Hybridization (CGH)

Motivation

- Comparative genomic hybridization (CGH) data measure the DNA copy number along the genome
- Very useful, in particular in cancer research to observe systematically variants in DNA content



Cancer prognosis: can we predict the future evolution?



Aggressive (left) vs non-aggressive (right) melanoma

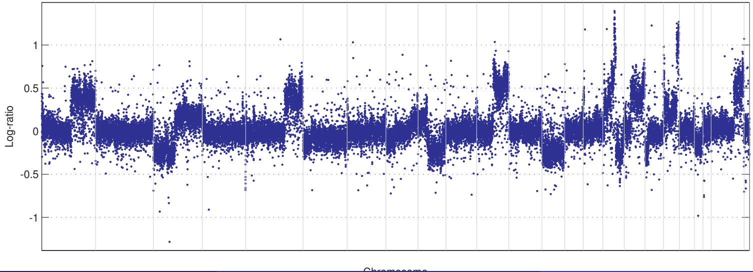
CGH array classification

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



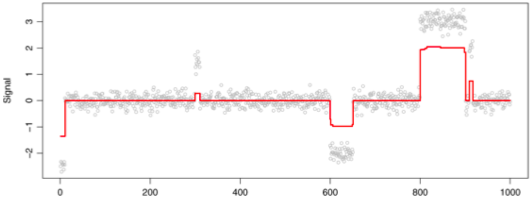
Promoting piecewise constant profiles with

- Total variation (Rudin et al., 1992; Land and Friedman, 1996):

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

- Fused lasso (Tibshirani et al., 2005; Tibshirani and Wang, 2008)

$$\min_{\beta \in \mathbb{R}^p} \| Y - \beta \|^2 + \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_{TV}$$



Fused lasso for supervised classification

- Idea: find the vector of weights β that best discriminates the aggressive vs non-aggressive, subject to the constraints that it should be sparse and piecewise constant
- Mathematically:

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|$$

- Computationnally: this is convex optimization problem that can be solved very efficiently (V. and Bleakley, 2012)

Fused lasso for supervised classification

- Idea: find the vector of weights β that best discriminates the aggressive vs non-aggressive, subject to the constraints that it should be sparse and piecewise constant
- Mathematically:

$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|$$

- Computationnally: this is convex optimization problem that can be solved very efficiently (V. and Bleakley, 2012)

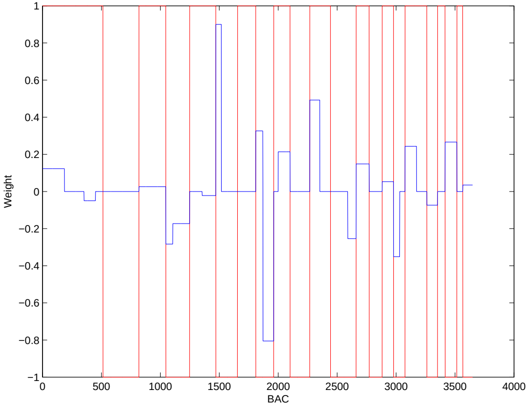
Fused lasso for supervised classification

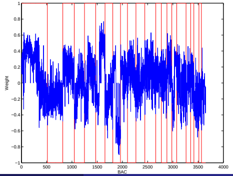
- Idea: find the vector of weights β that best discriminates the aggressive vs non-aggressive, subject to the constraints that it should be sparse and piecewise constant
- Mathematically:

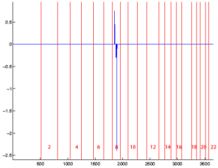
$$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i|$$

- Computationnally: this is convex optimization problem that can be solved very efficiently (V. and Bleakley, 2012)

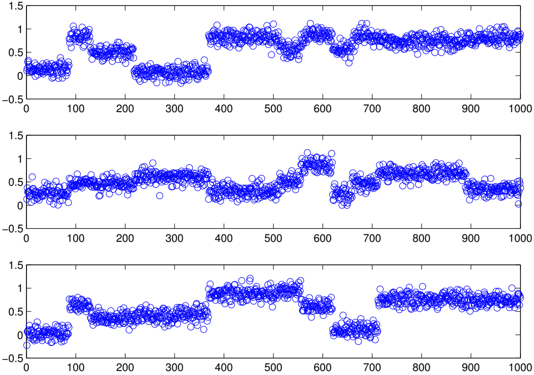
Prognostic in melanoma (Rapaport et al., 2008)



