Machine learning and feature selection in bioinformatics

Jean-Philippe Vert

Jean-Philippe.Vert@mines.org

Mines ParisTech / Curie Institute / Inserm

Machine Learning for Neuroimaging workshop,
Marseille, Nov 8-9, 2011.

Outline

1. Inference of gene regulatory networks

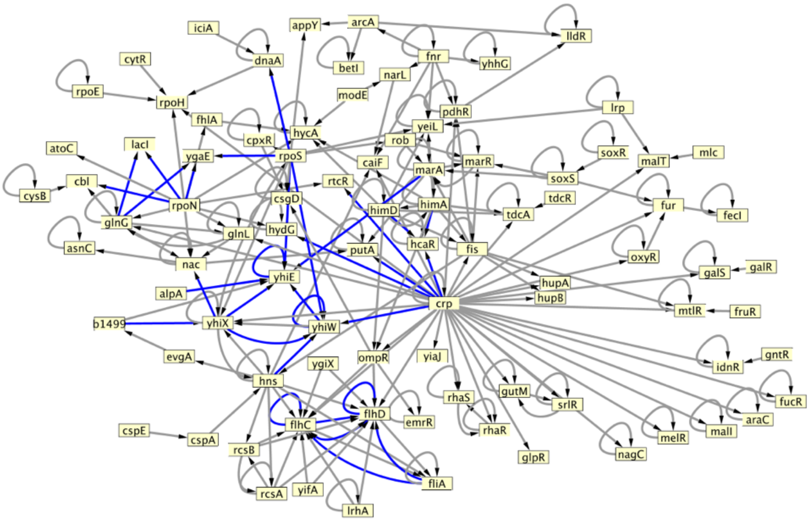
2. Diagnosis and prognosis from gene expression data

Outline

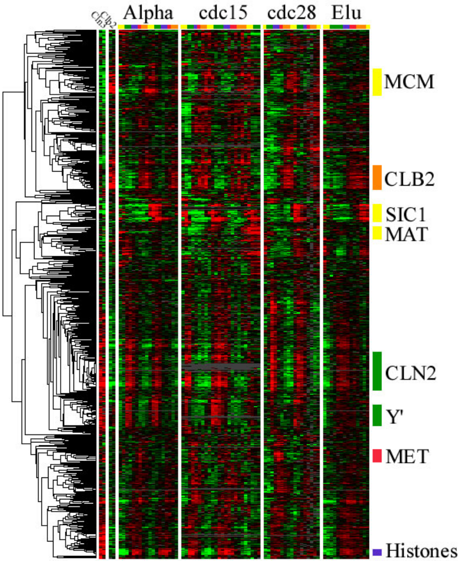
1. Inference of gene regulatory networks

2. Diagnosis and prognosis from gene expression data

Gene regulatory network (GRN) of E. coli

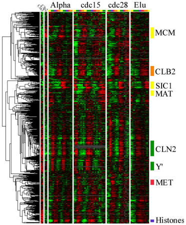


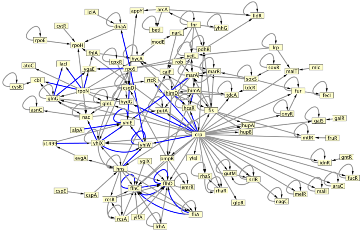
Gene expression data



GRN inference (*de novo*)

Given a set of gene expressions, infer the regulations.



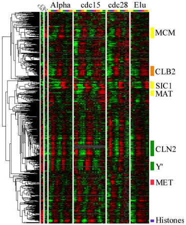


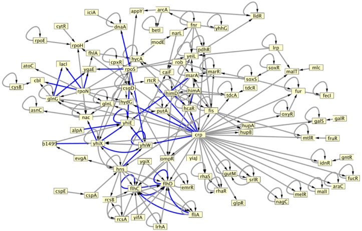
How?

- Model-based (dynamic systems)
- (Dynamic) Bayesian networks
- Similarity-based
- Feature selection

GRN inference (*de novo*)

Given a set of gene expressions, infer the regulations.

