

Integration Strategies of Cross-Platform Microarray Data Sets in Multiclass Classification Problem

Sebastian Student[1(✉)], Alicja Płuciennik[1,2], Krzysztof Łakomiec[1], Agata Wilk[1], Wojciech Bensz[1], and Krzysztof Fujarewicz[1]

[1] Institute of Automatic Control, Silesian University of Technology, Gliwice, Poland
sebastian.student@polsl.pl
[2] Wasko S.A., Gliwice, Poland

Abstract. Despite the increasing amount of available gene expression data, integrative analysis is still hindered by its high susceptibility to microenvironment fluctuations, resulting in inter-experiment variability known as batch effects. Therefore the development of data integration strategy is now more necessary than ever. Although several normalization algorithms have already been proposed, we believe that an effective model must rely on data migration between schemes. In this paper we apply this approach to a set of microarray data from core needle biopsy of breast cancers spanning different microarray platforms, and demonstrate its effectiveness in data preparation for unsupervised analysis and multiclass classification tasks. We propose a custom tool dedicated to defining the model structure. Additionally, we compare several pipelines of data processing, combining data normalization with different batch effect correction methods.

Keywords: Data integration · Batch effect · Multiclass classification

1 Introduction

Batch effects due to sample preparation or array variation are commonly observed across multiple series of microarray experiments. They are artifacts not related to the biological variation of scientific interest. The batch effects can influence the whole analysis, especially meta-analysis across studies. In high throughput gene expression experiments one of the biggest challenges is the data integration of different microarray platforms.

From the very beginning of microarray era the considerable impact of experiment microenvironment, such as position of the sample, time of day when the assay is performed [1] or the atmospheric composition [2], has been recognized. In large-scale studies, owing to the limitations concerning both the microarray scanner capacity and the prolonged time period of specimen acquisition, the readings cannot be conducted simultaneously. This induces further variability

© Springer Nature Switzerland AG 2019
S. Misra et al. (Eds.): ICCSA 2019, LNCS 11623, pp. 602–612, 2019.
https://doi.org/10.1007/978-3-030-24308-1_48

related to experiment conditions including, but not limited to the utilized hardware, reagents, or the responsible technician [3,4,8]. Furthermore, technological advances over the experiment timeline often result in microarray type or platform change mid-study [9].

While batch effect bias is inevitable even under the most rigorous experiment protocols, integrative analysis is not unfeasible. In particular, when analyzing data originating from several experiments, the two possible approaches are meta-analysis and merging. In the first one, each sub-experiment is evaluated independently, and the integration of obtained results (for instance in the form of accuracy measures) only occurs subsequently [10]. In the second strategy the available data is combined into a larger dataset, which is, after certain processing countering batch effect, used in further analysis.

Over the years since the problem identification, several computational methods have been proposed addressing inter-experimental variability. In 2000, the singular value decomposition (SVD) method based on linear space transformations and eigenarrays was introduced [5,6,11]. Later developed was an algorithm of Distance Weighted Discrimination (DWD) based on support vector machines [7], replacing the minimal distance criterion with a sum of inverse distances [9]. Other methods include batch mean-centering [13] and Surrogate Variable Analysis [12]. Finally, the Empirical Bayes method (also referred to as ComBat) in a parametric and non-parametric version, is now commonly used, which can be attributed to its applicability even for small sample sizes [16–18]. The rest of the paper is organized as follows: First, we give an overview of the methodology used, system architecture and data description. Then we describe the numerical experiment assumption. We describe the numerical experiment results in details. Finally, we summarize and discuss the obtained results.

2 Methodology

2.1 Data Normalization Schemes

We used three different strategies to combine data from two different microarrays platforms. First, in all used strategies we normalize each of microarray platform using FRMA algorithm [15]. The FRMA is quantile based microarray normalization method which is considered as robust and platform independent method. To normalize the input microarrays we used the custom (Entrez v19) chip definition file downloaded from Brainarray website [14]. Next we merge the normalized data by choosing the common genes in both platforms. And after that we used different algorithms to remove the batch effect. Used strategies are depicted in Fig. 1.

