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.
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
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?

- **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

<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)
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

Jean-Philippe Vert (ParisTech)

Machine learning in bioinformatics

ML for Neuroimagine 2011 13 / 28
Prognosis

A  Gene-Expression Profiling

B  St. Gallen Criteria

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

 Reporter Genes

B  Years

C

Years

Probability of Remaining Metastasis-free

P<0.001

Good signature

Poor signature

NO. AT RISK

Good signature 50 57 45 31 22 12
Poor signature 91 72 55 41 26 17 9

Low risk 22 22 21 17 9 5 2
High risk 120 107 88 60 48 34 19

P=0.05

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

- Glycolysis / Gluconeogenesis
- N-Glycan biosynthesis
- Protein kinases
- DNA and RNA polymerase subunits
- Glycolysis / Gluconeogenesis
- Porphyrin and chlorophyll metabolism
- Sulfur metabolism
- Riboflavin metabolism
- Folate biosynthesis
- Biosynthesis of steroids, ergosterol metabolism
- Lysine biosynthesis
- Phenylalanine, tyrosine and tryptophan biosynthesis
- Nitrogen, asparagine metabolism
- Oxidative phosphorylation, TCA cycle
- DNA and RNA polymerase subunits
- Purine metabolism
- N-Glycan biosynthesis
- Protein kinases
- DNA and RNA polymerase subunits
- Glycolysis / Gluconeogenesis
- Porphyrin and chlorophyll metabolism
- Sulfur metabolism
- Riboflavin metabolism
- Folate biosynthesis
- Biosynthesis of steroids, ergosterol metabolism
- Lysine biosynthesis
- Phenylalanine, tyrosine and tryptophan biosynthesis
- Nitrogen, asparagine metabolism
- Oxidative phosphorylation, TCA cycle
- DNA and RNA polymerase subunits
- Purine metabolism
- N-Glycan biosynthesis
- Protein kinases
- DNA and RNA polymerase subunits
- Glycolysis / Gluconeogenesis
- Porphyrin and chlorophyll metabolism
- Sulfur metabolism
- Riboflavin metabolism
- Folate biosynthesis
- Biosynthesis of steroids, ergosterol metabolism
- Lysine biosynthesis
- Phenylalanine, tyrosine and tryptophan biosynthesis
- Nitrogen, asparagine metabolism
- Oxidative phosphorylation, TCA cycle
- DNA and RNA polymerase subunits
- Purine metabolism
- N-Glycan biosynthesis
- Protein kinases
- DNA and RNA polymerase subunits
- Glycolysis / Gluconeogenesis
- Porphyrin and chlorophyll metabolism
- Sulfur metabolism
- Riboflavin metabolism
- Folate biosynthesis
- Biosynthesis of steroids, ergosterol metabolism
- Lysine biosynthesis
- Phenylalanine, tyrosine and tryptophan biosynthesis
- Nitrogen, asparagine metabolism
- Oxidative phosphorylation, TCA cycle
- DNA and RNA polymerase subunits
- 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
  \[ \Omega(\beta) = \sum_{i \sim j} (\beta_i - \beta_j)^2 \]

- 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
  \[ \Omega(\beta) = \sum_{i \sim j} |\beta_i - \beta_j| + \sum_{i=1}^p |\beta_i| \]
Hypotheses 2: select genes which are connected and belong to the same functional groups.

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

<table>
<thead>
<tr>
<th>METHOD</th>
<th>$\ell_1$</th>
<th>$\Omega_{\text{OVERLAP}}(\cdot)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERROR</td>
<td>0.38 ± 0.04</td>
<td>0.36 ± 0.03</td>
</tr>
<tr>
<td>MEAN # PATH.</td>
<td>130</td>
<td>30</td>
</tr>
</tbody>
</table>

Graph on the genes.

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