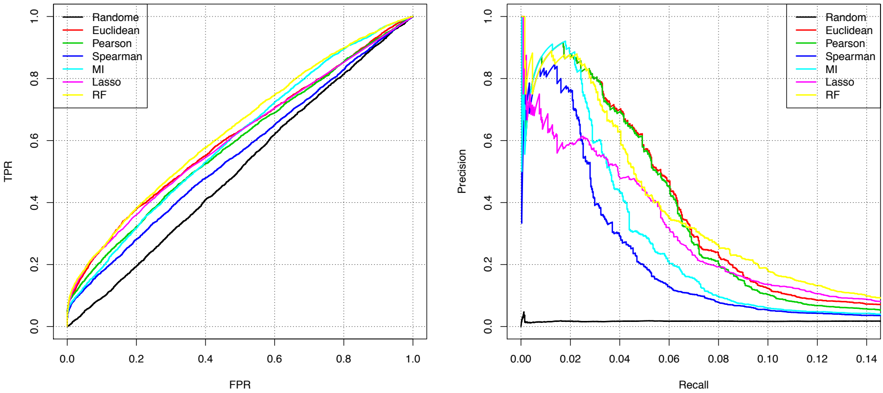




How?

- Model-based (dynamic systems)
- (Dynamic) Bayesian networks
- Similarity-based
- Feature selection

Evaluation (DREAM challenge)

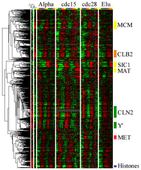


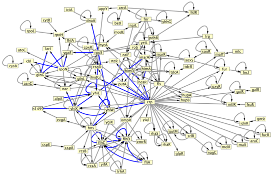
- Best results obtained by feature selection methods
- Bootstrap-based methods (RF, stability selection)
- Overall performance very disappointing (difficult problem...)

Supervised inference

The problem

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





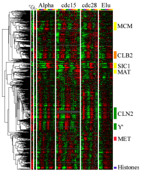
How?

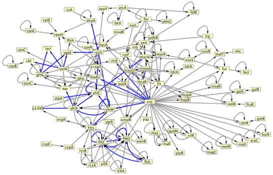
- Local models: for each TF, learn to discriminate the regulated vs non-regulated genes
- Global models: learn to discriminate connected vs non-connected TF-target pairs

Supervised inference

The problem

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



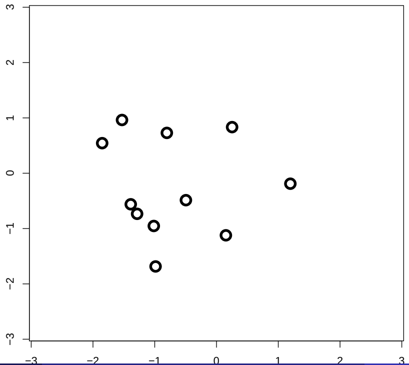


How?

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

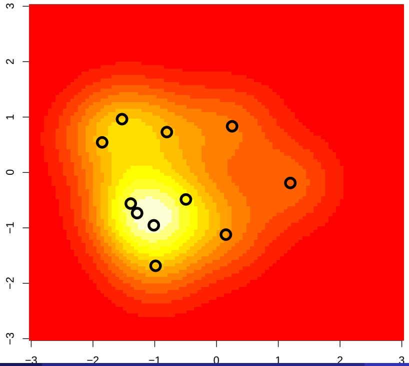
Example: one-class learning approach for local model

- 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



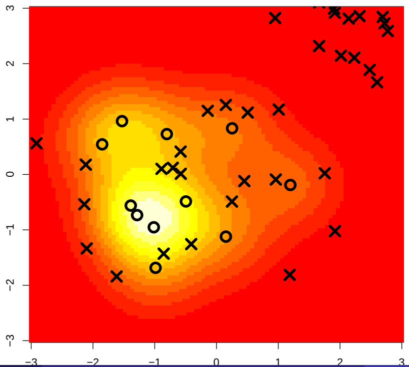
Example: one-class learning approach for local model

- 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

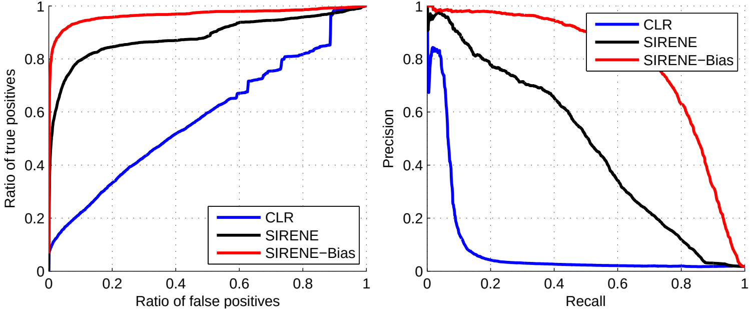


Example: one-class learning approach for local model

- 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

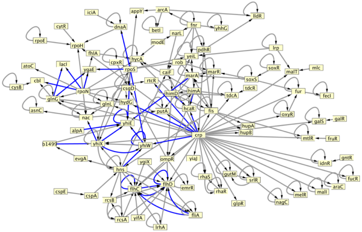


Method	Recall at 60%	Recall at 80%
SIRENE	**44.5%**	**17.6%**
CLR	7.5%	5.5%
Relevance networks	4.7%	3.3%
ARACNe	1%	0%
Bayesian network	1%	0%

