Reproducible Statistical Analysis in Microarray Profiling Studies

Ulrich Mansmann[1], Markus Ruschhaupt[2], and Wolfgang Huber[2]

[1] University of Heidelberg, Department for Medical Biometry and Informatics
INF 305, 69120 Heidelberg, Germany
mansmann@imbi.uni-heidelberg.de
[2] German Cancer Research Center, Division of Molecular Genome Analysis
INF 580, 69120 Heidelberg, Germany

Abstract. Reproducibility of calculations is a longstanding issue within the statistical community. Due to the complexity of the algorithms, the size of the data sets, and the limitations of the medium printed paper it is usually not possible to report all the minutiae of the data processing and statistical computations. Like the critical assessment of a mathematical proof it should be possible to check the software behind a complex data analysis. To achieve reproducible calculations and to offer an extensible computational framework the tool of a compendium is discussed.

1 Introduction

Microarray technology allows simultaneous measurement of thousands of transcripts within a homogeneous sample of cells [1]. It is of interest to relate these expression profiles to clinical phenotypes to improve the diagnosis of diseases and prognosis for individual patients [2]. A number of publications presented clinically promising results by combining this new kind of biological data with specifically designed algorithmic approaches. A selection out of these papers will be discussed [3–6] with respect to different aspects of reproducibility.

The most evident aspect of reproducibility is that of reproducing a calculation on the same data. Reproducing published results in the domain of microarray profiling studies is harder than it may seem. As example we look at a study of van ’t Veer et al. [3] which was reanalysed by Tibshirani and Efron [7]. Both state in their paper: *We re-analyzed the breast cancer data from van ’t Veer et al. ... Even with some help of the authors, we were unable to exactly reproduce this analysis.*

We do not know any example where classification results gained with one microarray technology and a special algorithm were reproduced using an alternative microarray platform and algorithm. Papers with diverging results on profiles for the prognosis of tumour recurrence for breast cancer patients are [3, 4]. How does this observation relate to the idea of a common underlying disease process? Should profiles have something common which are developed for the same disease? Is it of significance if they do not?

The confounding of algorithmic problems with biotechnology and biology creates a gordic knot. The paper applies the tool of a *compendium* as interactive strategy to settle the algorithmic backbone of a profiling study and to derive reproducible results

J. Dongarra, K. Madsen, and J. Waśniewski (Eds.): PARA 2004, LNCS 3732, pp. 939–948, 2005.

with a state-of-the-art methodology. A compendium is a document that bundles primary data, processing methods (computational code), derived data, and statistical output with the textual documentation and conclusions. It is interactive in the sense that it allows to modify the processing options, plug in new data, or insert further algorithms and visualisations.

2 Examples and Questions

2.1 Example 1

Van 't Veer et al. [3] classify breast cancer patients after curative resection with respect to the risk of tumour recurrence. The study includes 78 patients. Forty four patients had a *good prognosis* and did not suffer from a recurrence during the first 5 years after resection. Thirty four patients had a bad prognosis and experienced a recurrence during the first 5 years after resection. Agilent microarray technology was used to quantify the transcripts probed by 24,881 oligonucleotides. Additional prognostic factors like tumour grade, ER status, PR status, tumour size, patient age, angioinvasion were also documented. The authors were interested to develop a classifier based on the gene expression and to compare the relevance of the genomic signature to the prognostic value of standard clinical predictors. They used to following algorithm to establish the signature which contains 70 genes:

1. Starting with 24,881 genes, filtering on fold-change and a p-value criterion reduced the number of relevant genes to 4,936.
2. Based on an absolute correlation of at least 0.3 between gene expression and group indicator (0,1) a further reduction on 231 genes
3. Calculation of the 231 dimensional centroid vector for the 44 good prognosis cases.
4. Correlation of each case with this centroid is calculated, cut-off of 0.38 is chosen to exactly misclassify 3 with poor prognosis
5. Case is classified to good prognosis if correlation calculated for some number n of genes ($1 \leq n \leq 231$) with the centroid vector is ≥ 0.38, otherwise the case is classified to the *bad prognosis* group.
6. Starting with the top 5, 10, 15,... genes, classification procedure is carried out with leave-one-out cross-validation, to pick the optimal number of genes → 70

Based on this algorithm van 't Veer et al. achieved a correct classification for 26 of 44 patients without recurrence and for 31 of 34 with recurrence. Tibshirani and Efron [7] tried to reproduce this results and report: *Even with some help of the authors, we were unable to exactly reproduce this analysis*. In fact, the differences between both calculations were not dramatic. But, the example shows that algorithmic reproducibility is not a trivial issue and may have subjective elements. The algorithm is quite popular and is also used in [5] and other papers.

