Gene Expression Analysis : Tamoxifen Resistance in Breast Cancer

B. Haibe-Kains[1,2] G. Bontempi[2] C. Sotiriou[1]

[1]Unité Microarray, Institut Jules Bordet

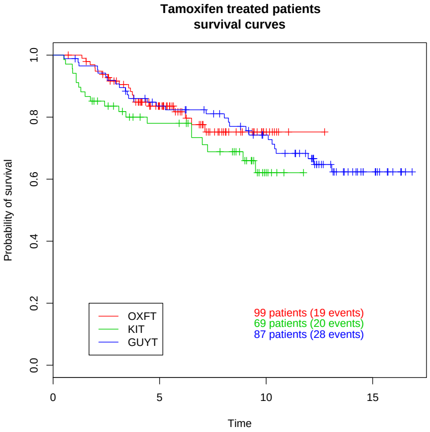
[2]Machine Learning Group, Université Libre de Bruxelles

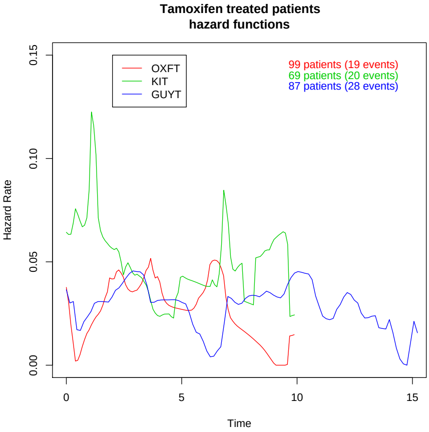
September 29, 2006

Biological Question and Datasets

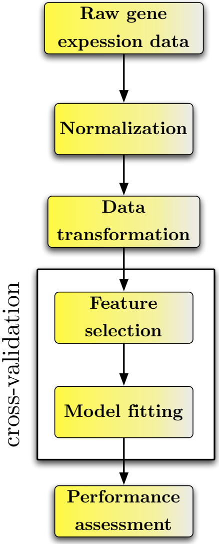
- Biological question : "Can we predict which patients will resist to the Tamoxifen treatment in an adjuvant setting ?"
- Tamoxifen treated patients coming from 3 different institutions, can we pool the data ?
    - gene-expressions : normalization
    - survival data : model fitting.
- 255 eligible patients (samples)
- 44928 probes (variables)

Survival Data





Tamoxifen Analysis



In order to keep the design simple, we fix :

- the number of models to aggregate (see feature selection step)
- the cutoff selection (see the model fitting step).

Normalization

- Goal : reduce the systematic variability between samples.
- Method : normalization.

Procedure

1. Background correction, expression quantification and normalization were performed using Robust Multichip Average [Irizarry et al., 2003, Bolstad et al., 2003].
2. RMA performed separately per population
3. Gene median centering per population.

➥ No artifact highlighted by unsupervised clustering.

Data Transformation

- Goal : reduce the dimensionality of the problem.
- Method : cluster highly correlated variables.

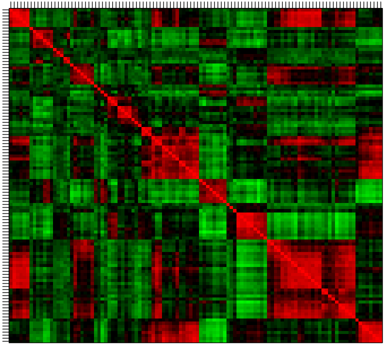
Procedure

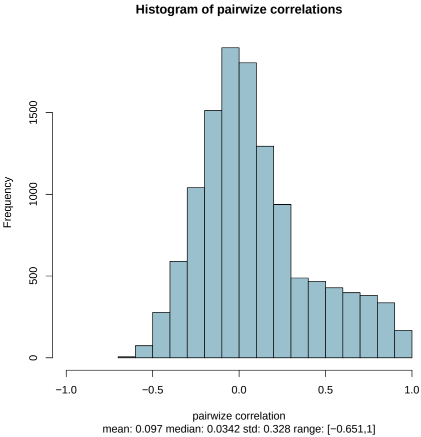
1. Use of an independent dataset of untreated patients (137 patients, 44929 probes)
2. Filtering of the less variant probes.
3. Hierarchical clustering (average linkage, uncentered Pearson correlation).
4. Cut the tree at height 0.5.
5. Keep only clusters with at least 5 known UniGenes.

➥ 110 clusters where half are significantly associated with survival (on the Tamoxifen dataset).

Data Transformation

Feature Pairwise Correlation





Feature Selection

- Goal : fast selection of the relevant features.
- Method : ranking based on univariate model significance.

Procedure

1. For each feature, compute the likelihood ratio test of the univariate Cox model.
2. Perform the ranking based on the p-value.
3. Select the top n features (n is fixed).

➥ n features are selected.

Model Fitting

- Goal : fit a simple model with low variance.
- Method : aggregation of *n* univariate classifiers.

