How we learned to cope with molecular biology data

Annalisa Barla
molecular biology

- a typical scenario is $n \ll d$
- number of samples cannot always be increased (rare diseases and expensive technology)
- (mostly) high-throughput data
  - new technologies (DNA microarrays, SNP, CGH etc.)
  - possibility to measure the whole genome
  - most of the times the data are noisy (getting better any day now..)
learning from example paradigm

the GOAL is not to memorize but to GENERALIZE, e.g. predict

given a set of examples:
\[ \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\} \]

find a function:
\[ f(x) \sim y \]

such that \( f \) is a good predictor on new data as well as on the given dataset

and possibly identify the most discriminating variables (genes)

--> gene signature
why going multivariate?

- search for **DIFFERENTIALLY EXPRESSED GENES** is not always sufficient! univariate approaches may not be flexible enough...
variable selection method

- Empirical Risk minimization combined with a mixed penalty:
  - $l_1$ term enforcing sparsity
  - $l_2$ term preserving correlation

$$\phi_{\tau,\mu} = \|Y - X\beta\|^2 + \tau\|\beta\|_1 + \mu\|\beta\|_2$$


variable selection method

- Empirical Risk minimization combined with a mixed penalty:
  - $l_1$ term enforcing sparsity
  - $l_2$ term preserving correlation

- Consistency guaranteed (the more samples available the better the estimator)
- Not univariate: takes into account behavior of many genes at once.

$$\phi_{\tau, \mu} = \|Y - X\beta\|^2 + \tau \|\beta\|_1 + \mu \|\beta\|_2^2$$
variable selection method

\[ \phi_{\tau,\mu} = \| Y - X\beta \|^2 + \tau \| \beta \|_1 + \mu \| \beta \|_2 \]

- **output**: One-parameter family of nested lists with equivalent prediction ability and increasing correlation among genes.
- \( \mu \to 0 \): minimal list of prototype genes
- \( \mu_1 < \mu_2 < \mu_3 < \ldots \): longer lists including correlated genes

Since we have a *correlation parameter* we can tune and vary the list length.
two stage approach (De Mol, Mosci, Traskine, Verri 2009)

variable selection step (\(l_{1l2}\)):

\[
\|Y - X\beta\|^2 + \tau \|\beta\|_1 + \mu \|\beta\|^2_2
\]

correlation parameter

classification step (rls):

\[
\|Y - X\beta\|^2_2 + \lambda \|\beta\|^2_2
\]

for each \(\mu\) we have to choose \(\lambda\) and \(\tau\)
the optimal pair \((\lambda^*, \tau^*)\) is one of the \(A \cdot B\) possible pairs \((\lambda, \tau)_{ij}\)

computational time in the LOO case (for one task):

\[
\text{time}_{1-\text{optim}} = (2.5s \div 25s)
\]

depending on the correlation parameter

Total Time = \(A \cdot B \cdot \text{n.samples} \cdot \text{time}_{1-\text{optim}} \approx 20 \cdot 20 \cdot 30 \cdot \text{time}_{1-\text{optim}} \approx 2 \cdot 10^4 s ÷ 2 \cdot 10^6\)
Identifying the Hypoxia Signature of Neuroblastoma via Regularization

- Partner: IGG Molecular Biology lab
- Dataset: 9 neuroblastoma (NB) cell lines cultured under normoxic (normal O\textsubscript{2}) and hypoxic conditions (low O\textsubscript{2}).
- Technology: Affymetrix GeneChip U133 plus 2.0. (~54000 variables)
- t-test: no genes selected!
- L1L2 protocol: 11 genes for the minimal list (frequency > 30%)

Fardin P, Barla A, Mosci S, Rosasco L, Verri A, Varesio L
BMC Genomics 2009, 10:474 (15 October 2009)
a step forward: prior (GO)

- Specific prior knowledge can be used to better understand the biological phenomenon under study

- Possible sources:
  - Digital online data libraries
  - Textbook knowledge
  - MDs / experts

We selected subset of variables from the Gene Ontology and performed the variable selection on those sub-matrices

- hypoxia related groups
- MYCN related groups
- neuroblastoma related groups

another step forward: Function-based analysis of microarray data via l1-l2 regularization

• combine the selection protocol with the GO structure automatically

• provide a way to easily interpret the output of feature selection protocol

joint work with University of Padova

Function-based analysis of microarray data via l1-l2 regularization
poster @ ECCB’09
case study: breast cancer data (GSE7390)

Selected GO nodes after l1l2 feature selection step

232 GO nodes selected with test error $\leq 20\%$

Shrinkage

Selected GO nodes after l1l2 feature selection step
case study: breast cancer data

After pruning, 76 enriched GO nodes remain

Pruning
Selected GO nodes after enrichment test
(selected genes/total genes in the node)
case study: breast cancer data

The remaining nodes are grouped by average linkage hierarchical clustering based on semantic similarity.

The 76 GO nodes are clustered into 12 clusters.
what we learned

- beware of the selection bias
- go multivariate
- make use of the vast prior knowledge

- learn how to distribute the computation (grid/cloud/cluster)
- use open source software (in order to distribute on a cloud)

- some biology
- a common language with biologists and MDs
work in progress: methods to incorporate prior knowledge

- Kernel design
- Group lasso/Graph lasso (Jacob, L Obozinski, G Vert, JP)
- Semantic learning