Journal of Statistical Software

July 2014, Volume 59, Issue 1. http://www.jstatsoft.org/

cancerclass: An R Package for Development and Validation of Diagnostic Tests from High-Dimensional Molecular Data

Jan Budczies
Charité Hospital, Berlin

Daniel Kosztyla
Olgahospital, Stuttgart

Christian von Törne
Siemens Healthcare Diagnostics

Albrecht Stenzinger
University Hospital Heidelberg

Silvia Darb-Esfahani
Charité Hospital, Berlin

Manfred Dietel
Charité Hospital, Berlin

Carsten Denkert
Charité Hospital, Berlin

Abstract

Progress in molecular high-throughput techniques has led to the opportunity of a comprehensive monitoring of biomolecules in medical samples. In the era of personalized medicine, these data form the basis for the development of diagnostic, prognostic and predictive tests for cancer. Because of the high number of features that are measured simultaneously in a relatively low number of samples, supervised learning approaches are sensitive to overfitting and performance overestimation. Bioinformatic methods were developed to cope with these problems including control of accuaracy and precision. However, there is demand for easy-to-use software that integrates methods for classifier construction, performance assessment and development of diagnostic tests. To contribute to filling of this gap, we developed a comprehensive R package for the development and validation of diagnostic tests from high-dimensional molecular data. An important focus of the package is a careful validation of the classification results. To this end, we implemented an extended version of the multiple random validation protocol, a validation method that was introduced before. The package includes methods for continuous prediction scores. This is important in a clinical setting, because scores can be converted to probabilities and help to distinguish between clear-cut and borderline classification results. The functionality of the package is illustrated by the analysis of two cancer microarray data sets.

Keywords: diagnostic test, molecular data, microarrays, supervised learning, classification, R.

1. Introduction

Ongoing progress in the molecular analysis techniques for biological samples has led to the advancement of high-throughput methods for genomics, transcriptomics, proteomics, metabolomics and other -omics fields (Ellis *et al.* 2007; Service 2008; Shi *et al.* 2006; Stratton *et al.* 2009). Microarrays, for example, allow the simultaneous monitoring of virtually all genes that may be expressed in a tumor sample. Data sets generated by high-throughput methods are typically high-dimensional in the number of molecules (some hundreds to some ten thousands), but often include only a limited number of samples (some tens to some hundreds).

Since the early days of microarrays, much attention has been paid to the problem of class prediction by gene expression profiles. In 1999, a landmark study on gene expression of haematological malignancies showed discrimination of acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) with high precision (Golub *et al.* 1999). More recent studies focus on the identification of molecular signatures that are prognostic for disease outcome in breast, lung, colorectal, and others cancers (Denkert *et al.* 2009; Fritzmann *et al.* 2009; Chen *et al.* 2007; Roepman *et al.* 2005; Van de Vijver *et al.* 2002). Further molecular signatures were identified that predict the response to therapies, for example the success of neoadjuvant treatment for breast cancer (Ayers *et al.* 2004; Farmer *et al.* 2009; Liedtke *et al.* 2009). The aim of such investigations is the development of molecular tests that refine cancer stratification and lead to individual, patient-tailored therapies with effective treatment and a minimum of adverse side effects from the therapy. For breast cancer, gene expression based prognostic tests are already used in clinical practice. Two commercially available assays, Mammaprint® and OncotypeDX®, are currently being validated in large clinical trials (Sparano 2006; Cardoso *et al.* 2007). More recently, Endopredict® was developed and shown to be reproducible between molecular pathology laboratories (Filipits *et al.* 2011; Denkert *et al.* 2012).

The methodical challenges associated with class prediction from high-dimensional data include (i) selection of the features (genes, proteins, metabolites, ...) relevant for the problem and (ii) avoidance of overfitting and overestimation of results by using appropriate methods for classifier construction and error estimation. Feature selection identifies the informative markers out of thousands or ten-thousands of measurements. As a consequence, the signal-to-noise ratio in the data is increased, leading to more precise test results. Further, a molecular test that is sparse in the number of features is easier and cheaper to implement.

The critical point for correct interpretation of a classification study is an appropriate methodology for error estimation and validation (Lottaz *et al.* 2008). Many molecular studies apply cross-validation and in particular the leave-one-out approach for estimation of the error rate. This approach gives a nearly unbiased estimate of the errors rate, but can be imprecise: it has a high variability and can produce large outliers (Braga-Neto and Dougherty 2004; Braga-Neto *et al.* 2004; Efron and Tibshirani 1997). In molecular studies, often an optimal classifier is selected after comparing the performance of a large number of classifiers that include different sets of genes. Combination of such a classifier optimization procedure with cross-validation suffers from the variance of error rates and can lead to overfitting and overestimation of the classifier performance. Another common validation method is based on a fixed split of the data in training and test sets.

However, a recent meta-analysis of studies on cancer outcome has shown a critical dependence of classification results on the split of the data in training and test sets (Michiels *et al.* 2005).

In this work, a new multiple random validation strategy was introduced that is based on repeated drawing of training sets of different size. In n-fold cross-validation of N samples, the size of the training set is $(1 - 1/n)N$. Different than in cross-validation protocols where the training set size is fixed, in the multiple random validation protocol the performance is studied for a multitude of training sets of different size. In this way, the multiple random protocol can be considered as generalization of cross-validation. Studying the classification performance for several training set sizes can help to detect how many samples are necessary for good or for optimal classification results. Therefore, we included multiple random validation in the software.

Many of the classification methods that are common in the field of machine learning have been applied to high-dimensional molecular data. These include nearest-centroid-classification, linear discriminant analysis, classification trees, and more complex methods such as support vector machines, neural networks, or partial least squares. Many of the classification methods are available in R (R Core Team 2014), a programming language and statistical environment shown to be extremely convenient for the analysis of high-dimensional molecular data. It turned out that good results can often be obtained by simple methods, such as nearest-centroid-classification. Indeed, simple classifiers have been shown to perform equally or better in direct comparison to more complex methods (Wessels *et al.* 2005; Dudoit *et al.* 2002). In recent years, methods for pattern recognition have been adapted to the analysis of high-dimensional molecular data. For example, nearest shrunken centroid classification (Tibshirani *et al.* 2002) combines a soft thresholding method for feature selection with nearest centroid classification. However, there is need for a simple to use software for predictor optimization, reliable estimation of error rates and trade-off between sensitivity and specificity in a uniform framework.

The MicroArray Quality Control (MAQC)-II (Shi *et al.* 2010) study showed that, for generation of predictive models, good modeling practice was more important than the actual choice of a particular algorithm. To contribute to the availability of state-of-the-art modeling tools to the biomedical community, we developed **cancerclass** (Budczies and Kosztyla 2011a), a comprehensive R package for development and validation of diagnostic tests. An important focus of the package is a careful validation of the classification results and it therefore includes the recently introduced multiple random validation protocol. Multiple random validation can be considered as more comprehensive than cross-validation, because it allows to study the prediction accuracy in dependence of the training set size. A second important focus of the package is inclusion of continuous prediction scores. Continuous scores are important in a clinical setting, because they are convertible to class membership probabilities and help to distinguish between clear-cut and borderline classification results. The functionality of the package is illustrated by the analysis of two cancer microarray data sets.

