VERSITA

STUDIES IN LOGIC, GRAMMAR AND RHETORIC 35 (48) 2013

DOI: 10.2478/slgr-2013-0040

The Stability of Gene Selection in Microarray Experiments

Magdalena Wietlicka-Piszcz[1]

[1] Department of Theoretical Backgrounds of Biomedical Science and Medical Informatics, Collegium Medicum in Bydgoszcz, Nicolaus Copernicus University, Poland

Abstract. This paper addresses the issue of the stability of lists of genes identified as differentially expressed in microarray experiments. The similarities between gene rankings yielded by various gene selection methods performed with resampled datasets were assessed. The mean percentage of overlapping genes for two rankings varied from 10 to 90% depending on the applied gene selection method and the size of the list. The assessment of the stability of obtained gene rankings seems to be relevant in the analysis of microarray data.

Introduction

Microarrays are a new technology applied in the genetics field, enabling simultaneous investigation of expression levels of thousands or tens of thousands of genes. In many cases, the main aim of a microarray experiment is the identification of genes involved in the aetiology of a particular disease or the selection of genes enabling the differentiation between disease subtypes as well as the prediction of future events, such as response to applied therapy, survival times, relapse of a disease or cancer recurrence. Thus, the proper identification of genes differentially expressed, e.g. between the diseased and normal tissue, is crucial because it often determines, to a certain extent, the direction of further research, which is frequently focused on selected genes.

In the vast majority of microarray experiments, the selection procedures of differentially expressed genes restrict the number of investigated genes from the total of tens of thousands to tens and "the results of microarray studies are usually the starting point for further more expensive and time-consuming experiments, which involve only a small number of candidate genes" (Aerts et al., 2006).

Typically, in a microarray experiment the number of arrays is in the range of a few microarrays to over 200 or 300, while the number of exam-

ISBN 978–83–7431–392–6 ISSN 0860-150X

ined genes is in the range between a few thousand and tens of thousands. Therefore, the selection of differentially expressed genes is a very important stage of microarray data analysis and involves the use of methods that can be used when the number of features (genes) is much bigger than the number of samples (microarrays). There have been many methods developed for the selection of active genes in microarray settings. However, the lists of genes identified as differentially expressed produced by those methods may differ substantially.

Another problem associated with the gene selection procedure is the stability of gene lists obtained with a particular method but with slightly modified versions of the dataset (Boulesteix et al., 2009), e.g. subsampled dataset. The gene rankings obtained with a method performed, e.g. with bootstrap samples of the original dataset, may significantly differ from the ranking returned by the same method applied to the whole dataset.

In the vast majority of cases, the standard procedure of microarray data elaboration involves the use of one active gene selection method applied to the whole data set and further analysis and reasoning is partly based on that list.

This work addresses the issue of gene selection stability and its dependence on the applied method of active gene identification. The similarities between gene rankings yielded by various methods and between rankings obtained for the perturbed dataset were considered. The analysis was performed for three datasets. Four methods of gene selection were applied and compared.

Material and Methods

The analysis was performed based on three high density oligonucleotide microarray datasets downloaded from the public repositories. The first dataset contained data from 167 samples of oral squamous cell carcinoma (OSCC) and 17 samples of oral dysplasia tissues (Chen et al., 2008; public repository GEO – GSE30784). The data were used to identify genes enabling differentiation between OSCC and dysplasia. In further considerations this dataset will be called *dataset 1*.

Another dataset concerned the problem of early detection of colorectal cancer (CRC) and consisted of expression data from 100 whole blood samples from patients with CRC and from 100 samples from patients without any symptoms of CRC, inflammatory bowel diseases or polyps (Xu et al., 2013; public repository ArrayExpress – E-MTAB-1532). The

gene expression profiles from this dataset were used to pick up genes useful in the early recognition of CRC. This dataset will be referred to as *dataset 2* from this point forward.

The third dataset comprised expression profiles from a total of 58 cirrhotic tissue samples from liver tissue with HCV infection. Seventeen of them were from patients with hepatocellular carcinoma and 41 were from patients without hepatocellular carcinoma (Mas et al., 2009; public repository GEO – GSE1423). The analysis of this dataset aimed to identify genes enabling the differentiation between cirrhotic tissue and cirrhotic tissue with concomitant hepatocellular carcinoma. This dataset will be referred to as *dataset 3*.

The data were previously pre-processed, so for each probe expression summaries were available. The gene selection was performed by the use of two methods based on the t-test, i.e. Parametric Empirical Bayes Method (Limma) (Smyth, 2004) and Significance Analysis of Microarrays (SAM) (Tusher et al., 2001), and the Nonparametric Empirical Bayes Method based on Wilcoxon rank sums (EbamWilcoxon) (Efron et al., 2002) and the Wilcoxon rank sum test (Wilcox) were also applied.

The above mentioned methods of variable selection return ordered lists of candidate genes, where the genes are ordered according to the criterion used to rank variables, i.e. the absolute value of a particular test statistic. The highest ranked genes are considered for further analysis. The top k-list of candidate genes is the list of k genes with the highest rank values, so there are the genes from the top of the ordered list of candidate genes.

