Machine learning in bioinformatics

Jean-Philippe Vert
Jean-Philippe.Vert@mines.org

Mines ParisTech / Curie Institute / Inserm

Biointelligence symposium, Sophia-Antipolis, July 4, 2011.

Institut Curie - Mines ParisTech - INSERM U900

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

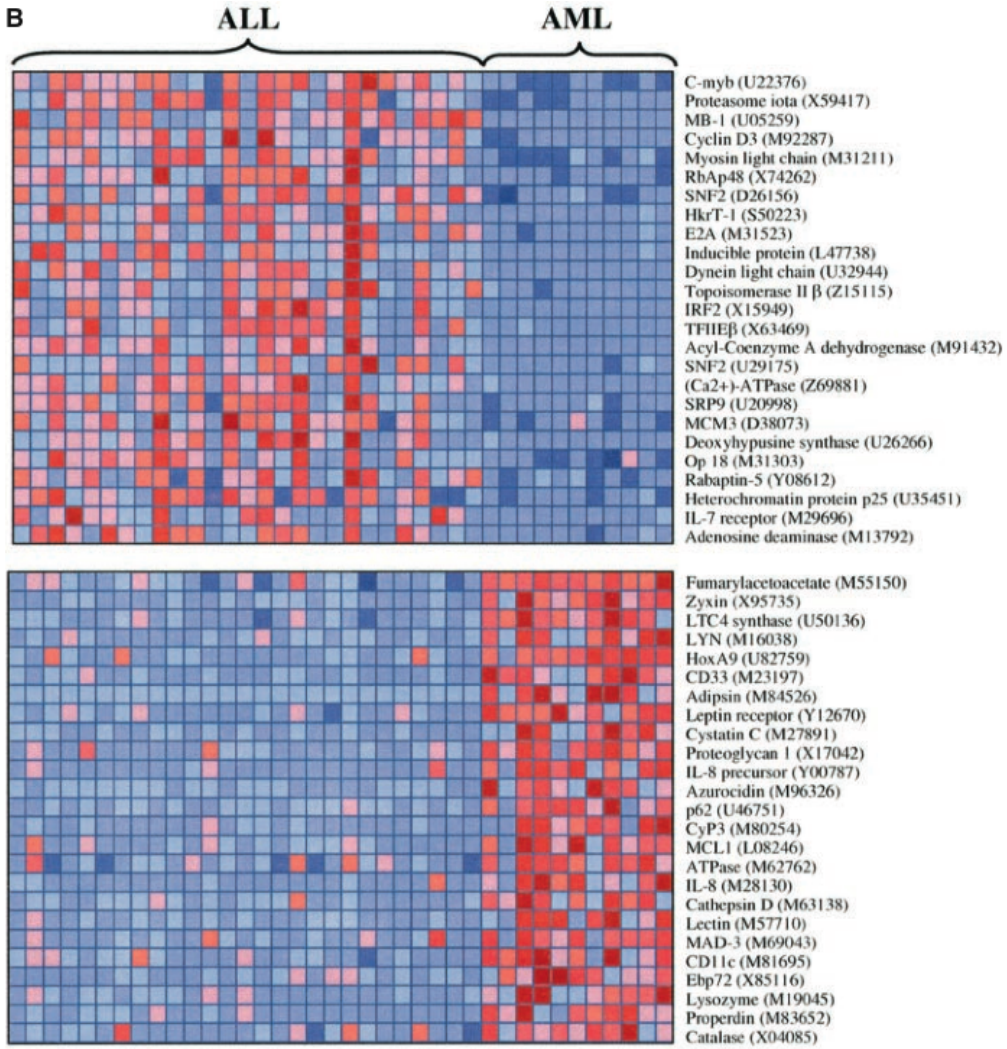
Main topics

- 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.

