Predicting breast cancer chemotherapeutic response using a novel tool for microarray data analysis

Jie Cheng[1,*], Joel Greshock[2], Jeffery Painter[1], Xiwu Lin[1], Kwan Lee[1], Shu Zheng[3], Alan Menius[1]

[1]Quantitative Sciences, GlaxoSmithKline, Collegeville, PA 19426, USA

[2]Cancer Research, GlaxoSmithKline, Collegeville, PA 19426, USA

[3]Cancer Institute, Zhejiang University, Hangzhou, 310009, China

Summary

We developed a novel tool for microarray data analysis that can parsimoniously discover highly predictive genes by finding the optimal trade off between fold change and t-test p value through rigorous cross validation. In addition to find a small set of highly predictive genes, the tool also has a procedure that recursively discovers and removes predictive genes from the dataset until no such genes can be found. We applied our tool to a public breast cancer dataset with the goal to discover genes that can predict patient's response to a preoperative chemotherapy. The results show that estrogen receptor (ER) gene is the most important gene to predict chemotherapeutic response and no gene signatures can add much clinical benefit for the whole patient population. We further identified a clinically homogenous subgroup of patients (ER-negative, PR-negative and HER2-negative) whose response to the chemotherapy can be reasonably predicted. Many of the discovered predictive markers for this subgroup of patients were successfully validated using a blinded validation set.

1 Introduction

Feature selection is one of the most important topics in microarray data analysis. The goal of feature selection is to find informative genes in high dimensional data from a set of examples with known clinical outcome. Feature selection serves two distinct purposes: (a) identify a parsimonious set of genes that yield a predictive model with good performance in independent cases; (b) identify all significantly differentially expressed genes between two outcome groups in order to gain insight into biological processes which differentiate the groups by making use of pathway analysis or gene ontology analysis tools.

In many microarray data analysis algorithms, features need to be ranked using a certain statistic, either before or inside the classifier learning process. For features that follow a normal distribution, a classical t-test is an excellent choice for ranking. However, in many datasets, predictive features are far from being normally distributed. For example, many oncogenes only show large elevation in a small portion of the samples of one phenotype. Feature selection using t-test cannot pick up such genes if stringent p value is used. For such features a simple fold change or mean difference test is a better choice than a classical t-test. For this reason, various modified t-tests, e.g., SAM [1], Efron-t (equation (2.8) of [2]), "shrinkage-t" [3], have been proposed for feature selection in high dimensional array data, primarily to balance the trade-off between a mean difference test and a t-test in order to efficiently detect different types of predictive features.

* To whom correspondence should be addressed. Email: jie.j.cheng@gsk.com

To address these issues, we have developed a Java based data analysis tool which selects features by searching for the desired thresholds of both mean difference test and t-test p value. For any pair of thresholds, the features that satisfy both thresholds are used to build a diagonal linear discriminate analysis (*DLDA*) classifier [4], which is a simple linear classifier similar to weighted voting [5]. By varying the thresholds of these two statistics in certain steps within their acceptable ranges, we can achieve various trade-offs and control the size of the feature sets. The pair of the thresholds of the two statistics solely determines the feature set and thus the *DLDA* model, as there is no parameter to be tuned for *DLDA* modelling. The tool can automatically find the optimal balance based on cross-validated model performance and generate an optimal model. In a sense, our tool can be viewed as a generalization of various modified t-tests – instead of using a fixed trade-off, the proposed adaptive method will automatically discover optimal trade-offs for a given dataset.

As different types of biomarker signals can coexist in the same data set, an iterative wrapper procedure was developed to enable the discovery of more features by finding a number of effective trade-offs between the two statistics.

To demonstrate the clinical value of the proposed method, we performed detailed analyses on a publicly available breast cancer chemotherapeutic response dataset [6].The results show that our tool is very effective in identifying predictive features. The majority of the identified biomarkers from a clinically homogeneous subpopulation were successfully validated using an independent set of patients.

2 Methods

Our Java tool has three main functions.

The procedure ***'findGeneSignature'*** uses a grid search procedure that searches through various trade-offs of mean difference test and t-test. A prediction model is generated using the optimal trade-off in terms of cross-validated model performance. In our tool, we use area of the ROC (receiver operating characteristic) curve (AUC) to measure model performance.

A nested cross-validation (CV) procedure called ***'estimatePerformance'*** is used to estimate model performance. In this procedure, procedure ***'findGeneSignature'*** is called to tune the parameters using the inner cross-validation. The outer cross-validation is used to estimate model performance.