2.2 Classification System Description

The system described in this work is built around Spicy, an R language library (ver. 3.5.1), originally developed as a part of remote platform for hypothesis

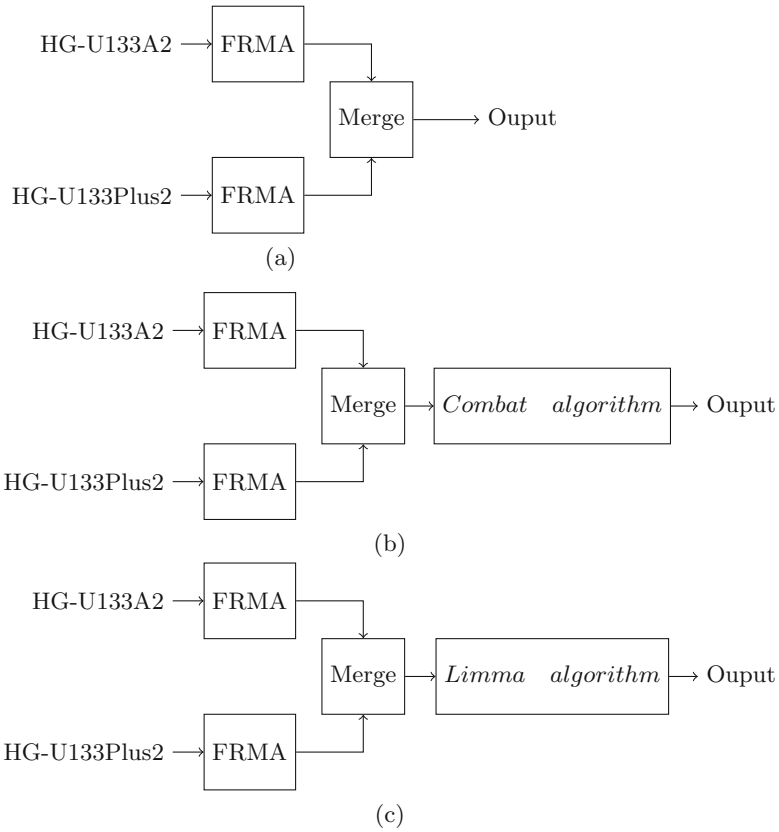

Fig. 1. Different strategies used to combine data from two different microarray platforms: (A) Without additional batch effect removing step, (B) Scheme with remove the batch effect using *Combat* algorithm, (C) Scheme with remove the batch effect using *Limma* algorithm.

testing and analysis of multiclass biomedical data [19] described in [20,21]. The instance of Spicy central to the described system is installed on high performance computing cluster Ziemowit. We used R environment and a Galaxy server tool with additional custom visual interface that constitutes a convenient alternative to the standard, form based method of defining model structure. Galaxy Server was installed on CentOS 7 virtual machine. The system is available up on request.

3 Numerical Experiment

We performed a numerical experiment using Principal Component Analysis (PCA) to check if different array platforms will cause batch effect. As a result, we made scatterplots for the first two PC and we marked the spots from different platforms with distinct colors.

For validation of investigated data combining strategies we performed both, an unsupervised and a supervised methods. For the unsupervised analysis we used hierarchical clustering (with euclidean distances) for both, samples and preselected features. We chose the features with the highest mean variance (the threshold was the 0.75 quantile). We annotated samples with the proper microarray platform. We also checked the classification accuracy for a multi-class problem. We used a popular genomic feature selection method—fold change(FC) and SVM classifier with bootstrap validation (500 iterations). The classifier was used in one versus rest (OvR) mode to manage multiple cancer types. We computed the mean quality for models in the range from 2 up to 20 features with step equal to 2. We calculated the classification error with a confidence interval for all tested variants.

3.1 Data Description

Our data set contains 241 microarrays with samples form core needle biopsy of breast cancers. Samples belong to 5 cancer types - luminal A (LumA), luminal B (LumB), basal-like (Basal), HER2-enriched (Her2) and normal-like (Normal) breast cancer. The microarrays were made using two microarray types - 200 microarrays were performed on Affymetrix HG-U133 Plus 2 arrays and 41 microarrays were performed with Affymetrix HG-U133 A2 arrays.

