Machine learning with prior knowledge for genomic data

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

Mines ParisTech / Curie Institute / Inserm

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

<table>
<thead>
<tr>
<th></th>
<th>ALL</th>
<th>AML</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>C-myb (U22376)</td>
<td>Fumarylacetoacetate (M55150)</td>
</tr>
<tr>
<td></td>
<td>Proteasome iota (X59417)</td>
<td>Zyxin (X95735)</td>
</tr>
<tr>
<td></td>
<td>MB-1 (U035259)</td>
<td>LT54-synthase (U50136)</td>
</tr>
<tr>
<td></td>
<td>Cyclin D3 (N99287)</td>
<td>LYT (M16038)</td>
</tr>
<tr>
<td></td>
<td>Myosin light chain (M31211)</td>
<td>HoxA9 (M382759)</td>
</tr>
<tr>
<td></td>
<td>RhoA (X74262)</td>
<td>CD33 (M23197)</td>
</tr>
<tr>
<td></td>
<td>SNF2 (D26156)</td>
<td>Adipin (M98526)</td>
</tr>
<tr>
<td></td>
<td>HNF1-1 (S50223)</td>
<td>Leptin receptor (Y12670)</td>
</tr>
<tr>
<td></td>
<td>EZA (M31523)</td>
<td>Cystatin C (M278911)</td>
</tr>
<tr>
<td></td>
<td>Inducible protein (L47738)</td>
<td>Prostacyclin synthase (X17042)</td>
</tr>
<tr>
<td></td>
<td>Dynamin light chain (U32944)</td>
<td>IL-8 precursor (Y00287)</td>
</tr>
<tr>
<td></td>
<td>Topoisomerase II B (Z15115)</td>
<td>Azurocidin (M90326)</td>
</tr>
<tr>
<td></td>
<td>(RF2 (X15949)</td>
<td>α2s (U460751)</td>
</tr>
<tr>
<td></td>
<td>TFIIE (X63469)</td>
<td>CyP73 (M400254)</td>
</tr>
<tr>
<td></td>
<td>Acyl-Coenzyme A dehydrogenase (M91432)</td>
<td>MCL1 (L08266)</td>
</tr>
<tr>
<td></td>
<td>SNF2 (U29175)</td>
<td>ATPase (M62762)</td>
</tr>
<tr>
<td></td>
<td>(Ca2+-)ATPase (Z99881)</td>
<td>IL-8 (M28130)</td>
</tr>
<tr>
<td></td>
<td>SRP9 (U209906)</td>
<td>Cathepsin D (M63138)</td>
</tr>
<tr>
<td></td>
<td>MCM3 (D38073)</td>
<td>Lectin (M55771)</td>
</tr>
<tr>
<td></td>
<td>Deoxyhypusine synthase (U26266)</td>
<td>MAD-3 (M69043)</td>
</tr>
<tr>
<td></td>
<td>Op-15 (M31103)</td>
<td>CD11c (M81695)</td>
</tr>
<tr>
<td></td>
<td>Rab9p-5 (U08612)</td>
<td>Ebp72 (X85116)</td>
</tr>
<tr>
<td></td>
<td>Histone acetylase protein p25 (U35451)</td>
<td>Lysozyme (M19045)</td>
</tr>
<tr>
<td></td>
<td>IL-7 receptor (M59696a)</td>
<td>Propolin (M85582)</td>
</tr>
<tr>
<td></td>
<td>Adenosine deaminase (M13792)</td>
<td>Catalase (X04085)</td>
</tr>
</tbody>
</table>
Virtual screening for drug discovery

Given a **training set** of labeled data with...

*learn* a discrimination rule...

... in order to *predict* the label of new data
Given a training set of labeled data with...

learn a discrimination rule...

... in order to predict the label of new data
Given a **training set** of labeled data with...

2. **learn** a discrimination rule...

3. ... in order to **predict** the label of new data
Given a training set of labeled data with...

learn a discrimination rule...

... in order to predict the label of new data
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 regulation

- Gene regulatory sequences
- Spacer DNA
- Gene regulatory proteins
- General transcription factors
- RNA polymerase
- TATA box
- Upstream
- Promoter
- Start of transcription
Gene regulatory network

Protein 3

mRNA 3

Gene 3

Protein 2

mRNA 2

Gene 2

protein complex

Gene 1

Protein 1

mRNA 1
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...)
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 $\beta$ is sparse, i.e., only a few $\beta_i$ are non-zero.

We can estimate the sparse regression model from a matrix of expression data.

Non-zero $\beta_i$’s correspond to predicted regulators.
Feature selection with the lasso

\[
\min_{\beta \in \mathbb{R}^p} \| Y - X\beta \|^2 + \lambda \| \beta \|_1 \quad \text{where} \quad \| \beta \|_1 = \sum_{i=1}^p |\beta_i|
\]

- 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 \( p = 2 \)
For $t = 1$ to $T$ do
- Bootstrap a random sample $S_t$ from the training set
- Randomly reweight each feature (uniform on $[\alpha, 1]$)
- 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

![Graph showing frequency of selection over the number of LARS steps](image-url)
**Influence of \( \alpha \) and scoring method**

![Graphs showing AUPR and AUROC](image)

*Area Original*

**DREAM5 in silico network.**
DREAM5: GENIE and TIGRESS ranked 1st and 2nd out of 29 on the \textit{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
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)

\[ \text{Ratio of false positives} \]
\[ \text{Ratio of true positives} \]

