Computational Intelligence for Predicting the Chemotherapy Outcomes in Breast Cancer

René Natowicz[1], Antônio P. Braga[2], Roberto Incitti[3], Marcelo A. Costa[2], Patrick Siarry[4], Arben Çela[1], Euler G. Horta[2], Carmén DM. Pataro[2], Thiago Souza[2], Roman Rouzier[5]

[1] University of Paris-Est, Esiee-Paris - {rene.natowicz, arben.cela}@univ-paris-est.fr

[2]Federal University of Minas Gerais, Belo Horizonte
{apbraga,azevedo,ehorta,cdmp}@ufmg.br, thiago@dcc.ufla.br

[3]INSERM, Mondor Institute for Molecular Medicine, Créteil - roberto.incitti@inserm.fr

[4]University of Paris-Est, LiSSi, Créteil - siarry@univ-paris12.fr

[5]Pierre & Marie Curie University, Tenon Hospital, Paris - roman.rouzier@tnn.aphp.fr

Abstract. *Predicting the outcome of chemotherapy treatments in oncology is an important clinical issue for better allocating the patients to the treatments, and in pharmacology because an accurate characterization of the non responders could be of great help for designing new treatments dedicated to these cases. We present an ongoing franco-brazilian research for the design of efficient multi-genomic predictors in breast cancer. This research is supported by a four years* CAPES-COFECUB *program (2008-2011)*[1].

Keywords: *bioinformatics, computational intelligence, genomic predictors, chemotherapy treatments, breast cancer.*

1. Statement of the problem

Since the advent of high throughput genomic technologies a decade ago, massive information at the genomic level is available, which can be exploited in medical studies. Microarrays allow to measure simultaneously the expression levels of a large fraction of all known genes (ranging from 5 000 to 30 000). Relying on these data, it has become possible to design efficient predictors that significantly outperform the previous clinico-biologic ones. Our research is concerned with pre-operative chemotherapy treatments for breast cancer.

In the process of designing predictors of the outcomes of chemotherapy treatments, the main issues are that of identifying the genes actually involved in the response to the chemotherapy treatment; combining the expression levels of these genes for predicting the response to the treatment; and assessing the robustness of the predictors, i.e. the statistical independence of the method – gene selection and computational model – relatively to the learning set of patient cases. Our dataset comes from clinical trials which were jointly conducted at the Gustave Roussy Institute, Villejuif (France), and at the Anderson

[1] Project #591/08, Multigenic Predictors of the Responses to Preoperative Chemotherapy Treatment in Breast Cancer, shared by : University of Paris-Est, France; University Federal of Minas Gerais in Belo Horizonte, Brazil, School of Engineering, Computational Intelligence Laboratory; University Pierre and Marie Curie, Paris, France, Hospital Tenon, Obstetrics and Gynecology department; INSERM, Créteil, France, Mondor Institute for Molecular Medicine, Genomic Platform.

Cancer Center, Houston-Texas (USA) in 2004 and 2005 [Hess et al. 2006]. For each patient case, the data are the outcome of the pre-operative chemotherapy treatment, either responder or non responder, and the expression level of more than 22 000 genes, measured on the tumor tissues. The gene expression profiling was performed using oligonucleotide microarrays (Affymetrix U133A).

In its essence, designing a predictor is a supervised learning process aiming at recognizing the class of patient cases who are responders to the treatment and that of the non responders. The novelty comes from the nature of the high throughput genomic data and clinical trials. The number n of cases of the clinical trial is less than 200 while the number of variables, p, which are the expression levels measured, is more than 22 000. Otherwise stated, the problem is a two classes supervised learning where p greatly exceeds n [Simon 2003]. Furthermore, the data are very noisy because of the present state of the technology of DNA microarrays and because we are dealing with biological data. Hence, selecting a very small subset of relevant genes is crucial for designing robust predictors.