For each of the considered datasets in this work, the identification of differentially expressed genes was performed by the use of the four various methods of feature selection. Additionally, to address the issue of the stability of obtained gene lists, the feature selection methods were applied with the re-sampled datasets. The Jackknife subsampling technique was applied. The dataset was split into 10 disjoint folds of approximately equal size and the samples were created by removing the consecutive folds from the whole dataset. This procedure was repeated 10 times, so 100 samples were created, each comprising approximately 90% of the whole dataset.

To assess the similarity of two gene rankings, the proportion of common genes in the two top k-lists was calculated. This measure is also denoted as a percentage of overlapping genes (POG) (Zhang et al., 2009). To visualize the similarity of two gene rankings (the two top k-lists) versus the size k of the list, a descriptive plot called correspondence at the top (CAT-plot) was applied (Irizarry et al., 2005). CAT curves show the proportion of common genes plotted against the size of the lists.

To assess the stability of a gene ranking obtained with a particular gene selection method, the comparison of the ranking obtained for the original dataset and rankings derived for subsamples was performed. The POG for pairs of rankings (the ranking for the original dataset and for the subsample) was calculated and then the mean value of POG for all these pairs was also determined.

To consider the similarities between rankings derived with various gene selection methods, the POG was calculated for pairs of rankings obtained with different methods. Then, for each pair of methods, the mean value of POG for subsamples was computed.

Additionally, an attempt to aggregate rankings based on subsamples was made and then the comparison of classification, based on the single ranking and the ranking derived from aggregation, was carried out.

To perform the selection of differentially expressed genes and classification based on identified gene sets, the whole dataset was split into 10 disjoint folds of approximately equal size. Each fold was used as a testing set while the remaining arrays were used as a training set. This procedure was repeated 10 times (ten ten-fold cross-validation). For each training set (TS), the selection of genes was performed in two ways. At first, active gene identification was performed for the whole TS, which resulted in a single ranking. Then, the selection procedure was applied to the re-sampled TS. The bootstrap samples were created by drawing samples (arrays) with replacements from the TS. For each bootstrap sample, the selection of genes was performed. The re-sampling was repeated 100 times, so for each TS 100 gene rankings were obtained. Then the rankings were aggregated and for each gene the final score was estimated as a sum of ranks from all the lists. A single rank of a gene g in a list was estimated as $1/r$ (to assign bigger weight to genes on the top of the list) where r is the position of the gene on a single list, so the final score s of a gene g was estimated as $s_g = \sum_{i=1}^{n} 1/r$, where n denotes the number of lists. Therefore, for each TS two gene rankings were derived: the ranking from the single selection of genes from the TS (*ranking 1*) and an aggregated ranking from the resampling of the TS (*ranking 2*). Then, classification based on the two rankings was performed. The classification was carried out for the consecutive subsets of the first $2, 3, \ldots, 100$ highest ranked genes from both rankings.

The procedures of gene selection and classification were repeated for all pairs of training and testing sets. For classification, the classifiers widely used in microarray data analysis (Boulesteix et al., 2008; Van Sanden et al., 2008), such as the Support Vector Machines with linear kernel (SVMl), Support Vector Machines with radial kernel (SVMr), Diagonal

Linear Discriminant Analysis (DLDA) and Diagonal Quadratic Discriminant Analysis (DQDA) were used.

Results and Discussion

The considered gene selection methods were applied to identify the differentially expressed genes in the three analyzed datasets. Each method was applied both with the original dataset and with the re-sampled dataset.

To assess the stability of gene rankings produced by a particular gene selection method, the similarities between the rankings obtained with the original dataset and the rankings obtained with the re-sampled dataset were examined. The values of POG for the top k-list for the whole dataset and the top k-lists corresponding to the sub-samples were calculated and averaged.

The mean values of POG for the consecutive top k-lists, for k = 10, $20, \ldots, 300$ were calculated. Figures 1–3 present the CAT-curves for the top k-lists for the whole and sub-sampled dataset, for various gene selection methods and investigated datasets.

The highest number of common genes was observed for *dataset 2*. The POG for the top 100 genes was over 80%; however, the POG for *dataset 1*