The van 't Veer examples rises the following questions: Why was a heuristic classification algorithm chosen and not a standard algorithm from machine learning? Are its computational aspects well understood? How important is the choise of the parameters for correct classification? Is the result easy to interpret? What justifies the popularity of the algorithm? Is the leave-one-out CV strategy appropriate? What is the effect of other CV strategies on the classification result?

2.2 Example 2

Huang et al. [4] investigated primary tumour samples from 52 patients with breast tumours and 1-3 positive lymph nodes. 18 patients had a recurrence within three years after surgery, and 34 patients did not. The authors concluded that they could predict tumour recurrence with misclassification rates of 2/34 and 3/18, respectively.

The authors presented a tree classifier with split decisions based on an a posteriori distribution over all possible splits. A description is available in the form of a technical report on the webpage of one of the authors. Due to ambiguities we did not succeed in translating the statistical ideas into software with which we could reproduce the analysis. More importantly, there is no *official* software version of the algorithm. The authors perform two dimension reduction steps before they apply the Bayesian tree classifier. First, the 12,625 probe sets on the HGU95Av2 Affymetrix GeneChips are reduced to 7,030 by excluding probe sets with a maximum intensities below 2^9 and a low variatiability across the samples. The second reduction creates 496 *metagenes* out of the 7,030 probe sets by performing k-means clustering and using the first principal component of each cluster as expression measure for the *metagene*.

As in example 1 this study is concerned with the prognosis of tumour recurrence of breast cancer patients after curative resection of the tumour. The authors state that they could not find any of the 70 genes used in the classifier of van ’t Veer et al. in any of the metagenes which come up in the Bayesian tree classifier.

The Huang example rises the following questions: Why is it impossible for us to reproduce the good classification result? Why is the classification based on the idiosyncratic method so good? The preprocessing is not part of the CV loop, which influence may this have on the misclassification rate? How is it possible to disentangle computation, technology, and biology? Can we find a link between the Huang and van ’t Veer classifiers?

3 The Compendium

Publications on microarray profiling studies often present one new microarray data set and one new classification method. Is it necessary to develop an idiosyncratic classification approach for each specific data set? Which classification result could be achieved with standard approaches? What loss in accuracy has to be traded for a rise in interpretability? How can I use new data to validate former results? If validation creates discrepancies how is it possible to assess the contribution made by algorithmic aspects: error in implementation or no success with validation? Do I have sufficient instructions and details to reproduce the method under validation in an exact way? How dependent is the classification result on the method used? What can be learned from a published profiling study for future projects? To answer these and other questions we introduce the compendium of computational diagnostic tools (CCDT)

3.1 Compendium for Computational Diagnostic Tools – CCDT

The compendium is an interactive document that calculates the misclassification rate (MCR) for different classification methods. The validation strategy is fixed but may be modified by setting specific parameters. Therefore, an outer and an inner cross-validation is mandatory: the inner CV tunes the algorithm specific parameters (including also parameters for the preprocessing) and the outer CV to estimate the misclassification rate. We see the preprocessing as part of the classification algorithm. The CV strategy is sketched in figure 1.

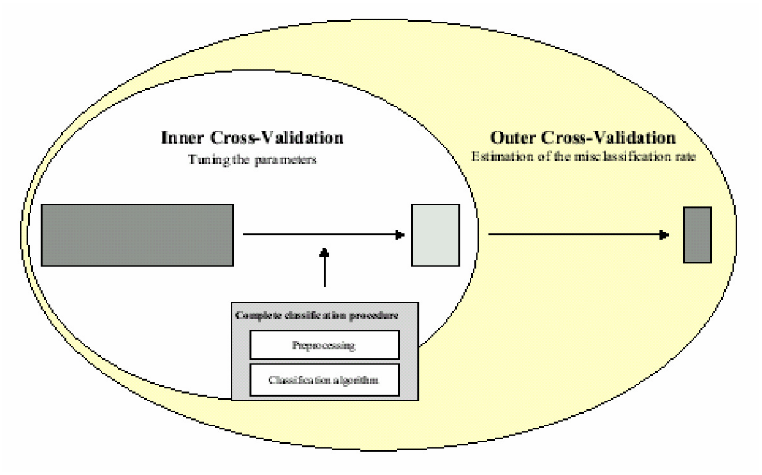


Fig. 1. The cross-validation stategy (CV - cross validation, MCR - missclassification rate)

Things that can be changed easily are: Classification methods, preprocessing steps, parameters for classification method and validation, and data set.