2. Selecting the genes

The performances of a predictor are measured by its accuracy, sensitivity and specificity. The accuracy, Ac, is the proportion of well predicted cases, regardless of their classes. The sensitivity, Se, is the proportion of well predicted responder cases, and the specificity of the predictor, Sp, is that of the well predicted non responder cases. Because only a very small number of genes can enter the predictor, each of them should give an information about both classes. Such genes are *bi-informative* [Natowicz R. et al. 2008b]. It has appeared to us that this main characteristics of the gene selection had been neglected in the previous works. A lot of methods have been proposed so far for selecting the genes, but the most widely used method still consists in selecting them according to a very basic statistical criterion, namely the the p-value to a t-test. According to this approach, one considers the mean value of the expression levels of each gene[2] measured on each class, then one selects the genes whose mean difference is the most likely not to have been obtained *by chance.* According to this criterion, the expression levels of the most relevant genes are typically as depicted in figure 1, which is the box-plot of the expression levels of the gene MAPT (*microtubule-associated protein tau*, probe 203929_s_at). The gene MAPT is the most relevant gene according to the p-value of a t-test [Rouzier R. et al. 2005]. From this figure one can see that the information given by the gene MAPT is essentially about the non-responder cases. Such genes are *mono-informative*, and almost all the genes selected this way are *mono-informative.*

[2] The expression levels are those of the microarray's DNA probes. A lot of genes are represented by several DNA probes on the Affymetrix microarrays. Hence we should talk of the level of expression of the DNA probes rather than that of a gene. In the following, for making short, we will nevertheless talk of the expression level of a gene.

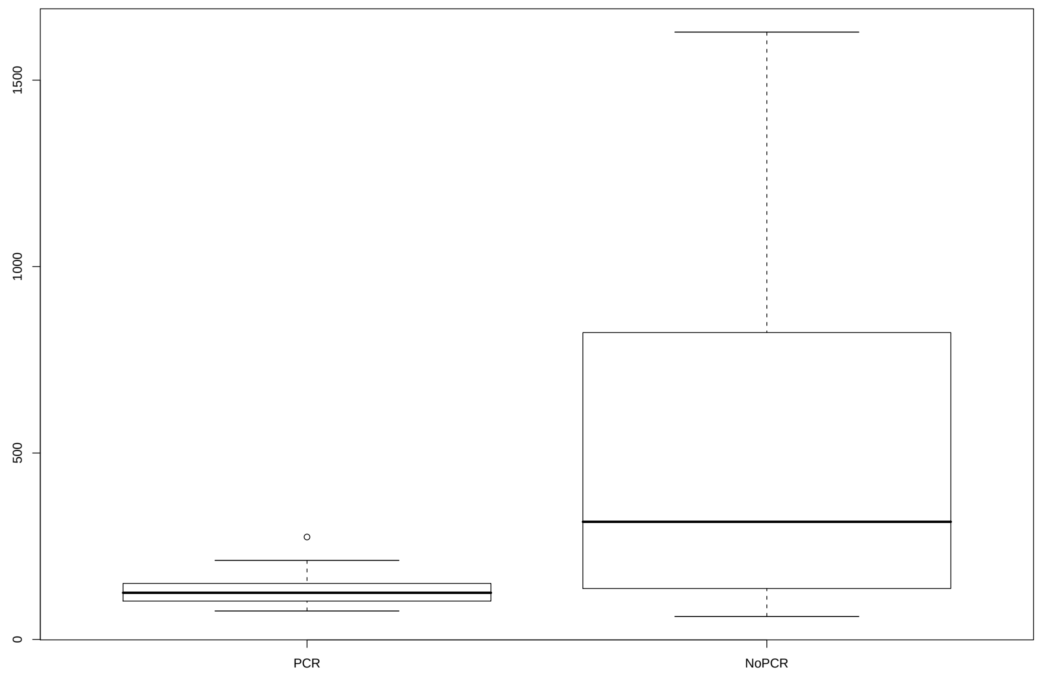

Figure 1. Boxplots of the expression levels of the gene MAPT (probe 203929_s_at). MAPT is a mono-informative gene. Left : expression levels of the responder cases (PCR : pathologic complete response); right : those of the non responder cases (NoPCR).