The procedure ***'findImportantGenes'*** is developed to find more important genes for pathway analysis. It is a wrapper procedure that iteratively collects generated gene signatures and removes those genes from further runs. This process continues until procedure '***estimatePerformance'*** returns close to random performance (i.e., most informative genes have been identified and removed).

High level pseudo code is listed as follows.

Procedure ***trainModel*** (training data, a pair of thresholds for mean difference test and t-test respectively) {

1. Collect mean difference and t-test p value for each gene based on training data
2. Using both statistics to filter genes according to the thresholds.
3. Build *DLDA* model using the genes that passed the filtering process.
4. Return *DLDA* model

}

Procedure ***testModel*** (test data, model) {

Use the *DLDA* model to return continuous scores for test cases

}

Procedure ***findGeneSignature*** (training data, ranges and steps of thresholds for statistic A and statistic B) {

1. For each combination of the two thresholds, call (repeated) cross-validation procedure to get averaged performance and averaged size of feature set. Procedure ***trainModel*** and procedure ***testModel*** are called repeatedly within the cross-validation procedure
2. Select the optimal pair of thresholds based on model performance. (Users can also hand pick a pair of thresholds based on performance, size of feature size, and fold change etc.)
3. Call procedure ***trainModel***(training data, selected pair of thresholds)

}

Procedure ***estimatePerformance*** (training data) {

1. Run (repeated) cross-validation to estimate model performance. Procedure ***findGeneSignature*** is called repeatedly within the cross-validation procedure using an internal training data. Models returned from procedure ***findGeneSignature*** are repeatedly evaluated by procedure ***testModel*** using internal validation data.
2. Internal validation results are collected and the average performance is returned.

}

Procedure ***findImportantGenes*** (training data, performance threshold) {

While (procedure ***estimatePerformance*** (training data) > performance threshold) {

Call procedure ***findGeneSignature***

Add the genes of the returned model to the list of important genes

Remove these genes from the training data

}

Return the list of important genes

}

3 Experimental results

The chemotherapeutic response dataset we analyzed in this paper is publicly available from NCBI GEO data repository (http://www.ncbi.nlm.nih.gov/geo/; GEO accession GSE20194). It is one of the six datasets analyzed in the FDA led MicroArray Quality Control project (MAQC) [6]. The dataset contains pretreatment gene expression data (Affymetrix HG-U133A) and clinical information from 230 patients with stage I-III breast cancer. The goal of this study was to develop a gene expression based model that can predict pathologic complete response (*pCR*) to a preoperative *paclitaxel* + 5-fluoruracil, doxorubicin, and cyclophosphamide (T/FAC) chemotherapy regimen. The dataset was generated at two stages. In the first stage, 130 samples were generated and analyzed. Additional 100 samples were subsequently generated to validate the findings of the first stage. The original analysis result from the samples of the first stage is published in [7], where a 30 probe set gene signature was developed. In our experiment, we follow the practice of MAQC project, that is, to use the first 130 samples as the training data and the additional 100 samples as the blinded validation data.

3.1 Biomarker discovery using the training data

To discover the most predictive features, we applied procedure ***'findGeneSignature'***, which resulted in a signature of 16 probe sets (13 genes), including the estrogen receptor gene (*ESR1*), which regulates cell division and DNA replication in ~60% of breast cancers, serves as a pathological marker for diagnosis and treatment. In this data set *ESR1* has the second largest fold change among the 13 genes. This analysis confirms what is already known, i.e., estrogen receptor (ER) status is a very important factor for predicting chemotherapy response in breast cancers. Interestingly, ESR1 gene is missing from the 30 probe set signature developed in the original publication.

To identify more of the highly predictive genes for further functional analysis, the iterative procedure '***findImportantGenes'*** is applied. It proved to be more effective to apply this procedure twice: one time for finding features with high average expression values in one class, and the second time for finding features with high average expression value in the other class. Figure 1 shows how the prediction performance drops after important features being iteratively removed. This process returned over 300 probe sets. Through pathway analysis, we found that most of these genes are related to ER gene.