The compendium allows to combine guidelines with software, and to embed good statistical analysis in a text which is accessible for medical or biological researchers. It is a document that bundles primary data, processing methods (computational code), derived data, and statistical output with the textual documentation and conclusions it. Especially it contains specific tools to represent results. The inclusion of an implemented classification method is via a wrapper. Writing a wrapper function for new classification algorithms is simple. So far, five classification approaches are already implemented: Shrunken centroids (PAM) [18], support vector machines (SVM) [19], random forest [20], general partial least squares, and penalized logistic regression [21]. The compendium can be found as a package for the statistical language R at

`http://www.biometrie.uni-heidelberg.de/technical_reports.`

The compendium does not implement new algorithms but offers a framework to apply existing implementations of classification algorithms in a correct way.

3.2 Static Versus Interactive Approaches

The answer to questions like: *Which classification result could be archived with standard approaches?* or *What loss in accuracy has to be traded for a rise in interpretability?* is generally given in a paper which compares N different classification algorithms

(C) and M pre-processing (P) strategies on K different data sets (D). The N x M x K performance measures are tabulated and discussed. Dudoit et al. [12] published such a study which is extended by Lee et al. [13]. It is difficult to use their results for guidance because they certainly do not implement all algorithms of interest, the pre-processing strategies may change, and the data will become irrelevant when a new generation of microarrays is introduced. Therefore, it may be wise to replace the static by an interactive approach by offering the machinery which allows the researcher herself to perform such a study on the algorithms and pre-processing strategies of interest together with relevant data.

The compendium offers different levels of interactivity. It can be used to produce a textual output comparable with the static approach. On an intermediate level one interacts with the compendium by specifying parameters and data sets. For example, one could change the kernel of a support vector machine by simply changing the parameter `poss.pars`:

```
> poss.pars = c(list(cost = cost.range, kernel = "linear"),poss.k)
```

This is the level of sensitivity analyses or of comparing the performance of implemented algorithms on different data sets. The advanced level of interaction consists in introducing new ideas like new classification algorithms or new tools for the presentation of the results. Writing wrapper functions for new classification methods is simple. The following example shows a wrapper for diagonal discriminant analysis. The essential part is the specific definition of the predict.function by user specific needs and ideas:

```
> DLDA.wrap = function(x, y, pool = 1, ...) {
+ require(sma)
+ predict.function = function(testmatrix) {
+ res = stat.diag.da(ls = x, as.numeric(y), testmatrix,
+       pool = pool)$pred
+ return(levels(y)[res])
+ }
+ return(list(predict = predict.function, info = list()))
+ }
```

3.3 Sweave

Sweave [24] is a specific approach for generation of a dynamic report. We use Sweave as the technology for the compendium. It mixes S (R) and LaTeX in a sequence of code and documentation chunks in a Rnw file. It uses S (R) for all tangling and weaving steps and hence has a very fine control over the S (R) output. Options can be set either globally to modify the default behaviour or separately for each code chunk to control how the output of the code chunks is inserted into the LaTeX file.

The working principle of Sweave is sketched in figure 2 and demonstrated in the following example. The parts between `<<...>>= ...@` describe the code chunks which will be evaluated by S (R) and whose results will be woven if required into the output (again a LaTeX of postscript document) or only available for later evaluation steps. The paragraphs between the code chunks will be processed as LaTeX code for the output document.