2. Implementation

cancerclass is freely available from the **Bioconductor** repository (`http://www.Bioconductor.org/`).

2.1. Architecture

Table 1 gives an overview of the main classes and methods. The package was implemented in

Prediction or validation method	Description	Input class(es)	Output class	Plot methods for output class
`validate()`	Construction and validation of predictors in dependence of the training set size.	'ExpressionSet'	'validation'	xy, genes, samples
`nvalidate()`	Construction and validation of predictors in dependence of the number of genes.	'ExpressionSet'	'nvalidation'	xy, genes, samples
`fit()`	Construction of a predictor from a training data set.	'ExpressionSet'	'predictor'	genes
`predict()`	Calculation of continuous prediction scores in a test data set using a predictor.	'ExpressionSet', 'predictor'	'prediction'	histogram, curves, roc, logistic, samples
`loo()`	Combination of `fit()` and `predict()` using leave-one-out cross-validation.	'ExpressionSet'	'prediction'	histogram, curves, roc, logistic, samples

Table 1: Methods for predictor construction and validation. All methods need an object of the class 'ExpressionSet' as input that includes an expression data set and information about class membership, for example clinical outcome. The functions `validate()` and `nvalidate()` perform classification in a multiple random validation protocol. The functions `fit()` and `validate()` implement training and validation using a fixed split in training and test data. The function `loo()` performs leave-one-out cross-validation. For most of the output classes, there are several `plot` methods that can be selected by the parameter `type`.

an object-orientated programming style. Most of the functions have instances of S4 classes as input and output objects. S4 is the up-to-date method for object-oriented programming in R, described and implemented in the package **methods** (R Core Team 2014).

The molecular data are stored in 'ExpressionSet' objects, as they are defined in the R package **Biobase** (Gentleman *et al.* 2004). These objects contain a matrix of expression data integrated with phenotype and array probe annotations. For the example data sets delivered with the package (see below), the data frame describing the phenotype contains a column `"class"`, a binary variable that is used for classification. Most of the output objects can be visualized using the generic function `plot` that has specialized methods for the results of a classification task. Often, the user can select between different kinds of plots by an additional parameter `type` that is passed to the `plot` function.

2.2. Classification protocol

The classification protocol includes the following three steps: feature selection, predictor construction using the nearest centroid method, and validation. Various statistics are available

to rank the features and to select a set of top array probes. These include fold change, t test, Wilcoxon test, or different versions of outlier statistics like os, ort, and copa (Tibshirani and Hastie 2007; Wu 2007). Feature selection is always conducted using only the training data, and not the complete data. This procedure is known to be necessary to obtain reliable misclassification rates (Simon *et al.* 2003; Ambroise and McLachlan 2002; West *et al.* 2001). For high-dimensional data, feature selection is the computational most demanding part of the classification procedure. Therefore, we implemented this step using external C code that is integrated into the package using the `.C` interface.

The implemented methods for validation include the multiple random validation protocol (functions `validate` and `nvalidate`) that was introduced in Michiels *et al.* (2005). Following this protocol, a large number of training sets of different size is drawn. For each of these training sets, a predictor is constructed and evaluated on the remaining samples. The protocol works with balanced training data sets that contain an equal number of patients from each of the classes. For a fixed size of training data sets, a mean misclassification rate and confidence intervals are estimated from the distribution of the classification results. Further, we have implemented functions for validation using a fixed training-test split (`fit` and `validate`) as well as leave-one-out cross-validation (`loo`).

2.3. Nearest centroid classification

Classification is done with the nearest centroid method: First, the centroids in the space of selected features are calculated for both classes. Then, the distance of the test sample to the centroids is calculated and the test sample is predicted to belong to the class with the nearest centroid. **cancerclass** offers four different similarity measures for calculation of the distance that can be specified by `dist = "euclidean", "center", "angle", "cor"`. All four similarity measures can be derived from the euclidean metric by centering and/or scaling of centroids and test sample before calculating the distance. In detail, starting from an expression vector x, the centered, scaled and the centered-scaled vectors are denoted by $x_c = x - \bar{x}$, $x_s = x/\|x\|$ and $x_{cs} = (x - \bar{x})/\|x - \bar{x}\|$. Then, the four similarity measures are defined by

$$d_{euclidean}(x, y) = \|x - y\|, \tag{1}$$

$$d_{center}(x, y) = \|x_s - y_s\|, \tag{2}$$

$$d_{angle}(x, y) = \|x_c - y_c\| = \sqrt{1 - \cos \text{angle}(x, y)}, \tag{3}$$

$$d_{cor}(x, y) = \|x_{cs} - y_{cs}\| = \sqrt{1 - \text{cor}(x, y)}. \tag{4}$$

As stated above, the last two similarity measures can be expressed in terms of the geometric angle or the Pearson correlation coefficient, respectively.

2.4. Continuous prediction score

In context of clinical applications, it can be advantageous to work with a continuous prediction score rather than a clear-cut answer "yes" or "no" for class membership. **cancerclass** offers three different continuous prediction scores z, ζ and ξ. When c_1 and c_2 are the centroids of the two classes and x is a test sample,

$$z = \frac{d(x, c_1) - d(x, c_2)}{d(c_1, c_2)} \tag{5}$$

is the normalized difference between the distances of the test sample to the centroids of the two classes,

$$\zeta = \left(x - \frac{1}{2}(c_1 + c_2)\right) \cdot \left(\frac{c_2 - c_1}{\|c_2 - c_1\|}\right) \tag{6}$$

is the projection of the test sample on the difference vector between the centroids of the two classes, and

$$\xi = \log_2 \frac{d(x, c_1)}{d(x, c_2)} \tag{7}$$

is the logarithm of the ratio of the distances of the test sample and the centroids of the two classes. All three scores are negative, if the test sample is closer to the centroid of class 1 than class 2, and positive if this is not so.

2.5. Data sets

cancerclass is delivered with an example data set of leukemia. GOLUB (72 patients, 7129 genes) includes gene expression data of blood and bone marrow of AML and ALL patients (Golub *et al.* 1999). Gene expression data of breast cancer tissues are available from an additional R package **cancerdata** (Budczies and Kosztyla 2011b). VEER (78 patients, 24481 genes) includes the training data set of Mammaprint® classifier for breast cancer (Van't Veer *et al.* 2002). VIJVER (295 patients, 24481 genes) is a larger breast cancer expression data set that includes some of the samples of the earlier series (Van de Vijver *et al.* 2002). VEER1 and VIJVER1 (4948 genes) are versions of the above data sets after unsupervised filtering of genes as described in Van't Veer *et al.* (2002). All patients of the breast cancer studies underwent surgery as initial treatment. As in Van de Vijver *et al.* (2002), patients are divided in a good prognosis group (NODM) that did not receive adjuvant systemic therapy and remained free of distant metastases for at least 5 years and a poor prognosis group (DM) that received adjuvant systemic therapy or not and developed distant metastases within 5 years.