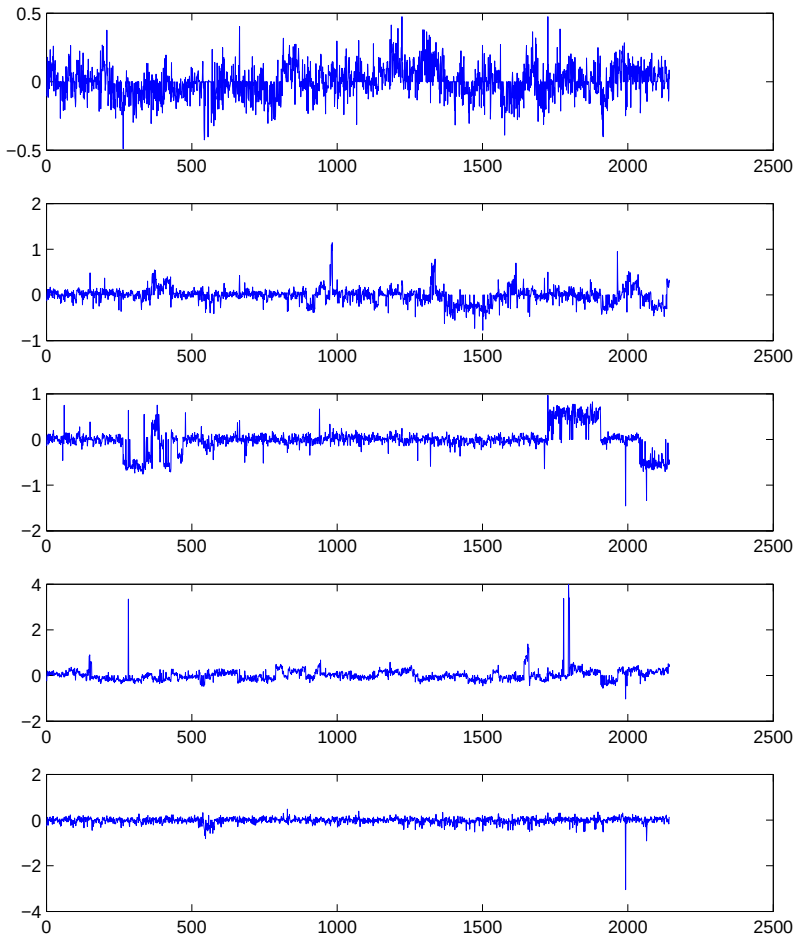
Outline

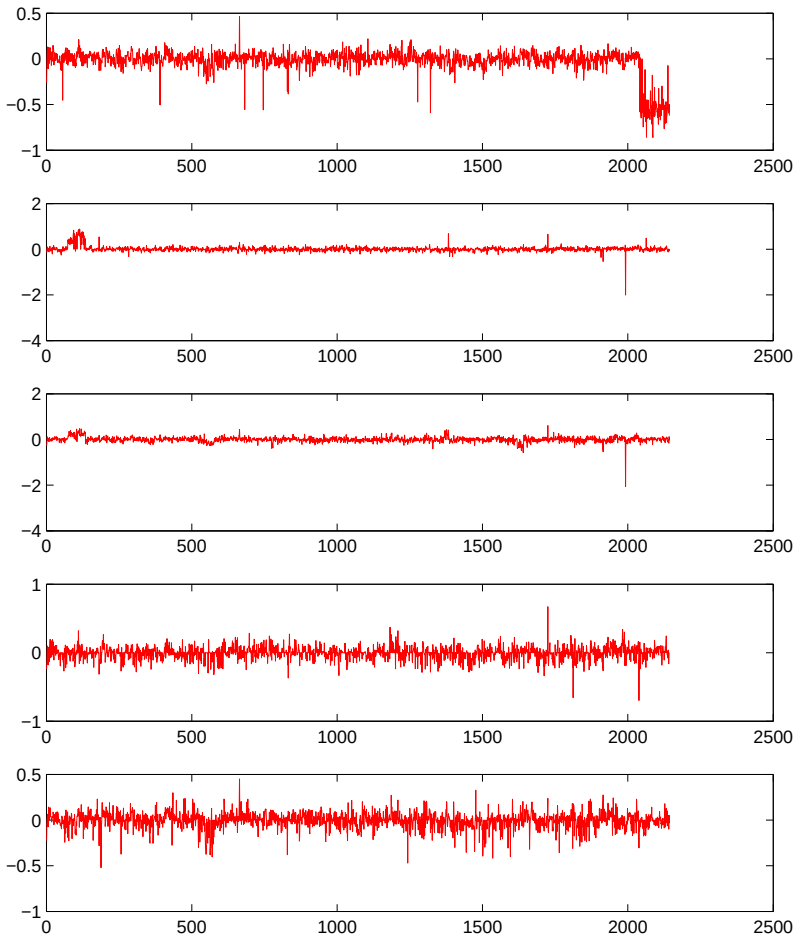
Outline

Cancer diagnosis

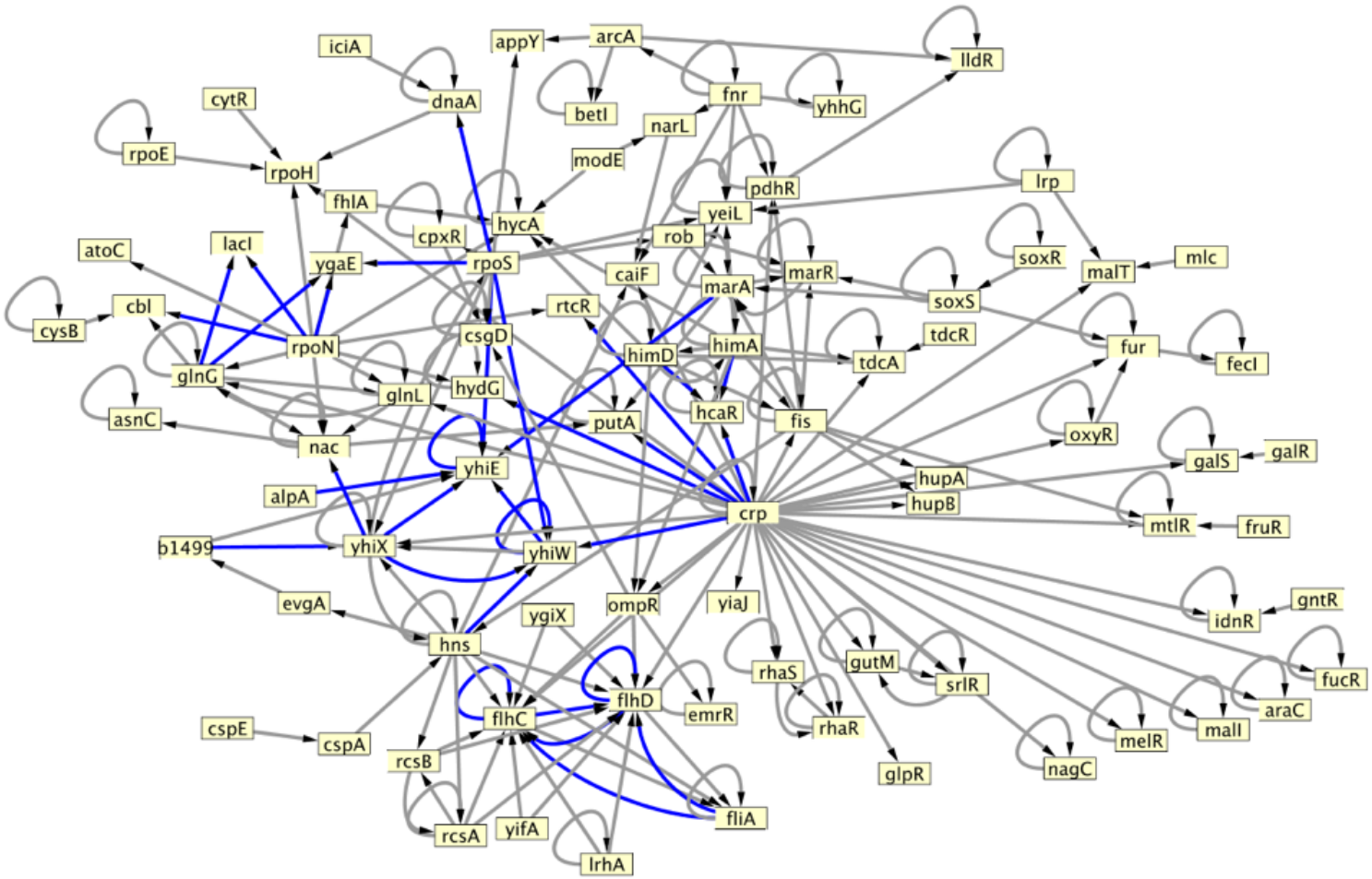


Cancer prognosis

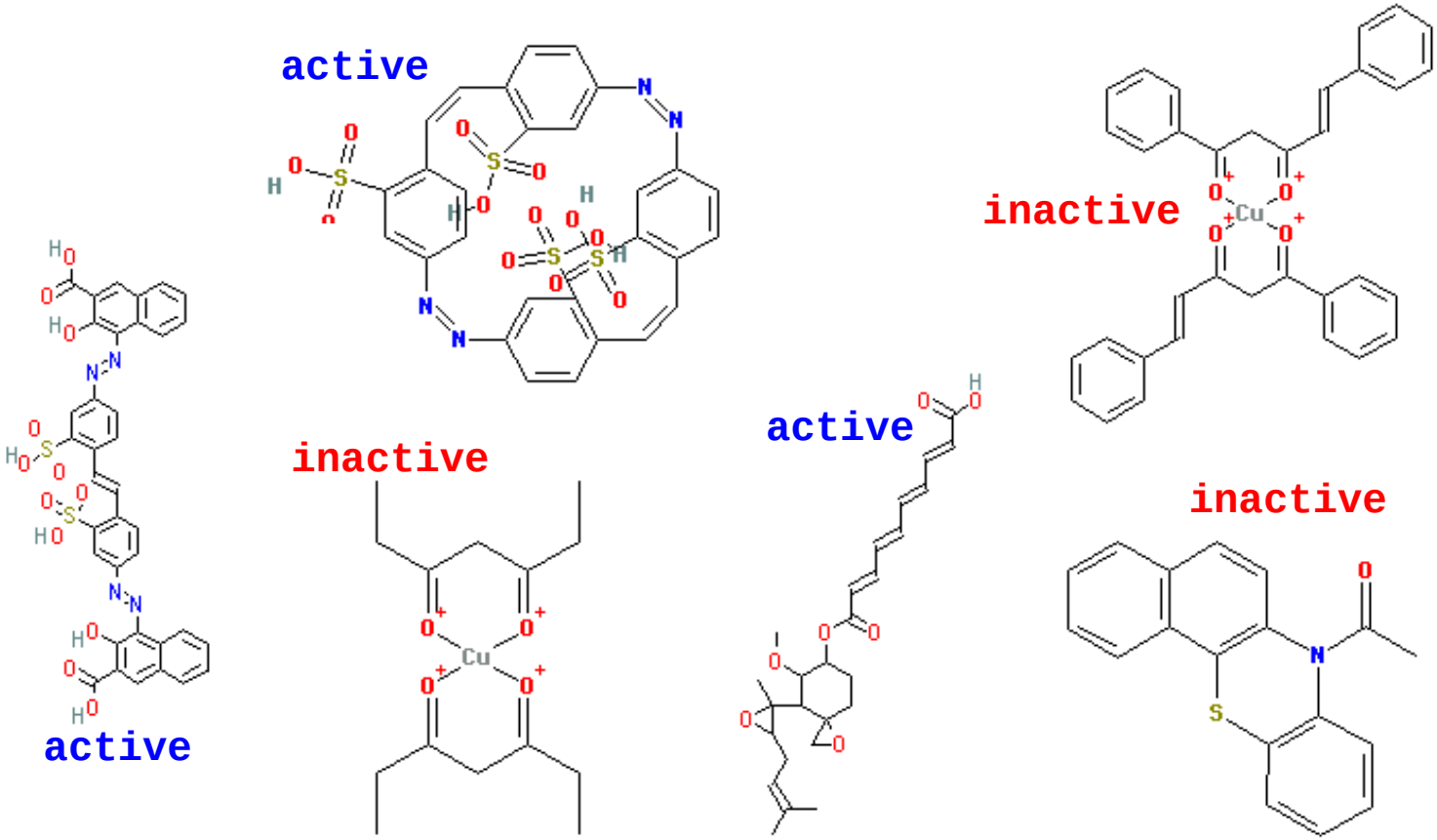




Gene network inference

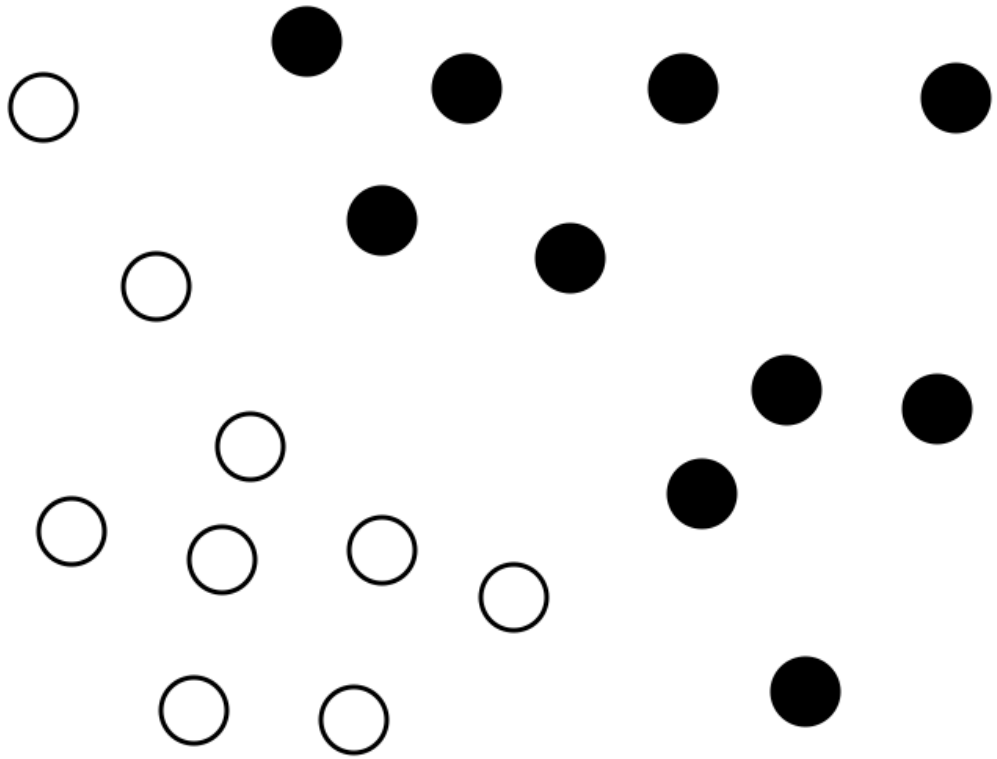


Virtual screening for drug discovery



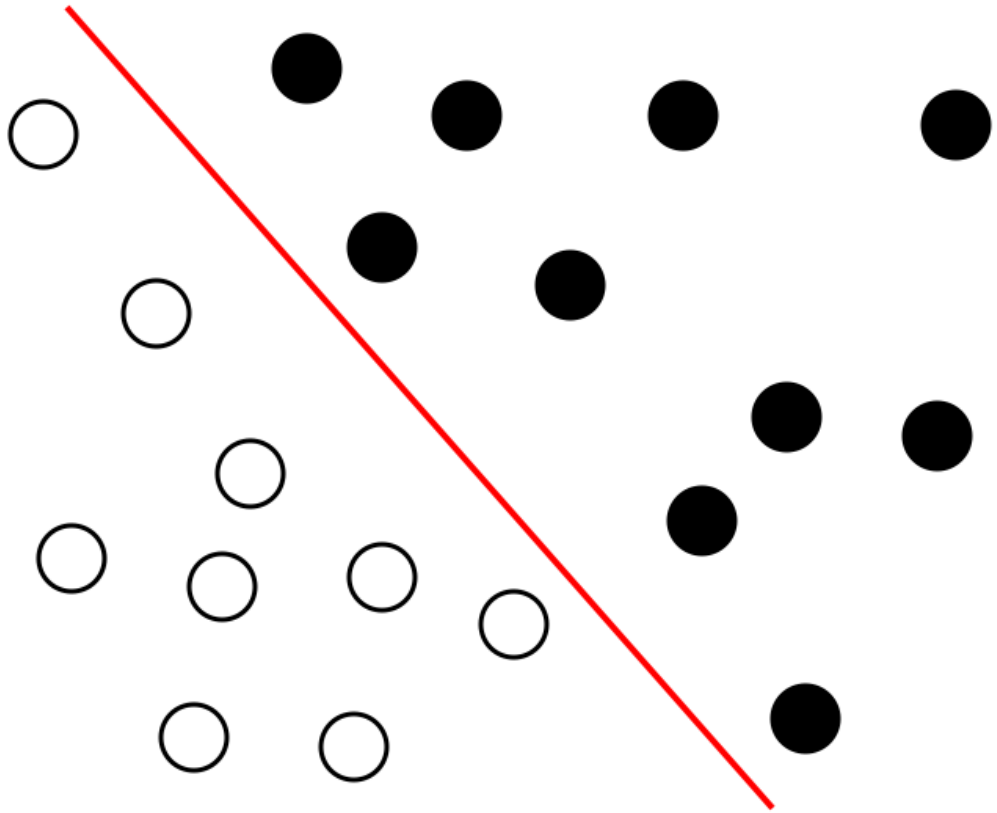
NCI AIDS screen results (from http://cactus.nci.nih.gov).

Machine learning



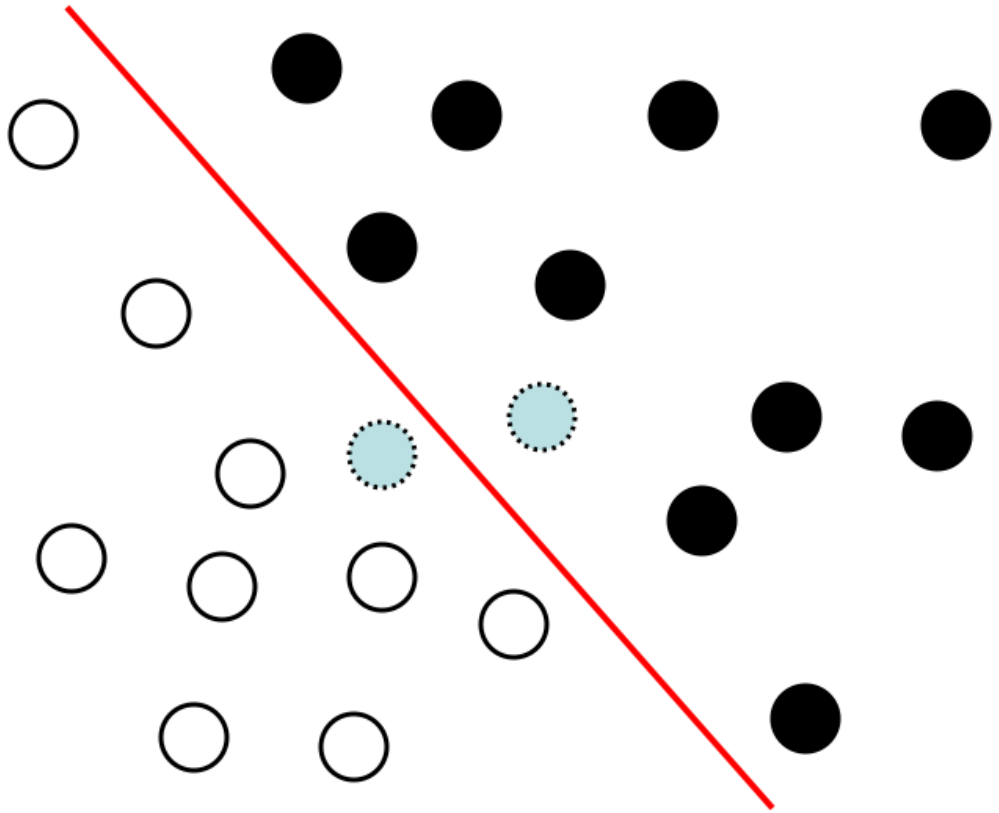
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning



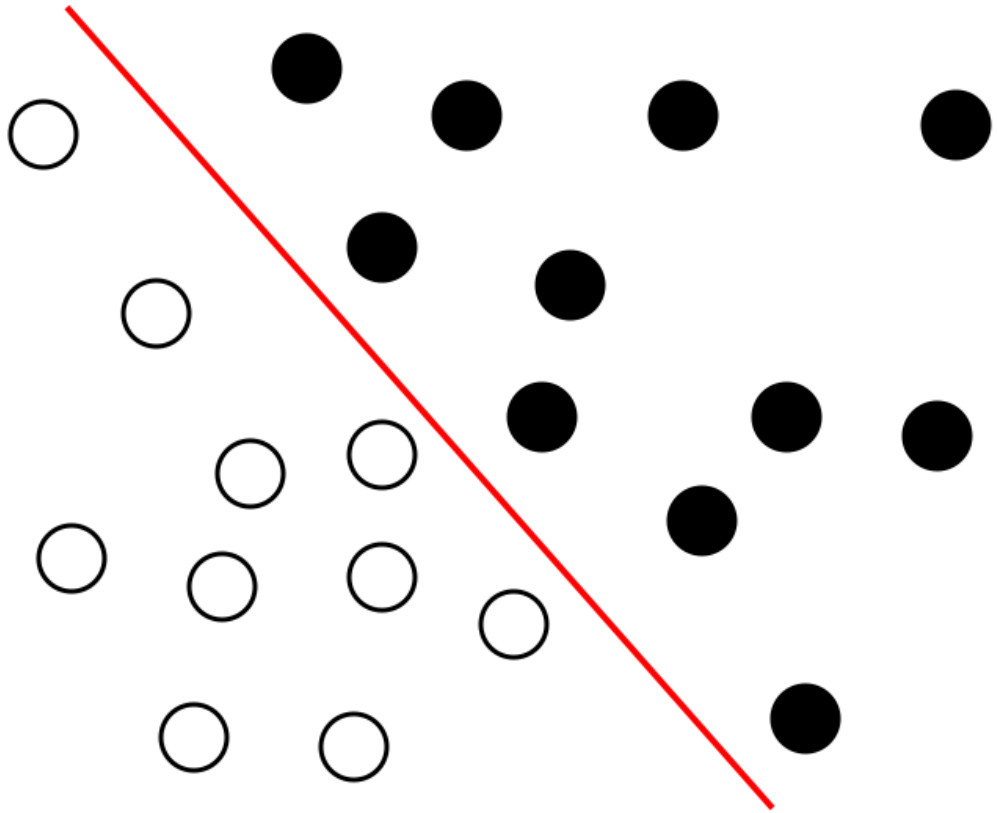
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

Machine learning



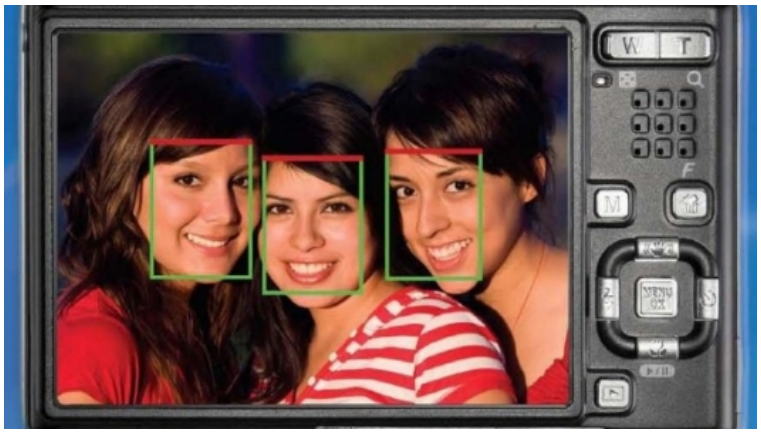
1. Given a training set of labeled data with...
2. learn a discrimination rule...
3. ... in order to predict the label of new data

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



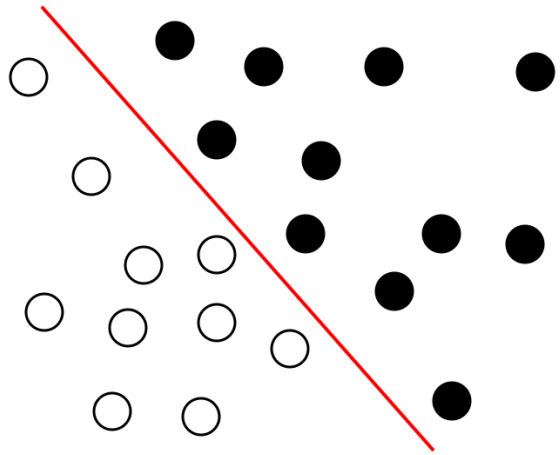
Many applications

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



Genome annotation, systems biology, personalized medicine...

Challenges

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

Outline

Gene expression

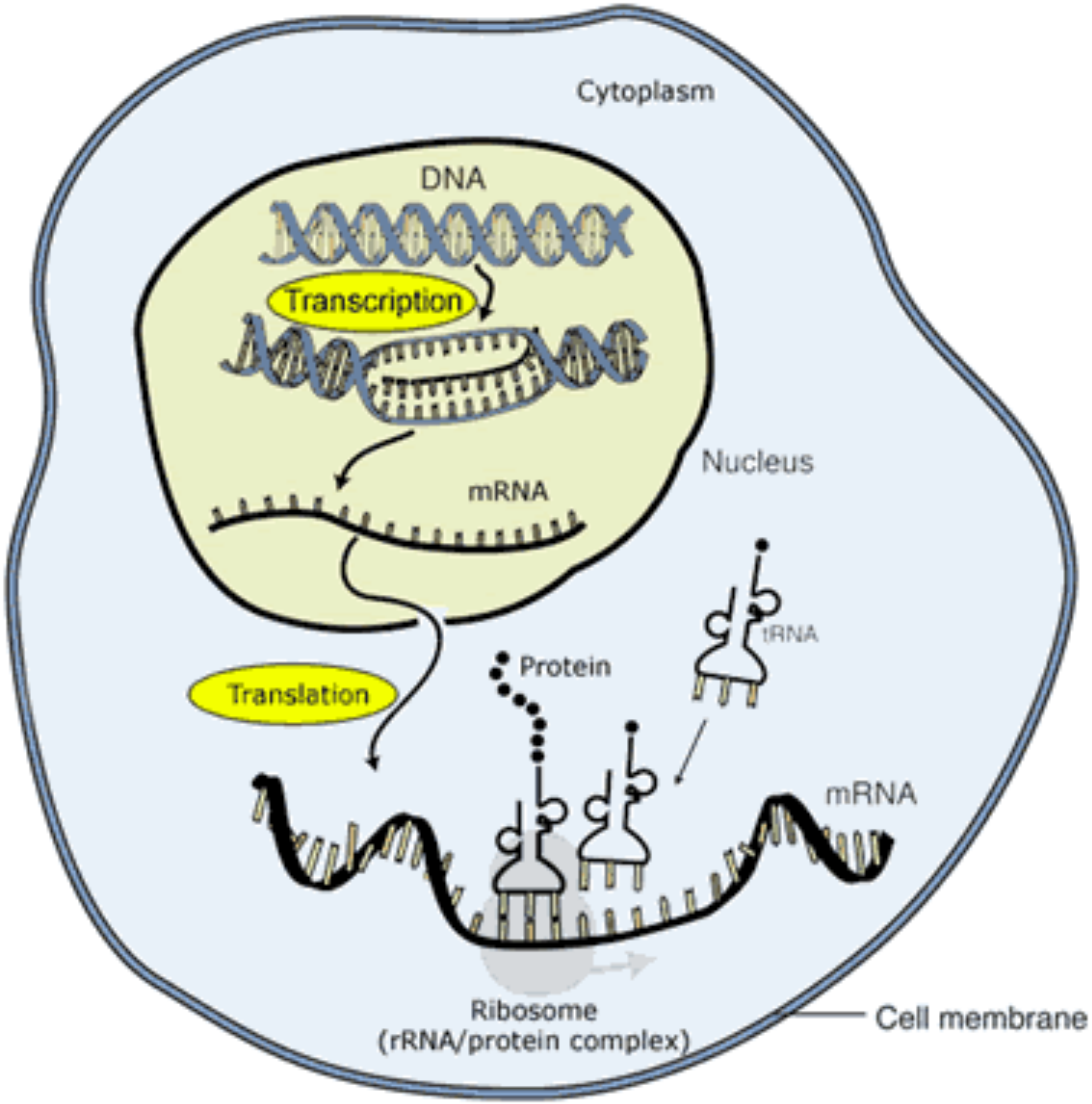
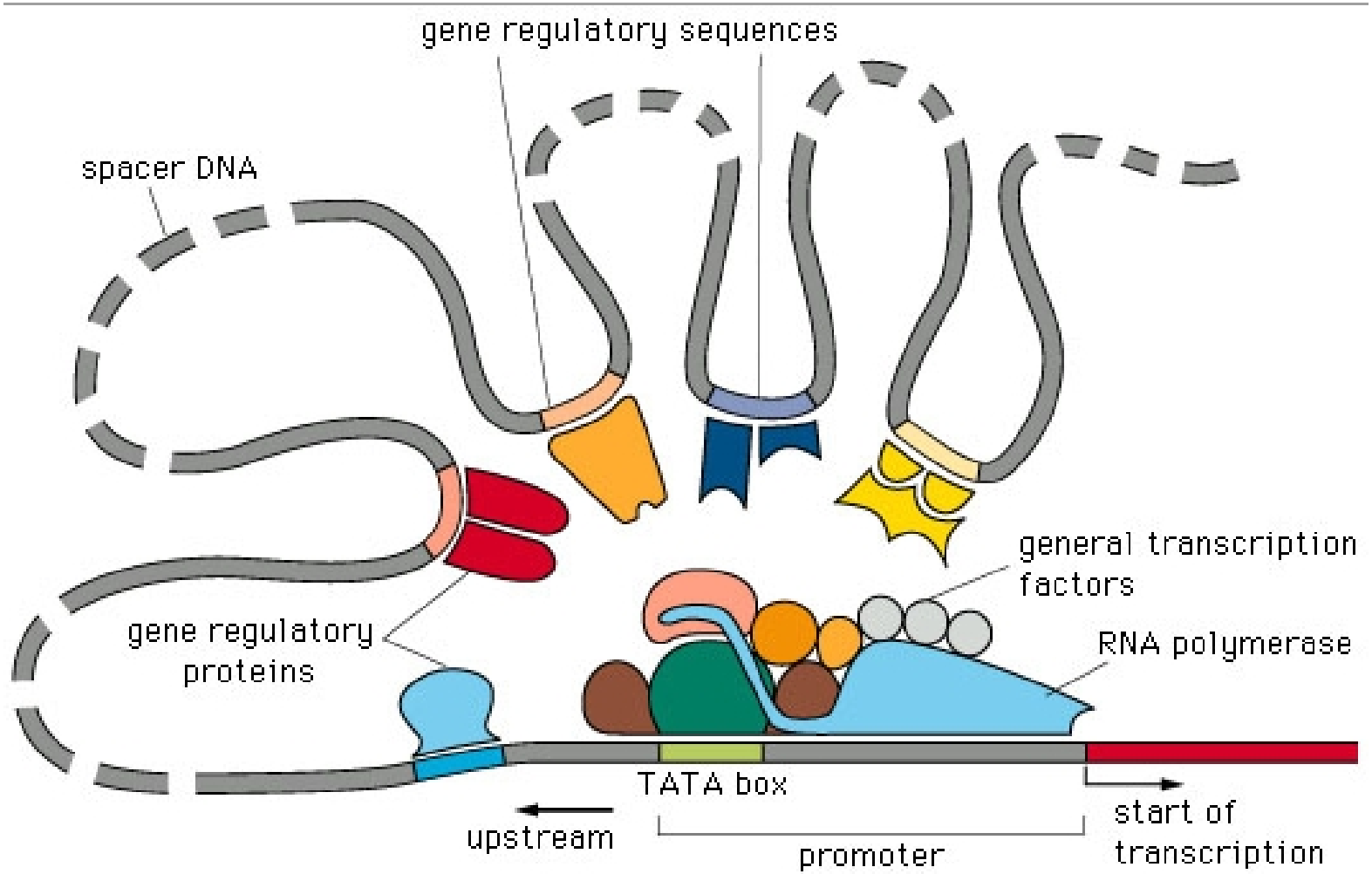
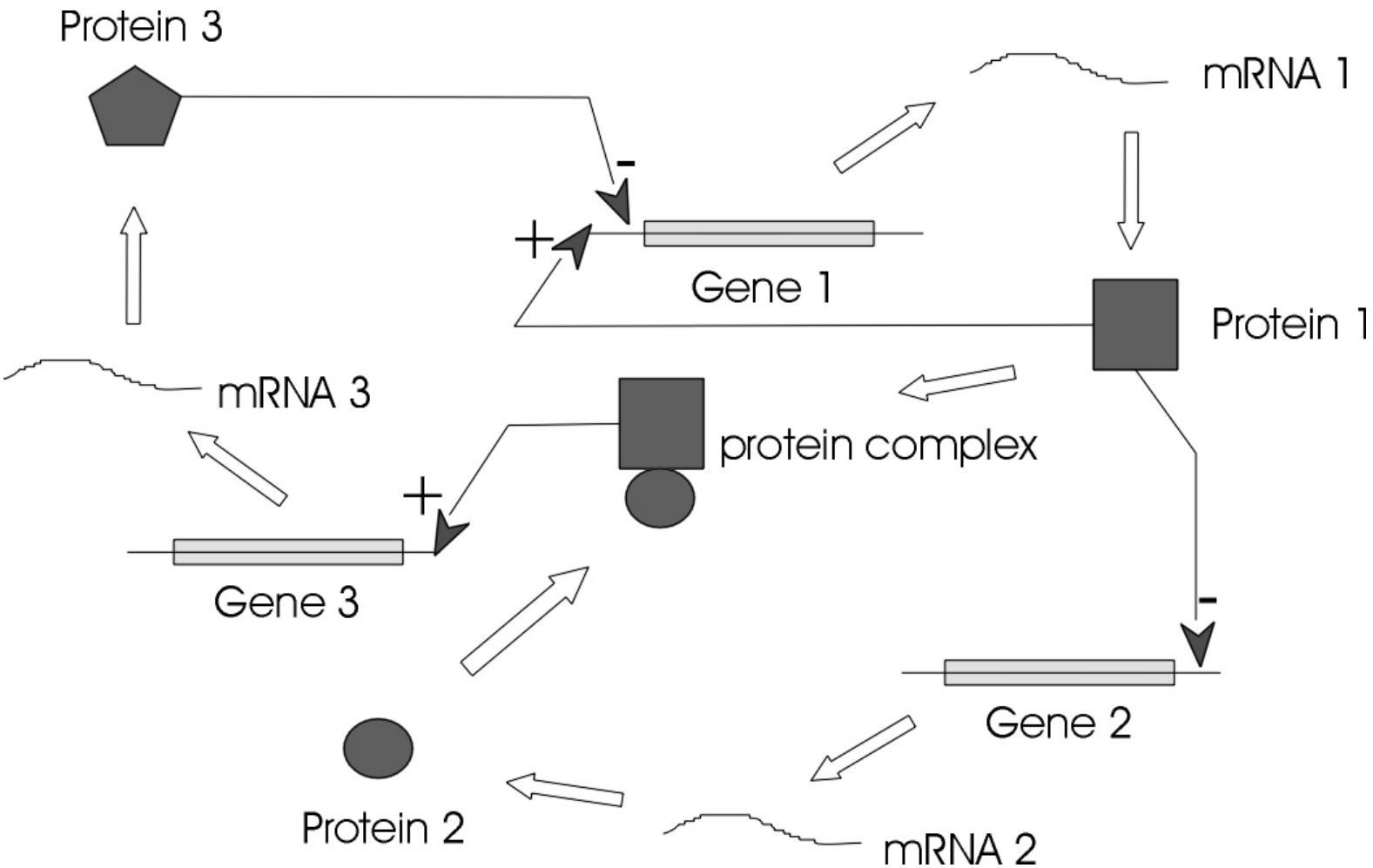