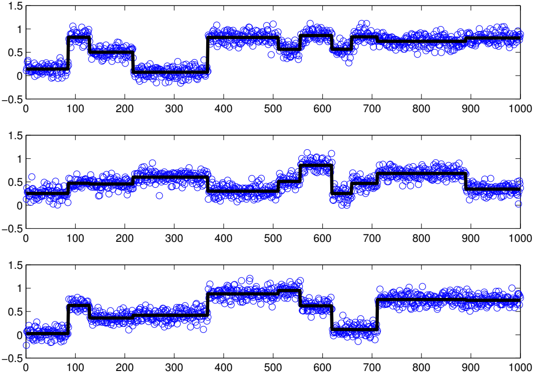




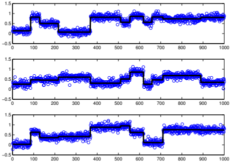
Extension: finding multiple change points shared by several profiles



Extension: finding multiple change points shared by several profiles



"Optimal" segmentation by dynamic programming



- Define the "optimal" piecewise constant approximation $\hat{U} \in \mathbb{R}^{p \times n}$ of Y as the solution of

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}\left(U_{i+1,\bullet} \neq U_{i,\bullet}\right) \leq k$$

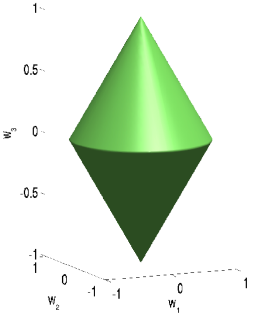
- DP finds the solution in $O(p^2kn)$ in time and $O(p^2)$ in memory
- But: does not scale to $p = 10^6 \sim 10^9$...

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*:

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



$$\Omega(w_1, w_2, w_3) = \|(w_1, w_2)\|_2 + \|w_3\|_2$$
$$= \sqrt{w_1^2 + w_2^2} + \sqrt{w_3^2}$$

GFLseg (Bleakley and V., 20011)

- Replace

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} \mathbf{1}\left(U_{i+1,\bullet} \neq U_{i,\bullet}\right) \leq k$$

by

$$\min_{U \in \mathbb{R}^{p \times n}} \| Y - U \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} w_i \| U_{i+1,\bullet} - U_{i,\bullet} \| \leq \mu$$

- We can solve it efficiently in $O(np)$
- It converges to the true segmentation when the number of profiles increases

Speed trial

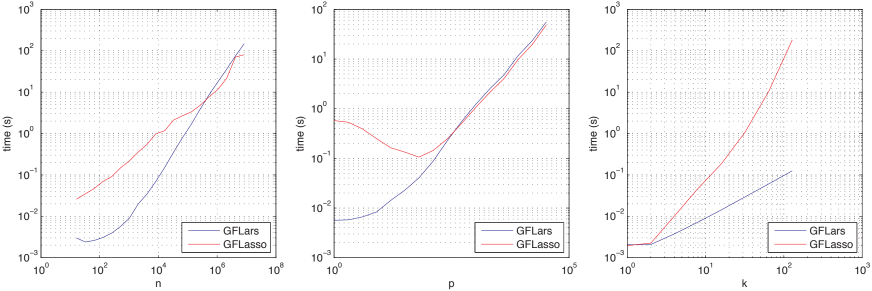


Figure 2: **Speed trials for group fused LARS (top row) and Lasso (bottom row).** *Left column:* varying n, with fixed $p = 10$ and $k = 10$; *center column:* varying p, with fixed $n = 1000$ and $k = 10$; *right column:* varying k, with fixed $n = 1000$ and $p = 10$. Figure axes are log-log. Results are averaged over 100 trials.

Performance

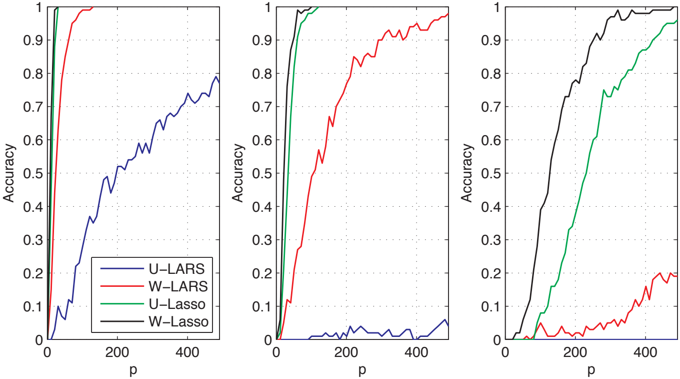


Figure 4: **Multiple change-point accuracy.** Accuracy as a function of the number of profiles p when change-points are placed at the nine positions $\{10, 20, \dots, 90\}$ and the variance σ^2 of the centered Gaussian noise is either 0.05 (left), 0.2 (center) and 1 (right). The profile length is 100.