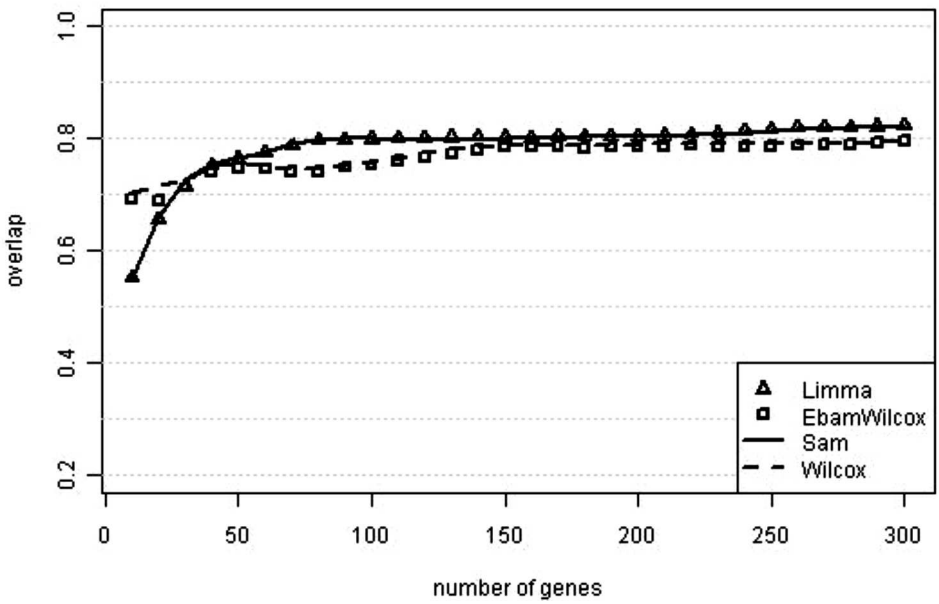

Figure 1. Mean percentage of overlapping genes for *dataset 1*

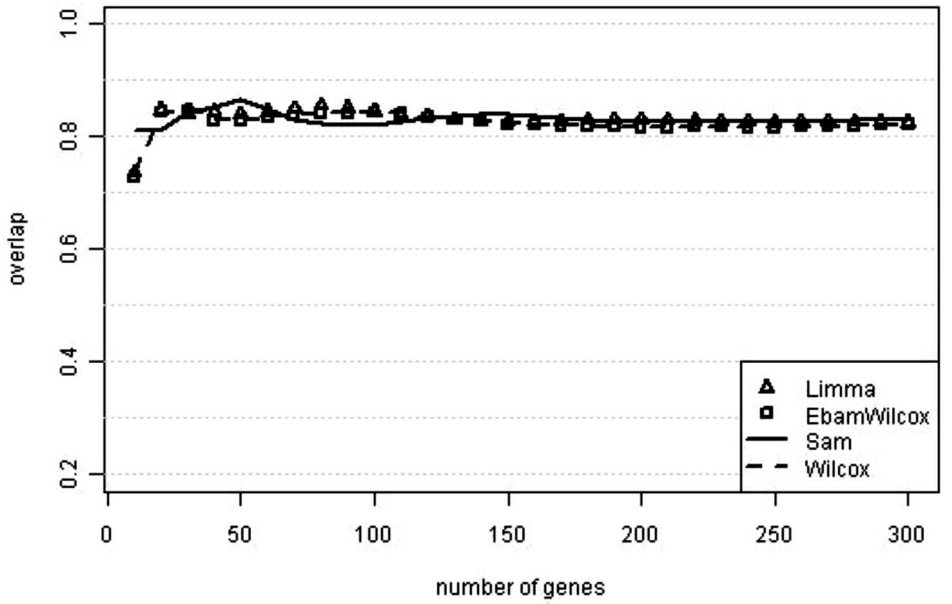

Figure 2. Mean percentage of overlapping genes for *dataset 2*

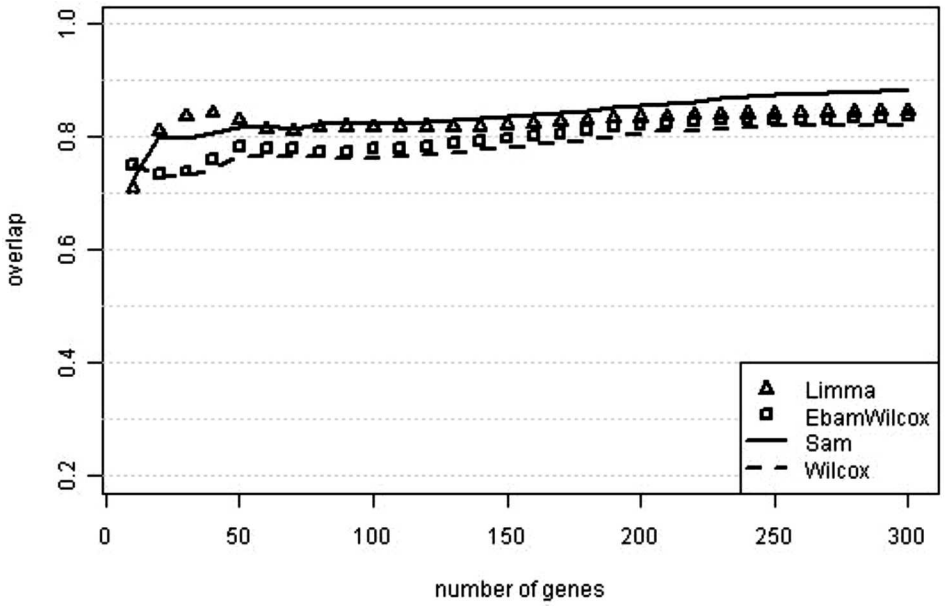

Figure 3. Mean percentage of overlapping genes for *dataset 3*

and *dataset 3* for the same number of genes was from 75 to 80%, depending on the method. The standard errors for the mean values of POG were in the range between 0.1 and 1%.