2.6. Statistical methods

The following statistical methods are applied during generation of the plots: Confidence intervals for proportions are estimated using Wilson's method as implemented by the function `binom.confint()` from the R package **binom** (Dorai-Raj 2014). To convert the continuous prediction score to probabilities, logistic regression is executed using the function `glm()` from the R package **stats** (R Core Team 2014).

3. Results

In the examples below, we show how the expression data sets GOLUB and VIJVER can be analyzed using the methods `validate`, `nvalidate`, `loo`, `fit` and `predict` of the package **cancerclass**. A short description of the methods is given in Table 1.

3.1. Performance of predictors in dependence of the number of genes

The function `nvalidate()` allows studying the performance of predictors in dependence of the numbers of genes. For each number of genes, multiple splits in training and test sets

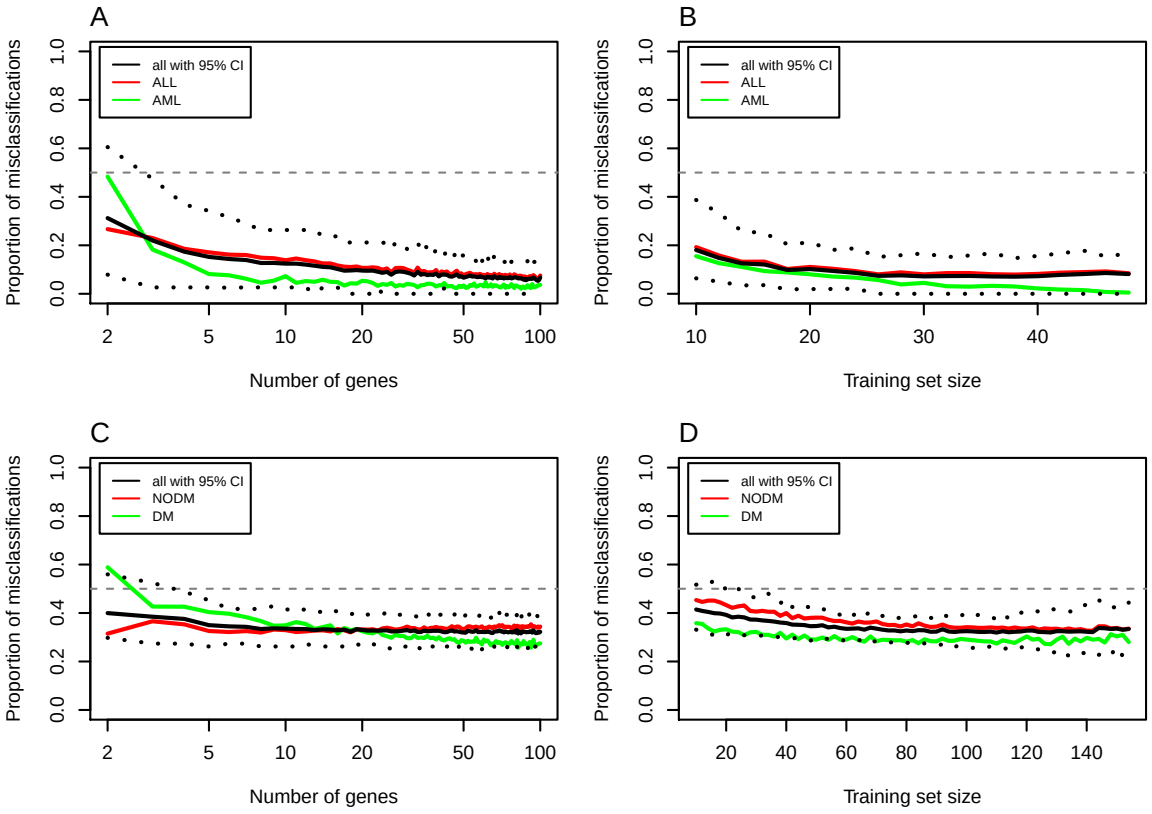

Figure 1: Performance of nearest centroid predictors in dependence of the number of genes and the training set size. Percentage of misclassifications for all patients (with two-sided 90% confidence intervals, black dotted lines) as well as individually for each of the classes (red and green lines). Prediction of the leukemia subtype using the GOLUB expression data (A−B). Prediction of breast cancer prognosis (NODM = no distant metastasis within five years, DM = distant metastasis within 5 years) using the VIJVER expression data (C−D). Classification performance in dependence of the number of genes in the predictor (A+C). Learning curves showing the improvement of 50-gene-predictors with growing training set size (B+D). For each number of genes and each training set size, misclassification rates and confidence intervals were estimated from 200 randomly drawn training sets.

are drawn randomly. For each split in training and test data set, the top genes that are included in the predictor are selected by Welch's t test. By the commands below, predictors are constructed training sets including 2/3 of the patients and evaluated in test sets including the remaining 1/3 of patients:

```
R> library("cancerclass")
R> data("GOLUB", package = "cancerclass")
R> nvalidation <- nvalidate(GOLUB, ngenes = 2:100, ntrain = "balanced",
+    method = "welch.test", dist = "cor")
R> plot(nvalidation, type = "xy")
```

While this training set size and balanced training sets are the default, the number of patients