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

Lessons learned

- Many ways to formalize the GRN inference problem (structure learning)
- De novo inference is best solved by feature selection
- Supervised inference better when the structure is partially known
- Simple local models outperform structured output learning
- Performance remains low. Still an open problem!

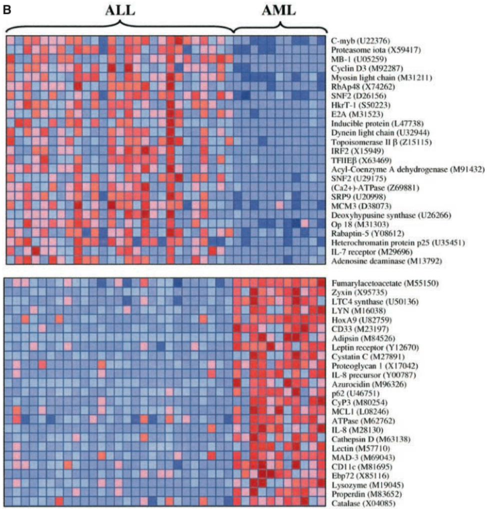


Outline

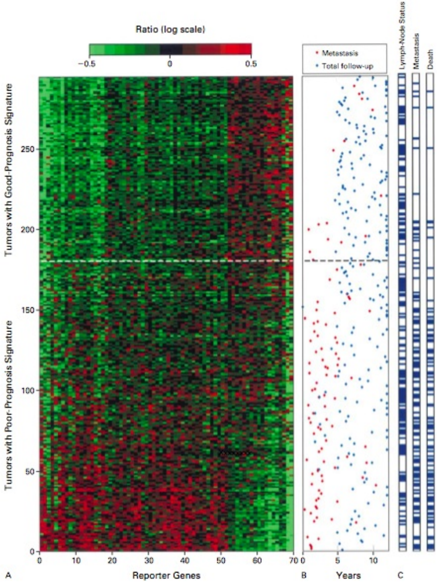
1. Inference of gene regulatory networks

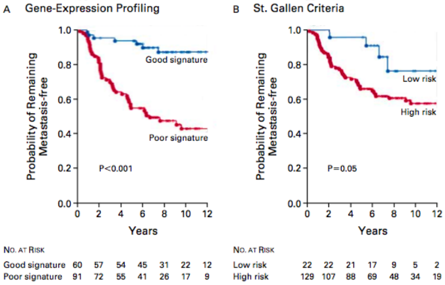
2. Diagnosis and prognosis from gene expression data

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





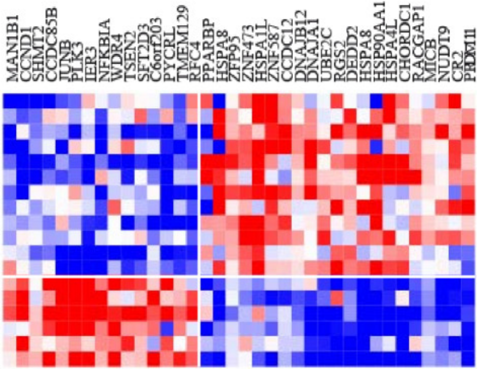
No. at Risk							
Good signature	60	57	54	45	31	22	12
Poor signature	91	72	55	41	26	17	9