For *dataset 3*, the Significance Analysis for Microarrays and Parametric Empirical Bayes Method seemed to yield a higher overlap of selected genes than the Wilcoxon rank sum test or the Nonparametric Empirical Bayes Method, however the differences were small. For *dataset 1* and *dataset 2*, the mean overlap of gene rankings was comparable for all considered methods.

Additionally, to illustrate the variability between the top k-lists, the proportion of genes common for all the samples using each method was calculated. Figures 4–6 show the percentages of common genes for the top k-lists obtained for all subsamples. The percentage of overlapping genes varied from 20–25% for the top 100-list for *dataset 1* and 20–40% for *dataset 3* to 35–40% for *dataset 2*. This means that for 100 lists of the top 100 genes, for the analyzed datasets, 20–40 genes, depending on the dataset and the method, might be encountered on each list. For the top ranked list containing 300 genes, the overlap was also between 20 and 40%, so 60 to 120 genes were common for all the rankings.

To assess the similarities of gene rankings yielded by the considered gene selection methods, for each subsample, the POG for pairs of rankings produced by different methods were calculated. Then for each pair of methods the mean values of POG for all subsamples were computed. The mean values of POG were calculated for the consecutive top k-lists, $(k = 10, 20, \ldots, 300)$. To visualize the similarities of the gene lists yielded by various feature selection methods, the CAT-plots were created. Figures 7–9 present the CAT-curves for pairs of methods. The highest percentages

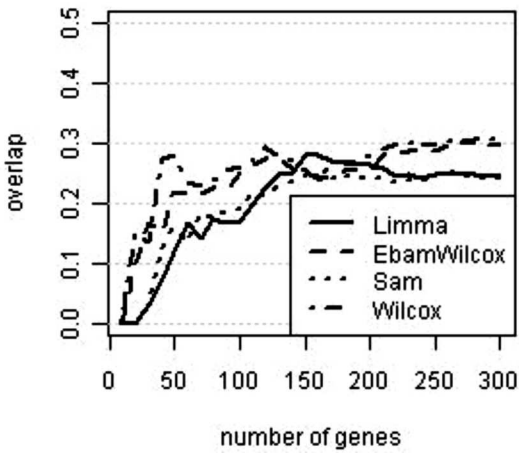

Figure 4. Percentage of overlapping genes for 100 sub-samples of *dataset 1*

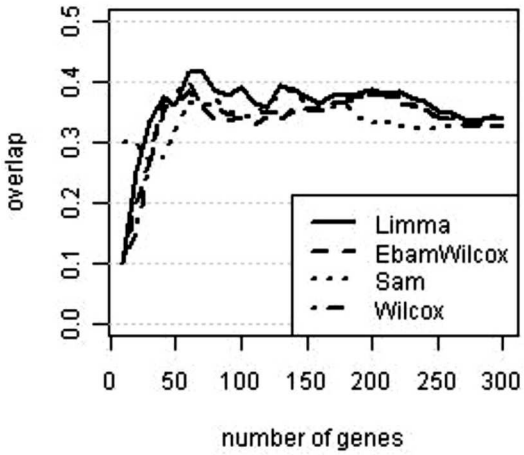

Figure 5. Percentage of overlapping genes for 100 sub-samples of *dataset 2*

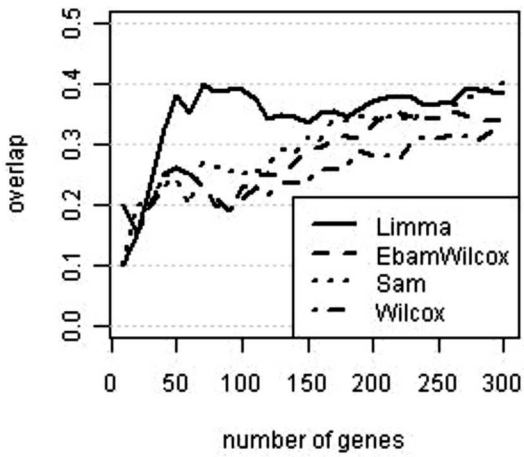

Figure 6. Percentage of overlapping genes for 100 sub-samples of *dataset 3*

of overlapping genes were obtained by using the Wilcoxon rank sum test and the Nonparametric Empirical Bayes Method – based on this test, the percentage was over 90%. Also, the overlap of rankings produced by Parametric Empirical Bayes Method and Significance Analysis of Microarrays was about 85 and over 90% for *dataset 1* and *dataset 2*, respectively, for the list of 100 features, but for *dataset 3*, for the same list size, it was about 50%. For the other methods, the results obtained for different datasets varied depending on the dataset. The highest values of POG for all

