Gene Expression Analysis of Tamoxifen Resistance in Breast Cancer

B. Haibe-Kains[1,2], S. Loi[3], C. Desmedt[1], F. Lallemand[1], A. M. Tutt[4], C. Gillet[4], P. Ellis[4], K. Ryder[4], A. Harris[6], J. Bergh[5], D. Larsimont[1], M. Buyse[7], G. Bontempi[2], M. J. Piccart[1], and C. Sotiriou[1]

[1]Unité Microarray, Institut Jules Bordet, Brussels, Belgium
[2]Machine Learning Group, Université Libre de Bruxelles, Brussels, Belgium
[3]Peter MacCallum Cancer Center, Melbourne, Australia
[4]Guys Hospital, London, United Kingdom
[5]Karolinska Institute, Stockholm, Sweden
[6]John Radcliffe Hospital, Oxford, United Kingdom
[7]International Drug and Development Institute, Brussels, Belgium
[8]NCCR Molecular Oncology and Swiss Institute of Bioinformatics, Epalinges, Switzerland

1. Introduction

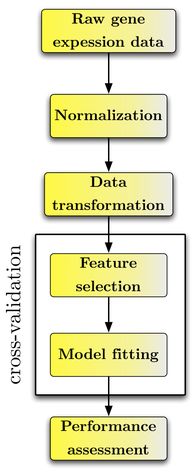
MAJORITY of early-stage breast cancers express estrogen receptors (ER) and receive tamoxifen in the adjuvant setting. Yet up to 40% of these patients will relapse on tamoxifen and develop incurable metastatic disease. Recent evidence from large randomised controlled trials exploring the role of aromatase inhibitors (AI) in the adjuvant setting shows a benefit from the novel strategy, however, the optimal sequence and duration of AI/Tamoxifen treatment is unknown. Therefore, it is vital that we learn to identify those women at higher risk of tamoxifen resistance. Our aim is to identify genes that could predict for this subset of women.

2. Materials

GENE expression profiles are determined from 255 tamoxifen-only treated ER positive early stage breast cancer using Affymetrix U133 A, B and Plus2 chips (44928 probes in common). The patients come from 3 different institutions (called OXFT, KIT and GUYT). 27% of them developed distant recurrence within 13 years.

3. Methods

WE introduce a complete gene expression analysis design including data transformation, feature selection, model fitting and validation steps.



3.1 Data Transformation

After the RMA normalization procedure [5] of the raw gene expressions, a data transformation is performed in order to reduce the dimensionality of the problem and to cluster highly correlated variables. This transformation consists in a hierarchical clustering [3, 2] based on an independent dataset (137 untreated breast cancer patients, 44928 probes). Each cluster of probes is summarized in a new variable called *metagene*.
The computed metagenes (110) are well conserved across datasets (see Figure 1) and are poorly correlated (see Figure 2).

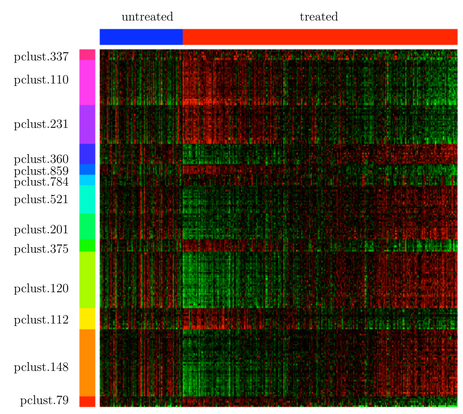


Figure 1: *Heatmap of probes composing some key metagenes in the two datasets.*

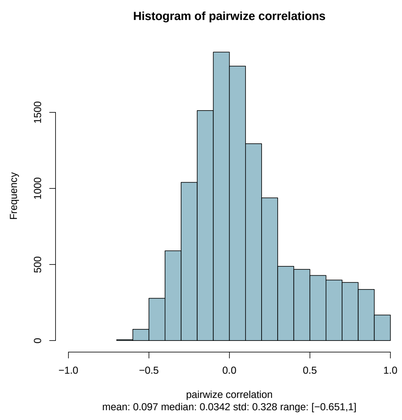


Figure 2: *Pairwise correlation between metagenes in the tamoxifen treated dataset.*

3.2 Feature Selection

The goal of feature selection is the fast selection of the relevant features. The method consists in a partial ranking based on univariate Cox model [1] assessing the power of survival prediction of a metagene. Only the top n metagenes are used to fit a model (n is called the signature size).

3.3 Model Fitting

The model is the average of the univariate Cox models computed in the feature selection step. Such a model is simple to interpret and its variance is low [6]. Finally, the continuous score generated by the model represents the risk of a patient to develop a distant metastasis.

4. Results

WE assess the performance of the method on the basis of cross-validation and its stability in selection of the relevant metagenes used to fit a classification model.