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

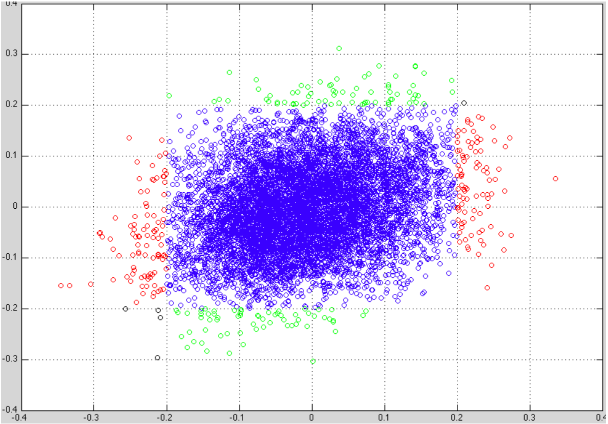
The idea

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

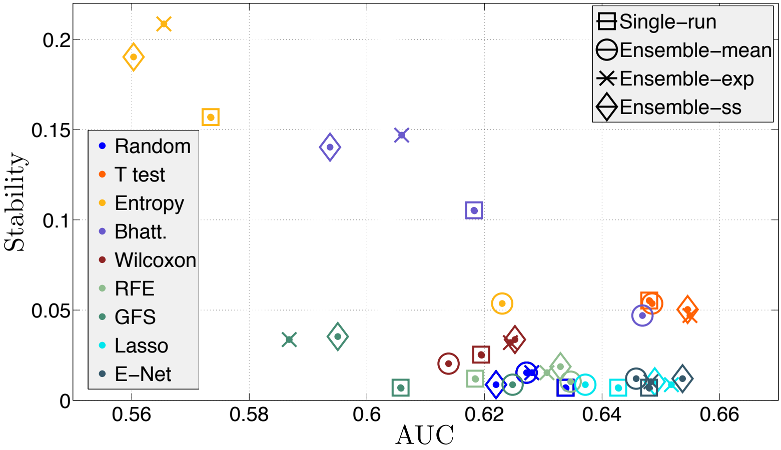


But... unstability of molecular signatures

- Wang dataset: $n = 286$, $p = 8141$
- Pearson correlation with the output on 2 random subsamples of 143 samples:

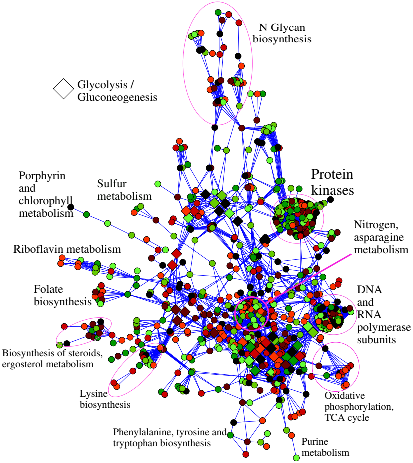


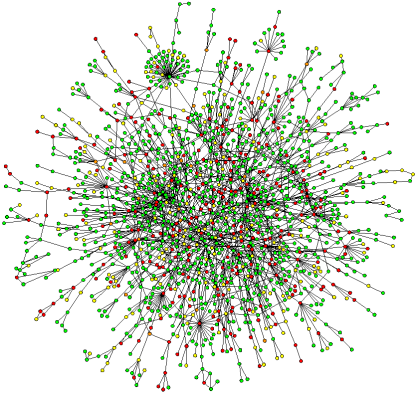
Comparison of feature selection methods...



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

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

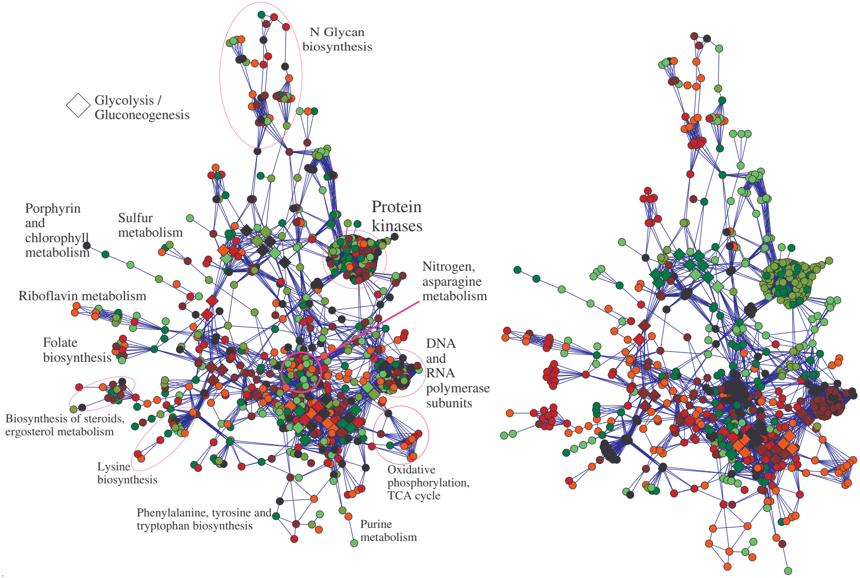
- Gene selection + smooth on the graph

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

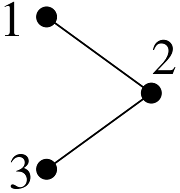
- Gene selection + Piecewise constant on the graph