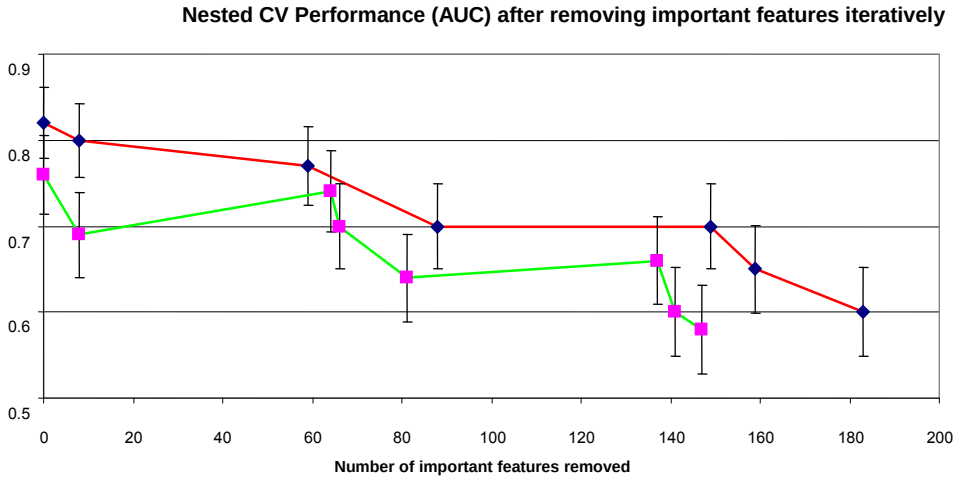

Figure 1: Decrease of nested CV performance (AUC) after removing important features iteratively. Procedure *findImportantGenes* is called twice. The first round for detecting genes with high average expression values in the non-pCR group (in red line). The second round for detecting genes with high average expression values in pCR group (in green line). The process stops when the nested CV performance (AUC) is below 0.60. Over 180 features were discovered in the first round and over 140 features were discovered in the second round. The nested CV performance is based on 10 times 5 fold CV for the inner CV (model tuning) and 5 times 5 fold CV for the outer CV (performance estimation).

To see if there are any predictive biomarkers that are independent of ER status, we separated the dataset into an ER-positive (n=80) set and an ER-negative (n=50) set. Applying procedure '***findImportantGenes***' to the *ER* positive set returned no significant genes, which was probably due to the small number of pCR cases (6 out of 80) in this cohort that has made the analysis underpowered. Applying the same procedure to the ER-negative set only returned four probe sets: three probe sets for the gene hydroxyprostaglandin dehydrogenase (*HPGD*), a prostaglandin metabolizing enzyme and one probe set for the human gene *AK056707* with unknown function.

By studying the pattern of *HPGD* expression in the 50 ER-negative samples, we found that the discriminating pattern is stronger in triple negative (ER-/PR-/HER2-) samples (see Figure 2). This led us to search for other biomarkers that can predict *pCR* in triple negative samples (n=26, 13 are *pCR*). By applying procedure '***findImportantGenes***', we were able to generate a number of significant models using the iterative procedure. In total, 70 probe sets were discovered. The top 31 of the 70 probe sets are listed in Table 1.