3.2 Results

In this section we present the influence of data migration strategies selection on microarray data unsupervised analysis and data classification. To explore the variability in analysed data and show the integration strategies differences we have used Principal Component Analysis methodology. The biggest difference we have observed between the simple integration variant A and the other variants (B and C). The platform difference variability is represented by the first principal component and the variability is much higher in the first variant A (Figs. 2, 3 and 4). As Fig. 5 indicates in the simplest integration scheme A the probe platform type groups are distinguishable in contrast to the other two methods (see Figs. 6 and 7). The results of the hierarchical clustering analysis confirm that the main variability source of 2998 most variable genes in the data integrated with strategy A is related with microarray platform group and not with the biological classes (see Figs. 8, 9 and 10).

Results of cancer type multiclass classification obtained using model built in Spicy tool are presented in Table 1. The best bootstrap based accuracy rate is obtained using the integration based on Combat algorithm (scheme B). Lowest accuracy is for the simplest integration variant based on scheme A. We denote significant difference between scheme A and the other two methods based on

Limma and Combat algorithm integration scheme (p-value < 0.05). Independently of gene number chose the data integration is necessary to obtain the best classification rate in multiplatform microarray data analysis (see Fig. 11).

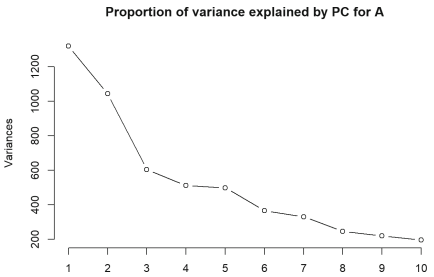

Fig. 2. The variability retained by each principal components in PCA analysis for data merging scheme A

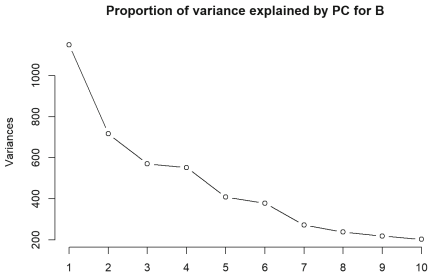

Fig. 3. The variability retained by each principal components in PCA analysis for data merging scheme B

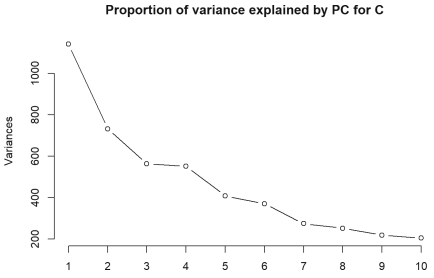

Fig. 4. The variability retained by each principal components in PCA analysis for data merging scheme C

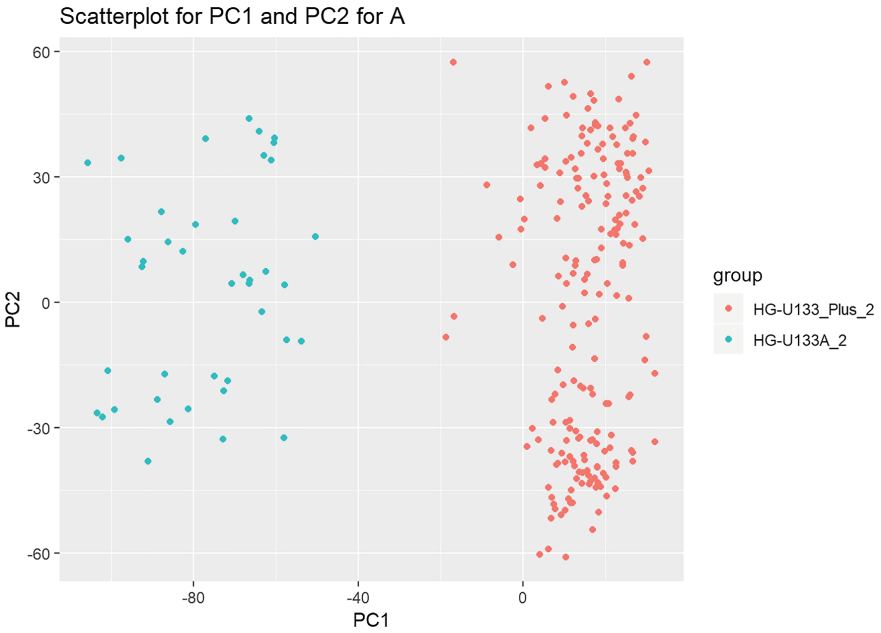

Fig. 5. PCA based probe groups analysis for data merging scheme A (Color figure online)