Image adapted from: National Human Genome Research Institute.

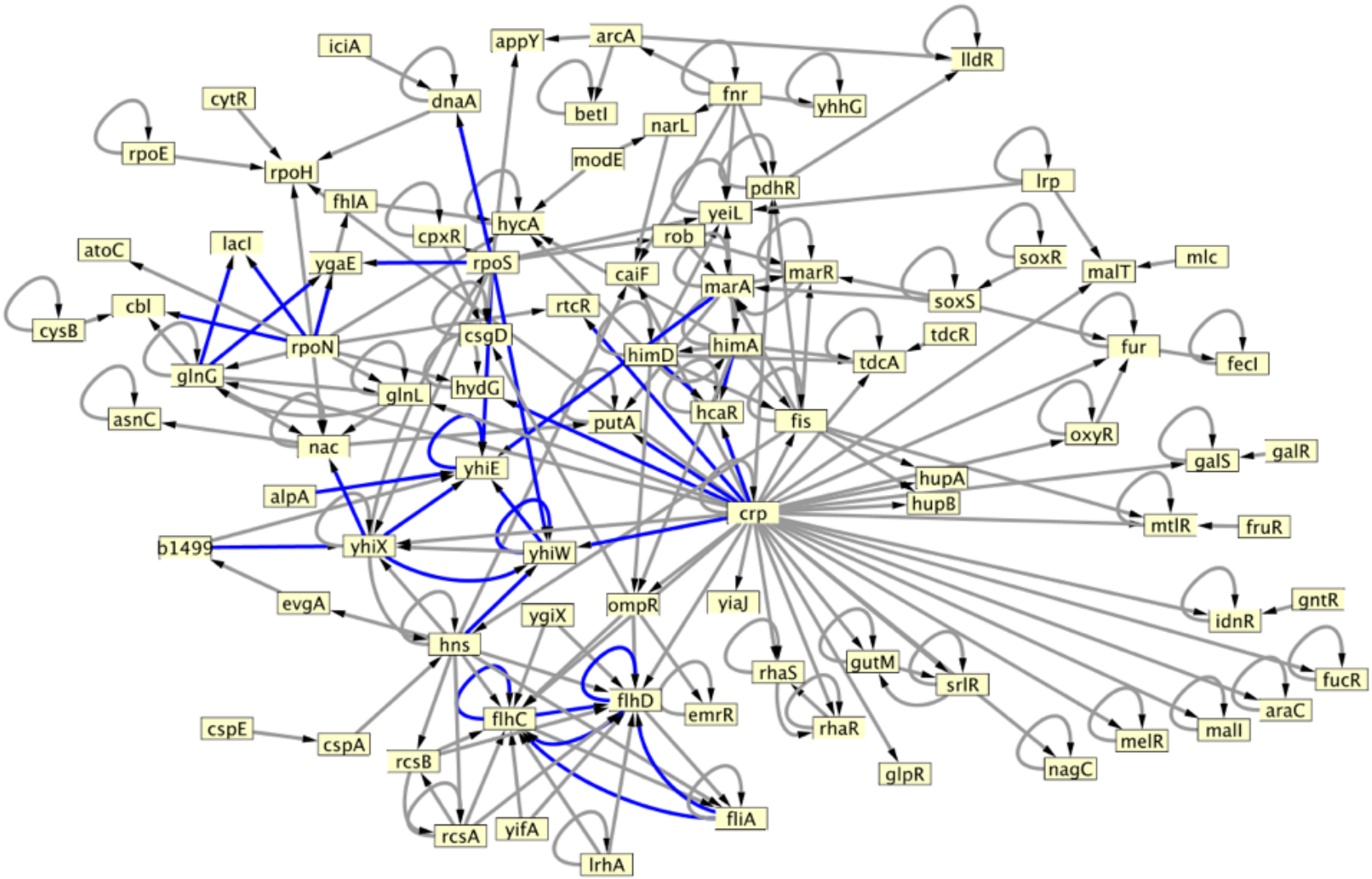
Gene expression regulation



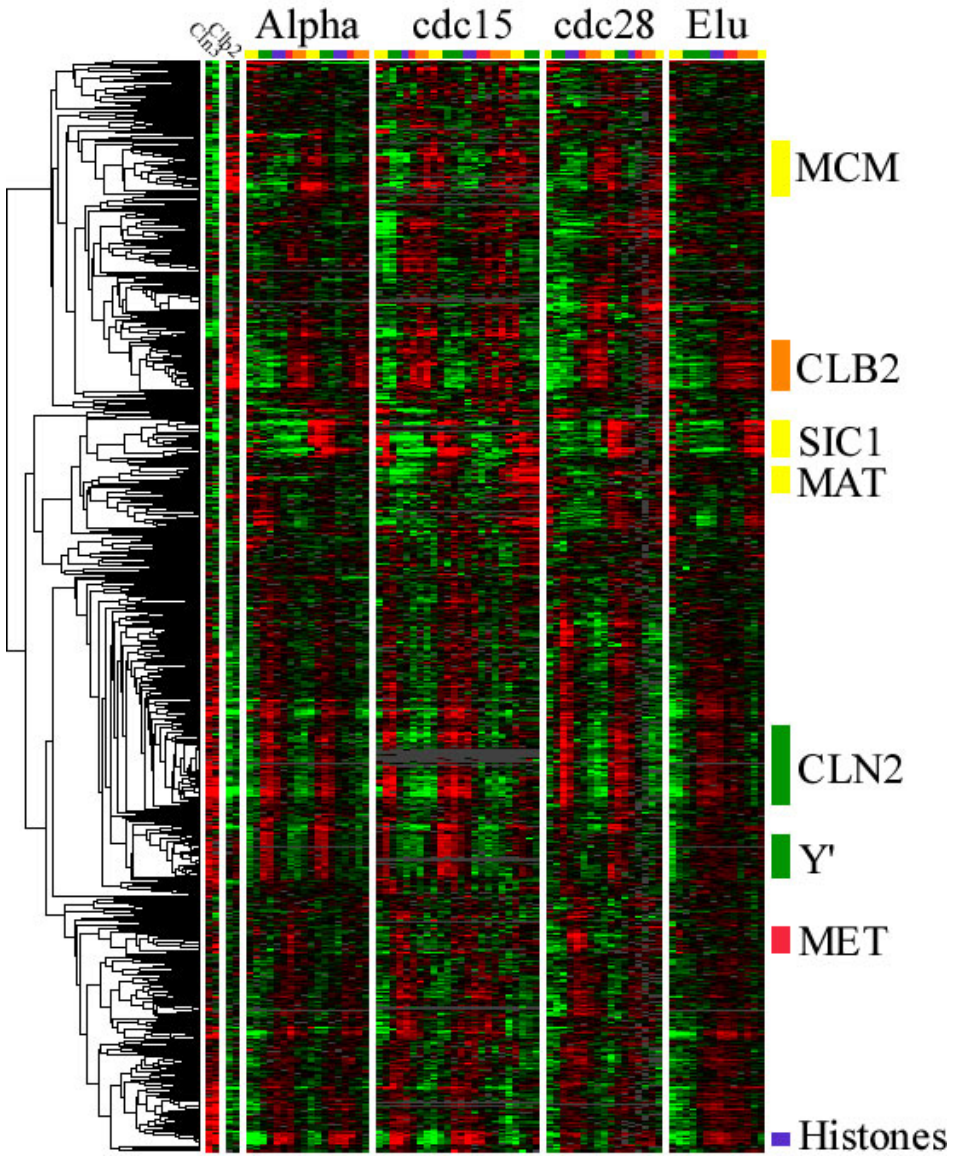
Gene regulatory network



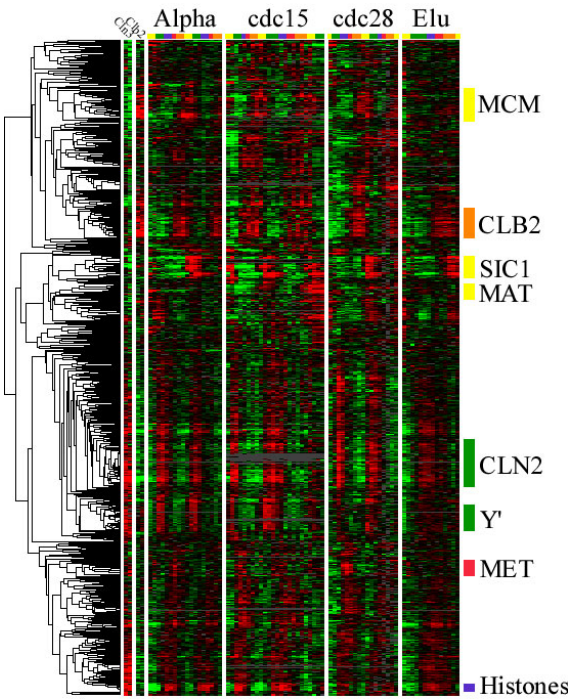
Gene regulatory network of E. coli

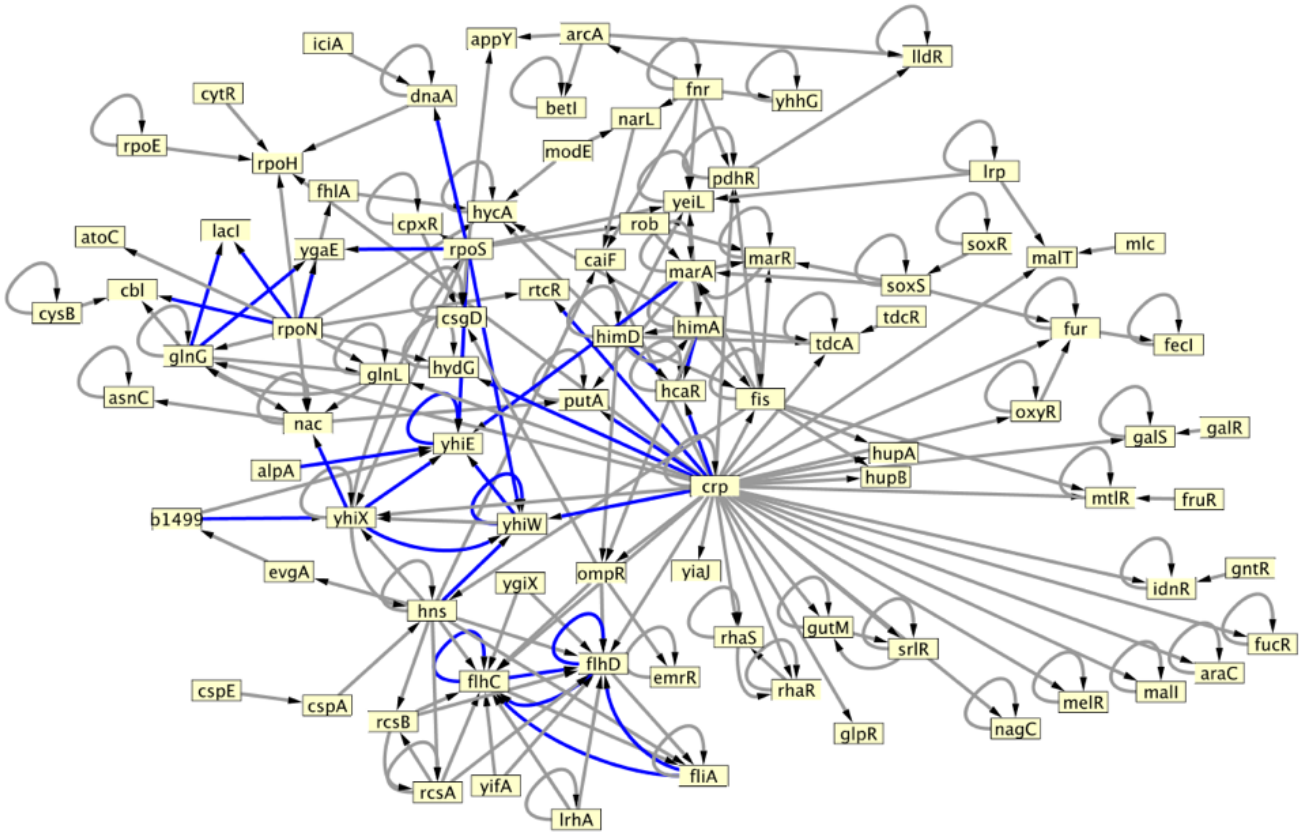


Gene expression data



Reconstruction of gene regulatory network from expression data

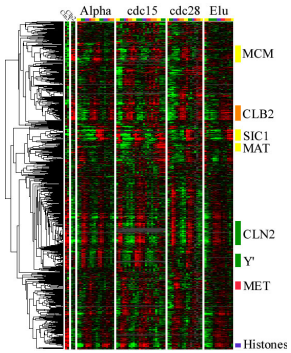


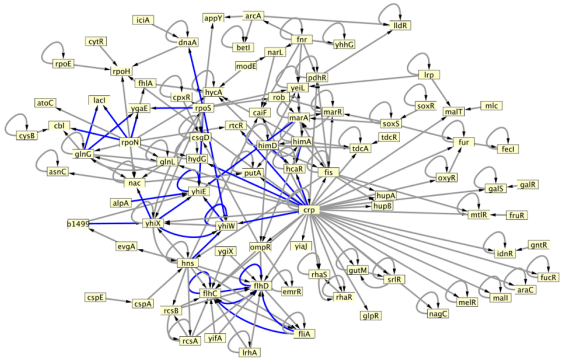


De novo inference

The problem

Given a set of gene expressions, infer the regulations.