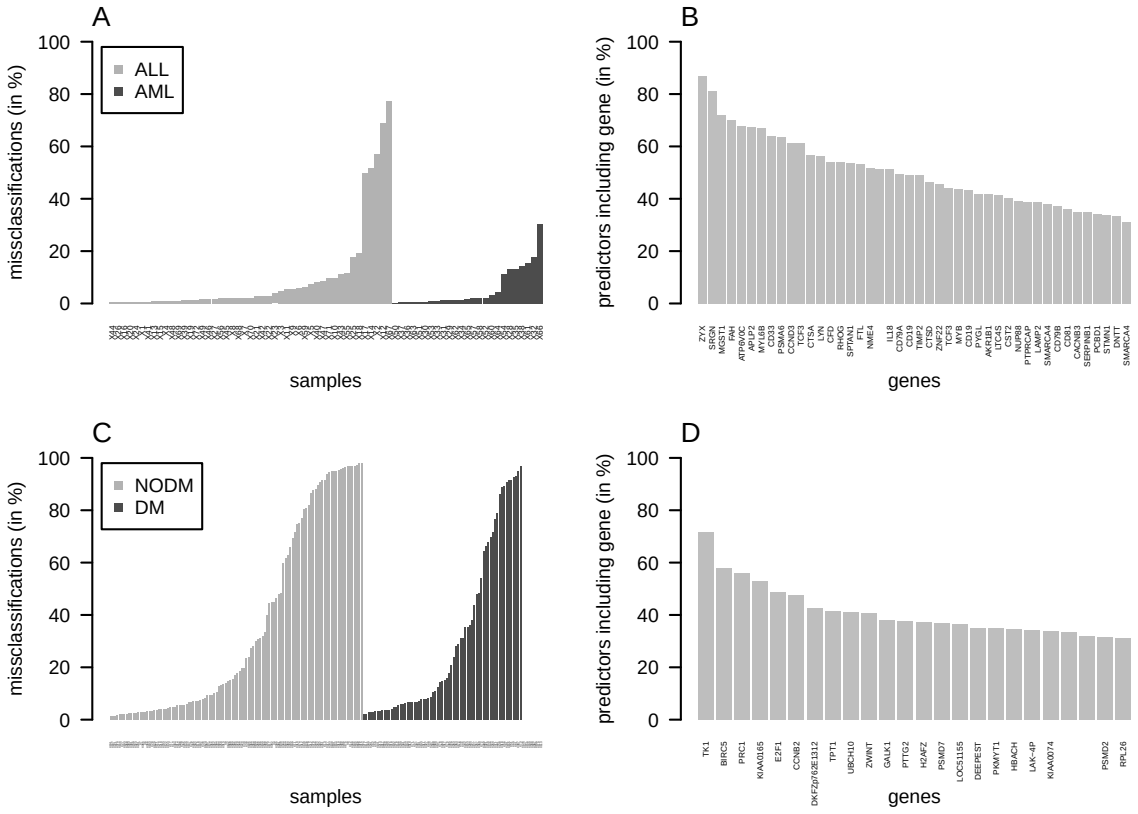

Figure 2: Detailed analysis of 50-gene-predictors. Prediction of the leukemia subtype using the GOLUB expression data (A−B). Prediction of breast cancer prognosis using the VIJVER expression data (C−D). Performance of predictors for each of the patients with the height of bars indicating the percentage of predictors that fail in classifying a particular patient (A+C). The top genes being most frequently included in the predictor with the height of the bars indicating the percentage of predictors that include a particular gene (B+D). For each training size set ranging from ten to the maximal possible size, predictors were trained in 200 randomly drawn training sets.

of the two classes in the training sets can be explicitly specified by the parameter `ntrain`. Figures 1A and C show the misclassification rates in dependence of the number of genes in the predictor. Using such kind of diagrams, predictors can be optimized in such a way that they perform well and are as sparse as possible (in the number of features) at the same time.

3.2. Performance of predictors in dependence of the training set size

Next, we chose a fixed number of genes and study the performance of 50 gene predictors in dependence of the training set size. This is the original method of multiple random validation (Michiels *et al.* 2005).

```
R> validation <- validate(GOLUB, ngenes = 50, ntrain = "balanced",
+    method = "welch.test", dist = "cor")
R> plot(validation, type = "xy")
```

	Class 1 (real)	Class 2 (real)	
Class 1 (pred.)	TP	FP	PPV = TP / (TP + FP)
Class 2 (pred.)	FN	TN	NPV = TN / (TN + FN)
	sensitivity = TP / (TP + FN)	specificity = TN / (TN + FP)	

Table 2: Performance characteristics for a diagnostic test. The basic quantities in the 2×2 table are the number of true positives (TP), the number of false positives (FP), the number of true negatives (TN), and the number of false negatives (FN). Starting from there, sensitivity, specificity as well as positive predictive values (PPV) and negative predictive values (NPV) of the diagnostic test can be estimated.

Figures 1B and D show the misclassification rate in dependence of the number of patients in the training data set. The rates are calculated for all patients in the test data sets as well as separately for each of the two classes.

3.3. Detailed analysis of predictors

Two further visualization methods for the '`validation`' and '`nvalidation`' objects allow a more detailed analysis of the predictors:

```
R> plot(validation, type = "samples")
R> plot(validation, type = "genes", min.percent = 30)
```

Figures 2A and C show the performance of 50-gene-predictors individually for each of the patients. Classification of a patient can depend on the training set where a predictor is trained, or it can fail consistently for most or all predictors, independent of the training set. For the breast cancer study, prediction of prognosis was feasible for half of the patients consistently with $> 80\%$ of the predictors. However, almost all predictors failed for approximately a quarter of the patients. For the leukemia study, the predictors performed excellent for almost all patients, except for 5 ALL cases that were predicted incorrectly by more than 30% of the predictors.

Figures 2B and D show the genes that were selected in more than 25% of the predictors. For the leukemia study, 23 top genes were included in more than 50% of the predictors. For the study on breast cancer, 4 top genes were included in more than 50% of the predictors: thymidine kinase 1 (TK1), survivin (BIRC5), protein regulator of cytokinesis (PRC1), and separase (KIAA0165). In addition to binary class prediction, **cancerclass** has methods for the construction and validation of continuous prediction scores that are discussed in the next three subsections.

3.4. Continuous prediction score: Leave-one-out cross-validation

Using the GOLUB data, we perform leave-one-out cross-validation:

```
R> cv <- loo(GOLUB, positive = "ALL", method = "welch.test", dist = "cor")
R> plot(cv, type = "histogram", score = "zeta")
R> plot(cv, type = "curves", score = "zeta")
R> plot(cv, type = "roc", score = "zeta")
```

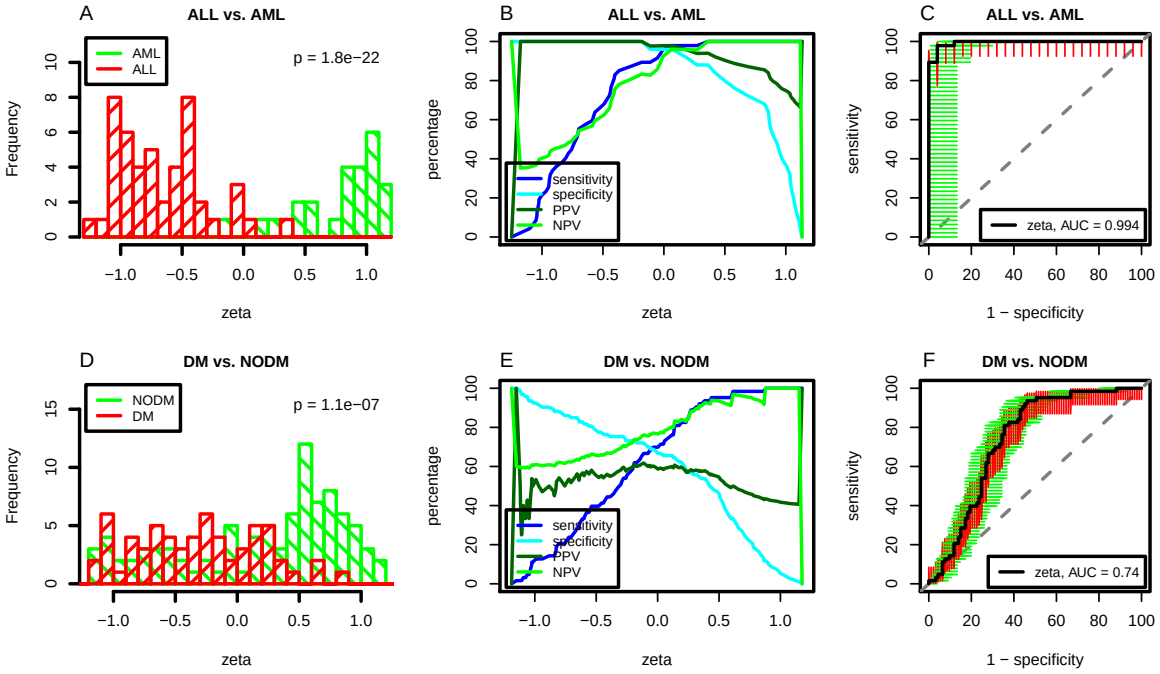

Figure 3: Performance of continuous 50-gene-predictors. The leukemia data were analyzed using leave-one-out cross-validation (A−C), the breast cancer data using a training-test approach (D−F). Histograms of the prediction score (A+D), sensitivity, specificity, PPV and NPV in dependence of the prediction score (B+E), ROC curves (C+F). The prediction score was significantly different between the two classes for both classification problems (histograms: p value from Welch's t test). In the ROC plots, 95% confidence intervals for sensitivity (red lines) and specificity (green lines) were calculated by Wilson's method.