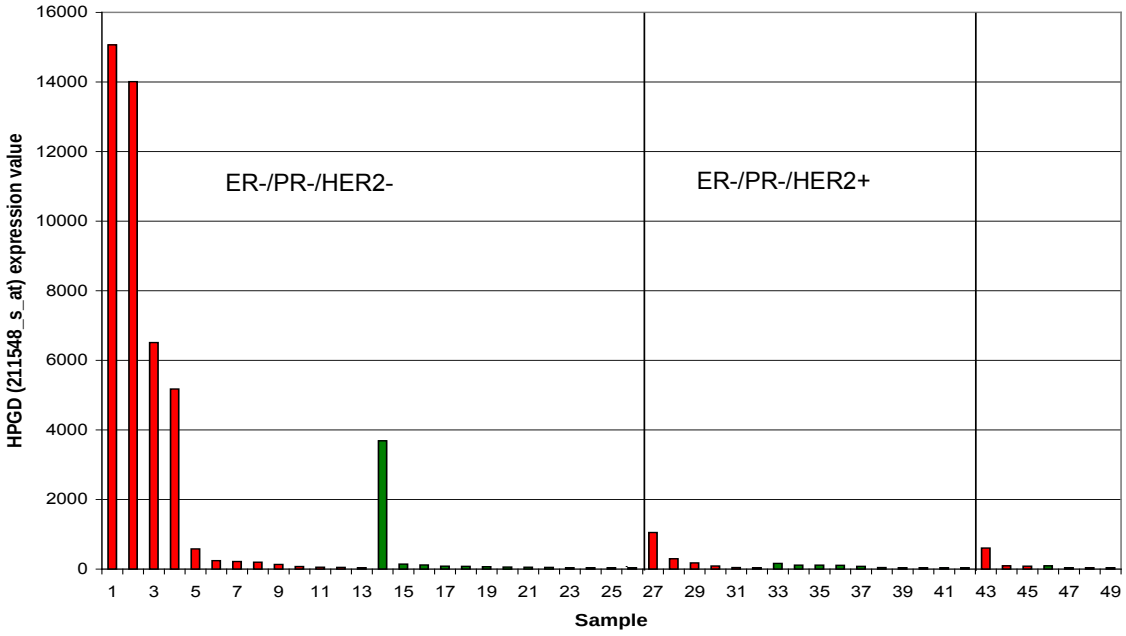

Figure 2: Gene expression values of HPGD (211548_s_at) in ER-negative patients. This graph partitions the ER- patients into three subgroups: ER-/PR-/HER2-, ER-/PR-/HER2+ and the rest. Non-pCR samples are shown in red bars and pCR samples are shown in green bars.

3.2 Biomarker validation using blinded validation set

We first validate the prediction performance (AUC) of our 16 probe set signature trained using the whole training set. The blinded validation AUC performance is 0.74. Although the performance is quite satisfactory, similar performance can be achieved by using the clinical variable ER alone or the ESR1 gene (probe set ID: *205225_at*) alone. From the results of different data analysis teams of the MAQC project [6], we also found that no team can generate models that outperform the ER gene, which means that gene signatures have little added benefit for predicting chemotherapeutic response if the ER-positive and ER-negative patients are analyzed together.

Within the 100 blinded validation samples, 30 samples were from triple negative patients. We used this subset to validate the genes discovered from the 26 triple negative samples of the training set. The result (in the last column of Table 1) shows that 19 of the 31 probe sets have reasonably good predictive power individually (*AUC* from 0.65 to 0.85). Twelve of them have low or no predictive power on this validation set (*AUC < 0.65*), including *HPGD* and *AK056707*. After checking the expressions of the *HPGD* gene of the validation set, we found that mean expression level is many times smaller than that of the training samples. We suspect that there may be some data quality issue in the validation set.

Table 1 Candidate markers identified in triple negative breast cancers for predicting chemotherapeutic response. The probe sets that can individually achieve reasonable validation performance (AUC > 0.65) are listed in bold face. The genes listed in the top portion of the table have higher average expression values in the non-pCR group; the genes listed in the bottom portion of the table have higher average expression values in the pCR group.

Gene symbol	*Probe set ID*	*Description*	*Independent validation AUC (std. err)*
C4A	**208451_s_at**	**complement component 4A**	**0.85 (0.09)**
C4A	**214428_x_at**	**complement component 4A**	**0.78 (0.10)**
SERHL2	**217284_x_at**	**serine hydrolase-like 2**	**0.77 (0.10)**
SERHL2	217276_x_at	serine hydrolase-like 2	0.59 (0.12)
COL1A2	**202404_s_at**	**collagen, type I, alpha 2**	**0.69 (0.11)**
COL1A2	**202403_s_at**	**collagen, type I, alpha 2**	**0.69 (0.11)**
CRAT	**209522_s_at**	**carnitine acetyltransferase**	**0.69 (0.11)**
CRAT	**205843_x_at**	**carnitine acetyltransferase**	**0.67 (0.11)**
DHRS2	**206463_s_at**	**dehydrogenase/reductase (SDR family) member 2**	**0.65 (0.11)**
AKR1C2	209699_x_at	aldo-keto reductase family 1, member C2	0.63 (0.12)
AKR1C2	211653_x_at	aldo-keto reductase family 1, member C2	0.56 (0.12)
HPGD	211548_s_at	hydroxyprostaglandin dehydrogenase 15-(NAD)	0.62 (0.12)
UGT2B28	211682_x_at	UDP glycosyltransferase 2 family, polypeptide B28	0.60 (0.12)
BUCS1	215432_at	butyryl Coenzyme A synthetase 1	0.52 (0.12)
ALDH3B2	204942_s_at	aldehyde dehydrogenase 3 family, member B2	0.48 (0.12)
ROPN1B	**220425_x_at**	**ropporin, rhophilin associated protein 1B**	**0.80 (0.10)**
EPHB3	**1438_at**	**EPH receptor B3**	**0.78 (0.10)**
BCL11A	**219498_s_at**	**B-cell CLL/lymphoma 11A (zinc finger protein)**	**0.78 (0.10)**
MFGE8	**210605_s_at**	**milk fat globule-EGF factor 8 protein**	**0.73 (0.11)**
TM4SF1	**209387_s_at**	**transmembrane 4 L six family member 1**	**0.72 (0.11)**
TM4SF1	**209386_at**	**transmembrane 4 L six family member 1**	**0.69 (0.11)**
TM4SF1	**215034_s_at**	**transmembrane 4 L six family member 1**	**0.68 (0.11)**
ANP32E	**221505_at**	**acidic (leucine-rich) nuclear phosphoprotein 32 family, member E**	**0.71 (0.11)**
S100A1	**205334_at**	**S100 calcium binding protein A1**	**0.70 (0.11)**
PDK1	**206686_at**	**pyruvate dehydrogenase kinase, isoenzyme 1**	**0.69 (0.11)**
ART3	**210147_at**	**ADP-ribosyltransferase 3**	**0.66 (0.11)**
TNFRSF21	218856_at	tumor necrosis factor receptor superfamily, member 21	0.60 (0.12)
SLC26A2	205097_at	solute carrier family 26 (sulfate transporter), member 2	0.58 (0.12)
TNPO3	214550_s_at	transportin 3	0.58 (0.12)
GAB2	203853_s_at	GRB2-associated binding protein 2	0.56 (0.12)
AK056707	212553_at	KIAA0460 protein	0.53 (0.12)