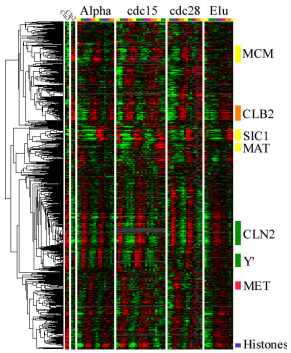
How?

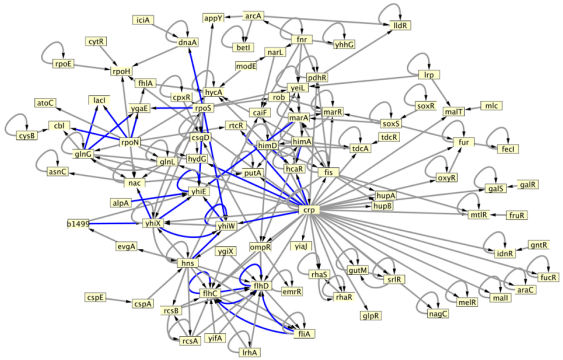
- Classical approach: connect "similar" genes
- Machine learning formulation: estimate regulators as the smallest set of TF necessary to predict the expression of the target (using, e.g., Lasso or random forest)

De novo inference

The problem

Given a set of gene expressions, infer the regulations.

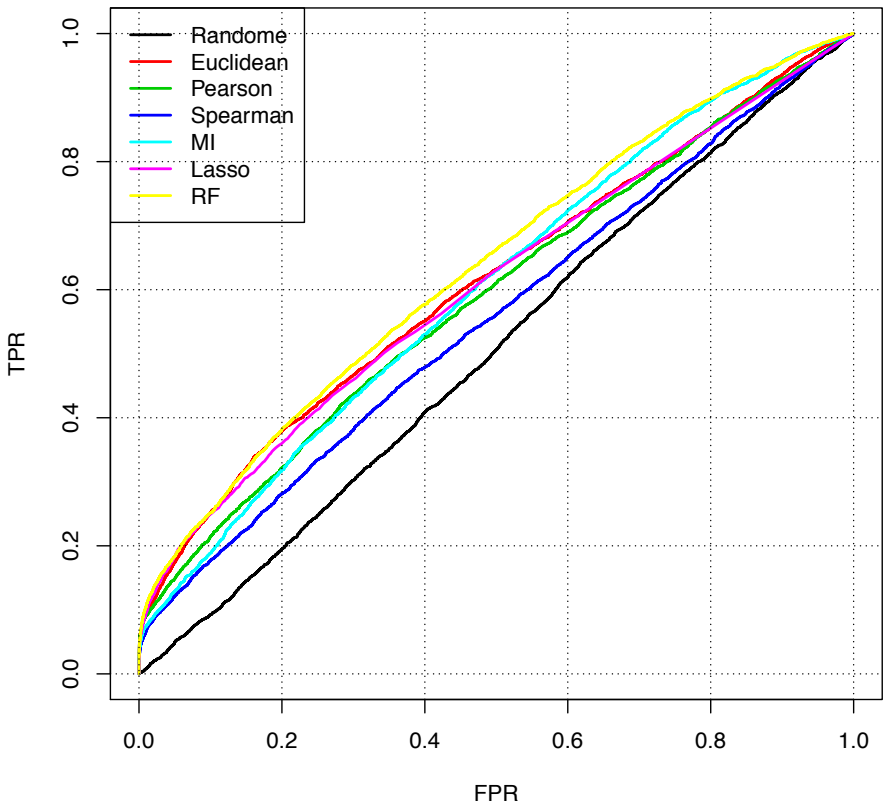


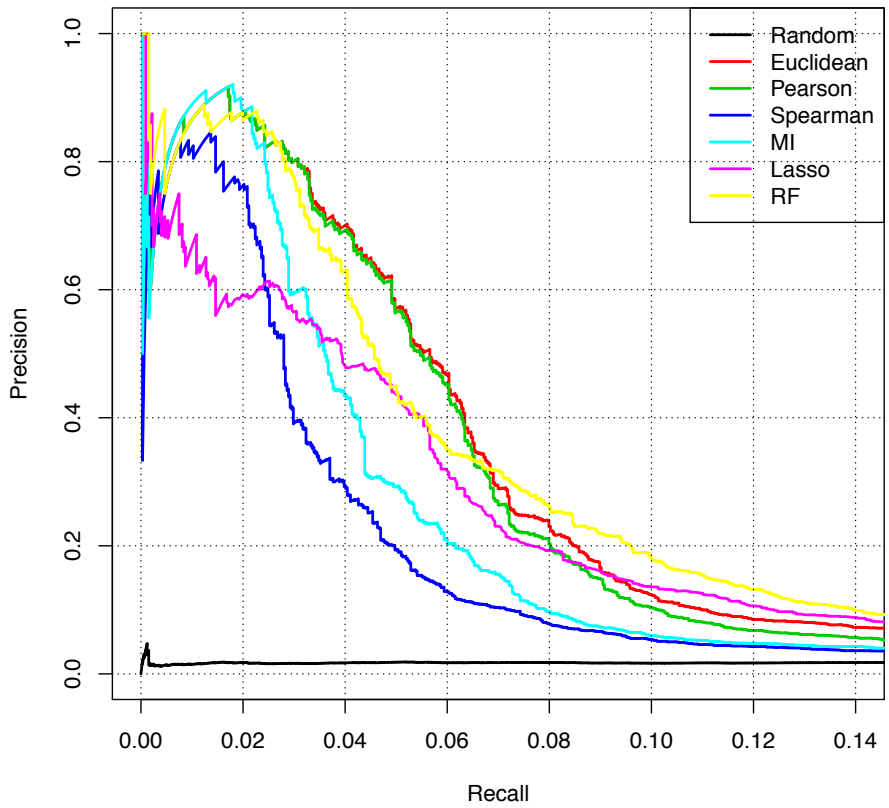


How?

- Classical approach: connect "similar" genes
- Machine learning formulation: estimate regulators as the smallest set of TF necessary to predict the expression of the target (using, e.g., Lasso or random forest)

Validation



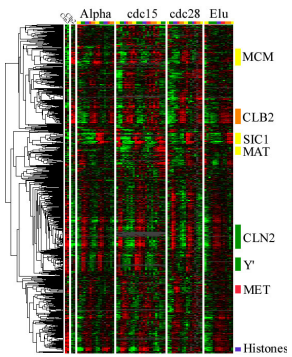


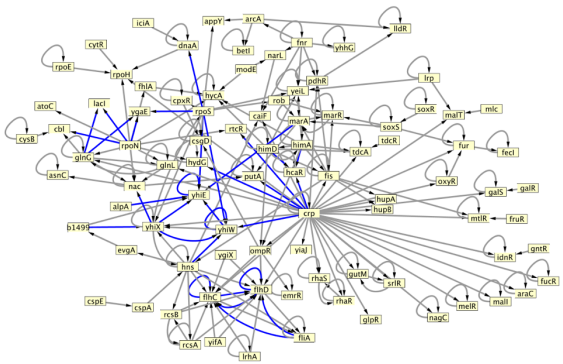
Random forests (Huynh-Thu et al., 2010) and Lasso regression (Haury et al., 2011) ranked 1st and 2nd at the 2010 DREAM5 in silico network inference challence

Supervised inference

The problem

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





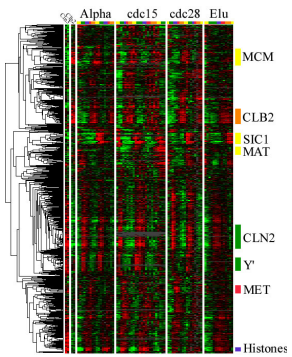
How?

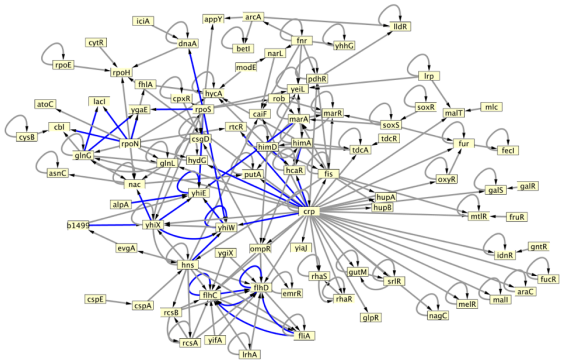
- Local models: for each TF, learn to discriminate the regulated vs non-regulated genes
- Global models: learn to discriminate connected vs non-connected TF-target pairs

Supervised inference

The problem

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



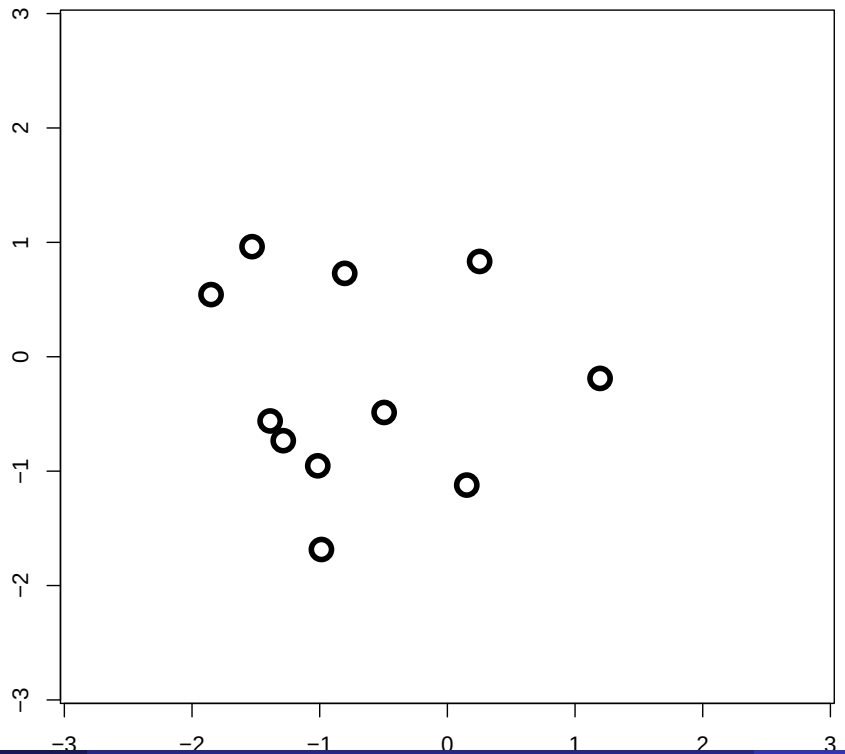


How?

- 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

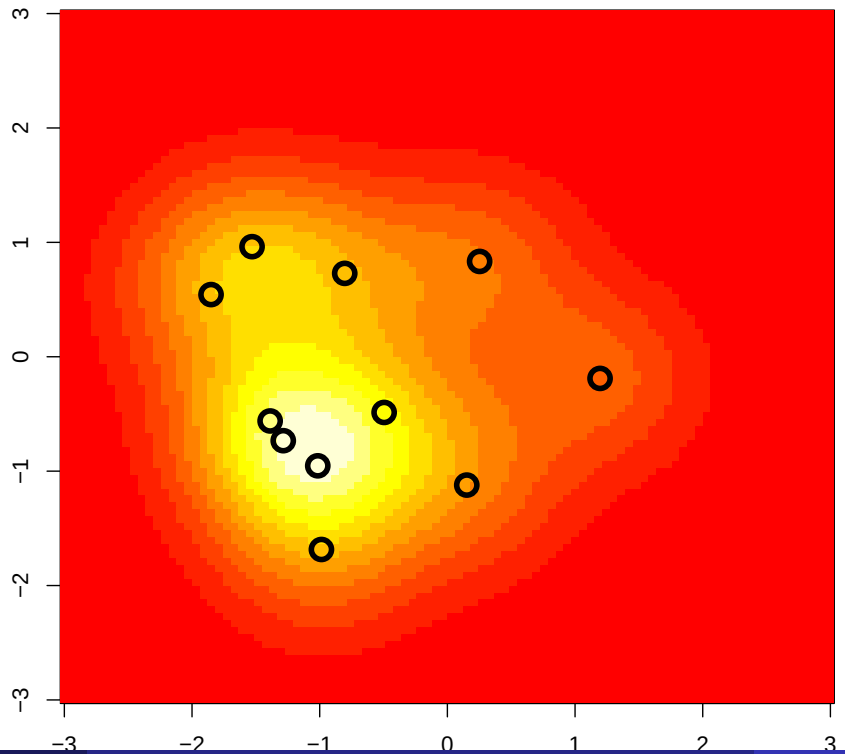
Example: one-class learning approach for local model

- 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



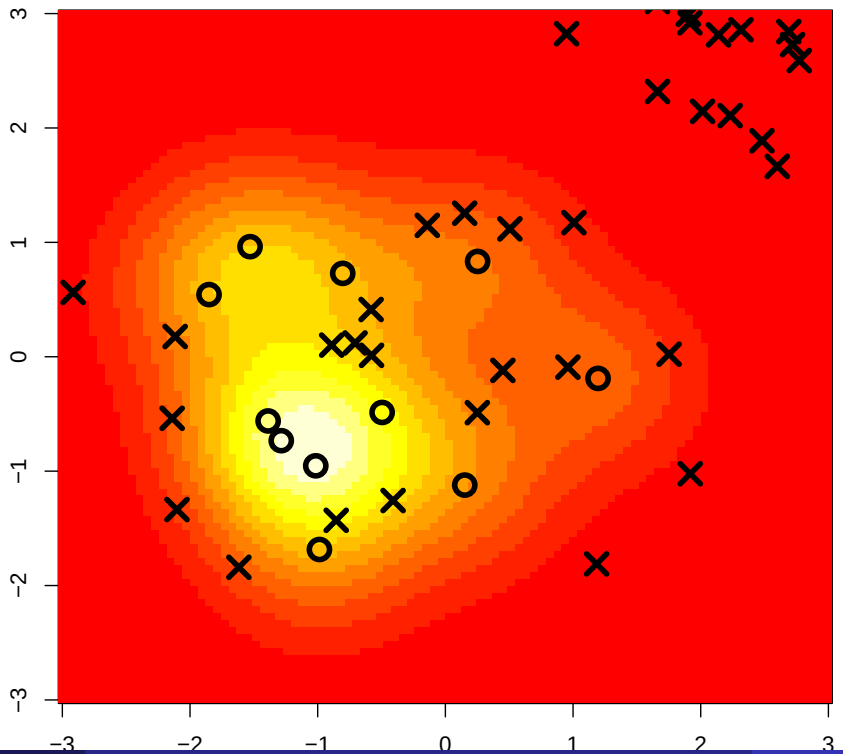
Example: one-class learning approach for local model