The distribution of the prediction scores ζ for ALL and AML patients is shown in Figure 3A. An overview on the common accuracy measures for diagnostic tests is given in Table 2. Sensitivity and specificity are usually of primary interest, because they do not depend on the prevalence of the classes in a population. The parameter `positive`, here set to `"ALL"`, allows to select the class that we would like to detect. Figure 3B shows sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) in dependence on the cutoff value for the prediction score. Furthermore, **cancerclass** can calculate receiver operating characteristic (ROC) curves to visualize sensitivity and specificity of diagnostic tests (Figure 3C).

3.5. Continuous prediction score: Training and test set

Using the breast cancer data, we show how **cancerclass** works in a training-test setting. The predictor is trained on VEER1, the same training set that was used in Van't Veer *et al.* (2002). The test set comprises the patients of VIJVER1 that are not in VEER1:

```
R> library("cancerclass")
R> data("VEER1", package = "cancerdata")
R> data("VIJVER1", package = "cancerdata")
```

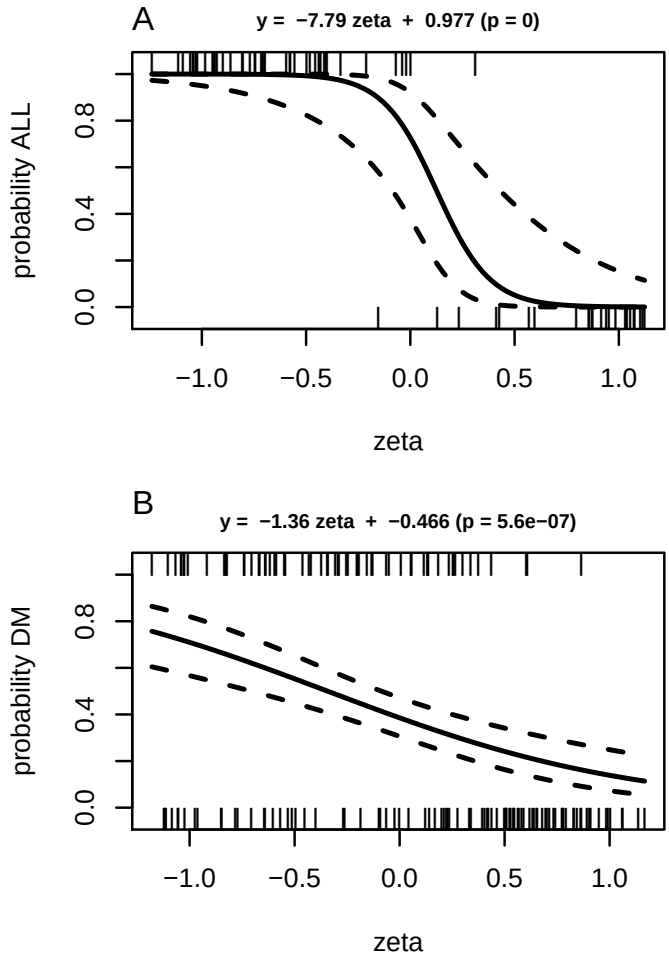

Figure 4: Probability of class membership calculated from continuous 50-gene predictors. The probability (with 95% confidence intervals) for membership in the poor prognosis class was estimated by logistic regression. Tracks at the top and the bottom of the graphics show the density of patients having a poor and good prognosis, respectively. Hazard for having ALL obtained from the GOLUB data (A). Hazard for developing a distant metastasis obtained from the VIJVER data (B).

```
R> VIJVER2 <- VIJVER1[, setdiff(sampleNames(VIJVER1), sampleNames(VEER1))]
R> predictor <- fit(VEER1, method = "welch.test")
R> prediction <- predict(predictor, VIJVER2, "DM", ngenes = 50, dist = "cor")
R> plot(prediction, type = "histogram", score = "zeta")
R> plot(prediction, type = "curves", score = "zeta")
R> plot(prediction, type = "roc", score = "zeta")
```

Again, the prediction score ζ and its correlation with the clinical outcome can be visualized in a histogram, plots of sensitivity and specificity, and a ROC curve (Figures 3D–F).

3.6. Conversion of prediction scores to risks

To assess the risk of a patient for a poorer course of the disease, it is desirable to convert the continuous prediction parameter into a probability for class membership. In case of the breast cancer study, this can be interpreted as hazard for a patient to develop a distant metastasis within 5 years. Having large data sets at hand, this hazard could be estimated by the incidence of distance metastases within a small interval around the value of ζ. However, usually this is not possible because the number of samples analyzed in high-throughput studies is small. For a limited number of patients, **cancerclass** allows the estimation of the hazard by logistic regression:

```
R> plot(cv, type = "logistic", score = "zeta")
```

Using this method, Figures 4A and B show the probability of ALL and the probability of a distance metastasis in dependence of ζ.

4. Discussion

The high-dimensionality of molecular data sets leads to methodic challenges for supervised learning. Overfitting should be avoided and misclassification rates should be estimated realistically in order to prevent an overoptimistic interpretation of results. **cancerclass** is a comprehensive R package for the development and validation of diagnostic tests from high-dimensional molecular data: Misclassification rates can be studied in dependence of the number of genes in the predictor and the size of the training set. ROC curves help to trade sensitivity off against specificity. As it is important for a realistic interpretation of results, misclassification rates, sensitivity and specificity are delivered as point estimates with confidence intervals.

A multitude of methods were developed to tackle the classification problem and many of them were applied to high-dimensional molecular data. The MicroArray Quality Control (MAQC)-II study (Shi *et al.* 2010) showed that the prediction performance depended on the proficiency of the analysis team, but not on the particular model choice, as many of the models performed equally. The models differed concerning algorithms and/or parameters and often simple methods performed as well as more complicated methods (Popovici *et al.* 2010). Good modeling practice appeared to be more important than the actual choice of a particular algorithm. In **cancerclass** we combined a simple classification algorithm with a comprehensive set of validation and visualization methods. In principle, it is possible to replace the nearest centroid classifier by any other classification algorithms, but this requires changes in the R code.