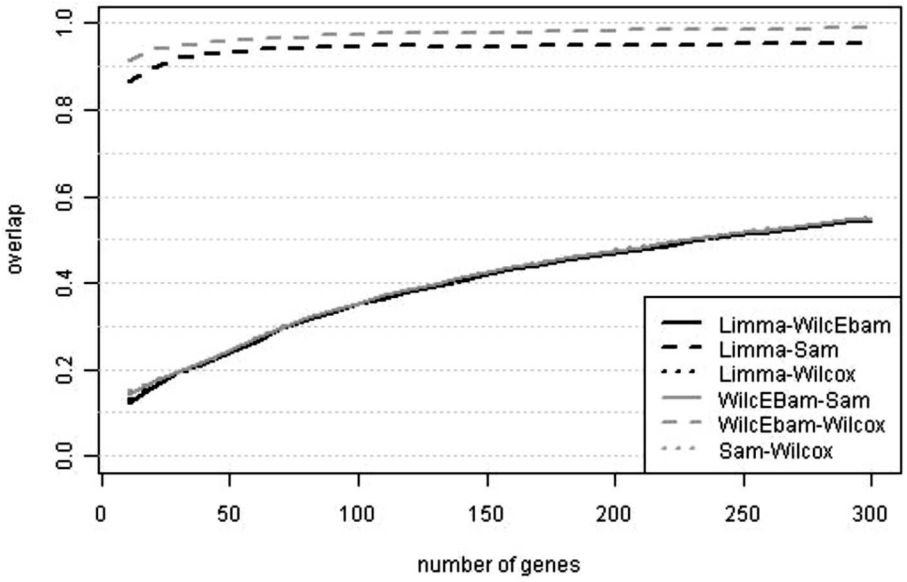

Figure 7. Mean percentage of overlapping genes for pairs of methods for *dataset 1*

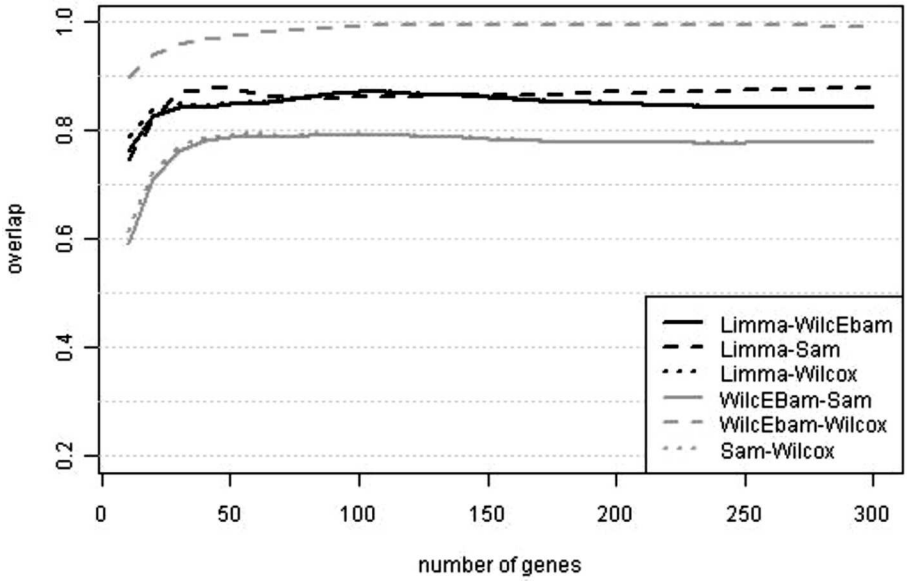

Figure 8. Mean percentage of overlapping genes for pairs of methods for *dataset 2*

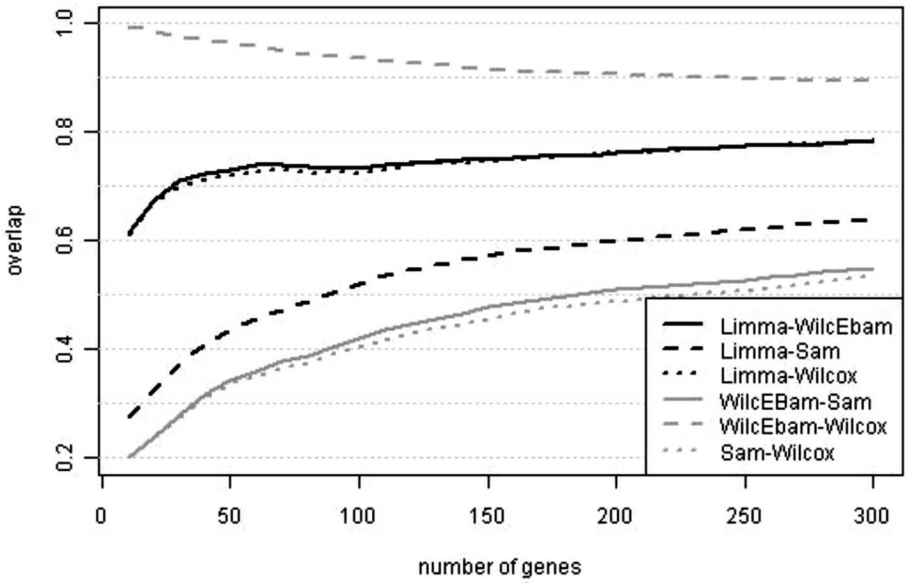

Figure 9. Mean percentage of overlapping genes for pairs of methods for *dataset 3*

methods were received for *dataset 2*, and for the list of 100 genes they varied from 80 to over 90%. This dataset comprised the biggest number of samples (200 arrays). The smallest overlap of gene rankings was obtained for *dataset 3*.