- 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

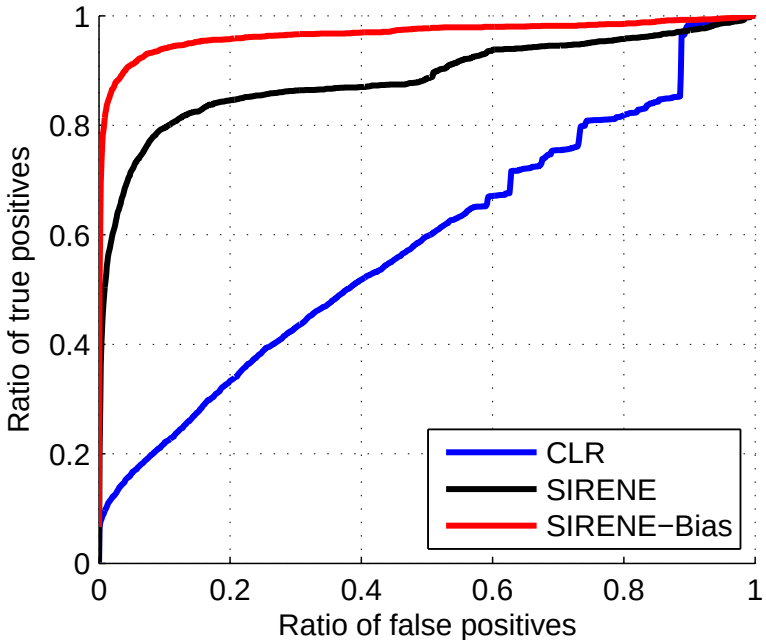


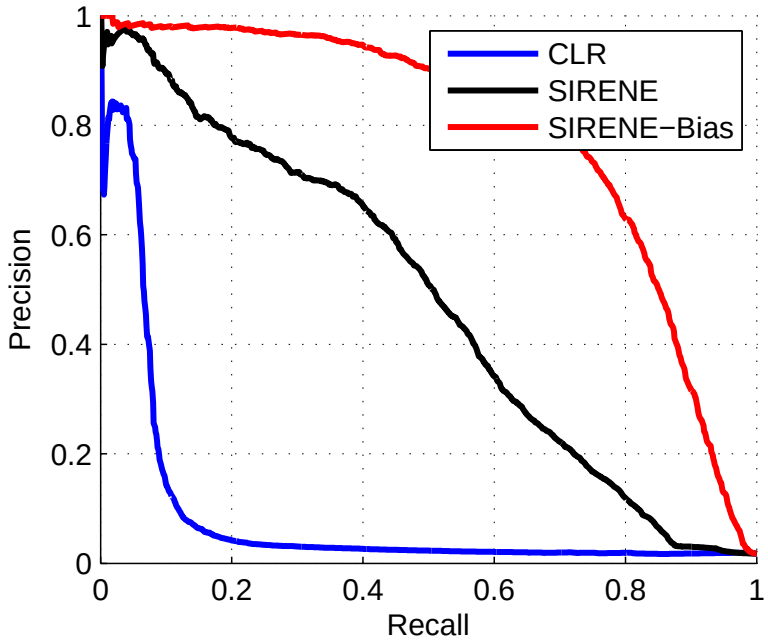
Example: one-class learning approach for local model

- 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

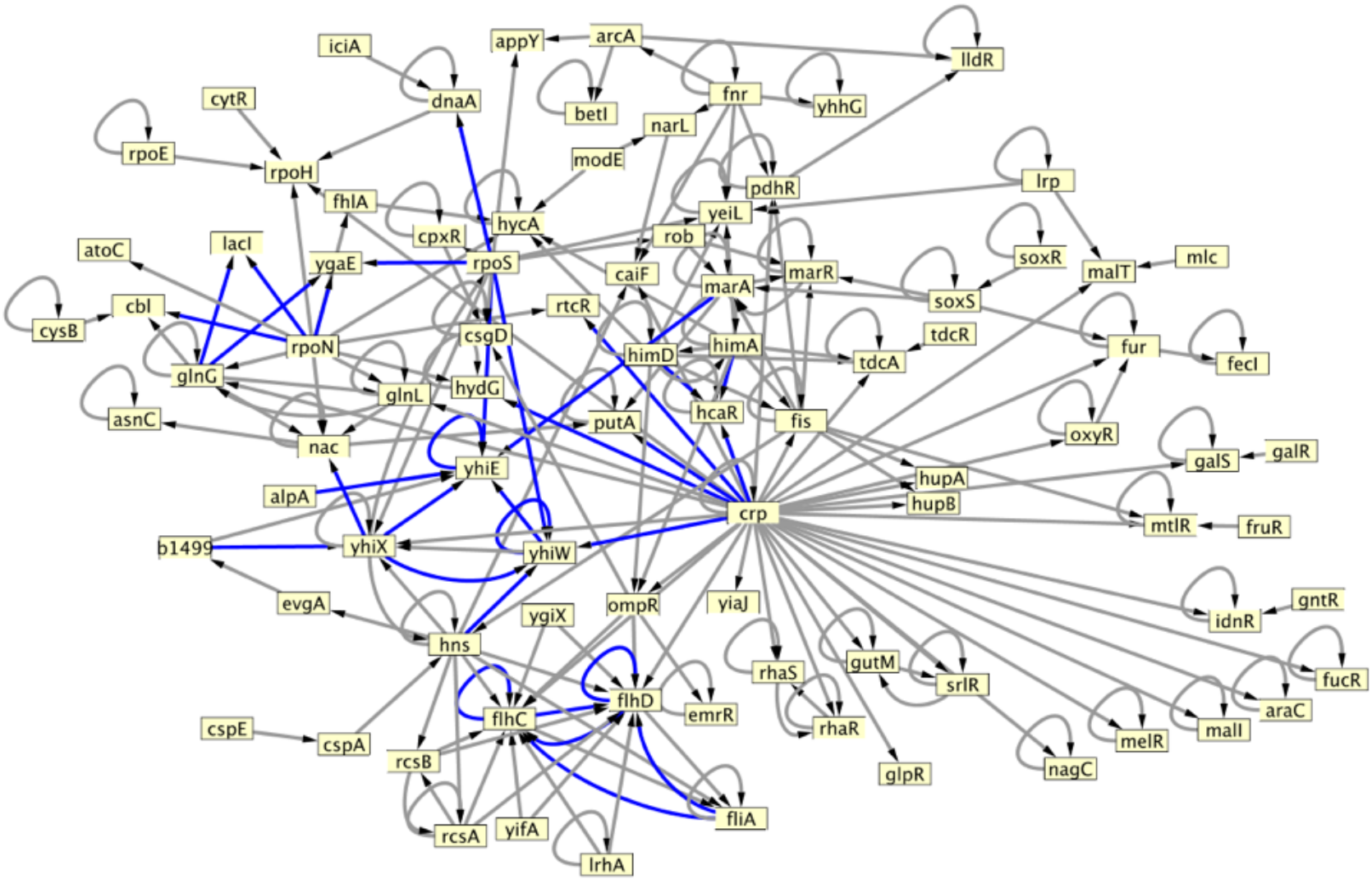




Method	Recall at 60%	Recall at 80%
SIRENE	**44.5%**	**17.6%**
CLR	7.5%	5.5%
Relevance networks	4.7%	3.3%
ARACNe	1%	0%
Bayesian network	1%	0%

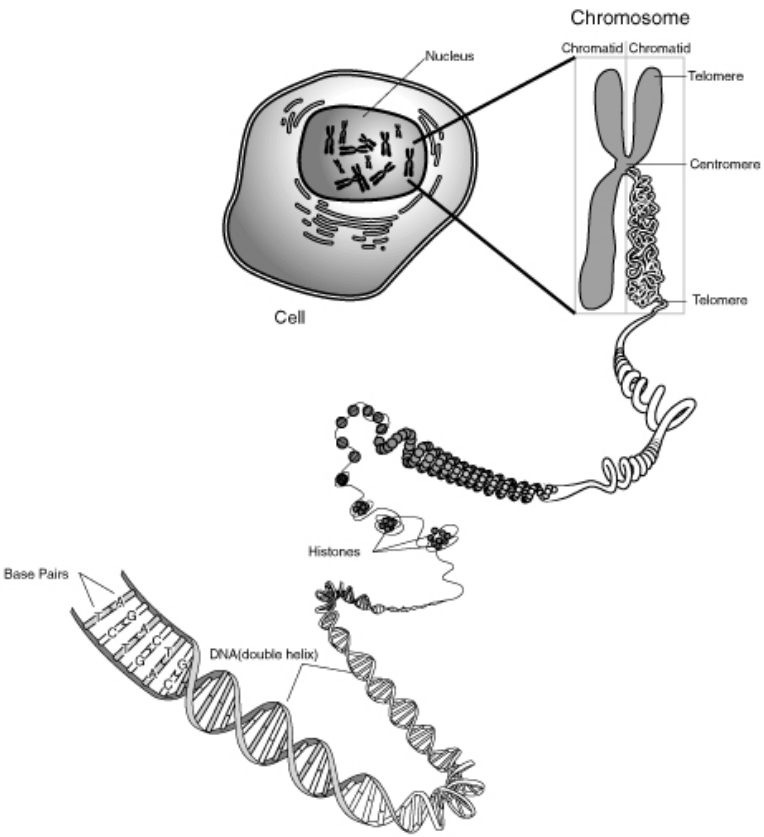
SIRENE = Supervised Inference of REgulatory NEtworks (Mordelet and V., 2008)

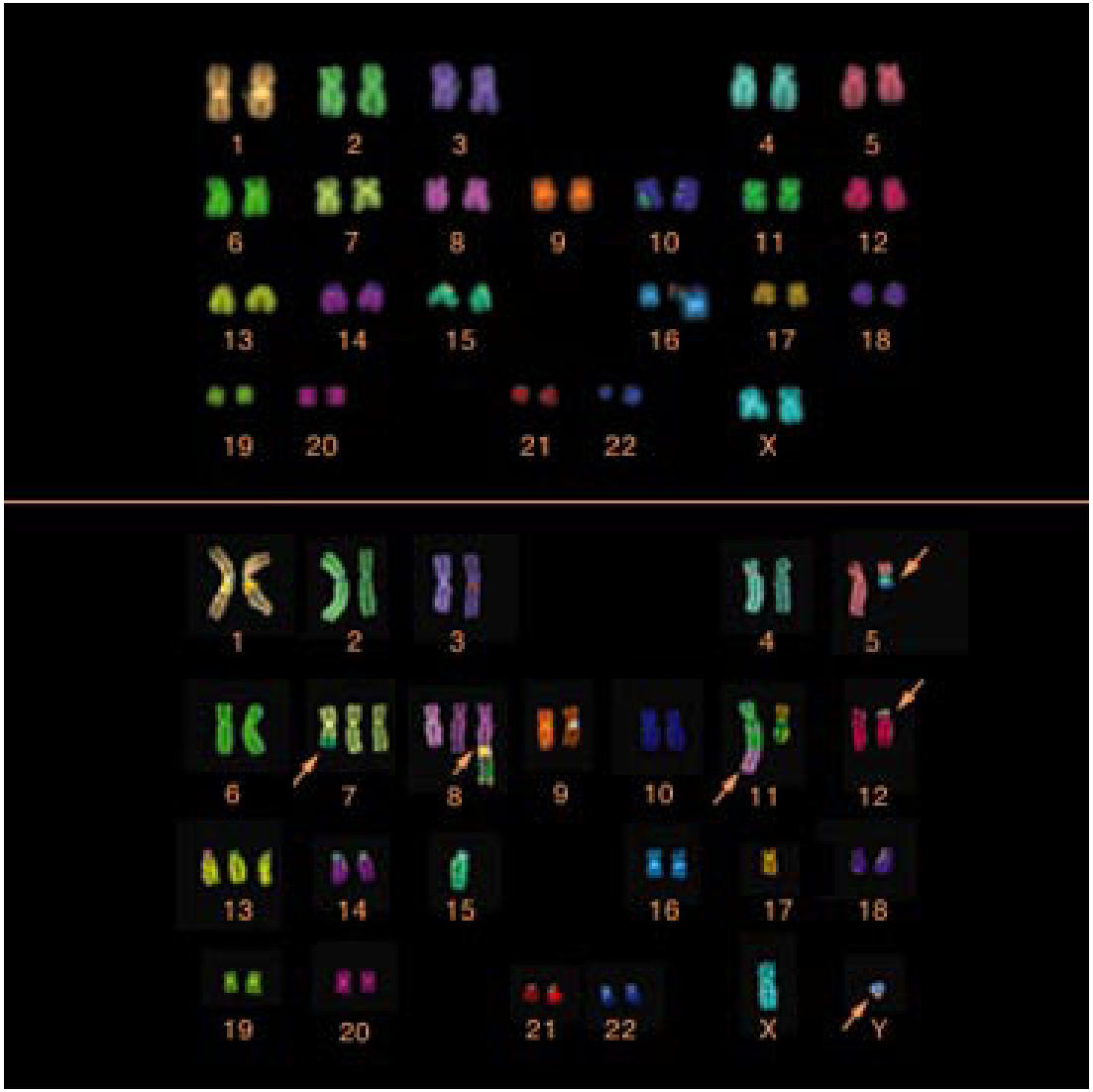
Application: predicted regulatory network (E. coli)



Outline

Chromosomic aberrations in cancer

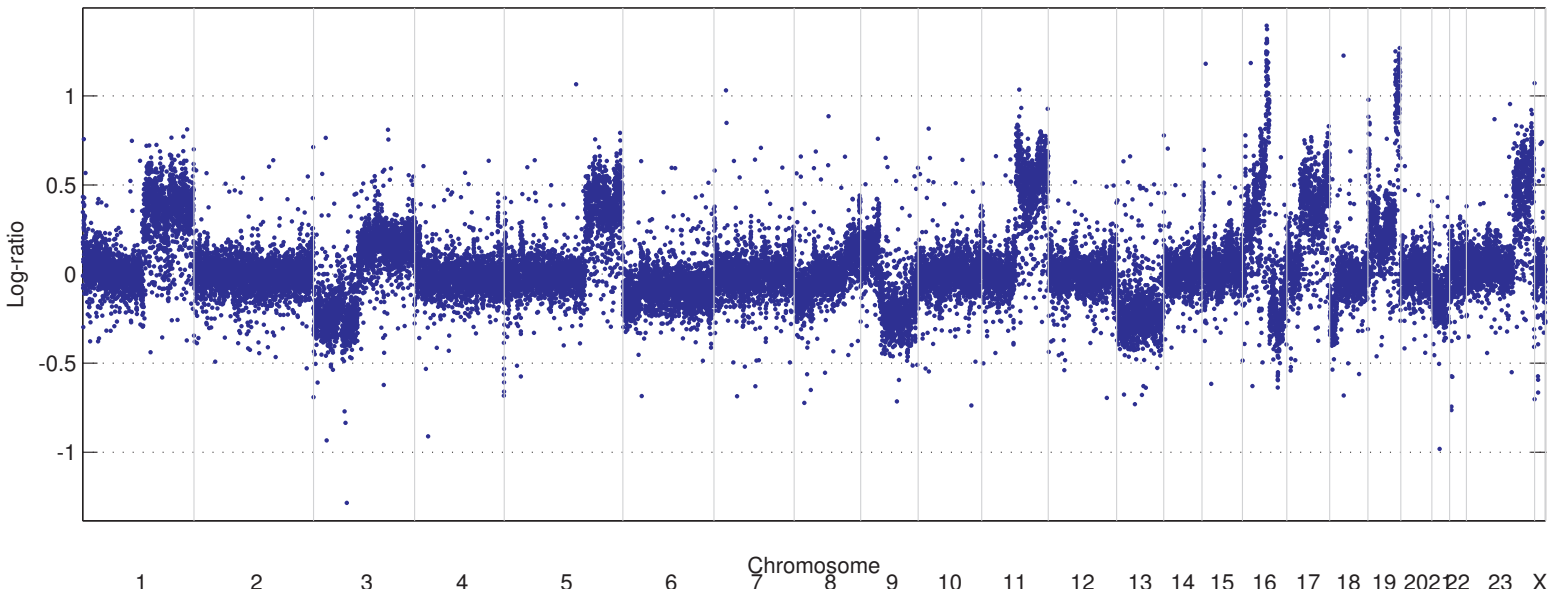




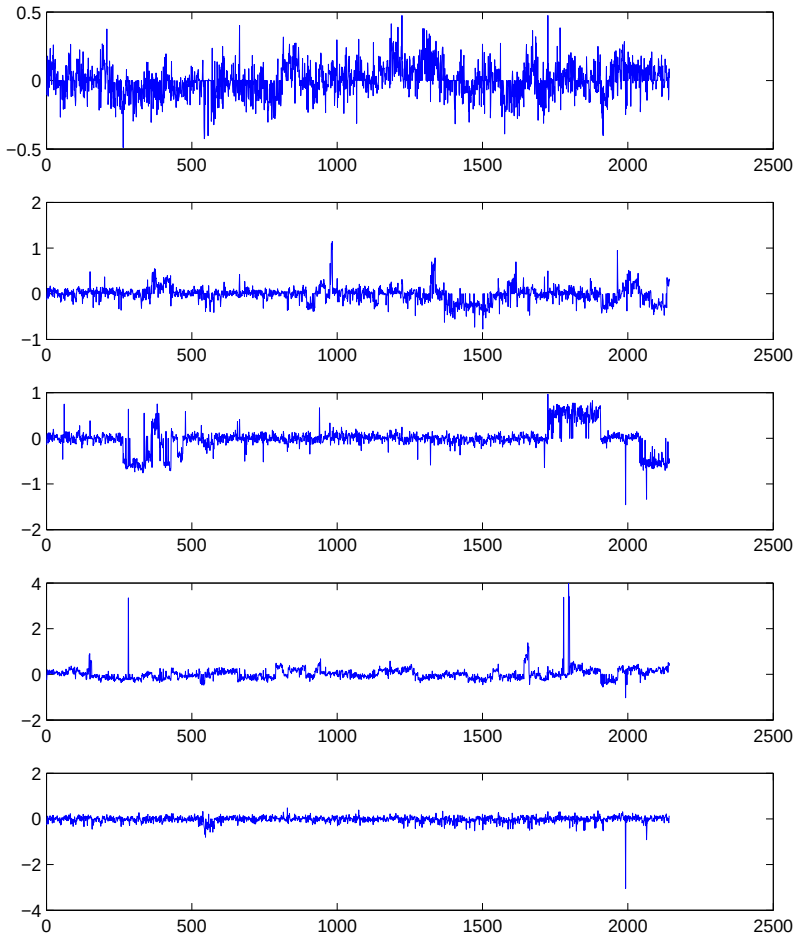
Comparative Genomic Hybridization (CGH)