It is an advantage of nearest centroid classification that it can handle balanced and unbalanced data sets without any problems. Estimation of centroids remains unbiased in the unbalanced case. However, using balanced training data can be considered as optimal when the biological variance is unknown or known and comparable for both classes. If the biological variance is equal for both classes, both centroids are estimated with the same precision, if the training set is balanced. Thus, we used balanced training sets in the example studies. However, the methods of **cancerclass** are capable to handle arbitrary kinds of training sets.

As examples, we analyzed two expression data sets on human cancer that were published before. Since the appearance of Golub's work on leukemia (Golub *et al.* 1999), it is well-known

that AML and ALL can be distinguished using gene expression profiling with high accuracy of about 95%. Here, we could reproduce the results of the AML vs. ALL classification with sensitivity and specificity around 95%. The ROC curve obtained from continuous scores in leave-one-out cross-validation yields an almost perfect separation (AUC = 0.994).

As a second example, we studied the prediction of breast cancer prognosis by expression data in the tumor tissue. Using the multiple random validation protocol, we were able to obtain sensitivities and specificities for the prediction of distant metastasis in the range of 60%–70%. Van't Veer *et al.* (2002) published classification accuracies of 83% in a cohort of 78 nodal-negative patients in a leave-one-out cross-validation and of 89% in an independent validation set of 19 young node-negative patients (Van't Veer *et al.* 2002). However, in a larger validation series of 180 patients, a sensitivity of 93%, a specificity of 53% and an overall accuracy of 62% were obtained (Van de Vijver *et al.* 2002). Using **cancerclass** a 50-gene-predictor was trained on the same data set as in the original publication (Van't Veer *et al.* 2002). Within statistical variation, the results for the larger validation series agreed with the previously published results, as indicated by the course of the ROC curve (Figure 3F).

Interestingly, the top 4 genes that are included in most of the 50-gene-predictors constructed from the VIJVER data set are all known cancer markers: Thymidine kinase 1 (TK1) levels are elevated in the serum of patients with breast cancer (Carlsson *et al.* 2009) and other cancers (Chen *et al.* 2010). Further, TK1 is a prognostic marker for ovarian cancer (Fujiwaki *et al.* 2002). Survivin (BIRC5) is known as inhibitor of metastasis, overexpressed in almost all cancers, and associated with advanced disease, high grade, abbreviated survival, resistance to therapy, and accelerated recurrences (Andersen *et al.* 2007). Protein regulator of cytokinesis (PRC1) is overexpressed in breast cancer cells while it is undetectable in most of the normal human tissues (Shimo *et al.* 2007). Overexpression of separase (KIAA0165) can induce aneuploidy and mammary tumorigenesis (Zhang *et al.* 2008).

In summary, we observed a substantially different performance of molecular tests comparing the leukemia and the breast cancer study. This was expected, as AML and ALL represent genetically different diseases, while distant metastasis formation in breast cancer is a complex biological process potentially driven by multiple genetic, epigenetic and environmental factors. Thus, although gene expression monitoring contributes to its solution, metastasis prediction is a multifactorial problem that is not easy to solve.

5. Conclusions

We presented an R package for the development and validation of diagnostic tests from high-dimensional molecular data. **cancerclass** is freely available from the **Bioconductor** repository (`http://www.Bioconductor.org/`). Important foci of the implementation are methods for a careful validation that include the multiple random validation protocol and methods for calculation and visualization of continuous prediction scores. The latter methods, including ROC curves, are in particular helpful when translating a molecular test into clinics, where sensitivity and specificity need to be balanced by choosing a suitable cutoff for the prediction score.

Acknowledgments

This work was funded by the BMBF, grant #01ES0725 (NEO-PREDICT) and by the European Commission, FP7 grant #200327 (METAcancer).

References

Ambroise C, McLachlan GJ (2002). "Selection Bias in Gene Extraction on the Basis of Microarray Gene Expression Data." *Proceedings of the National Academy of Sciences of the United States of America*, **99**(10), 6562–6566.

Andersen MH, Svane IM, Becker JC, Straten PT (2007). "The Universal Character of the Tumor-Associated Antigen Survivin." *Clinical Cancer Research*, **13**(20), 5991–5994.

Ayers M, Symmans WF, Stec J, Damokosh AI, Clark E, Hess K, Lecocke M, Metivier J, Booser D, Ibrahim N, Valero V, Royce M, Arun B, Whitman G, Ross J, Sneige N, Hortobagyi GN, Pusztai L (2004). "Gene Expression Profiles Predict Complete Pathologic Response to Neoadjuvant Paclitaxel and Fluorouracil, Doxorubicin, and Cyclophosphamide Chemotherapy in Breast Cancer." *Journal of Clinical Oncology*, **22**(12), 2284–2293.

Braga-Neto U, Hashimoto R, Dougherty ER, Nguyen DV, Carroll RJ (2004). "Is Cross-Validation Better than Resubstitution for Ranking Genes?" *Bioinformatics*, **20**(2), 253–258.

Braga-Neto UM, Dougherty ER (2004). "Is Cross-Validation Valid for Small-Sample Microarray Classification?" *Bioinformatics*, **20**(3), 374–380.

Budczies J, Kosztyla D (2011a). ***cancerclass**: Development and Validation of Diagnostic Tests from High-Dimensional Molecular Data*. R package version 1.8.0, URL http://www.Bioconductor.org/packages/release/bioc/html/cancerclass.html.

Budczies J, Kosztyla D (2011b). ***cancerdata**: Development and Validation of Diagnostic Tests from High-Dimensional Molecular Data: Datasets*. R package version 1.2.0, URL http://www.Bioconductor.org/packages/release/data/experiment/html/cancerdata.html.

Cardoso F, Piccart-Gebhart M, Van't Veer L, Rutgers E (2007). "The MINDACT Trial: The First Prospective Clinical Validation of a Genomic Tool." *Molecular Oncology*, **1**(3), 246–251.

Carlsson L, Larsson A, Lindman H (2009). "Elevated Levels of Thymidine Kinase 1 Peptide in Serum from Patients with Breast Cancer." *Upsala Journal of Medical Sciences*, **114**(2), 116–120.

Chen HY, Yu SL, Chen CH, Chang GC, Chen CY, Yuan A, Cheng CL, Wang CH, Terng HJ, Kao SF, Chan WK, Li HN, Liu CC, Singh S, Chen WJ, Chen JJW, Yang PC (2007). "A Five-Gene Signature and Clinical Outcome in Non-Small-Cell Lung Cancer." *The New England Journal of Medicine*, **356**(1), 11–20.

Chen Y, Ying M, Chen Y, Hu M, Lin Y, Chen D, Li X, Zhang M, Yun X, Zhou J, He E, Skog S (2010). "Serum Thymidine Kinase 1 Correlates to Clinical Stages and Clinical Reactions and Monitors the Outcome of Therapy of 1,247 Cancer Patients in Routine Clinical Settings." *International Journal of Clinical Oncology*, **15**(4), 359–368.

Denkert C, Budczies J, Darb-Esfahani S, Györffy B, Sehouli J, Könsgen D, Zeillinger R, Weichert W, Noske A, Buckendahl AC, Müller BM, Dietel M, Lage H (2009). "A Prognostic Gene Expression Index in Ovarian Cancer – Validation Across Different Independent Data Sets." *Journal of Pathology*, **218**(2), 273–280.