4 Discussion

Our tool performs feature selection by searching for the best trade off between mean difference test and t-test. The search is guided by cross validated model performance. To make the search effective, it is crucial that the cross-validation is done in an unbiased manner. For instance:

1. For model tuning, the feature selection must be performed within each run of cross-validation.
2. The cross-validation performance of the model tuning step cannot be used as an estimation of model performance. For estimating unbiased model performance, nested cross-validation must be performed. That is, the inner cross-validation for model tuning, the outer cross-validation for performance estimation. For stability, the inner cross-validation and outer cross-validation may need to be run multiple times.

Considering the computation cost of proper cross validation, we use *DLDA* classifier as our modelling tool, which is a simple linear classifier that has no parameters to be tuned. It is been shown that *DLDA* performance very well compared to other more complicated classifiers [8]. We believe the best way to gain better performance is through improving performance of feature selection, rather than tuning modeling parameters of complex models. Complicated learning schemes can make proper cross-validation too computational expensive to run.

Using the whole chemotherapy response data, we showed that our tool can effectively identify the most important biomarker for predicting chemotherapy response, i.e., the ER gene. We also showed that without grouping the patients into homogeneous subpopulations based on known clinical variable, it is hard to find gene signatures that have added clinical benefit over the clinical variable ER.

By checking the gene expression pattern of a gene (*HPGD*) selected from the ER-negative patients, we identified a clinically homogeneous set of patients (ER-/PR-/HER2-) whose chemotherapeutic response can be reasonably predicted. A set of genes were identified from this subpopulation. Many of these genes are shown to have good predictive power individually based on blinded validation data.

References

[1] V. G. Tusher, R. Tibshirani and G. Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences USA*, 98(9):5116–5121, 2001.

[2] B. Efron, R.Tibshirani, J. D.Storey and V.Tusher. Empirical Bayes Analysis of a Microarray Experiment. *Journal of the American Statistical Society*, 96(456):1151–1160, 2001

[3] R. Opgen-Rhein, and K. Strimmer. Accurate Ranking of Differentially Expressed Genes by a Distribution-Free Shrinkage Approach. *Statistical Applications in Genetics and Molecular Biology*, 6(1): 9, 2007.

[4] R. Dudoit, J. Fridly and T. P. Speed. Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, 97(457):77–87, 2002.

[5] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield and E. S. Lander. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science*, 286(5439):531–537, 1999.

[6] MAQC Consortium. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation. *Nature Biotechnology*, 28(8):827–838, 2010

[7] K. R. Hess, K. Anderson, W. F. Symmans, V. Valero, N. Ibrahim, J. A. Mejia, D. Booser, R. L. Theriault, A. U. Buzdar, P. J. Dempsey, R. Rouzier, N. Sneige, J. S. Ross, T. Vidaurre, H. L. Gómez, G. N. Hortobagyi and L. Pusztai. Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer. *Journal of Clinical Oncology*, 24(26):4236–4244, 2006.

[8] Y. Saeys, I. Inza and P. Larranaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.