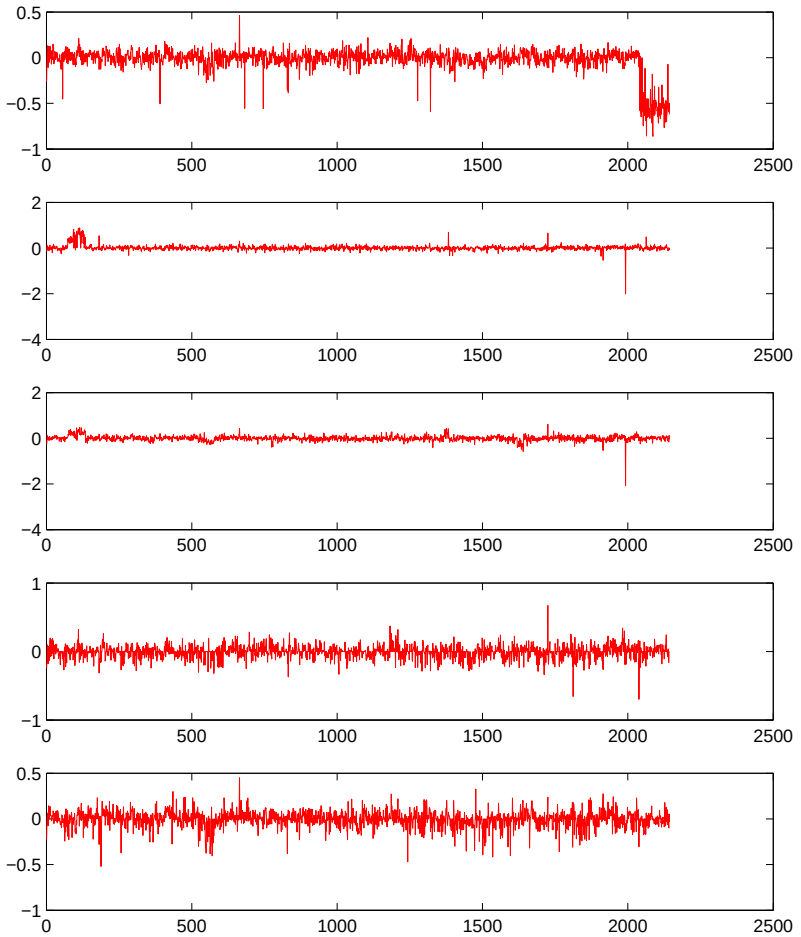
Motivation

- 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

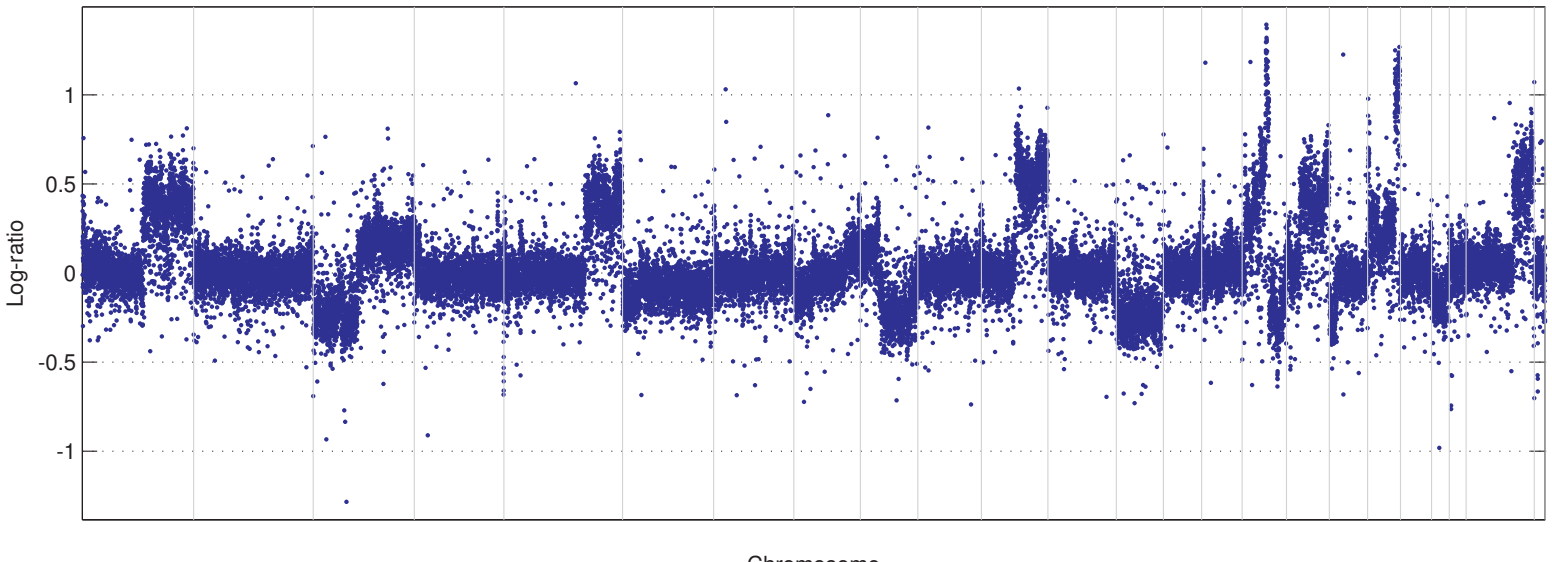
CGH array classification

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



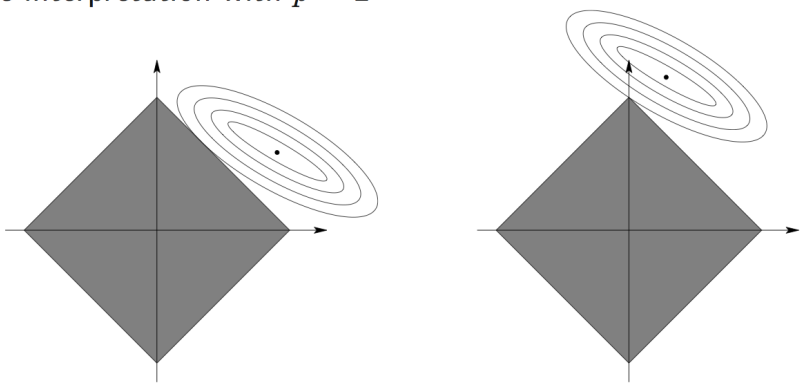
Fused lasso for supervised classification

- 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:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \max\left(1 - y_i \beta^\top x_i, 0\right) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \right\} .$$

- Computationnally: this is convex optimization problem that can be solved very efficiently

Geometric interpretation with $p = 2$



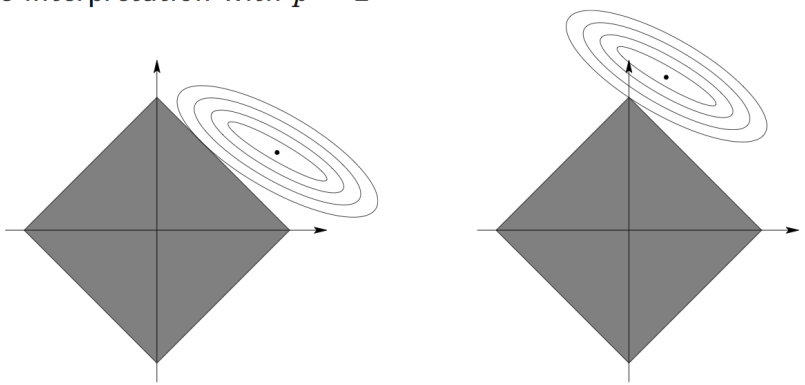
Fused lasso for supervised classification

- 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:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \max\left(1 - y_i \beta^\top x_i, 0\right) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \right\} .$$

- Computationnally: this is convex optimization problem that can be solved very efficiently

Geometric interpretation with $p = 2$



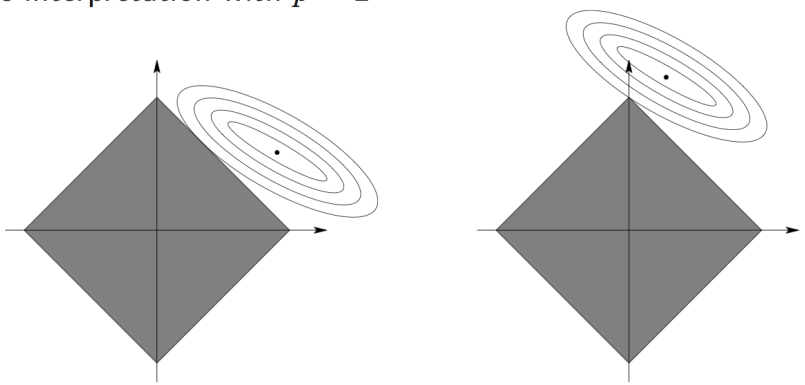
Fused lasso for supervised classification

- 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:

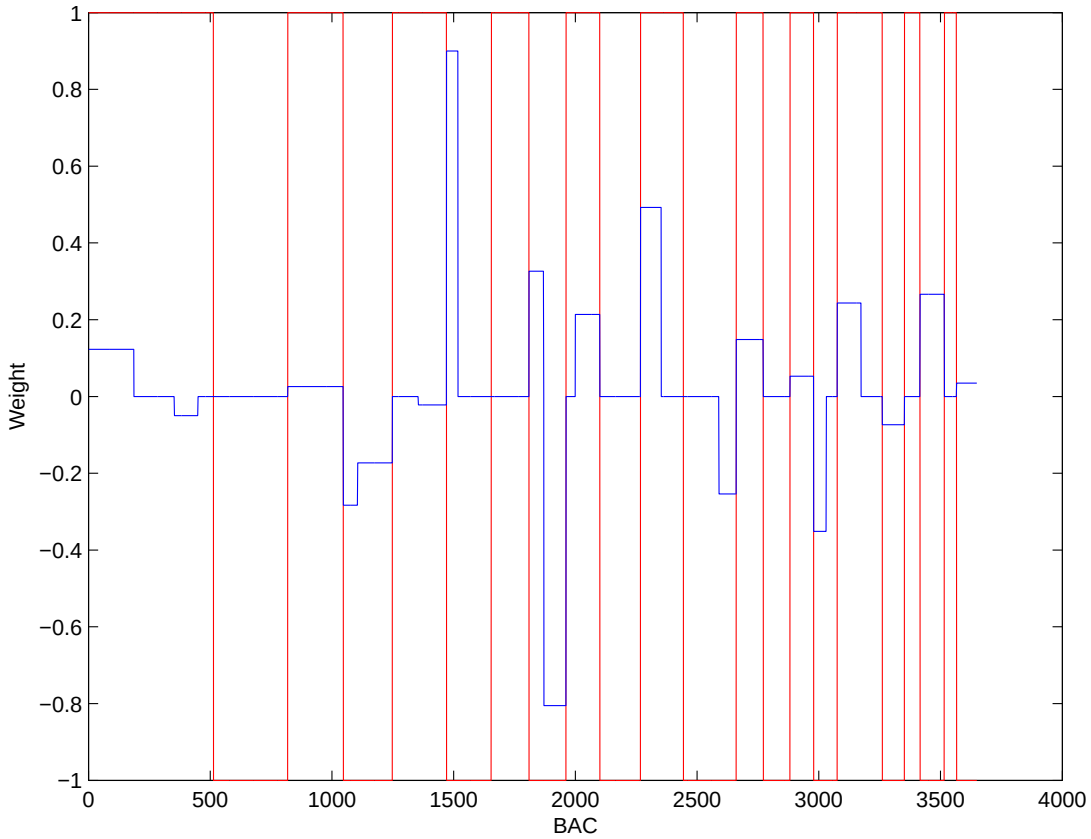
$$\min_{\beta \in \mathbb{R}^p} \left\{ \sum_{i=1}^{n} \max\left(1 - y_i \beta^\top x_i, 0\right) + \lambda_1 \sum_{i=1}^{p} |\beta_i| + \lambda_2 \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \right\} .$$

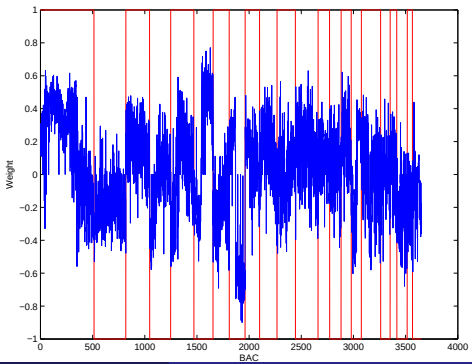
- Computationnally: this is convex optimization problem that can be solved very efficiently