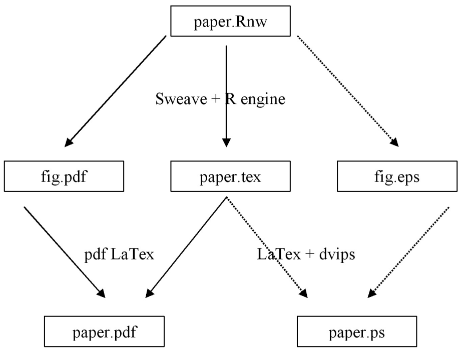


Fig. 2. The working principle of Sweave

To obtain normalized expression measures from the microarray data, Huang et al. used Affymetrix' Microarray Suite (MAS) Version 5 software.Additionally, they transformed the data to the logarithmic scale. Here, we use the function $\backslash Rfunction\{mas5\}$ in the $\backslash Rpackage\{affy\}$ library. $\backslash Robject\{eset\}$ is an object of $classexprSet$, which comprises the normalized expression values as well as the tumour sample data. All the following analyses are based on this object.

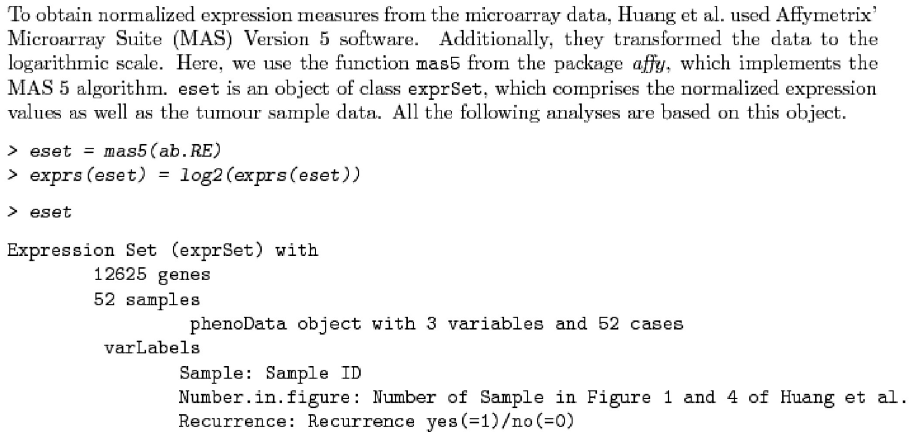
```
%%normalizing the affy batches
<<normalizing,eval=FALSE>>=
Huang.RE  <- mas5(affy.batch.RE)
exprs(Huang.RE) <- log2(exprs(Huang.RE))
@
So we have the following expression set for our further analysis
<<show1>>=
Huang.RE
@
```

The Sweave output of this part is presented in figure 3.

4 Results

The application of the compendium to data of microarray profiling study provides tools to answer crucial questions on the assessment of a new classification algorithm. A few aspects will be discussed.

A new classification algorithm can be implemented by writing the specific wrapper function which is an easy exercise. Misclassification results can be calculated and compared to the results of a set of competing algorithms. The compendium presents classification results on the basis of the confusion matrix and individual classification. Individual classification based on a specific pre-processing strategy and classification algorithm can be presented for the whole sample graphically. figure 4 shows a vote plot which presents for each subject the percentage of correct classification in the deter-



To obtain normalized expression measures from the microarray data, Huang et al. used Affymetrix' Microarray Suite (MAS) Version 5 software. Additionally, they transformed the data to the logarithmic scale. Here, we use the function `mas5` from the package *affy*, which implements the MAS 5 algorithm. `eset` is an object of class `exprSet`, which comprises the normalized expression values as well as the tumour sample data. All the following analyses are based on this object.

```
> eset = mas5(ab.RE)
> exprs(eset) = log2(exprs(eset))

> eset

Expression Set (exprSet) with
        12625 genes
        52 samples
                 phenoData object with 3 variables and 52 cases
         varLabels
                Sample: Sample ID
                Number.in.figure: Number of Sample in Figure 1 and 4 of Huang et al.
                Recurrence: Recurrence yes(=1)/no(=0)
```

Fig. 3. Sweave output of example

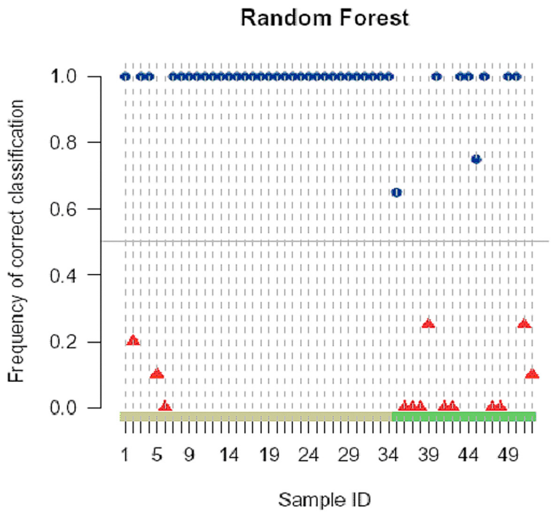


Fig. 4. Vote plot for classification result using: reanalysis of the Huang et al. data [4]

mined number of CV loops. The first 34 patients belong to the group with no tumour recurrence. The remaining 14 patients suffered a relapse.