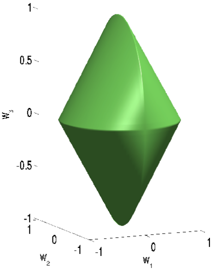
$$\Omega(\beta) = \sum_{i\sim j} \left| \beta_i - \beta_j \right| + \sum_{i=1}^{p} \mid \beta_i \mid$$

Illustration


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

Hypotheses 2: select genes which are connected of belong to the same functional groups





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

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

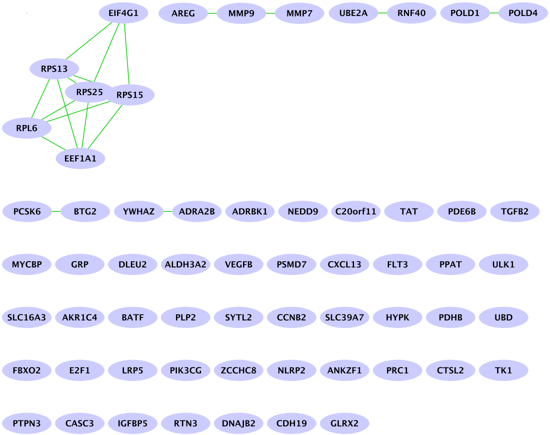
- Gene expression data for $8,141$ genes in 295 breast cancer tumors.
- Canonical pathways from MSigDB containing 639 groups of genes, 637 of which involve genes from our study.

Method	ℓ_1	$\Omega^{\mathcal{G}}_{\text{OVERLAP}}(.)$
Error	0.38 ± 0.04	0.36 ± 0.03
Mean ♯ path.	130	30

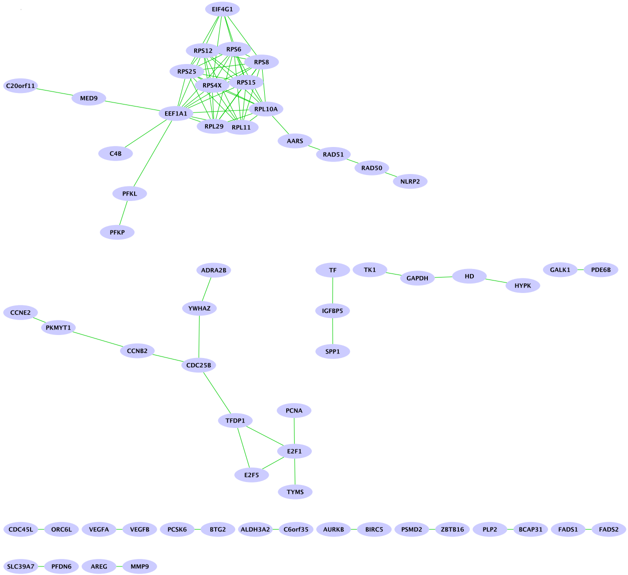
- Graph on the genes.

Method	ℓ_1	$\Omega_{graph}(.)$
Error	0.39 ± 0.04	0.36 ± 0.01
Av. size c.c.	1.03	1.30

Classical lasso signature



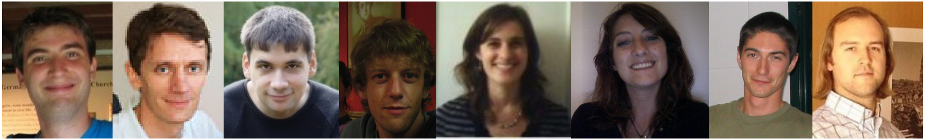
Graph Lasso signature



Discussion

- Very challenging problems: high dimensions, few samples, complex problems (supervised classification, structure inference)
- Methods that "work" in practice find the best trade-off between model complexity ("bias") and ability to learn from data ("variance")
- Methods that work in theory and on toy examples do not always work on real data (and vice-versa)...
- Shrinkage methods for structured sparsity is promising...
- ... but difficult to reconcile accuracy and interpretation
- Stability may be a useful empirical proxy to assess the trust we can have in selected features

Acknowledgements



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