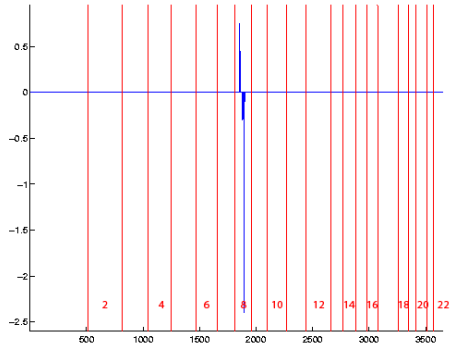
Geometric interpretation with $p = 2$



Example: predicting metastasis in melanoma

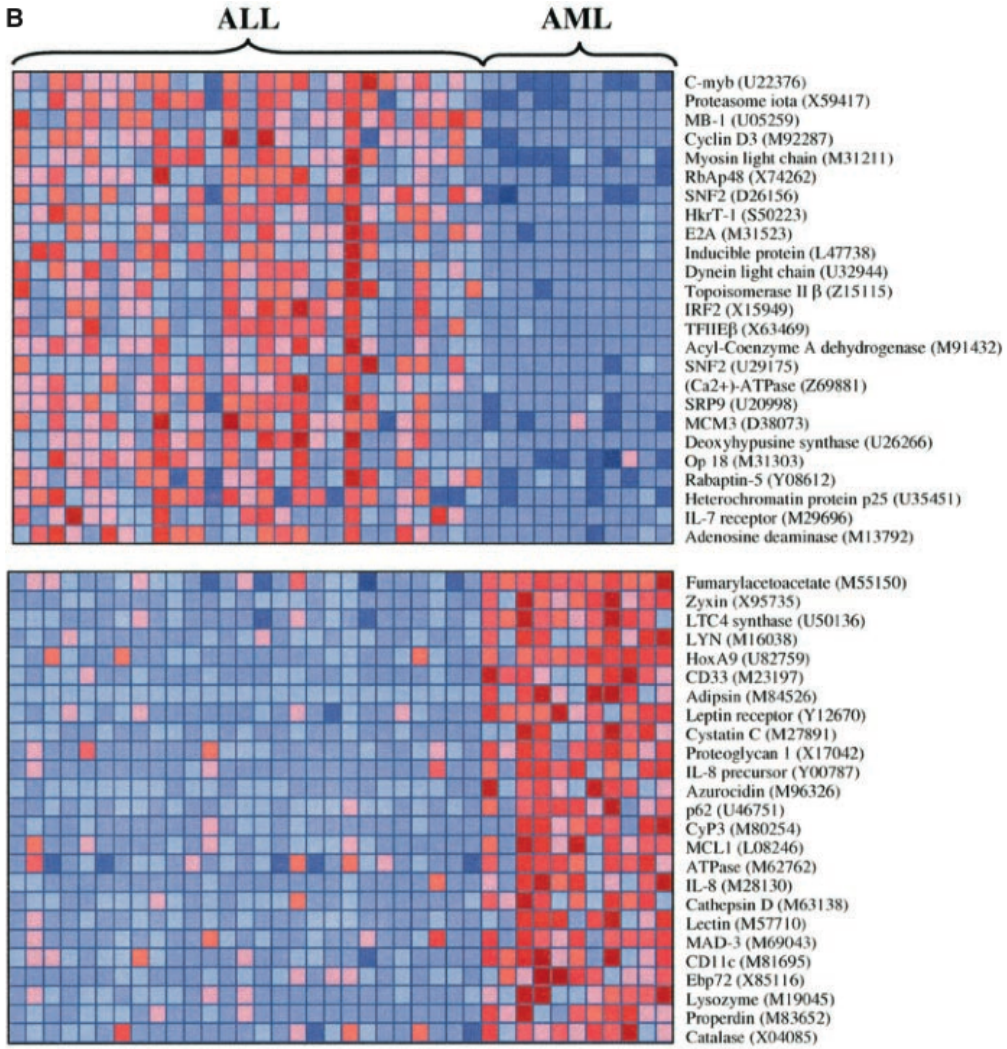




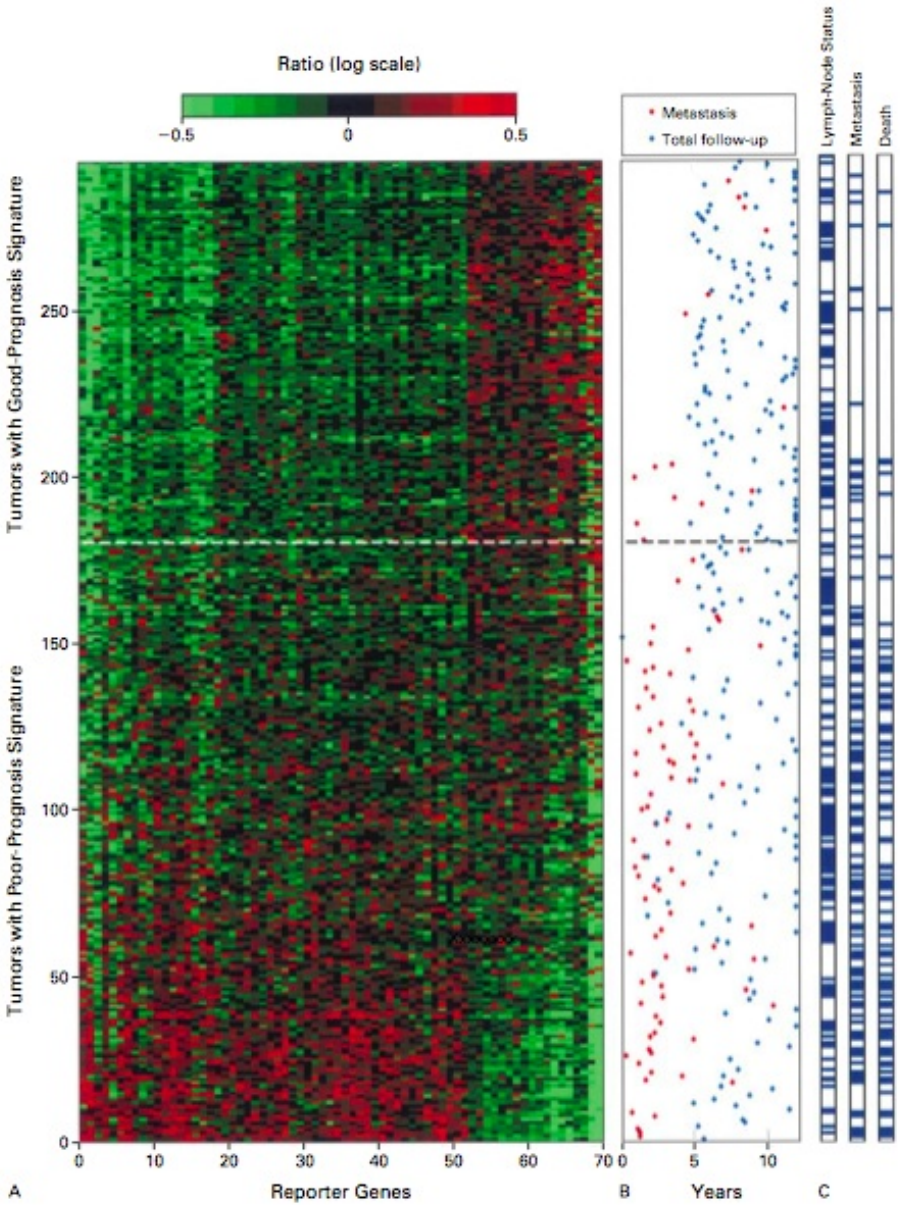


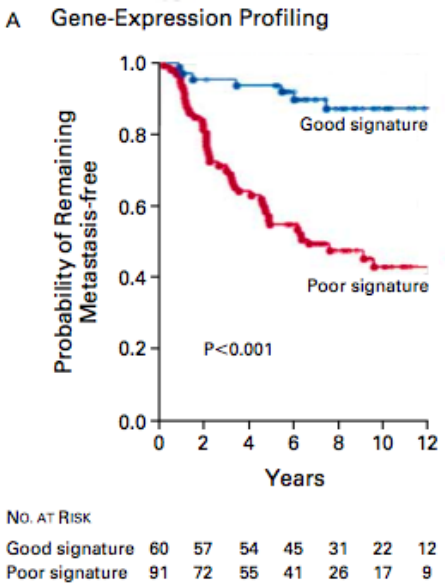
Outline

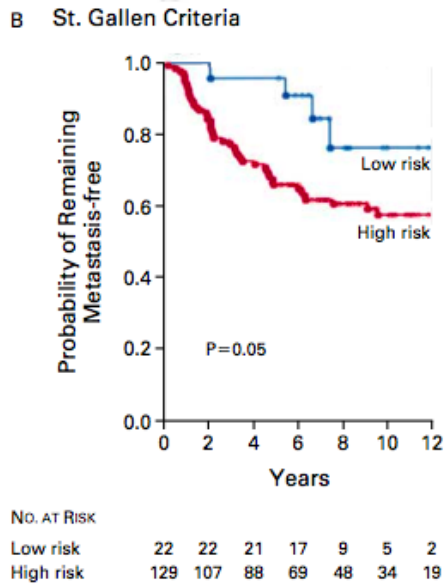
Diagnosis


B
ALL
AML
C-myb (U22376)
Proteasome iota (X59417)
MB-1 (U05259)
Cyclin D3 (M92287)
Myosin light chain (M31211)
RbAp48 (X74262)
SNF2 (D26156)
HkrT-1 (S50223)
E2A (M31523)
Inducible protein (L47738)
Dynein light chain (U32944)
Topoisomerase II β (Z15115)
IRF2 (X15949)
TFIIEβ (X63469)
Acyl-Coenzyme A dehydrogenase (M91432)
SNF2 (U29175)
(Ca2+)-ATPase (Z69881)
SRP9 (U20998)
MCM3 (D38073)
Deoxyhypusine synthase (U26266)
Op 18 (M31303)
Rabaptin-5 (Y08612)
Heterochromatin protein p25 (U35451)
IL-7 receptor (M29696)
Adenosine deaminase (M13792)
Fumarylacetoacetate (M55150)
Zyxin (X95735)
LTC4 synthase (U50136)
LYN (M16038)
HoxA9 (U82759)
CD33 (M23197)
Adipsin (M84526)
Leptin receptor (Y12670)
Cystatin C (M27891)
Proteoglycan 1 (X17042)
IL-8 precursor (Y00787)
Azurocidin (M96326)
p62 (U46751)
CyP3 (M80254)
MCL1 (L08246)
ATPase (M62762)
IL-8 (M28130)
Cathepsin D (M63138)
Lectin (M57710)
MAD-3 (M69043)
CD11c (M81695)
Ebp72 (X85116)
Lysozyme (M19045)
Properdin (M83652)
Catalase (X04085)

Prognosis


Ratio (log scale)
−0.5
0
0.5
Metastasis
Total follow-up
Lymph-Node Status
Metastasis
Death
Tumors with Good Prognosis Signature
Tumors with Poor Prognosis Signature
0
50
100
150
200
250
0
10
20
30
40
50
60
70
Reporter Genes
A
0
5
10
Years
B
C

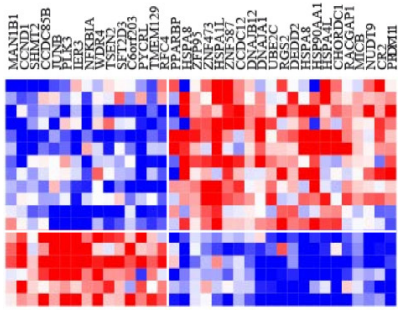

A Gene-Expression Profiling
Probability of Remaining Metastasis-free
1.0
0.8
0.6
0.4
0.2
0.0
Good signature
Poor signature
P<0.001
0
2
4
6
8
10
12
Years
No. at Risk
Good signature 60 57 54 45 31 22 12
Poor signature 91 72 55 41 26 17 9


B St. Gallen Criteria
Probability of Remaining Metastasis-free
1.0
0.8
0.6
0.4
0.2
0.0
Low risk
High risk
P=0.05
0
2
4
6
8
10
12
Years
No. at Risk
Low risk 22 22 21 17 9 5 2
High risk 129 107 88 69 48 34 19

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



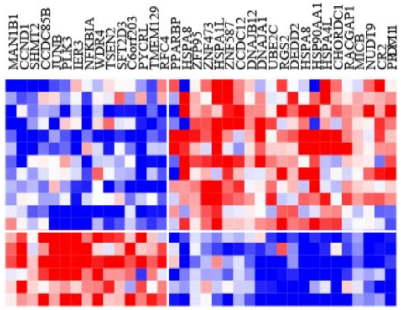
But:

- 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 selection, molecular signature

The idea

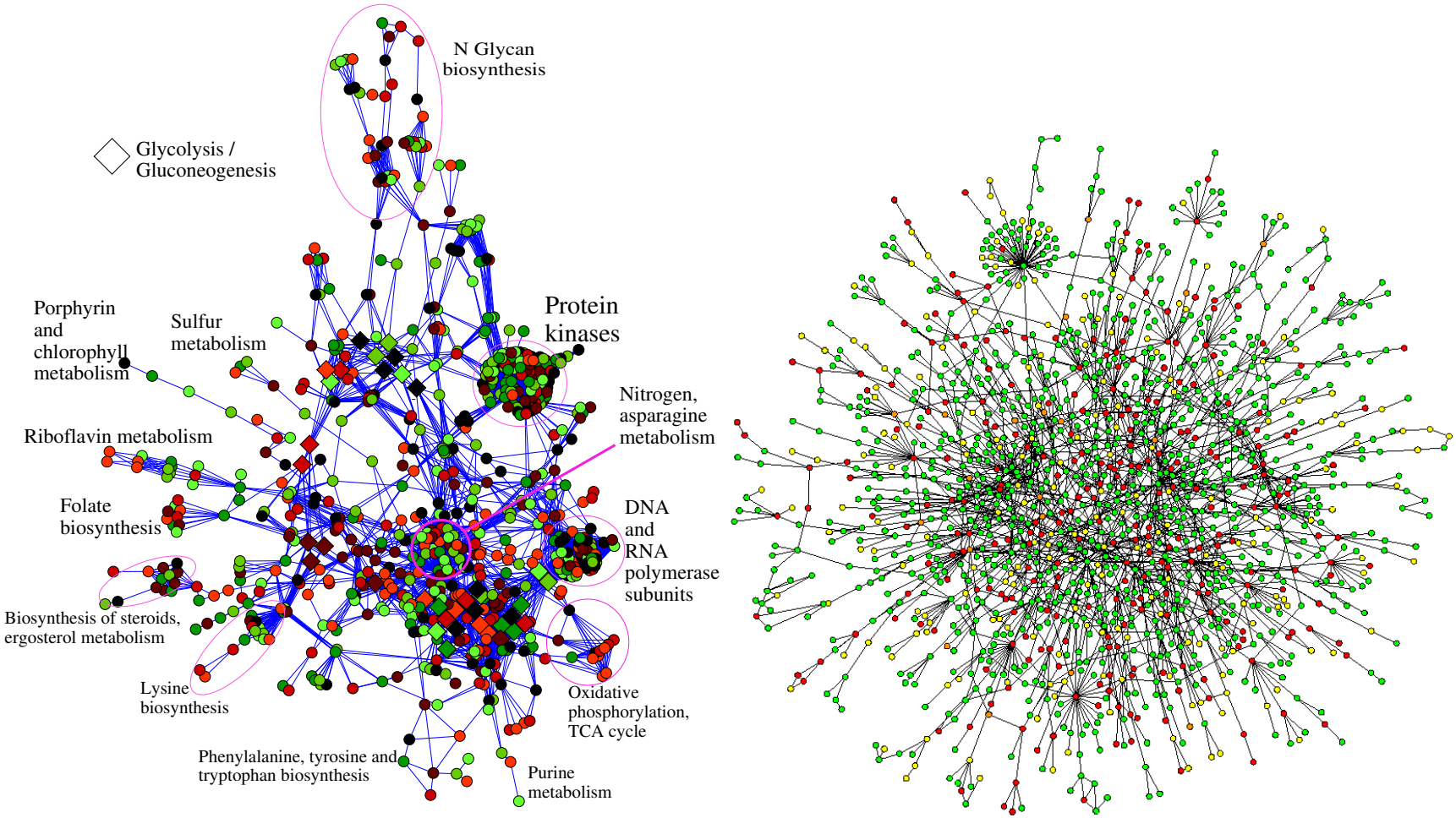
- We look for a limited set of genes that are sufficient for prediction.
- Selected genes should inform us about the underlying biology



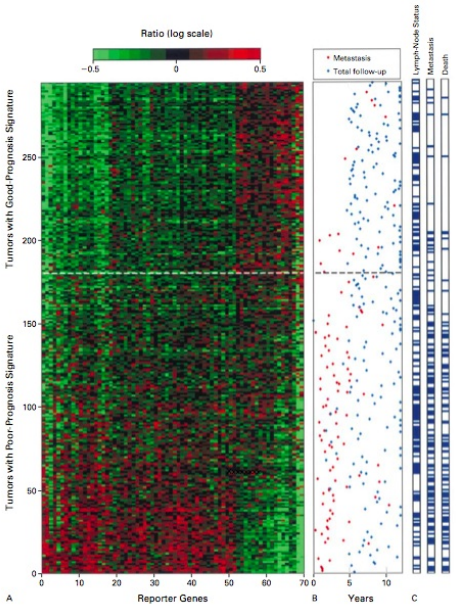
But:

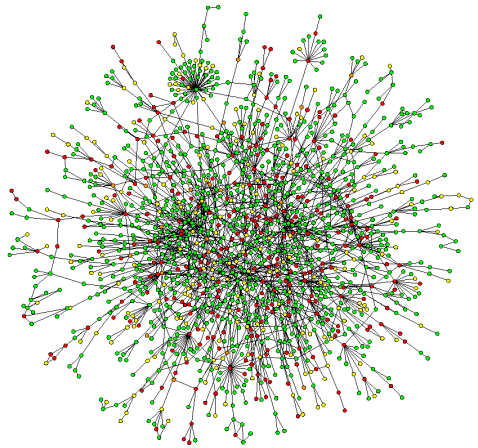
- We often observe little stability in the genes selected...
- Is gene selection the most biologically relevant hypothesis?
- What about thinking instead of "pathways" or "modules" signatures?

