RUHR-UNIVERSITÄT BOCHUM

BI006

RUB

mpc MEDIZINISCHES PROTEOM CENTER

Faculty of Medicine
Medizinisches Proteom-Center
WG Bioinformatics/Biostatistics

Medizinisches Proteom-Center
WG Bioinformatics/Biostatistics
Ruhr-Universität Bochum, Germany
michael.turewicz@rub.de
www.medizinisches-proteom-center.de

PAA - A New R Package for Autoimmune Biomarker Discovery with Protein Microarrays

Michael Turewicz, Maike Ahrens, Caroline May, Dirk Woitalla, Ralf Gold, Swaantje Casjens, Beate Pesch, Thomas Brüning, Jozsef I. Engelhardt, Katrin Marcus, Axel Mosig, Helmut E. Meyer, and Martin Eisenacher

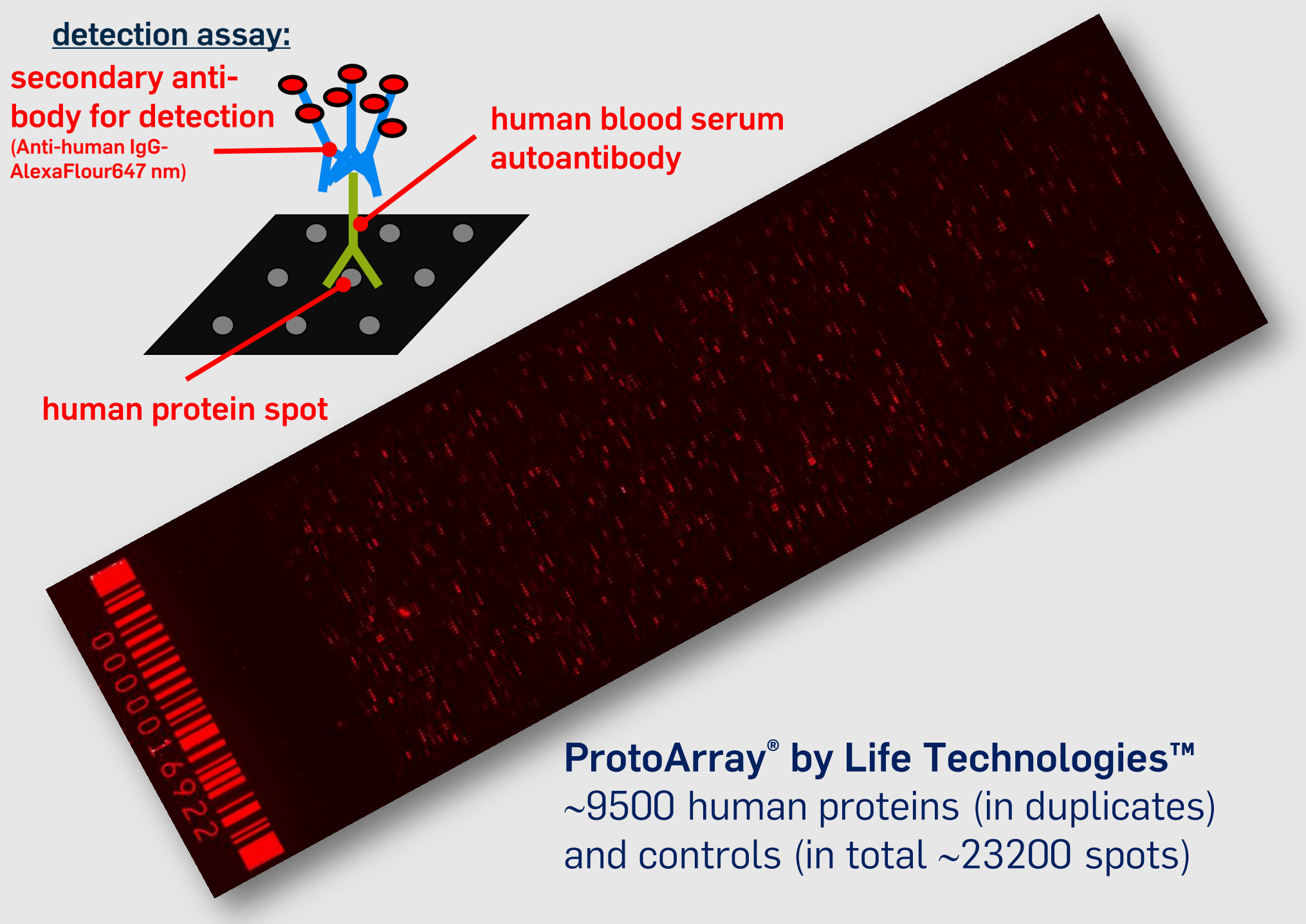
Abstract

Background: Protein microarrays like the ProtoArray (Life Technologies, Carlsbad, CA, USA) are used for autoimmune antibody screening studies to discover biomarker panels. For ProtoArray data analysis the software Prospector (Life Technologies) is often used because it provides an advantageous feature ranking approach ("M score"). Unfortunately, Prospector provides no capabilities regarding multivariate feature selection, classification, batch effect adjustment, and computational biomarker candidate validation.

Results: Therefore, we have adopted Prospector's M score approach and implemented a new R package called Protein Array Analyzer (PAA) that provides these features and a complete data analysis pipeline for ProtoArray and other single color microarray data that come in gpr file format. After optional data pre-processing and M score-based feature pre-selection a multivariate feature selection is performed. For this purpose, a backwards elimination (wrapper) approach ("gene shaving" with random forest) has been implemented. For the selection and validation of stable panels a frequency-based approach has been adopted. Furthermore, different plots and results files can be obtained to outline the analysis results.

Conclusions: We propose the new R package PAA for protein microarray data analysis. PAA has been used to successfully analyse several different ProtoArray data sets (e.g. "Parkinson", "Alzheimer", "Amyotrophic Lateral Sclerosis"). Thereby, its suitability for biomarker discovery with protein microarrays has been shown.

ProtoArray


detection assay:
secondary anti-body for detection
(Anti-human IgG-AlexaFlour647 nm)
human blood serum autoantibody
human protein spot
000001692

ProtoArray® by Life Technologies™
~9500 human proteins (in duplicates) and controls (in total ~23200 spots)

Fig. 1: The ProtoArray® protein microarray.

Raw Data Pre-Processing

PAA provides the following normalization methods: cyclic loess, quantile and vsn (wrapper to limma **(1)**) as well as robust linear **(2)**. Furthermore, PAA provides plots to compare the results of the different normalization approaches (see figure 2 and 3).