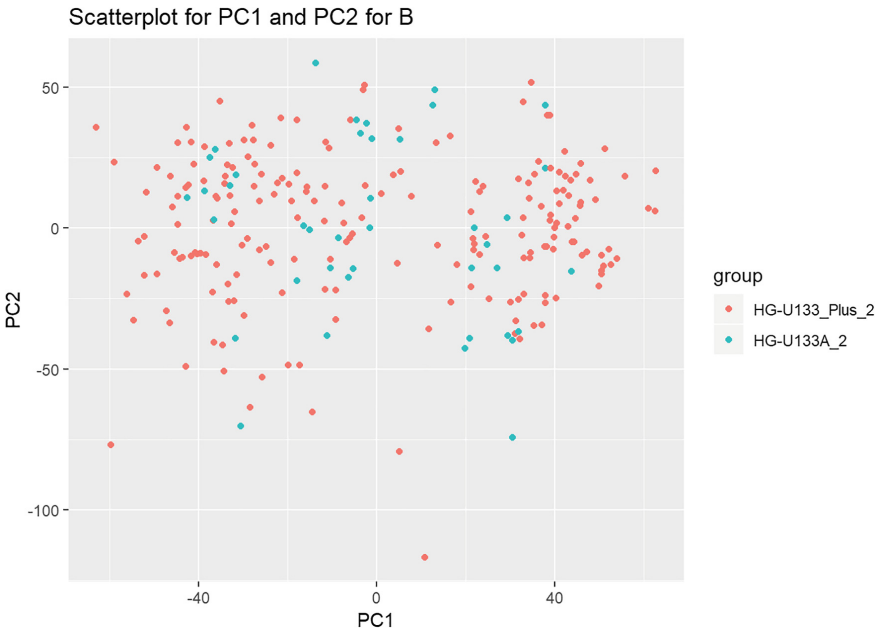

Fig. 6. PCA based probe groups analysis for data merging scheme B (Color figure online)

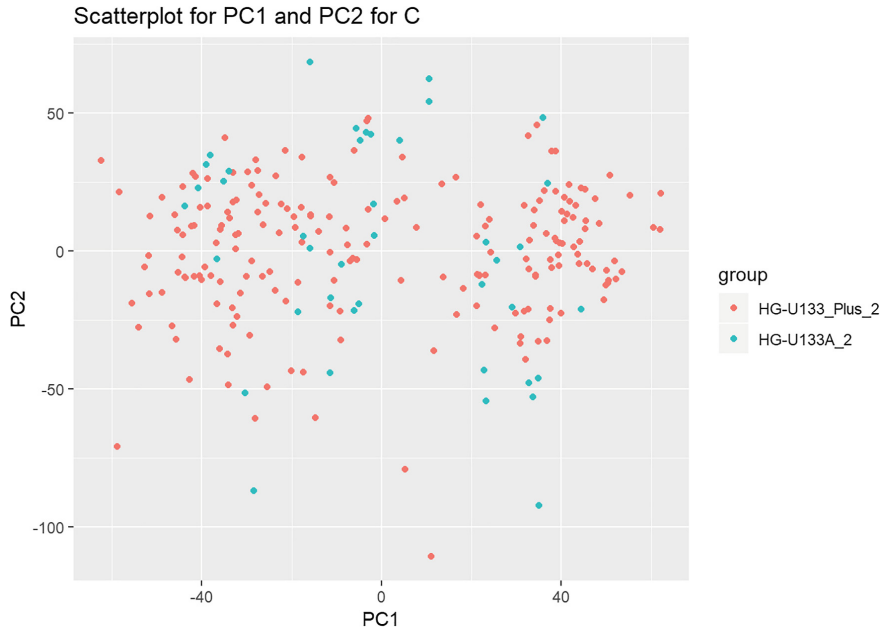

Fig. 7. PCA based probe groups analysis for data merging scheme C (Color figure online)