Figures 10–12 show the proportion of genes common for all subsamples for all pairs of considered gene selection methods. The highest overlap was obtained for *dataset 2* (about 25–35% for all pairs of methods for the list of 100 genes) and the lowest overlap was derived for *dataset 1*. For some methods, it was below 10% for the top 100-list.

There is a question regarding which genes should be used in further analyses. The answer is not straightforward and various strategies may be considered, depending on the main aim of the experiment. One of the possible solutions is the aggregation of ranks from multiple rankings.

The analysis of the three datasets shows that the smallest overlap of top k-lists was obtained for *dataset 1*. For this dataset, the comparison of classification based on rankings derived from the whole training set (*ranking 1*) and the rankings returned as a result of aggregation of lists obtained from re-sampling of the training set (*ranking 2*), was performed. The selection of genes was performed with the use of the Parametric Empirical Bayes Method.

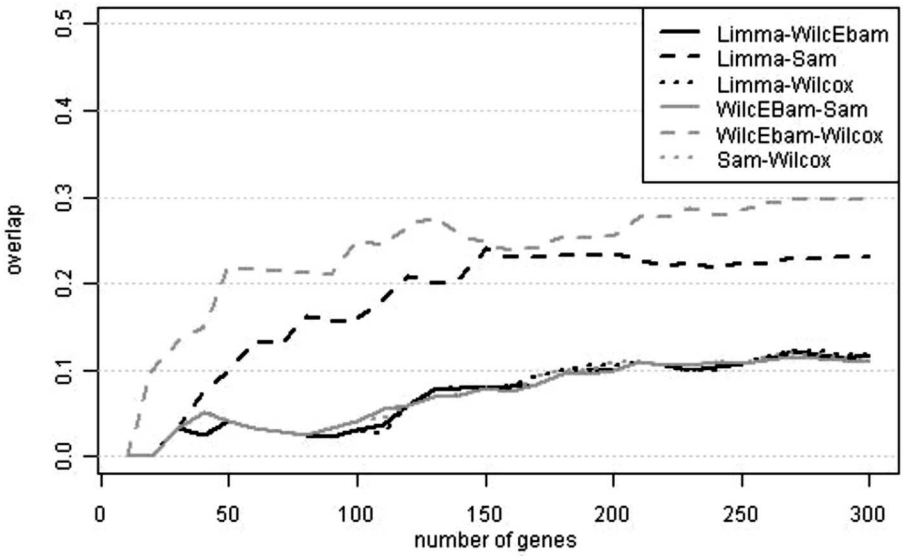

Figure 10. Percentage of overlapping genes for 100 sub-samples of *dataset 1*, for pairs of methods

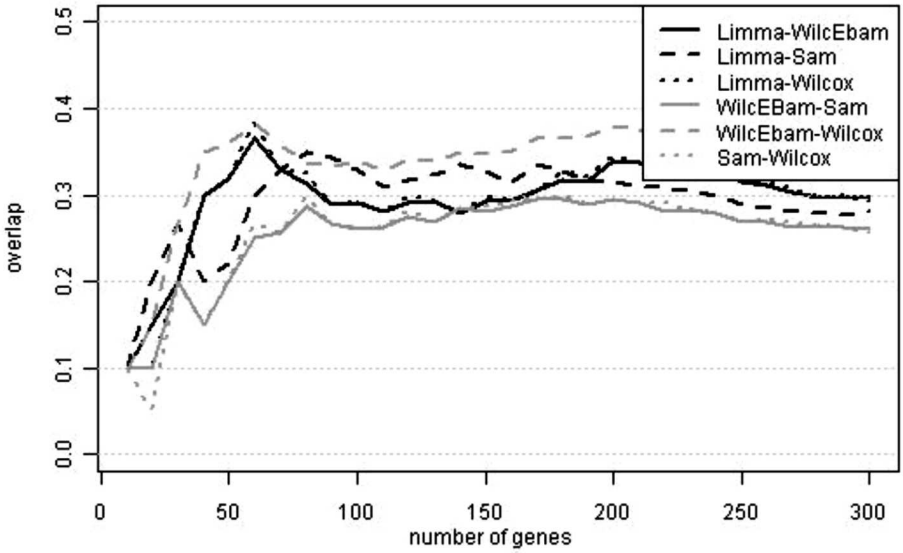

Figure 11. Percentage of overlapping genes for 100 sub-samples of *dataset 2*, for pairs of methods