Gene networks



Idea 1: graph-based SVM





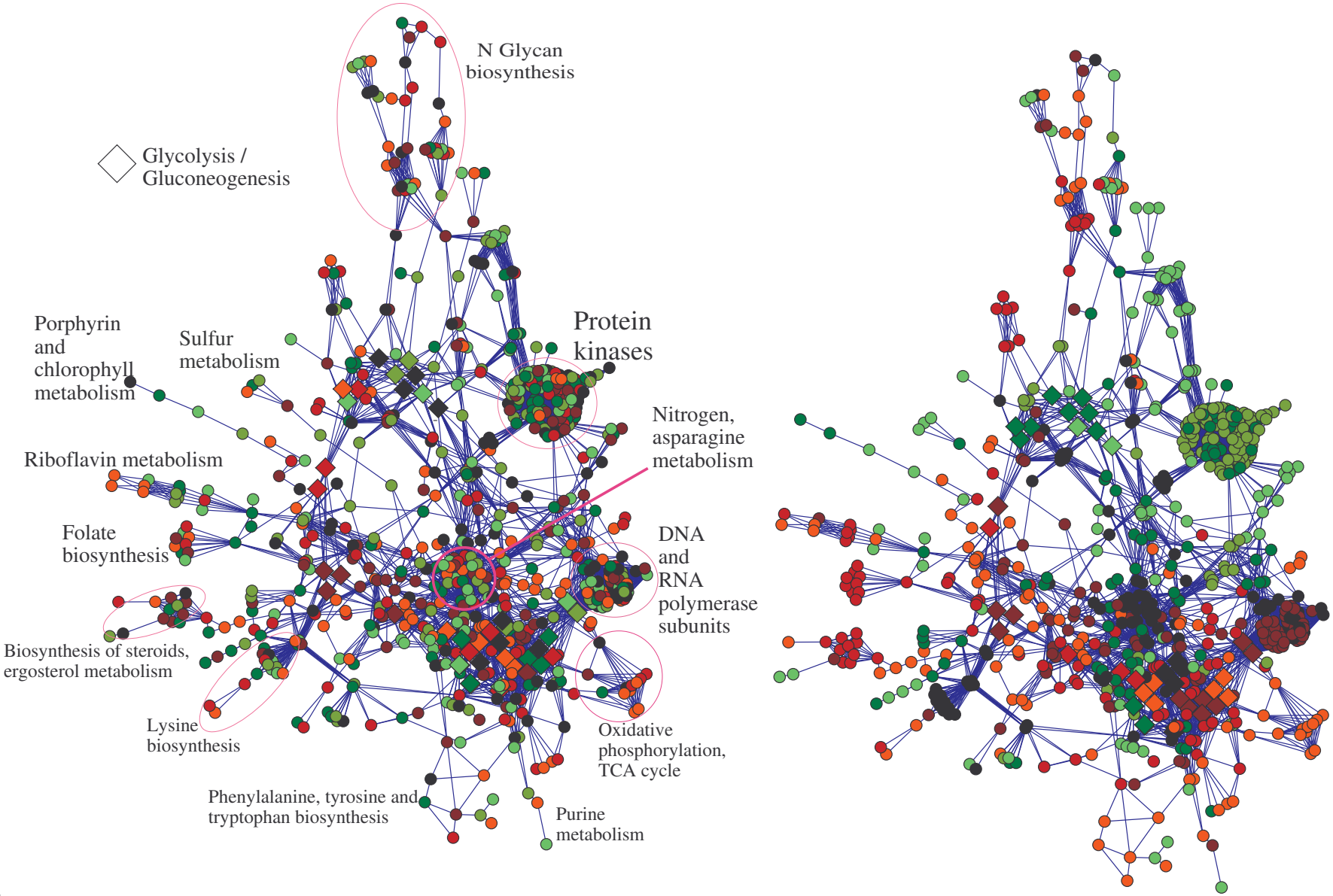
Hypothesis 1

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

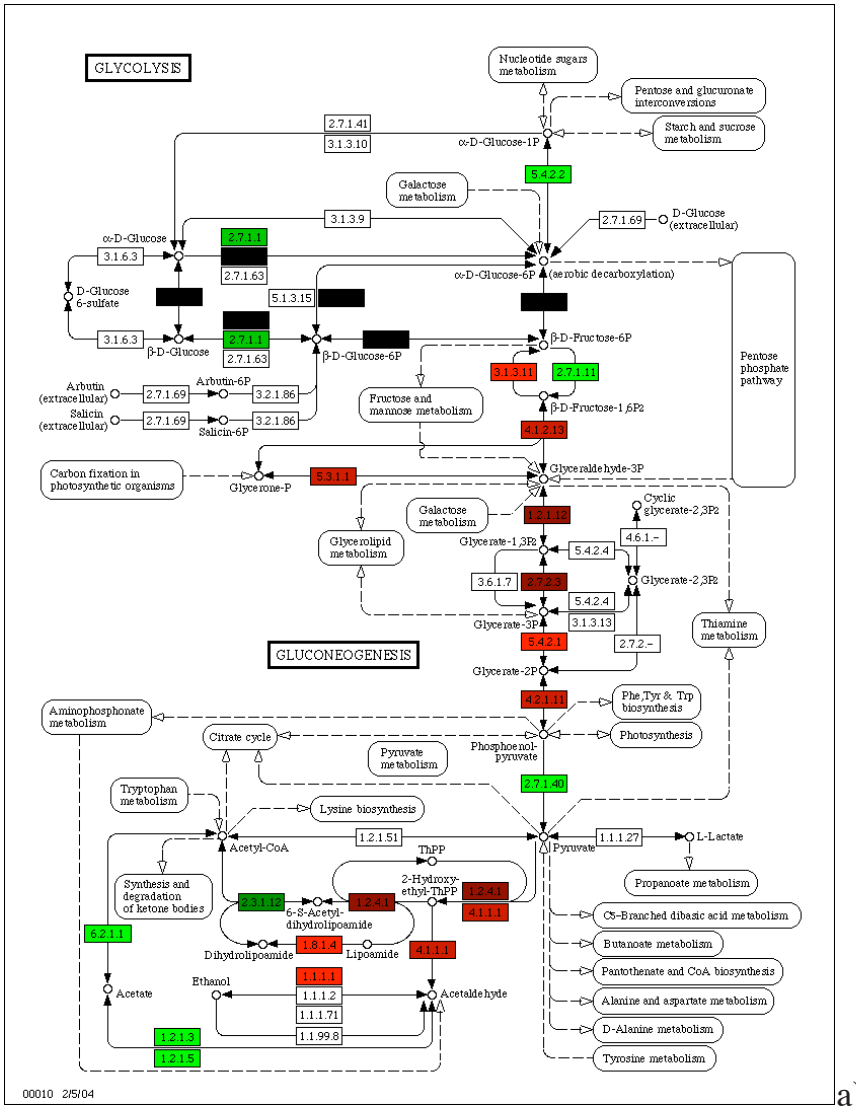
$$\min_{\beta} R(\beta) + \lambda \sum_{i \sim j} (\beta_i - \beta_j)^2$$

Rapaport et al. (2007)

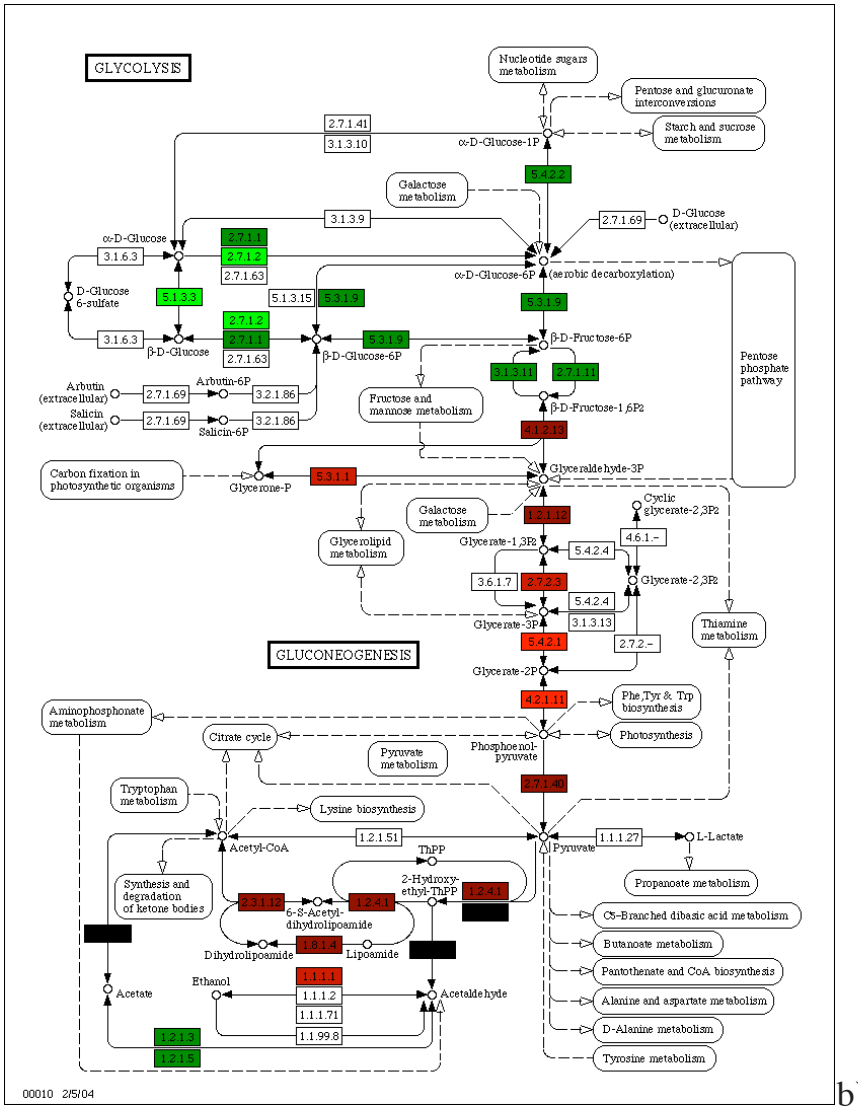
Classifiers



Classifiers



a)



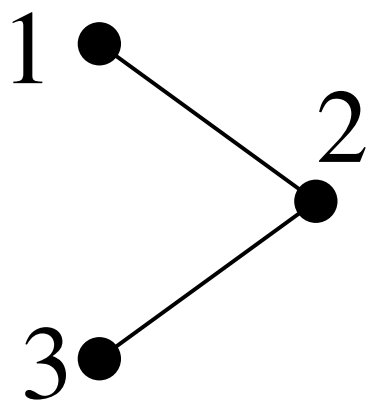
b)

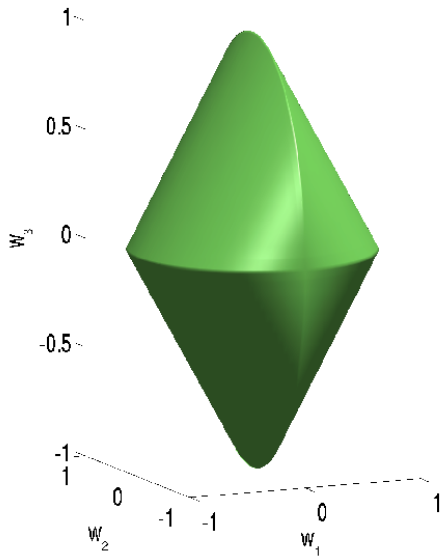
Idea 2: graph-based lasso

Hypothesis 2

Selecte genes which tend to be connected on the graph

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





Jacob et al. (2009)

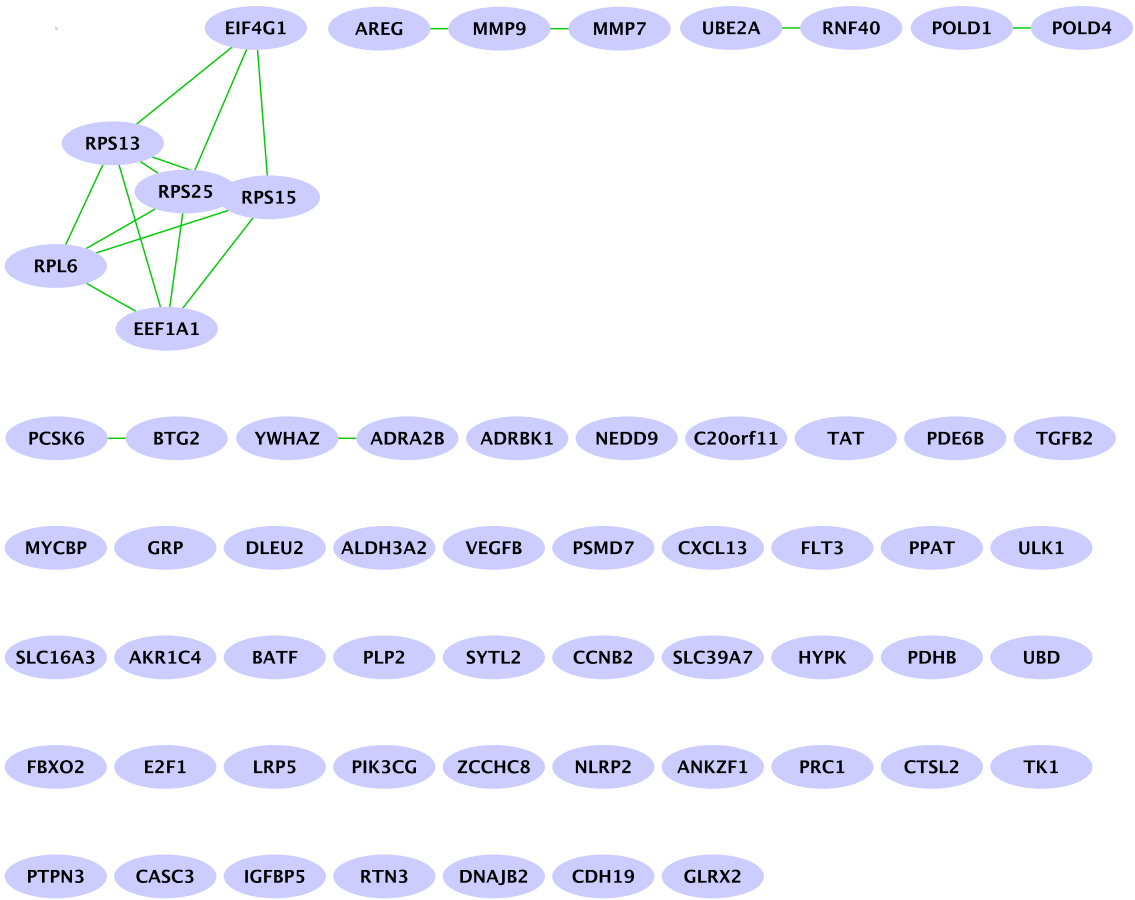
Preliminary results

Breast cancer data

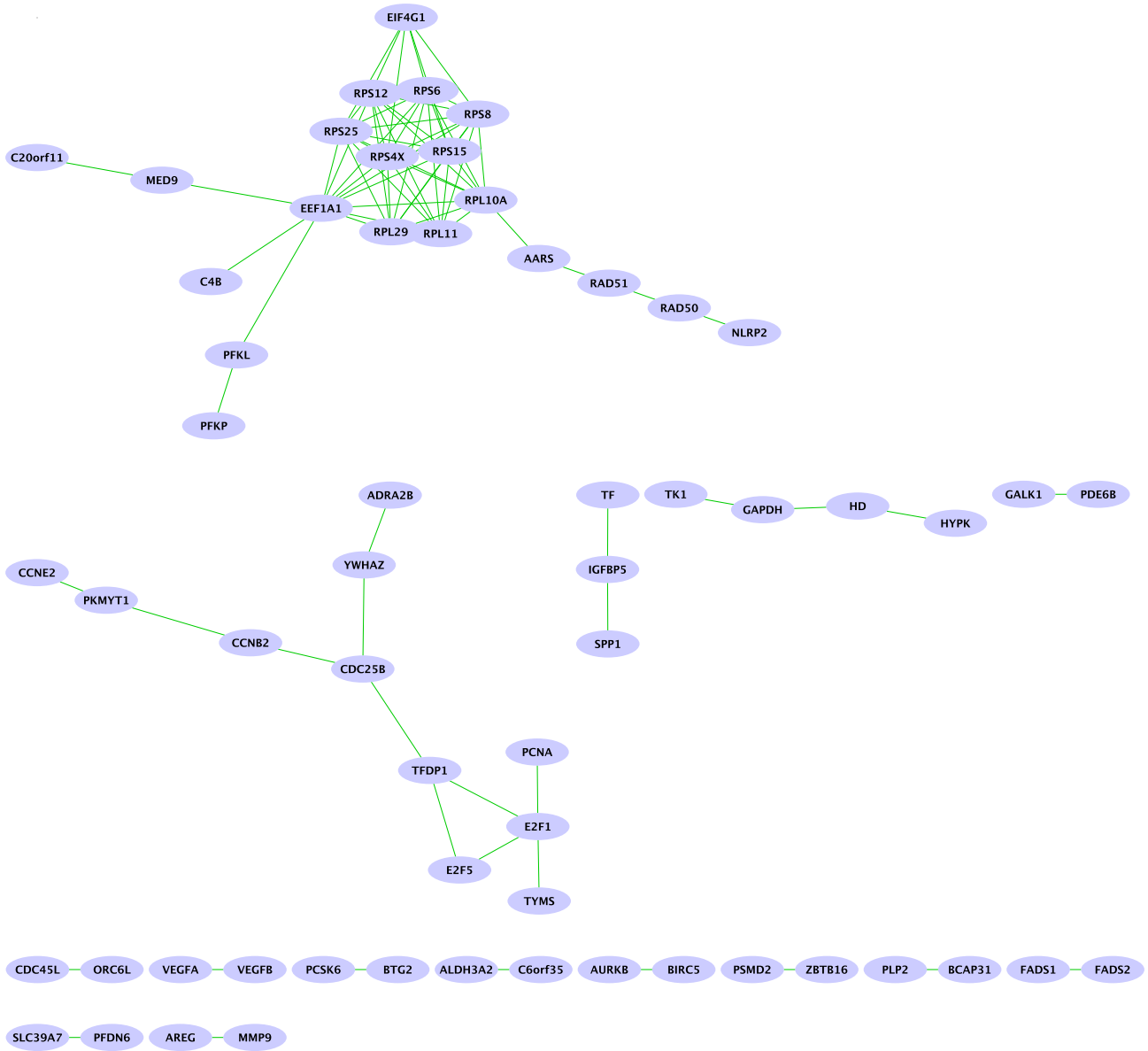
- Gene expression data for $8,141$ genes in 295 breast cancer tumors.
- Performance

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

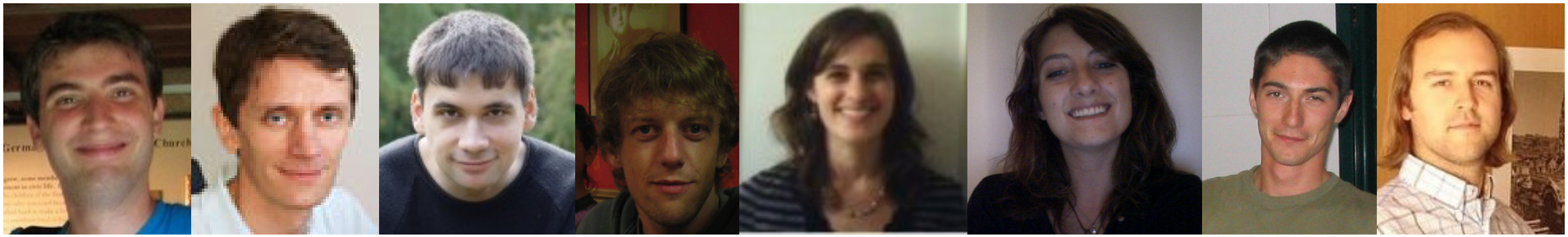


Outline

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
- Requires interdisciplinary collaborations to incorporate expert knowledge at the heart of learning algorithms
- Many other applications not covered in this presentation!

Acknowledgements!



Franck Rapaport (MSKCC), Emmanuel Barillot, Andrei Zynoviev, Kevin Bleakley (INRIA), Fantine Mordelet, Anne-Claire Haury, 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 AFFAIRS ETRANGERES

European Research Council
erc

AGENCE NATIONALE DE LA RECHERCHE
ANR

MEDICEN
PARIS REGION

Cancer-Bio-Santé
PÔLE DE COMPÉTITIVITÉ