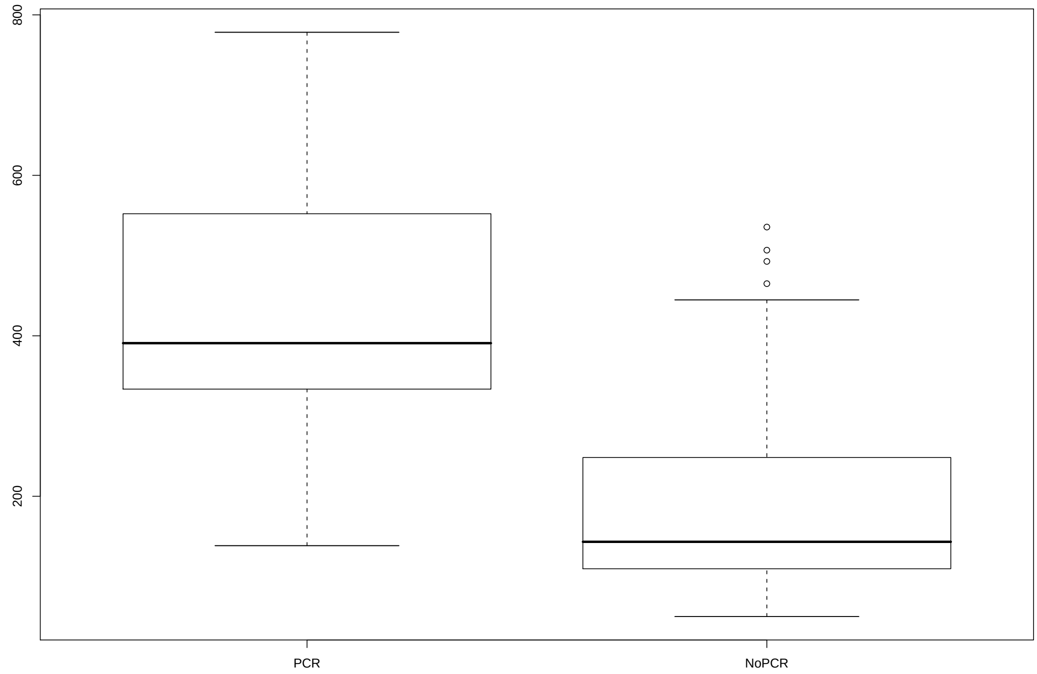

Figure 2. Boxplots of the expression levels of the gene BTG3 (probe 205548_s_at). BTG2 is a bi-informative gene. Left : expression levels of the responder cases; right : those of the non responder cases.

Because we wanted the predictors to be both highly sensitive and highly specific, we have developed a new method for characterizing and selecting *bi-informative genes*. In figure 2, the box-plot is that of the most relevant gene according to the criterion that we have proposed in [Natowicz R. et al. 2008b]. All the genes selected this way, were *bi-informative.* The method that we have developpped can be seen as a selection process of the genes according to their individual sensitivities and specificities. We have considered each gene as an elementary predictor or the response to the chemotherapy, the sensitivity and specificity values of which were the proportions of responder and non responder cases of the learning set of cases that were correctly predicted by the gene. This way, each gene could be plotted in the two dimensional space of sensitivity and specificity values. In this space, we have also plotted a hypothetic *ideal* gene, supposed to predict all the cases of the learning set (the sensitivity and specificity values of which were both equal to one). Then, we have selected the genes according to their euclidean distance to the *ideal* one, in the sensitivity-specificity space. In figure 3, all the genes of the DNA microarrays used for the clinical trial have been plotted (more than 20 000 genes), together with the *ideal* gene (at coordinates $(1, 1)$, at the upper right corner of the figure). In figure 4 are the 30 genes which were the closest to the *ideal* gene.