Outline

1 Introduction

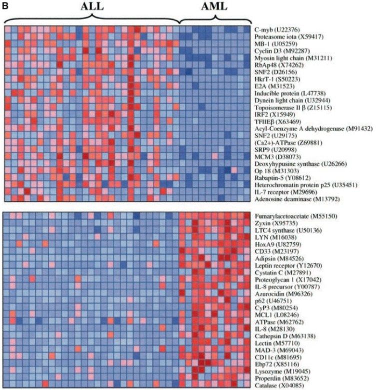
2 Inference of gene regulatory networks

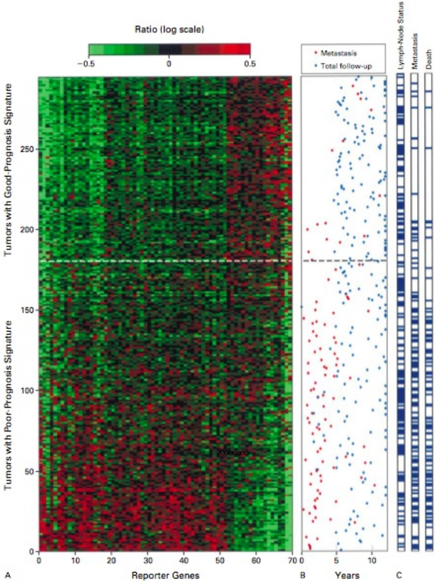
3 Cancer prognosis from DNA copy number variations

4 Diagnosis and prognosis from gene expression data

5 Conclusion

Molecular diagnosis / prognosis / theragnosis

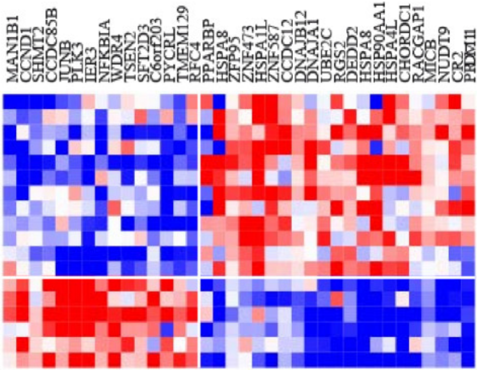




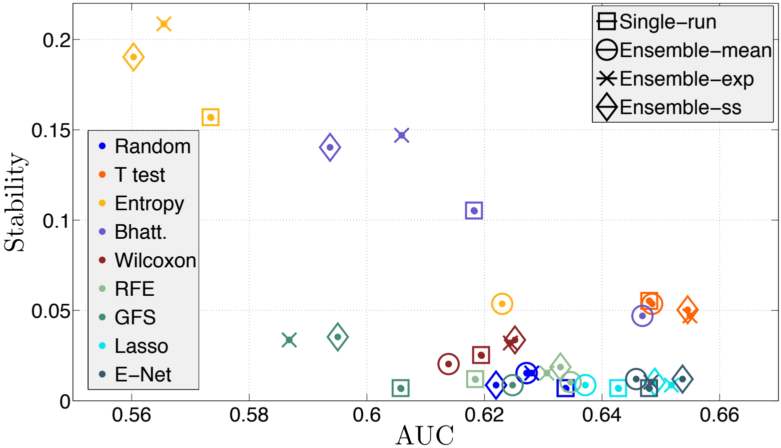
Gene selection, molecular signature

The idea

- We look for a limited set of genes that are sufficient for prediction.
- Selected genes should inform us about the underlying biology