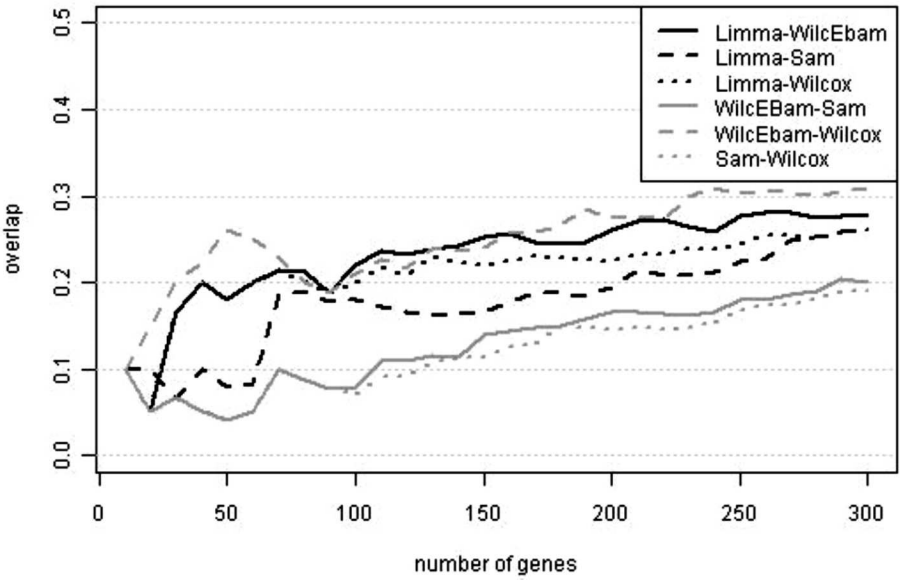

Figure 12. Percentage of overlapping genes for 100 sub-samples of *dataset 3*, for pairs of methods

Figure 13 presents the cross-validation misclassification error rates obtained for the classification based on the two types of rankings (*ranking 1* and *ranking 2*). The classification was carried out for the consecutive subsets of the first $2, 3, \ldots, 100$ genes from both rankings. DQDA, DLDA, SVMl and SVMr methods were used for classification. The lowest misclassification error rates were obtained for SVMr for *ranking 2* – for 33 genes the error was equal to 0.057, while for the same number of genes for *ranking 1*, the error was 0.071. The error rates obtained for the SVMl were comparable to those obtained for radial kernel. The lowest error was obtained for 83 genes and was equal to 0.058 for *ranking 2*, while for *ranking 1*, the error was 0.067. For DLDA, the error rates were in the range between 0.088 and 0.131 for *ranking 2* and, respectively, between 0.096 and 0.140 for *ranking 1*. The highest misclassification error rates were derived for DQDA. The errors varied from 0.116 to 0.151 for *ranking 2* and from 0.109 to 0.163 for *ranking 1*.

The misclassification errors were lower for rankings obtained from aggregation of many lists derived from resampling of the training set (*ranking 2*) for all applied methods and almost all considered subsets of genes.