3. Predicting the outcomes of the treatment

A lot of mathematical and computational models have been used so far for combining the expression levels into a prediction function [Natowicz R. et al. 2008a, Rouzier R. et al. 2009]. These models range from very simple linear regression, to highly non linear models, among which various models of neural networks. Whatever the model, the question of the robustness of the predictor is of first importance in the perspective of using it in clinical routine. Thanks to the gene selection step, we could rely on a small number of genes, the order of which was less than thirty in our first publications [Natowicz R. et al. 2008b], and presently less than twenty [Natowicz R. et al. 2008c]. In order to avoid the overfitting of the data, the expression levels of the selected genes were given as input to a multi-objective neural networks, a classifier model developed by Braga & al. [Braga AP. et al. 2006]. The learning process of this classifier consists in minimizing both the classification error and the overfitting of the data. In this approach, the search of for a *global optimum* is substituted by that of *Pareto-optimality*. After optimization, the Pareto-set contains the non-dominated solutions that cannot be improved in one of the objectives without degrading the others. The decision making procedures follow the Pareto-set generation: a solution is selected according to a pre-established criterion. The simplest selection approach is to minimize the error of a validation set. Other selection strategies that explore the Pareto set properties have been applied successfully [Kokshenev and Braga AP. 2008].

4. Results and developments to come

The predictors that we have designed have significantly outperformed the best predictors reported for the same problem and data [Hess et al. 2006]. The performances of our predictors, measured on an independent set of data (data neither used for selecting the genes nor designing the multi-objective classifier) are no less than: Ac=0.86, Se=0.92, Sp=0.84. The statistical validation was done by cross-validations and the stability of the method of gene selection was demonstrated on replicates.

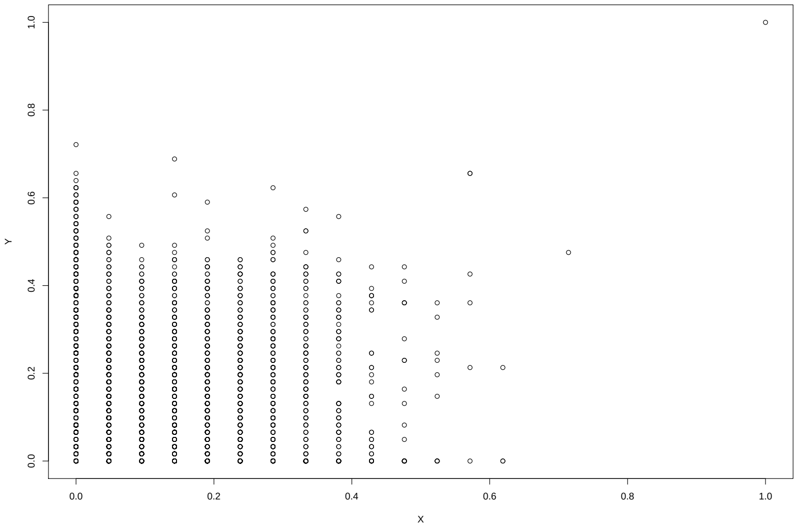

Figure 3. The genes of the DNA microarrays, plotted in the sensitivity-specificity space. X, Y axes: sensitivity and specificity values of the genes. The hypothetic *ideal* gene is at coordinates (1,1).

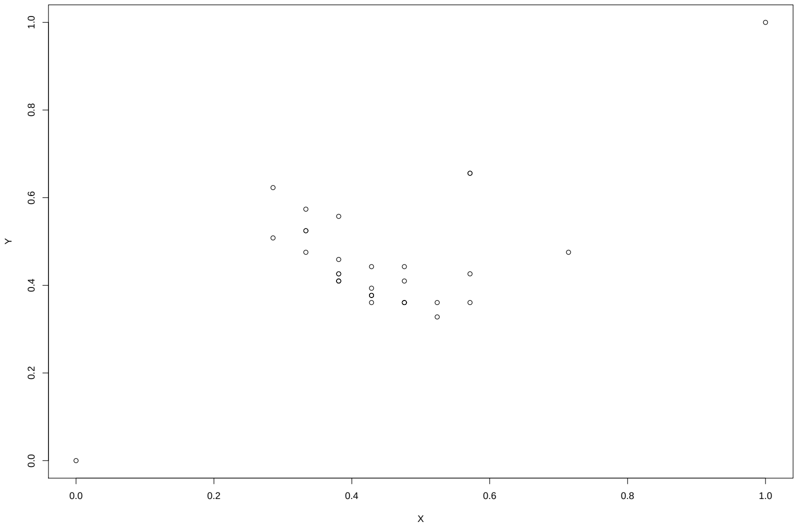

Figure 4. Coordinates of the 30 genes the closest to an *ideal* gene.