A table giving a synopsis of all individual misclassifications can also be produced. Figure 5 shows all subjects who are misclassified by at least one of the classification strategies under consideration in the data of Huang et al. [4]. The individuals are ordered with respect to the number of classification strategies which lead to misclassification. Other presentations of the comparison are possible, for example a scatter plot to contrast individual MCRs between the new and the standard approaches. This idea can also be

	RF	PAM	logReg	SVM	RF-M	PAM-M	logReg-M	SVM-M
6	X	X	X	X	X	X	X	X
36	X	X	X	X	X	X	X	X
37	X	X	X	X	X	X	X	X
38	X	X	X	X	X	X	X	X
41	X	X	X	X	X	X	X	X
42	X	X	X	X	X	X	X	X
47	X	X	X	X	X	X	X	X
48	X	X	X	X	X	X	X	X
39	X	X	X	X		X	X	X
51	X	X	X	X		X	X	X
45	X		X	X		X	X	X
35	X		X	X			X	X
52	X			X	X		X	X
2		X	X			X	X	
5		X			X	X		X
7		X						
40							X	
44							X	

Fig. 5. Reanalysis of Huang et al. data (RF - random forest, PAM - shrunken centroids, logReg - penealized logistic regression, SVM - support vector machine, M - using metagenes)

implemented on the advanced level of interaction by writing the code for the respective figure.

The Huang strategy misclassifies two of the 34 patients with good prognosis and three of the 18 patients with bad prognosis. The standard algorithms give results on correct classification below 80%. Why is it impossible for us to reproduce the good classification result with standard algorithms? No implementation for the Bayesian classification tree (BCT) is available and thus no direct comparison is possible. The algorithm of the BCT is not available and its description in a technical report does not give explicit advise for its implementation into software. Therefore, we tried a sensitivity analysis by using the intermediate interaction with the compendium. First, we excluded the the preproceeing form the inner CV loop, because the original paper [4] does not take care on the adjustment of the MCR for the pre-processing strategy. This did not improve classification quality in the expected way. Second, we reduced the data set to the two-hundred most discriminating genes which introduces a huge selection bias. Only this way we could come up with 6-7 misclassifications which is still worse as the results in the original paper. Can we trust the original result? Our analysis reminds us to be quite critical to the original claims.

Huang et al. [4] state that the genes found to be crucial for the classification are different from the genes found as crucial by van ’t Veer et al. [3]. What is the reason for the missing link between the Huang and van ’t Veer classifiers? How is it possible to disentangling computation, technology, and biology? The compendium takes care on the computational part. Additionally, the BCT and the van ’t Veer classifier need to be implemented and wrapper functions for both classification algorithms have to be written. Both wrapper and the respective pre-processing strategies will be applied to both data sets. For each data set correlation between the genes used in the classifiers can be calculated and visualised by a checkerboard figure. The checkerboard for genes of both classifiers on the same data set allows to identify genes with similar expression behaviour and to study common biological aspects behind both classifiers. Comparing the checkerboards between both data sets gives hints on sample differences between both studies and discrepancies introduced by the different microarray technologies used.

5 Discussion

The literature on the induction of prognostic profiles from microarray studies is a methodological wasteland. Ambroise and McLachlan [9] describe the unthorough use of cross-validation in a number of high-profile published microarray studies. Tibshirani and Efron [7] report the difficulty in reproducing a published analysis. Huang et al. [4] present results with the potential to revolutionize clinical practice in breast cancer treatment but use an ideosyncratic statistical method which is fairly complex and neither easy to implement nor to obtain as software. A series of papers published in Nature [3], NEJM [6], [17], and The Lancet [4], [5] base their impressive results on classification methods which were developed ad-hoc for the problem at hand. The global picture looks like: the MCRs reported are of questionable clinical relevance, reanalysis of a paper does not support the strong claims they made, results are not reproducible with respect to computation, validation, or biology.