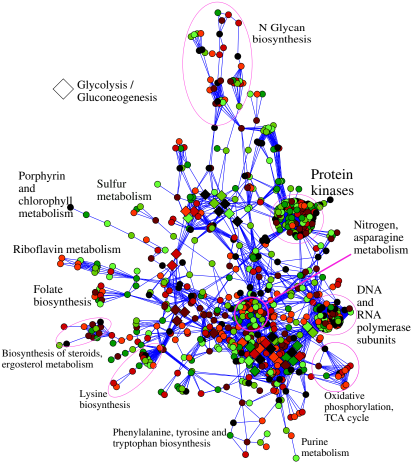


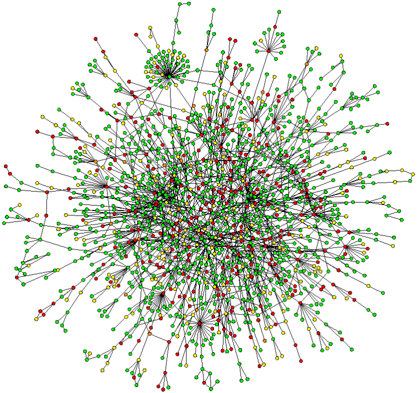
Lack of stability of signatures



Haury et al. (2011)

Gene networks


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



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
- We know these groups through functional groups and protein networks

Shrinkage estimators with prior knowledge

$$\min_{\beta} R(\beta) + \lambda\Omega(\beta)$$

How to design penalties $\Omega(\beta)$ to encode the following hypotheses:

1. Connected genes on a network should have similar weights
2. Select few genes that are connected or belong to same predefined functional groups

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
- We know these groups through functional groups and protein networks

Shrinkage estimators with prior knowledge

$$\min_{\beta} R(\beta) + \lambda \Omega(\beta)$$

How to design penalties $\Omega(\beta)$ to encode the following hypotheses:

1. Connected genes on a network should have similar weights
2. Select few genes that are connected or belong to same predefined functional groups

Hypothesis 1: connected genes on a network should have similar weights

- Smooth weights on the graph (or more generally graph kernels)

$$\Omega(\beta) = \sum_{i\sim j} \left(\beta_i - \beta_j\right)^2$$

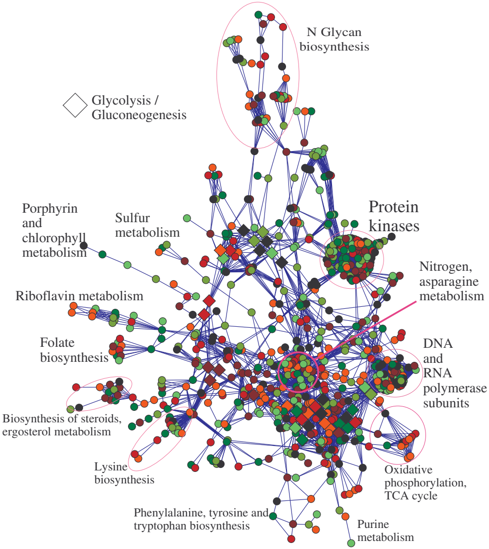
- Gene selection + smooth on the graph

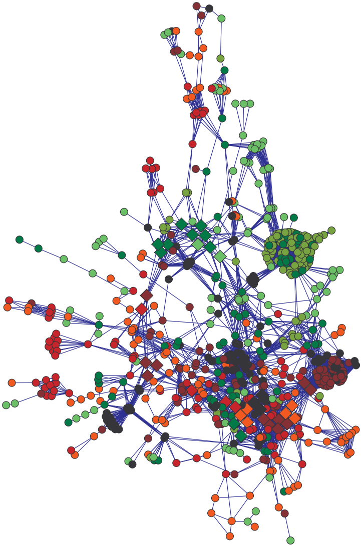
$$\Omega(\beta) = \sum_{i\sim j} \left(\beta_i - \beta_j\right)^2 + \sum_{i=1}^{p} \mid \beta_i \mid$$

- Gene selection + Piecewise constant on the graph (total variation)

$$\Omega(\beta) = \sum_{i\sim j} \left| \beta_i - \beta_j \right| + \sum_{i=1}^{p} \mid \beta_i \mid$$

Example (Rapaport et al., 2008)

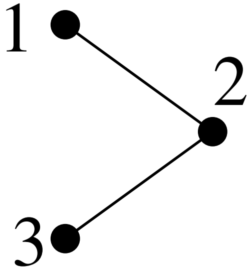

N Glycan biosynthesis
Glycolysis / Gluconeogenesis
Porphyrin and chlorophyll metabolism
Sulfur metabolism
Protein kinases
Nitrogen, asparagine metabolism
Riboflavin metabolism
Folate biosynthesis
DNA and RNA polymerase subunits
Biosynthesis of steroids, ergosterol metabolism
Lysine biosynthesis
Oxidative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