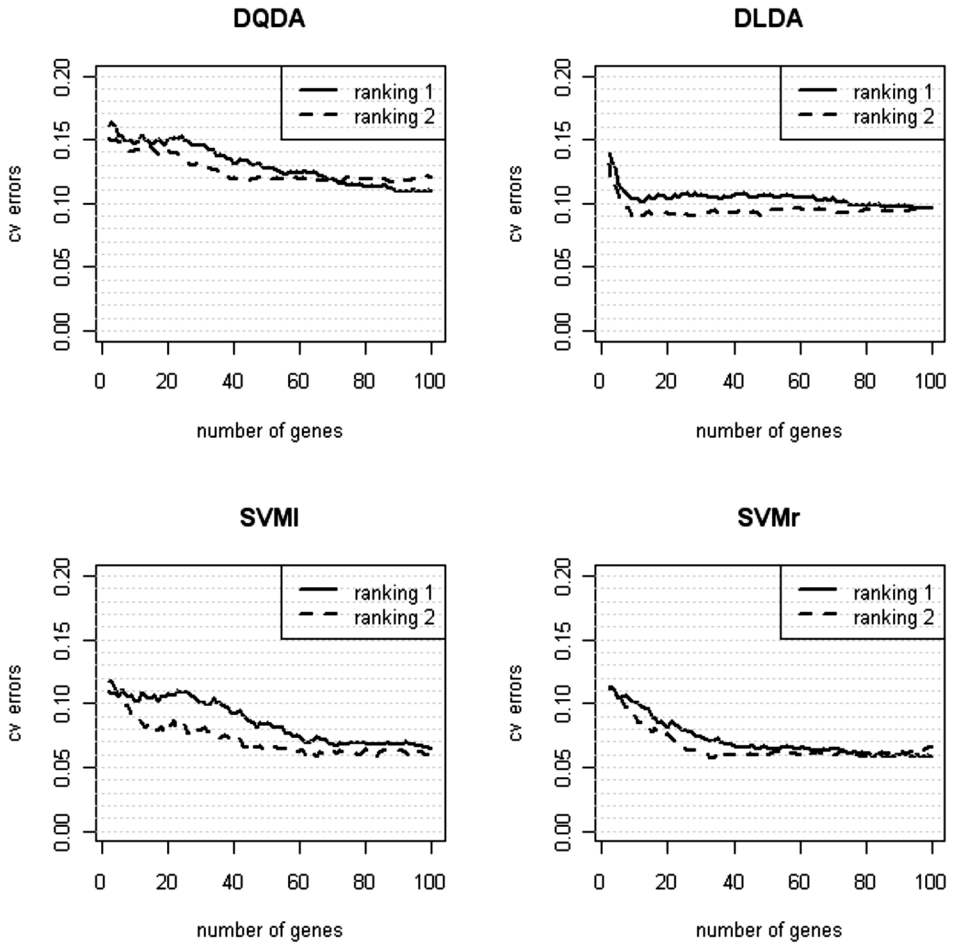

Figure 13. Comparison of cross-validation misclassification error rates for *dataset 1* for pairs of gene rankings: ranking based on the training set (ranking 1) and ranking derived from aggregation of rankings obtained from application of the gene selection method with subsamples of the training set (ranking 2)

Conclusion

The selection of features is a very important stage of elaboration of data from microarray experiments. The assessment of the stability of obtained gene rankings seems to be relevant and the careful analysis and comparison of gene lists obtained for perturbed datasets and/or various gene selection methods may help to get more reliable rankings. Certainly, further investigations in this area are necessary.

REFERENCES

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L. C., De Moor, B., Marynen, P., Hassan, B., Carmeliet, P., & Moreau, Y. (2006). Gene prioritization through genomic data fusion. *Nat Biotechnol*, 24, 537–544.

Boulesteix, A. L., & Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Brief Bioinformatics*, 10, 556–568.

Boulesteix, A. L., Strobl, C., Augustin, T., & Daumer, M. (2008). Evaluating Microarray-based Classifiers: An Overview. *Cancer Informatics*, 6, 77–97.

Chen, C., Mendez, E., Houck, J., Fan, W., Lohavanichbutr, P., Doody, D., Yueh, B., Futran, N. D., Upton, M., Farwell, D. G., Schwartz, S. M., & Zhao, L. P. (2008). Gene expression profiling identifies genes predictive of oral squamous cell carcinoma. *Cancer Epidemiol Biomarkers Prev*, 17(8), 2152–2162.

Efron, B., & Tibshirani, R. (2002). Empirical bayes methods and false discovery rates for microarrays. *Genet Epidemiol*, 23, 70–86.

Irizarry, R. A., Warren, D., Spencer, F., Kim, I. F., Biswal, S., Frank, B. C., Gabrielson, E., Garcia, J. G., Geoghegan, J., Germino, G., Griffin, C., Hilmer, S. C., Hoffman, E., Jedlicka, A. E., Kawasaki, E., Martinez-Murillo, F., Morsberger, L., Lee, H., Petersen, D., Quackenbush, J., Scott, A., Wilson, M., Yang, Y., Ye, S. Q., & Yu, W. (2005). Multiple-laboratory comparison of microarray platforms. *Nat Methods*, 2, 345–350.

Mas, V. R., Maluf, D. G., Archer, K. J., Yanek, K., Kong, X., Kulik, L., Freise, C. E., Olthoff, K. M., Ghobrial, R. M., McIver, P., & Fisher, R. A. (2009). Genes involved in viral carcinogenesis and tumor initiation in hepatitis C virus-induced hepatocellular carcinoma. *Mol Med*, 15(3–4), 85–94.

Public Repository ArrayExpress. (2013). Transcription profiling by array of whole blood samples from colorectal cancer patients and healthy individuals, accession number E-MTAB-1532. Retrieved from http://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-1532/.

Public Repository Gene Expression Omnibus. (2009). RMA expression data for liver samples from subjects with HCV, HCV-HCC, or normal liver, accession number GSE1423. Retrieved from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE14323.

Public Repository Gene Expression Omnibus. (2011). Gene expression profiling of oral squamous cell carcinoma (OSCC), accession number GSE30784. Retrieved from http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE30784.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*, 3(1). DOI: 10.2202/1544-6115.1027.

Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *PNAS*, 98(9), 5116–5121.

Van Sanden, S., Lin, D., & Burzykowski, T. (2008). Performance of gene selection and classification methods in a microarray setting: A simulation study. *Communications in Statistics – Simulation and Computation*, 37(2), 409–424.

Xu, Y., Xu, Q., Yang, L., Ye, X., Liu, F., Wu, F., Ni, S., Tan, C., Cai, G., Meng, X., Cai, S., & Du, X. (2013). Identification and Validation of a Blood-Based 18-Gene Expression Signature in Colorectal Cancer. *Clin Cancer Res*, 19(11), 3039–3049.

Zhang, M., Zhang, L., Zou, J., Yao, Ch., Xiao, H., Liu, Q., Wang, J., Wang, D., Wang, Ch., & Guo, Z. (2009). Evaluating reproducibility of differential expression discoveries in microarray studies by considering correlated molecular changes. *Bioinformatics*, 25(13), 1662–1668.