<table>
<thead>
<tr>
<th>Method</th>
<th>Recall at 60%</th>
<th>Recall at 80%</th>
</tr>
</thead>
<tbody>
<tr>
<td>SIRENE</td>
<td>44.5%</td>
<td>17.6%</td>
</tr>
<tr>
<td>CLR</td>
<td>7.5%</td>
<td>5.5%</td>
</tr>
<tr>
<td>Relevance networks</td>
<td>4.7%</td>
<td>3.3%</td>
</tr>
<tr>
<td>ARACNe</td>
<td>1%</td>
<td>0%</td>
</tr>
<tr>
<td>Bayesian network</td>
<td>1%</td>
<td>0%</td>
</tr>
</tbody>
</table>

*SIRENE = Supervised Inference of REgulatory NEtworks (Mordelet and V., 2008)*
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
Chromosomal 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
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 $\beta$ 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):

\[ \| \beta \|_{TV} = \| \nabla \beta \|_1 = \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \]

- 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 as dichotomic segmentation

Algorithm 1 Greedy dichotomic segmentation

**Require:** $k$ number of intervals, $\gamma(I)$ gain function to split an interval $I$ into $I_L(I), I_R(I)$

1. $I_0$ represents the interval $[1, n]$
2. $\mathcal{P} = \{I_0\}$
3. **for** $i = 1$ to $k$ **do**
4. $I^* \leftarrow \arg \max_{I \in \mathcal{P}} \gamma(I^*)$
5. $\mathcal{P} \leftarrow \mathcal{P} \setminus \{I^*\}$
6. $\mathcal{P} \leftarrow \mathcal{P} \cup \{I_L(I^*), I_R(I^*)\}$
7. **end for**
8. **return** $\mathcal{P}$

Theorem

**Fused lasso performs "greedy" dichotomic segmentation**

*(V. and Bleakley, 2010; see also Hoefling, 2009)*
Solving fused Lasso

\[
\min_{\beta \in \mathbb{R}^p} \| Y - \beta \|^2 \quad \text{such that} \quad \sum_{i=1}^{p-1} |\beta_{i+1} - \beta_i| \leq \mu
\]

- QP with sparse linear constraints in \( O(p^2) \) -> 135 min for \( p = 10^5 \) (Tibshirani and Wang, 2008)
- Coordinate descent-like method \( O(p) \)? -> 3s s for \( p = 10^5 \) (Friedman et al., 2007)
- For all \( \mu \) with the LARS in \( O(pK) \) (Harchaoui and Levy-Leduc, 2008)
- For all \( \mu \) in \( O(p \ln p) \) (Hoefling, 2009)
- For the first \( K \) change-points in \( O(p \ln K) \) (Bleakley and V., 2010)
Speed trial: 2 s. for $K = 100, \ p = 10^7$
**Fused lasso for supervised classification**

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

- **Mathematically:**
  \[
  \min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_{TV}
  \]

- **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 $\beta$ that best discriminates the aggressive vs non-aggressive, subject to the constraints that it should be sparse and piecewise constant.

- **Mathematically**:

  $$\min_{\beta \in \mathbb{R}^p} \mathcal{R}(\beta) + \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_{TV}$$

- **Computationally**: 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 $\beta$ that best discriminates the aggressive vs non-aggressive, subject to the constraints that it should be sparse and piecewise constant.

- **Mathematically**:

  $$\min_{\beta \in \mathbb{R}^p} R(\beta) + \lambda_1 \| \beta \|_1 + \lambda_2 \| \beta \|_{TV}$$

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

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

DP finds the solution in $O(p^2 kn)$ 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 $\ell_1/\ell_2$-norm induces sparse solutions *at the group level*:

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

$$
\Omega(w_1, w_2, w_3) = \|(w_1, w_2)\|_2 + \|w_3\|_2 = \sqrt{w_1^2 + w_2^2 + \sqrt{w_3^2}}
$$
GFLseg (Bleakley and V., 20011)

Replace

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

by

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

We can solve it efficiently in \( O(np) \)

It converges to the true segmentation when the number of profiles increases
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 $\{10, 20, \ldots, 90\}$ and the variance $\sigma^2$ 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

Jean-Philippe Vert (ParisTech)

Machine learning in genomics

Berkeley 2012 46 / 59
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
Protein kinases
DNA and RNA polymerase subunits
Glycolysis / Gluconeogenesis
Porphyrin and chlorophyll metabolism
Riboflavin metabolism
Folate biosynthesis
Biosynthesis of steroids, ergosterol metabolism
Lysine biosynthesis
Phenylalanine, tyrosine and tryptophan biosynthesis
Purine metabolism
Nitrogen, asparagine metabolism
DNA and RNA polymerase subunits
Oxidative phosphorylation, TCA cycle

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)

\[ \Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 \]

- Gene selection + smooth on the graph

\[ \Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 + \sum_{i=1}^p |\beta_i| \]

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

\[ \Omega(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_{i=1}^p |\beta_i| \]
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

<table>
<thead>
<tr>
<th>METHOD</th>
<th>$\ell_1$</th>
<th>$\Omega_{\text{graph}}(\cdot)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERROR</td>
<td>0.39 ± 0.04</td>
<td>0.36 ± 0.01</td>
</tr>
<tr>
<td>AV. SIZE C.C.</td>
<td>1.03</td>
<td>1.30</td>
</tr>
</tbody>
</table>
Classical 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)