Denkert C, Kronenwett R, Schlake W, Bohmann K, Penzel R, Weber KE, Höfler H, Lehmann U, Schirmacher P, Specht K, Rudas M, Kreipe HH, Schraml P, Schlake G, Bago-Horvath Z, Tiecke F, Varga Z, Moch H, Schmidt M, Prinzler J, Kerjaschki D, Sinn BV, Müller BM, Filipits M, Petry C, Dietel M (2012). "Decentral Gene Expression Analysis for ER+/HER2- Breast Cancer: Results of a Proficiency Testing Program for the EndoPredict Assay." *Virchows Archiv*, **460**(3), 251–259.

Dorai-Raj S (2014). ***binom**: Binomial Confidence Intervals For Several Parameterizations*. R package version 1.1-1, URL http://CRAN.R-project.org/package=binom.

Dudoit S, Fridlyand J, Speed TP (2002). "Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data." *Journal of the American Statistical Association*, **97**(457), 77–87.

Efron B, Tibshirani R (1997). "Improvements of Cross-Validation: The .632+ Bootstrap Method." *Journal of the American Statistical Assocoation*, **92**(438), 548–560.

Ellis DI, Dunn WB, Griffin JL, Allwood JW, Goodacre R (2007). "Metabolic Fingerprinting as a Diagnostic Tool." *Pharmacogenomics*, **8**(9), 1243–1266.

Farmer P, Bonnefoi H, Anderle P, Cameron D, Wirapati P, Becette V, André S, Piccart M, Campone M, Brain E, Macgrogan G, Petit T, Jassem J, Bibeau F, Blot E, Bogaerts J, Aguet M, Bergh J, Iggo R, Delorenzi M (2009). "A Stroma-Related Gene Signature Predicts Resistance to Neoadjuvant Chemotherapy in Breast Cancer." *Nature Medicine*, **15**(1), 68–74.

Filipits M, Rudas M, Jakesz R, Dubsky P, Fitzal F, Singer, F C, Dietze O, Greil R, Jelen A, Sevelda P, Freibauer C, Müller V, Jänicke F, Schmidt M, Kölbl H, Rody A, Kaufmann M, Schroth W, Brauch H, Schwab M, Fritz P, Weber KE, Feder IS, Hennig G, Kronenwett R, Gehrmann M, Gnant M (2011). "A New Molecular Predictor of Distant Recurrence in ER-Positive, HER2-Negative Breast Cancer Adds Independent Information to Conventional Clinical Risk Factors." *Clinical Cancer Research*, **17**(18), 6012–6020.

Fritzmann J, Morkel M, Besser D, Budczies J, Kosel F, Brembeck FH, Stein U, Fichtner I, Schlag PM, Birchmeier W (2009). "A Colorectal Cancer Expression Profile that Includes Transforming Growth Factor Beta Inhibitor BAMBI Predicts Metastatic Potential." *Gastroenterology*, **137**(1), 165–175.

Fujiwaki R, Hata K, Nakayama K, Moriyama M, Iwanari O, Katabuchi H, Okamura H, Sakai E, Miyazaki K (2002). "Thymidine Kinase in Epithelial Ovarian Cancer: Relationship with the other Pyrimidine Pathway Enzymes." *International Journal of Cancer*, **99**(3), 328–335.

Gentleman RC, Carey VJ, Bates DM, B BB, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, Hornik K, Hothorn T, Huber W, Iacus S, Irizarry R, Leisch F, Li C, Maechler M, Rossini AJ, Sawitzki G, Smith C, Smyth G, L LT, Yang JY, Zhang J (2004). "**Bioconductor**: Open Software Development for Computational Biology and Bioinformatics." *Genome Biology*, **5**, R80.

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, Bloomfield CD, Lander ES (1999). "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring." *Science*, **286**(5439), 531–537.

Liedtke C, Hatzis C, Symmans WF, Desmedt C, Haibe-Kains B, Valero V, Kuerer H, Hortobagyi GN, Piccart-Gebhart M, Sotiriou C, Pusztai L (2009). "Genomic Grade Index is Associated with Response to Chemotherapy in Patients with Breast Cancer." *Journal of Clinical Oncology*, **27**(19), 3185–3191.

Lottaz C, Kostka D, Markowetz F, Spang R (2008). "Computational Diagnostics with Gene Expression Profiles." In JM Keith (ed.), *Bioinformatics: Structure, Function and Applications*, volume 453 of *Methods in Molecular Biology*, pp. 281–296. Humana Press.

Michiels S, Koscielny S, Hill C (2005). "Prediction of Cancer Outcome with Microarrays: A Multiple Random Validation Strategy." *Lancet*, **365**(9458), 488–492.

Popovici V, Chen W, Gallas BG, Hatzis C, Shi W, Samuelson FW, Nikolsky Y, Tsyganova M, Ishkin A, Nikolskaya T, Hess KR, Valero V, Booser D, Delorenzi M, Hortobagyi GN, Shi L, Symmans WF, Pusztai L (2010). "Effect of Training-Sample Size and Classification Difficulty on the Accuracy of Genomic Predictors." *Breast Cancer Research*, **12**(1), R5.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Roepman P, Wessels LFA, Kettelarij N, Kemmeren P, Miles AJ, Lijnzaad P, Tilanus MGJ, Koole R, Hordijk GJ, Van der Vliet PC, Reinders MJT, Slootweg PJ, Holstege FCP (2005). "An Expression Profile for Diagnosis of Lymph Node Metastases from Primary Head and Neck Squamous Cell carcinomas." *Nature Genetics*, **37**(2), 182–186.

Service RF (2008). "Proteomics. Proteomics Ponders Prime Time." *Science*, **321**(5897), 1758–1761.

Shi L, Campbell G, Jones WD, Campagne F, Wen Z, Walker SJ, Su Z, Chu TM, Goodsaid FM, Pusztai L, Shaughnessy JD, Oberthuer A, Thomas RS, Paules RS, Fielden M, Barlogie B, Chen W, Du P, Fischer M, Furlanello C, Gallas BD, Ge X, Megherbi DB, Symmans WF, Wang MD, Zhang J, Bitter H, Brors B, Bushel PR, Bylesjo M, Chen M, Cheng J, Cheng J, Chou J, Davison TS, Delorenzi M, Deng Y, Devanarayan V, Dix DJ, Dopazo J, Dorff KC, Elloumi F, Fan J, Fan S, Fan X, Fang H, Gonzaludo N, Hess KR, Hong H, Huan J, Irizarry RA, Judson R, Juraeva D, Lababidi S, Lambert CG, Li L, Li Y, Li Z, Lin SM, Liu G, Lobenhofer EK, Luo J, Luo W, McCall MN, Nikolsky Y, Pennello GA, Perkins RG, Philip R, Popovici V, Price ND, Qian F, Scherer A, Shi T, Shi W, Sung J, Thierry-Mieg D, Thierry-Mieg J, Thodima V, Trygg J, Vishnuvajjala L, Wang SJ, Wu J, Wu Y, Xie Q, Yousef WA, Zhang L, Zhang X, Zhong S, Zhou Y, Zhu S, Arasappan