But, because our objective is to design efficient predictors to be widely used in clinical routine, we must increase further their sensitivity: the false negatives of a predictor being patient cases wrongly predicted not to benefit from the treatment, these situations should be avoided as much as possible, although the decision of not allocating a patient to the treatment will always belong to the clinicians.

At the present stage of our research, it has appeared to us that we must gain a deeper understanding of the statistical properties of the populations under study in the clinical trials (stratifying the data according to age and genetic criteria is a possible need) and of the statistics of the gene expressions themselves. This is a work in progress. Furthermore, the biological mechanisms underlying the responses to the chemotherapy treatments are those of gene interaction networks which, up to now, are scarcely known. The gene expressions that one measures are their resulting effects, but these interactions are not taken into account in the gene selection process itself. Hence, we would like to investigate the selection of genes subsets. Because the number of variables is around 20 000, one should rely on (and possibly develop) methods for searching efficiently the huge space of the gene subsets [Siarry P. and Michalewicz 2007]. It is a major development to come for this collaborative research, that motivates our wish to open the project to research teams in metaheuristics for optimization.

References

Braga AP., Takahashi, R., Costa MA., and Teixeira, R. (2006). *Multi-Objective Algorithms for Neural Networks Learning*. Studies in Computational Intelligence. Springer.

Natowicz R., Braga AP., Incitti R., Horta EG., Rouzier R., Rodrigues, T., and Costa MA. (2008a). A new method of dna probes selection and its use with multi-objective neural network for predicting the outcome of breast cancer preoperative chemotherapy. In *ESANN*, pages 71–76.

Natowicz R., Incitti R., Horta EG., Charles, B., Guinot, P., Yan, K., Coutant, C., Andre, F., Pusztai, L., and Rouzier R. (2008b). Prediction of the outcome of preoperative chemotherapy in breast cancer by dna probes that convey information on both complete and non complete responses. *BMC Bioinformatics*, 9:149+.

Natowicz R., Incitti R., Rouzier R., Çela A., Braga AB., Horta EG., Rodrigues, T., and Costa M. (2008c). *Downsizing Multigenic Predictors of the Response to Preoperative Chemotherapy in Breast Cancer*. Lecture Notes in Computer Science. Springer.

Rouzier R., Coutant, C., Lesieur, B., Mazouni, C., Incitti R., Natowicz R., and Pusztai, L. (2009). Direct comparison of logistic regression and recursive partitioning to predict chemotherapy response of breast cancer based on clinical pathological variables. *Breast Cancer Res. Treat.*

Rouzier R., Rajan, R., Wagner, P., Hess, K., Gold, D., Stec, J., Ayers, M., Ross, J., Zhang, P., Buchholz, T., Kuerer, H., Green, M., Arun, B., Hortobagyi, G., Symmans, W., , and Pusztai, L. (2005). Microtubule-associated protein tau: A marker of paclitaxel sensitivity in breast cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 102(23):8315–8320.

Siarry P. and Michalewicz, Z. (2007). *Advances in Metaheuristics for Hard Optimization*. Natural Computing Series. Springer Berlin Heidelberg.

Hess, K., Anderson, K., Symmans, W., Valero, V., Ibrahim, N., Mejia, J., Booser, D., Theriault, R., Buzdar, A., Dempsey, P., Rouzier R., Sneige, N., Ross, J., Vidaurre, T., Gomez, H., Hortobagyi, G., and Pusztai, L. (2006). Pharmacogenomic Predictor of Sensitivity to Preoperative Chemotherapy With Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide in Breast Cancer. *J Clin Oncol*, 24(26):4236–4244.

Kokshenev, I. and Braga AP. (2008). A multi-objective approach to rbf network learning. *Neurocomputing*, 71(7-9):1203–1209.

Simon, R. (2003). Supervised analysis when the number of candidate features (p) greatly exceeds the number of cases (n). *SIGKDD Explor. Newsl.*, 5(2):31–36.