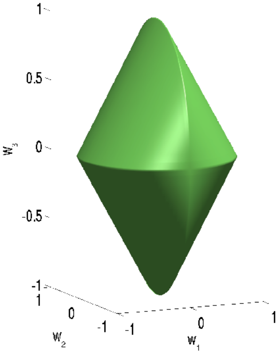


Hypothesis 2: select connected genes

- A difficult combinatorial problem
- A convex solution: the latent group Lasso (Jacob et al., 2009)

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





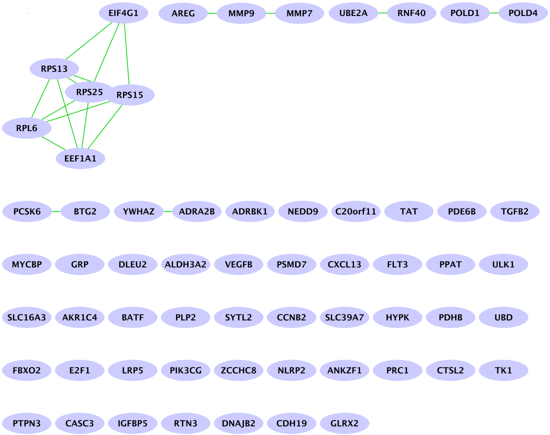
Preliminary results

Breast cancer data

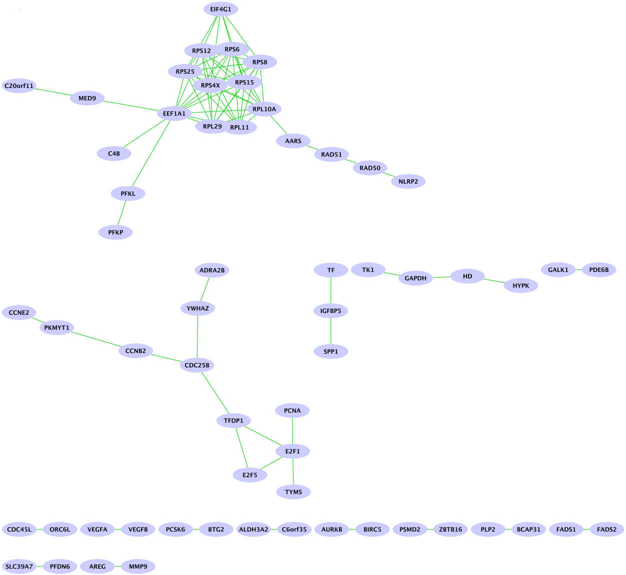
- Gene expression data for 8, 141 genes in 295 breast cancer tumors.
- Performance

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

Classical lasso signature



Graph Lasso signature



Outline

1 Introduction

2 Inference of gene regulatory networks

3 Cancer prognosis from DNA copy number variations

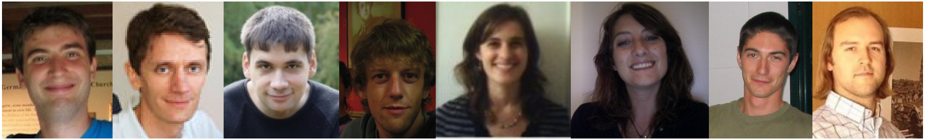
4 Diagnosis and prognosis from gene expression data

5 Conclusion

Conclusion

- Machine learning offers many powerful tools to learn predictive models from large sets of complex data
- Specific developments are required to solve complex problems that arise in bio-informatics
- Dedicated convex penalties in empirical risk minimisation offer a theoretically sound and computationally efficient framework
- Many other applications not covered in this presentation!

Acknowledgements!



Franck Rapaport (MSKCC), Emmanuel Barillot, Andrei Zynoviev (Curie), Kevin Bleakley (INRIA), Fantine Mordelet (Duke), Anne-Claire Haury (Mines), Laurent Jacob (UC Berkeley) Guillaume Obozinski (INRIA)

NATIONAL INSTITUTES OF HEALTH

JAPAN SOCIETY FOR THE PROMOTION OF SCIENCE FOUNDED 1932

FRANCE-BERKELEY FUND
FONDS FRANCE-BERKELEY
UNIVERSITY OF CALIFORNIA, BERKELEY | LE MINISTERE DES AFFAIRES ETRANGERES

European Research Council erc

AGENCE NATIONALE DE LA RECHERCHE ANR

medicen PARIS REGION

Cancer-Bio-Santé
PÔLE DE COMPÉTITIVITÉ

SEVENTH FRAMEWORK PROGRAMME