D, Bao W, Lucas AB, Berthold F, Brennan RJ, Buness A, Catalano JG, Chang C, Chen R, Cheng Y, Cui J, Czika W, Demichelis F, Deng X, Dosymbekov D, Eils R, Feng Y, Fostel J, Fulmer-Smentek S, Fuscoe JC, Gatto L, Ge W, Goldstein DR, Guo L, Halbert DN, Han J, Harris SC, Hatzis C, Herman D, Huang J, Jensen RV, Jiang R, Johnson CD, Jurman G, Kahlert Y, Khuder SA, Kohl M, Li J, Li L, Li M, Li QZ, Li S, Li Z, Liu J, Liu Y, Liu Z, Meng L, Madera M, Martinez-Murillo F, Medina I, Meehan J, Miclaus K, Moffitt RA, Montaner D, Mukherjee P, Mulligan GJ, Neville P, Nikolskaya T, Ning B, Page GP, Parker J, Parry RM, Peng X, Peterson RL, Phan JH, Quanz B, Ren Y, Riccadonna S, Roter AH, Samuelson FW, Schumacher MM, Shambaugh JD, Shi Q, Shippy R, Si S, Smalter A, Sotiriou C, Soukup M, Staedtler F, Steiner G, Stokes TH, Sun Q, Tan PY, Tang R, Tezak Z, Thorn B, Tsyganova M, Turpaz Y, Vega SC, Visintainer R, von Frese J, Wang C, Wang E, Wang J, Wang W, Westermann F, Willey JC, Woods M, Wu S, Xiao N, Xu J, Xu L, Yang L, Zeng X, Zhang J, Zhang L, Zhang M, Zhao C, Puri RK, Scherf U, Tong W, Wolfinger RD (2010). "The MicroArray Quality Control (MAQC)-II Study of Common Practices for the Development and Validation of Microarray-Based Predictive Models." *Nature Biotechnology*, **28**(8), 827–838.

Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen RV, Johnson CD, Lobenhofer EK, Puri RK, Schrf U, Thierry-Mieg J, Wang C, Wilson M, Wolber PK, Zhang L, Amur S, Bao W, Barbacioru CC, Lucas AB, Bertholet V, Boysen C, Bromley B, Brown D, Brunner A, Canales R, Cao XM, Cebula TA, Chen JJ, Cheng J, Chu TM, Chudin E, Corson J, Corton JC, Croner LJ, Davies C, Davison TS, Delenstarr G, Deng X, Dorris D, Eklund AC, hui Fan X, Fang H, Fulmer-Smentek S, Fuscoe JC, Gallagher K, Ge W, Guo L, Guo X, Hager J, Haje PK, Han J, Han T, Harbottle HC, Harris SC, Hatchwell E, Hauser CA, Hester S, Hong H, Hurban P, Jackson SA, Ji H, Knight CR, Kuo WP, LeClerc JE, Levy S, Li QZ, Liu C, Liu Y, Lombardi MJ, Ma Y, Magnuson SR, Maqsodi B, McDaniel T, Mei N, Myklebost O, Ning B, Novoradovskaya N, Orr MS, Osborn TW, Papallo A, Patterson TA, Perkins RG, Peters EH, Peterson R, Philips KL, Pine PS, Pusztai L, Qian F, Ren H, Rosen M, Rosenzweig BA, Samaha RR, Schena M, Schroth GP, Shchegrova S, Smith DD, Staedtler F, Su Z, Sun H, Szallasi Z, Tezak Z, Thierry-Mieg D, Thompson KL, Tikhonova I, Turpaz Y, Vallanat B, Van C, Walker SJ, Wang SJ, Wang Y, Wolfinger R, Wong A, Wu J, Xiao C, Xie Q, Xu J, Yang W, Zhang L, Zhong S, Zong Y, Slikker WJ (2006). "The MicroArray Quality Control (MAQC) Project Shows Inter- and Intraplatform Reproducibility of Gene Expression Measurements." *Nature Biotechnology*, **24**(9), 1151–1161.

Shimo A, Nishidate T, Ohta T, Fukuda M, Nakamura Y, Katagiri T (2007). "Elevated Expression of Protein Regulator of Cytokinesis 1, Involved in the Growth of Breast Cancer Cells." *Cancer Science*, **98**(2), 174–181.

Simon R, Radmacher MD, Dobbin K, McShane LM (2003). "Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification." *Journal of the National Cancer Institute*, **95**(1), 14–18.

Sparano JA (2006). "TAILORx: Trial Assigning Individualized Options for Treatment (Rx)." *Clinical Breast Cancer*, **7**(4), 347–350.

Stratton MR, Campbell PJ, Futreal PA (2009). "The Cancer Genome." *Nature*, **458**(7239), 719–724.

Tibshirani R, Hastie T (2007). "Outlier Sums for Differential Gene Expression Analysis." *Biostatistics*, **8**(1), 2–8.

Tibshirani R, Hastie T, Narasimhan B, Chu G (2002). "Diagnosis of Multiple Cancer Types by Shrunken Centroids of Gene Expression." *Proceedings of the National Academy of Sciences of the United States of America*, **99**(10), 6567–6572.

Van de Vijver MJ, He YD, Van't Veer LJ, Dai H, Hart AAM, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, Van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R (2002). "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer." *The New England Journal of Medicine*, **347**(25), 1999–2009.

Van't Veer LJ, Dai H, Van de Vijver MJ, He YD, Hart AAM, Mao M, Peterse HL, Van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH (2002). "Gene Expression Profiling Predicts Clinical Outcome of Breast Cancer." *Nature*, **415**(6871), 530–536.

Wessels LFA, Reinders MJT, Hart AAM, Veenman CJ, Dai H, He YD, Van't Veer LJ (2005). "A Protocol for Building and Evaluating Predictors of Disease State Based on Microarray Data." *Bioinformatics*, **21**(19), 3755–3762.

West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, Zuzan H, Olson JAJ, Marks JR, Nevins JR (2001). "Predicting the Clinical Status of Human Breast Cancer by Using Gene Expression Profiles." *Proceedings of the National Academy of Sciences of the United States of America*, **98**(20), 11462–11467.

Wu B (2007). "Cancer Outlier Differential Gene Expression Detection." *Biostatistics*, **8**(3), 566–575.

Zhang N, Ge G, Meyer R, Sethi S, Basu D, Pradhan S, Zhao YJ, Li XN, Cai WW, El-Naggar AK, Baladandayuthapani V, Kittrell FS, Rao PH, Medina D, Pati D (2008). "Overexpression of Separase Induces Aneuploidy and Mammary Tumorigenesis." *Proceedings of the National Academy of Sciences of the United States of America*, **105**(35), 13033–13038.

Affiliation:

Jan Budczies
Institute of Pathology

Charité – Universitätsmedizin Berlin
Charitéplatz 1,
10117 Berlin, Germany
E-mail: jan.budczies@charite.de

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume 59, Issue 1 *Submitted:* 2012-07-18
July 2014 *Accepted:* 2014-03-11