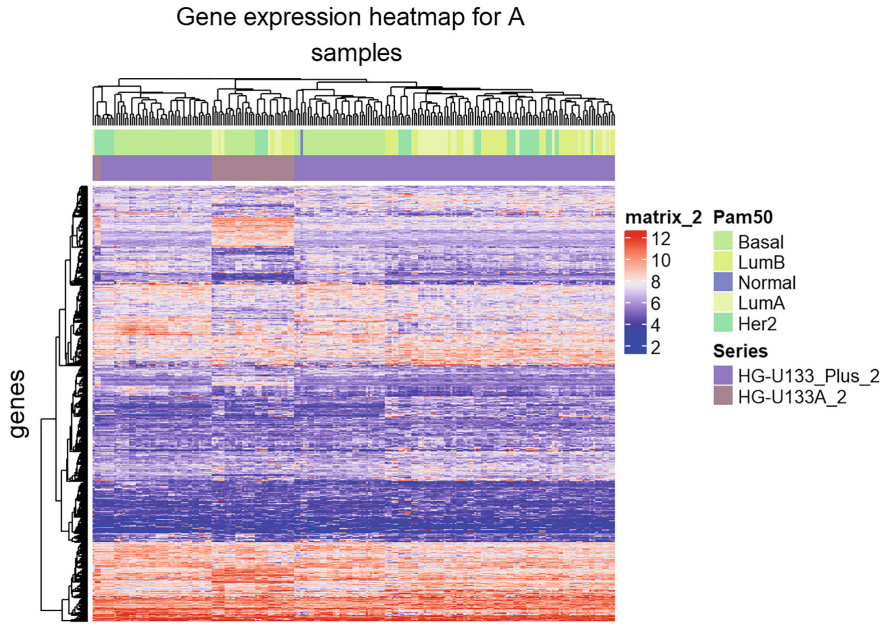

Fig. 8. Heatmap with 2998 most important features for data merging scheme A

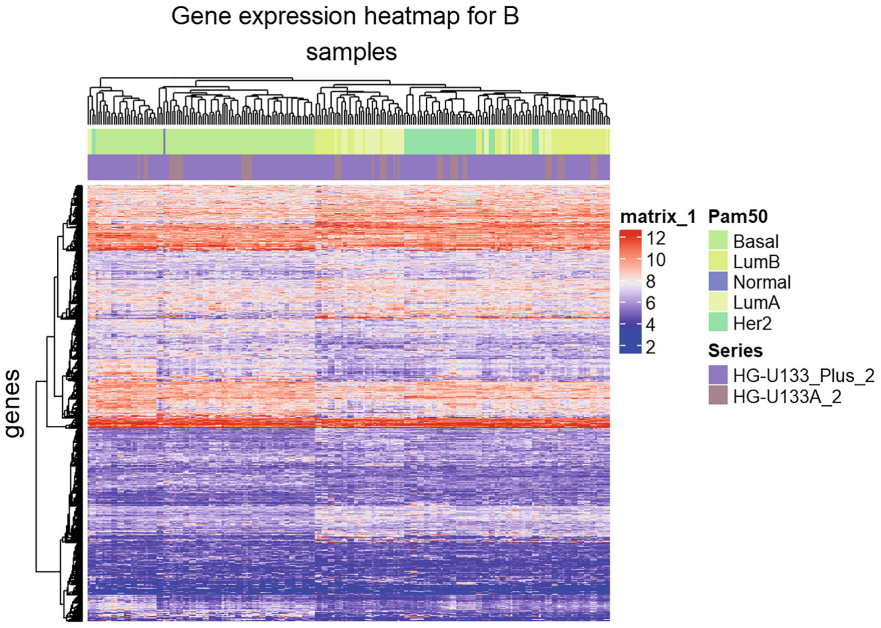

Fig. 9. Heatmap with 2998 most important features for data merging scheme B

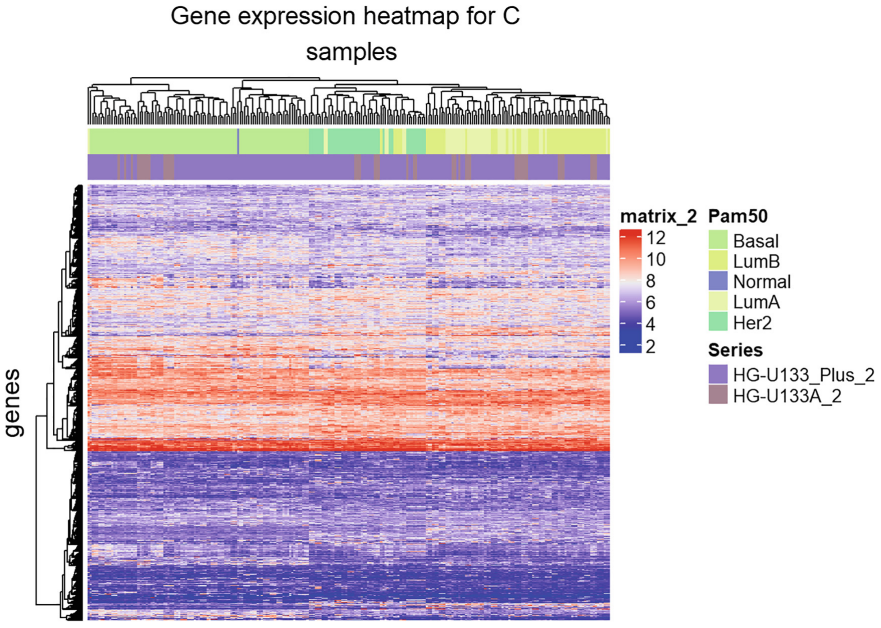

Fig. 10. Heatmap with 2998 most important features for data merging scheme C

Table 1. Bootstrap based classification accuracies with 95% confidence intervals and t-test statistics based p-values of data obtained using different combine strategies from two different microarray platforms. Strategy A - simple data merging, Strategy B - Combat algorithm, Strategy C - Limma algorithm

Algorithm	Number of features	acc	accL	accH	p-values A	p-values B	p-values C
Scheme A	20	80.93	80.59	81.28	–	0.021	0.029
Scheme B	20	81.46	81.11	81.81	0.021	–	0.903
Scheme C	20	81.39	81.04	81.75	0.029	0.903	–

on

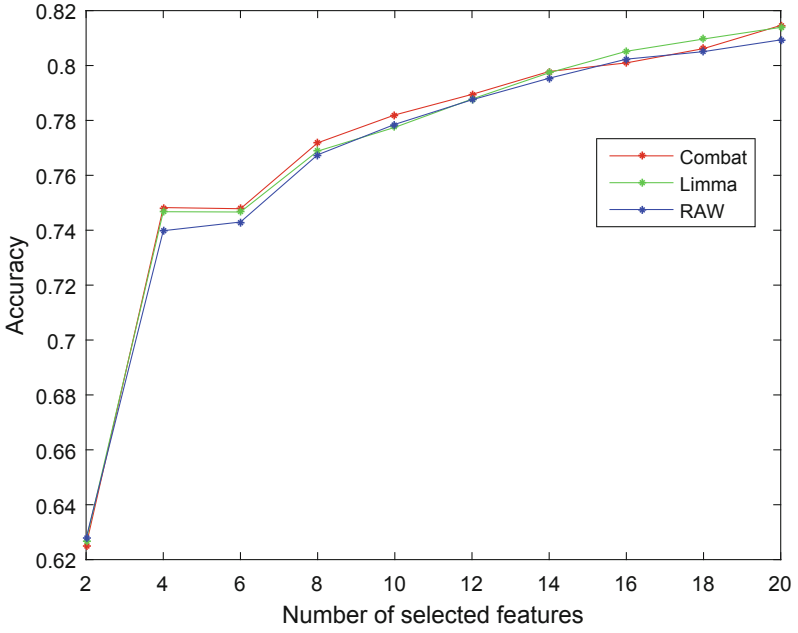

Fig. 11. Accuracy obtained for different number of selected features.

4 Conclusion

Our results show that for combining data from two different microarray platforms one need to take care about the batch effect. The general conclusion is that in case of multi-platform microarray analysis the standard data normalization is insufficient to remove the differences between compatible Affymetrix HG-U133 Plus 2 and Affymetrix HG-U133 A2 platforms. In our case the Combat algorithm was slightly better in the terms of multi-class classification problem.

Acknowledgement. This work was supported by Polish National Centre for Research and Development under Grant Strategmed2/267398/4/NCBR/2015 and Silesian University of Technology Grant 02/010/BK_18/0102 and by Polish Ministry of Science and Higher Education as part of the Implementation Doctorate program at the Silesian University of Technology, Gliwice, Poland (contract No 10/DW/2017/01/1) (AP). Data analysis was partially carried out using the Biotest Platform developed within Project n. PBS3/B3/32/2015 financed by the Polish National Centre of Research and Development (NCBiR). Calculations were performed using the infrastructure supported by the computer cluster Ziemowit (www.ziemowit.hpc.polsl.pl) funded by the Silesian BIOFARMA project No. POIG.02.01.00-00-166/08 and expanded in the POIG.02.03.01-00-040/13 in the Computational Biology and Bioinformatics Laboratory of the Biotechnology Centre at the Silesian University of Technology. This work was partially supported by the Polish Ministry of Science and Higher Education as part of the Implementation Doctorate program at the Silesian University of Technology, Gliwice, Poland (contract No 10/DW/2017/01/1).