Procedure

1. The univariate models for the *n* top features were computed during the previous step.
2. Compute a linear combination of these classifiers with weight of 1.

➥ Continuous score representing the risk of a patient.

Performance Assessment

- We do binary classification from survival data.
- We can not use traditional statistics (sensitivity, specificity, χ^2 test, ...)

➥ Adaptation of such estimators to deal with censoring.

Survival Statistics for 2 Groups

There exist several ways to assess difference in survival between 2 groups

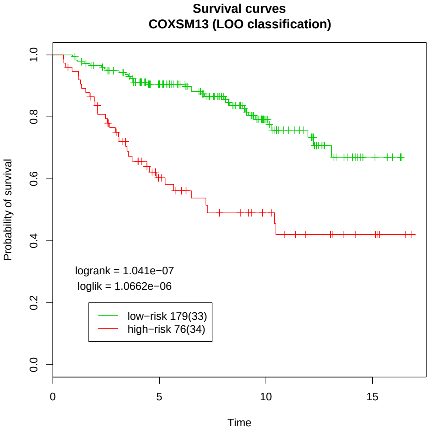
- **Kaplan-Meier estimator** and **Logrank test** :
  - KM method estimates survivor function such that

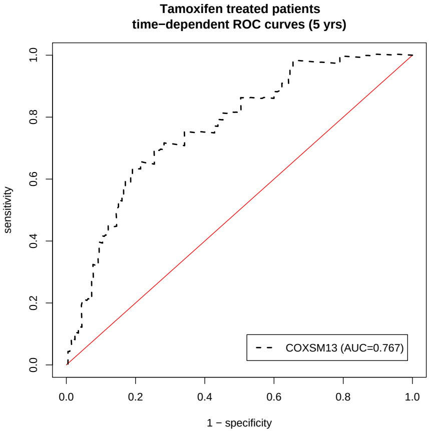
$$\widehat{S}(t) = \prod_{j:t_j \leq t} [1 - \frac{\overbrace{d_j}^{\#\ \text{death at time}\ t_j}}{\underbrace{n_j}_{\#\ \text{at risk at time}\ t_j}}].$$

  - logrank method tests $H_0 : S_1(t) = S_2(t)\ \forall t \geq 0$.
- **Hazard ratio** (HR) : relative hazard between 2 groups using Cox model with one dummy variable ($G = 0/1$ for low and high-risk groups).
- **Time-dependent ROC Curves** and area under ROC curves : extension of traditional ROC curves dealing with censoring data.

Performance in LOO

Logrank and Time-Dependent ROC Curve



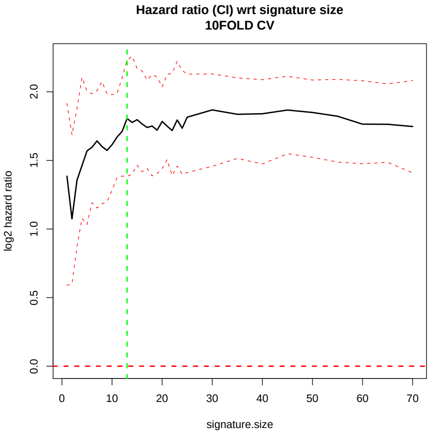


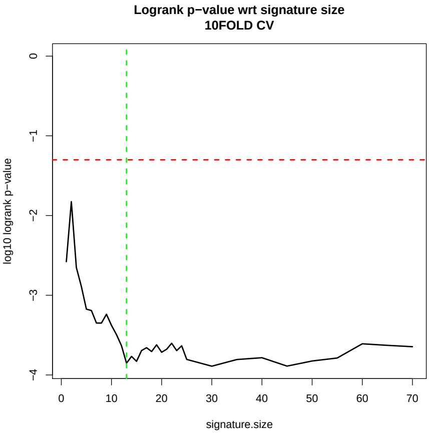
Performance

Hazard Ratio and Logrank

Training set	Test set	Hazard ratio	Log-rank p
OXFT (99/19)	KIT/GUYT (156/48)	2.17 [1.2,3.91]	8.46e-3
KIT (69/20)	OXFT/GUYT (186/47)	4.07 [2.23,7.41]	7.98e-7
GUYT (87/28)	OXFT/KIT (168/39)	5.93 [3,11.75]	1.24e-9
KIT/GUYT (156/48)	OXFT (99/19)	14.59 [5.38,39.52]	1.74e-11
OXFT/GUYT (186/47)	KIT (69/20)	3.44 [1.36,8.67]	5.27e-3
OXFT/KIT (168/39)	GUYT (87/28)	2.23 [1.05,4.71]	3.11e-2
Leave-one-out c.v.		3.85 [2.32;6.41]	1.04e-7
Multiple 10-fold c.v.		3.28 [2.66,3.84]	9.04e-7