4.1 Validation

The method performance (called *COXSM13*) is estimated using a cross-validation scheme [7]. A leave-one-out strategy is adopted for assessing the classification accuracy and a multiple 10-fold cross-validation is adopted to study the stability of the feature selection.

4.2 Classification

We use several performance estimators :

- Time-dependent ROC curves [4] : extension of traditional ROC curves dealing with survival data (see Figure 3).
- Kaplan-Meier estimator and logrank test to assess the difference in survival between low and high-risk group of patients (see Figure 4).
- Hazard ratio : relative hazard between two groups using Cox model with one dummy variable (see Table 1).

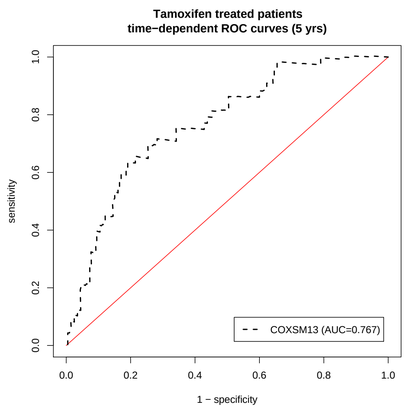


Figure 3: *Time-dependent ROC curve of the classification (using leave-one-out).*

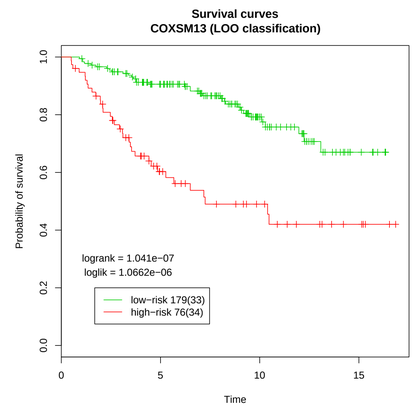


Figure 4: *Survival curves of the low and high-risk groups of patients.*

Training set	Test set	Hazard ratio	Logrank p-value
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

Table 1: *Hazard ratio and logrank p-values for different validation schemes.*

4.3 Further Analyses

The impact of the signature size on stability of the feature selection method is studied empirically (see Figure 5 and 6). A new statistic, called *area under partial ranking*, allows to measure the stability ranging from 0 (a set of features is never selected) to 1 (the same set of features is selected every time).

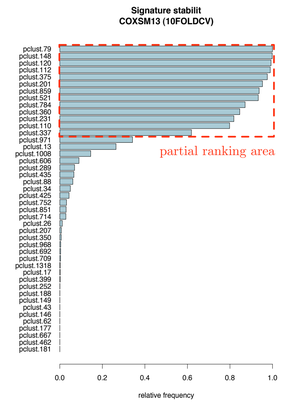


Figure 5: *Example of area under partial ranking. The bars represent the selection frequency for a metagene.*

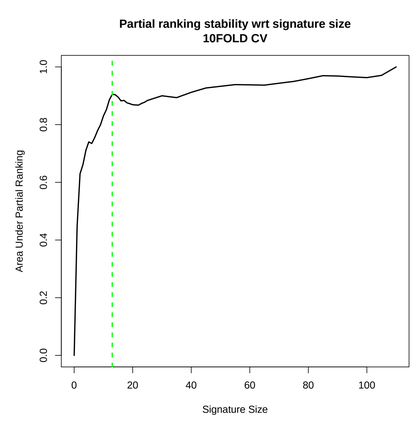


Figure 6: *Evolution of the feature selection stability wrt the signature size.*

5. Conclusion

THESE RESULTS suggest that a group of genes can identify breast cancer patients at risk of early distant relapse on Tamoxifen. These patients could be the ideal candidates for upfront AIs, while the others would be considered for sequential Tamoxifen/AI. This hypothesis will be tested in an upcoming prospective clinical trial MINDACT.

This work was presented in [8].

References

[1] D. R. Cox. Regression models and life tables. *Journal of the Royal Statistical Society Series B*, 34:187–220, 1972.

[2] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95:14863–14868, 1998.

[3] J. A. Hartigan. *Clustering Algorithms*. Wiley, 1975.

[4] P. J. Heagerty, T. Lumley, and M. S. Pepe. Time-dependent roc curves for censored survival data and a diagnostic marker. *Biometrics*, 56:337–344, 2000.

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

[6] J. Kittler, M. Hatef, R. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–238, 1998.

[7] R. Kohavi. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of IJCAI-95*, 1995. available at http://robotics.stanford.edu/users/ronnyk/ronnyk-bib.html.

[8] S. Loi, M. Piccart, B. Haibe-Kains, C. Desmedt, A. Harris, J. Bergh, P. Ellis, L. Miller, E. Liu, and C. Sotiriou. Prediction of early distant relapses on tamoxifen in early-stage breast cancer (BC): a potential tool for adjuvant aromatase inhibitor (AI) tailoring. In *Proceedings of the American Society of Clinical Oncology Meeting, Orlando 2005, abstract 509*, 2005.