References

1. Lander, E.S.: Array of hope. Nat. Genet. **21**, 3–4 (1999)
2. Fare, T.L., et al.: Effects of atmospheric ozone on microarray data quality. Anal. Chem. **75**(17), 4672–4675 (2003)
3. Luo, J., et al.: A comparison of batch effect removal methods for enhancement of prediction performance using MAQC-II microarray gene expression data. Pharmacogenomics J. **10**, 278–291 (2010)
4. Leek, J.T., et al.: Tackling the widespread and critical impact of batch effects in high-throughput data. Nat. Rev. Genet. **11**, 733–739 (2010)
5. Simek, K., et al.: Using SVD and SVM methods for selection, classification, clustering and modeling of DNA microarray data. Eng. Appl. Artif. Intell. **17**(4), 417–427 (2004)
6. Lisowska, K., et al.: Unsupervised analysis reveals two molecular subgroups of serous ovarian cancer with distinct gene expression profiles and survival. J. Cancer Res. Clin. Oncol. **142**, 1239–1252 (2016)
7. Fujarewicz, K., Kimmel, M., Rzeszowska-Wolny, J., Swierniak, A.: A note on classification of gene expression data using support vector machines. J. Biol. Syst. **11**(1), 43–56 (2003)
8. Chen, C., et al.: Removing batch effects in analysis of expression microarray data: an evaluation of six batch adjustment methods. PLoS One **6**(2), e17238 (2011)
9. Perou, C.M., et al.: Adjustment of systematic microarray data biases. Bioinformatics **20**(1), 105–114 (2004)
10. Coletta, A., et al.: Batch effect removal methods for microarray gene expression data integration: a survey. Brief. Bioinform. **14**(4), 469–490 (2012)
11. Alter, O., Brown, P.O., Botstein, D.: Singular value decomposition for genome-wide expression data processing and modeling. Proc. Nat. Acad. Sci. **97**(18), 10101–10106 (2000)
12. Leek, J.T., Storey, J.D.: Capturing heterogeneity in gene expression studies by surrogate variable analysis. PloS Genet. **3**(9), 1724–1735 (2007)
13. Sims, A.H., et al.: The removal of multiplicative, systematic bias allows integration of breast cancer gene expression datasets–improving meta-analysis and prediction of prognosis. BMC Med. Genomics **1**, 42 (2008)

14. Sandberg, R., Larsson, O.: Improved precision and accuracy for microarrays using updated probe set definitions. BMC Bioinform. **8**, 48 (2007)
15. McCall, M.N., Bolstad, B.M., Irizarry, R.A.: Frozen robust multiarray analysis (fRMA). Biostatistics (Oxford, England) **11**(2), 242–253 (2010)
16. Li, C., Johnson, W.E., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics **8**(1), 118–127 (2006)
17. Müller, C., et al.: Removing batch effects from longitudinal gene expression - quantile normalization plus combat as best approach for microarray transcriptome data. PLoS One **11**, e0156594 (2016)
18. Cai, H., et al.: Identifying differentially expressed genes from cross-site integrated data based on relative expression orderings. Int. J. Biol. Sci. **14**, 892–900 (2018)
19. Student, S., Fujarewicz, K.: Stable feature selection and classification algorithms for multiclass microarray data. Biol. Direct **7**(33), 1–20 (2012)
20. Fujarewicz, K., et al.: Large-scale data classification system based on galaxy server and protected from information leak. In: Nguyen, N.T., Tojo, S., Nguyen, L.M., Trawiński, B. (eds.) ACIIDS 2017. LNCS (LNAI), vol. 10192, pp. 765–773. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-54430-4_73
21. Pojda, K., Jakubczak, M., Student, S., Świerniak, A., Fujarewicz, K.: Comparing different data fusion strategies for cancer classification. In: Rocha, Á., Guarda, T. (eds.) ICITS 2018. AISC, vol. 721, pp. 417–426. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-73450-7_40