This situation has several implications: 1) It is nearly impossible to assess the value of the presented studies in terms of statistical quality and clinical impact. 2) Scientists looking for guidance to design similar studies are left puzzled by the plethora of methods. 3) It is left unclear how much potential there is for follow-up studies to incrementally improve on the results.

References

1. Microarray special. *Statistical Science*, **18**:1-117, 2003.
2. Simon R., Rademacher M.D., Dobbin K., McShane L.M.: Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *J Nat. Cancer Inst.*, **95**: 14-18, 2003.
3. van 't Veer L., Dai H., van de Vijver M.J., He Y.D., Hart A.A.M., Mao M., Petersen H.L., van de Kooy K., Marton M.J., Witteveen A.T., Schreiber G.J., Kerkhoven R.M., Roberts C., Linsley P.S., Bernards R. and Friend S.H.: Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**:530-536, 2002.
4. Huang E., Cheng S.H., Dressman H., Pittman J., Tsou M.H., Horng C.F., Bild A., Iversen E.S., Liao M., Chen C.M., West M., Nevins J.R. and Huang A.T.: Gene expression predictors of breast cancer outcomes. *The Lancet*, **361**:1590-1596, 2003.
5. Chang J., Wooten E., Tsimelzon A., Hilsenbeck S., Gutierrez C, Elledge R., Mohsin S., Osborne K., Chamness G., Allred C., O'Connell P.: Gene expression profiling for the prediction of therapeutic response to docetaxel in patients with breast cancer. *The Lancet*, **362**:362-369, 2003.
6. Bullinger L., Döhner K., Bair E., Fröhling S., Schlenk R.F., Tibshirani R., Döhner H., Pollack J.R.: Use of Gene-Expression Profiling to Identify Prognostic Subclasses in Adult Acute Myeloid Leukemia. *NEJM*, **350**:1605-1616, 2004.
7. Tibshirani R.J., Efron B.: Pre-validation and inference in microarrays. *Statistical Applications in Genetics and Molecular Biology*, **1**:1, 2002.

8. Breiman L.: Statistical Modelling: The Two Cultures. *Statistical Science*, **16**:199-231, 2001.
9. Ambroise C., McLachlan G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc Natl Acad Sci*, **99**:6562-6566, 2002.
10. Brenton J.D., Caldas C.: Predictive cancer genomics - what do we need? *The Lancet*, **362**:340-341, 2003.
11. Leisch F., Rossini A.J.: Reproducible statistical research. *Chance*, **16**:41-46, 2003.
12. Dudoit S., Fridlyand J., Speed T.P.: Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data. *Journal of the American Statistical Association*, **97**:77-87, 2002.
13. Lee J.W., Korea University, Department of statistics, personal communication.
14. Ihaka R., Gentleman R.: R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, **5**:299-314, 1996.
15. Gentleman R., Carey V.: Bioconductor. *R News*, **2(1)**:11-16, 2002.
16. Leisch F.: Dynamic generation of statistical reports using literate data analysis. *Compstat 2002 - Proceedings in Computational Statistics*, 575-580, 2002.
17. van de Vijver M.J., He Y.D., van 't Veer L.J. et al.: A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, **347**:1999-2009, 2002.
18. Tibshirani R., Hastie T., Narasimhan B., Chu G.: Class prediction by nearest shrunken centroids, with application to DNA microarrays. *Statistical Science*, **18**:104-117, 2003.
19. Vapnik V.: *The Nature of Statistical Learning Theory*. Springer, New York (1999)
20. Breiman L.: Random Forests. *Machine Learning Journal*, **45**:5-32, 2001.
21. Eilers P.H., Boer J.M., Van Ommen G.J., Van Houwelingen H.C.: Classification of Microarray Data with Penalized Logistic Regression. *Proceedings of SPIE volume 4266:progress in biomedical optics and imaging*, **2**:187-198, 2001.
22. Carey V.J.: Literate Statistical Programming: Concepts and Tools. *Chance*, **14**:46-50, 2001.
23. Sawitzki, G.: Keeping Statistics Alive in Documents. *Computational Statistics*, **17**:65-88, 2002.
24. Leisch, F.: Dynamic generation of statistical reports using literate data analysis. *Compstat 2002 - Proceedings in Computational Statistics*, 575-580, 2002.