Performance wrt Signature Size

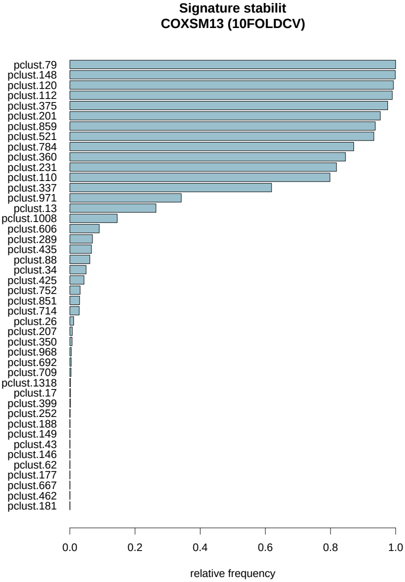




Performance vs Random

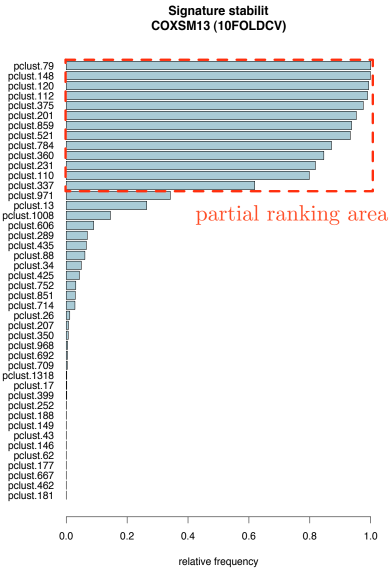
- At level of performance estimation :
  - ▶ 1000 random permutations of the labels and perform the whole procedure
  - ➥ only 1% of such classifications gives better discrimnation.
- At level of feature selection :
  - ▶ random selection of *n* features
  - ➥ most of the feature selection result in good classifiers because of the high proportion of relevant features.
  - ▶ use of the "anti"-ranking
  - ➥ very poor performance.

Feature Selection Stability



- Over the multiple 10-fold cross-validations, several top *n* features are selected
- If the same set of features are selected every time, we see high relative frequency for these features.

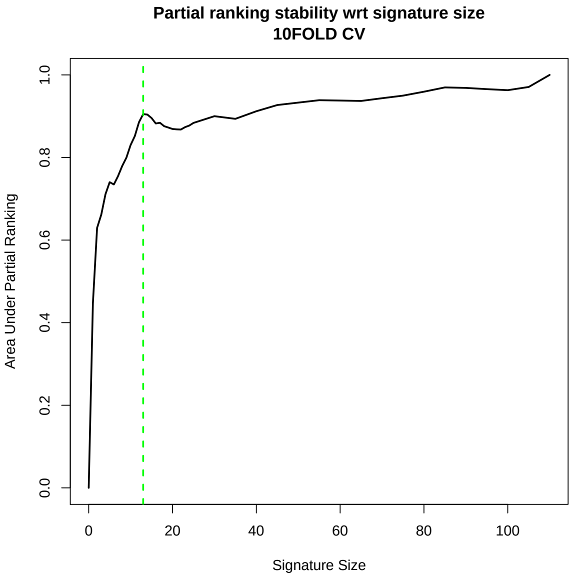
Feature Selection Stability



- Over the multiple 10-fold cross-validations, several top *n* features are selected
- If the same set of features are selected every time, we see high relative frequency for these features.

➥ We can compute the *area under partial ranking* w.r.t. the number of selected features.

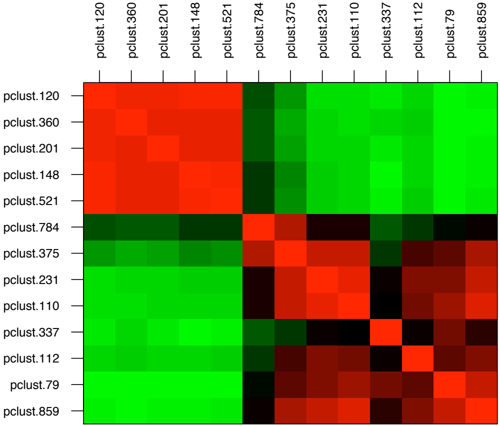
Stability wrt Signature Size



- $n = 13$ seems to be a good trade-off between the number of selected features (signature) and its stability.

Final Model

- Use the same method with all the samples.
- The result is a model with a set of features.
- The model and the features are expected from previous results.



Future Works

- Study the stability of the initial clustering (data transformation step).
- Use of Gene Ontology to study the signature in a biological point of view.
- Objective comparison with the traditional clinical variables.

Current Research Interests

- Meta-analysis.
- Ranking statistics.
- Input space transformation (using biological knowledge, such that gene list enrichment or GO).
- Optimization framework for binary classification of survival data.

Thank you for your attention.

Bolstad, B. M., Irizarry, R. A., Astrand, M., and Speed, T. P. (2003).
A comparison of normalization methods for high density oligonucleotide array data based on variance and bias.
Bioinformatics, 19(2):185–193.

Irizarry, R. A., Boldstad, B. M., Collin, F., Cope, L. M., Hobbs, B., and Speed, T. R. (2003).
Summaries of affymetrix genechip probe level data.
Nucleic Acids Research, 31(4).