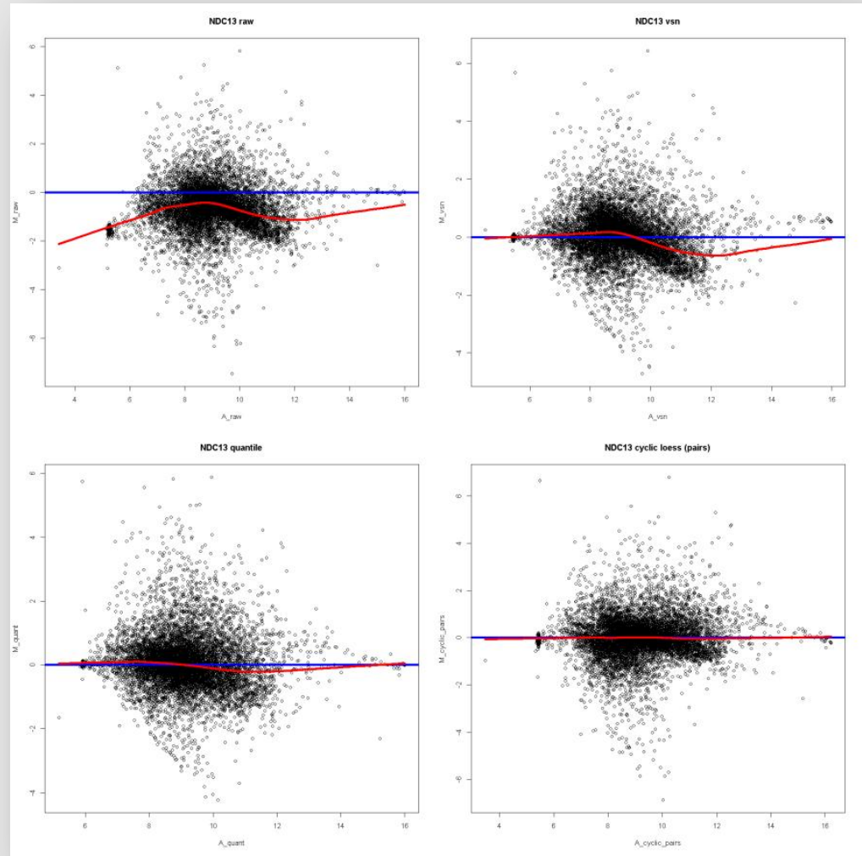


Fig. 2: Comparing normalization results with MA plots.

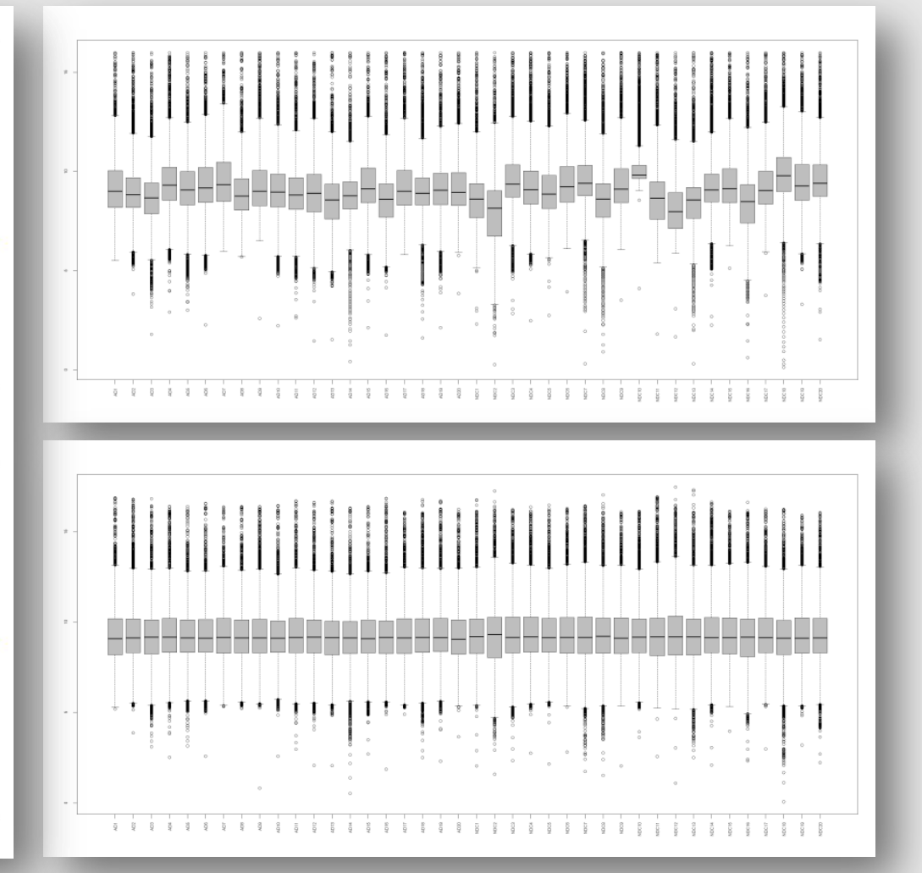


Fig. 3: Comparing normalization results with box plots.

PURE

This work was funded from P.U.R.E. (Protein Unit for Research in Europe), a project of Nordrhein-Westfalen, a federal state of Germany.

Univariate Pre-Selection

For univariate feature pre-selection PAA provides the "minimum M Statistic" ("M Score", **(3)**), a measure that is sensitive for significant subgroups:

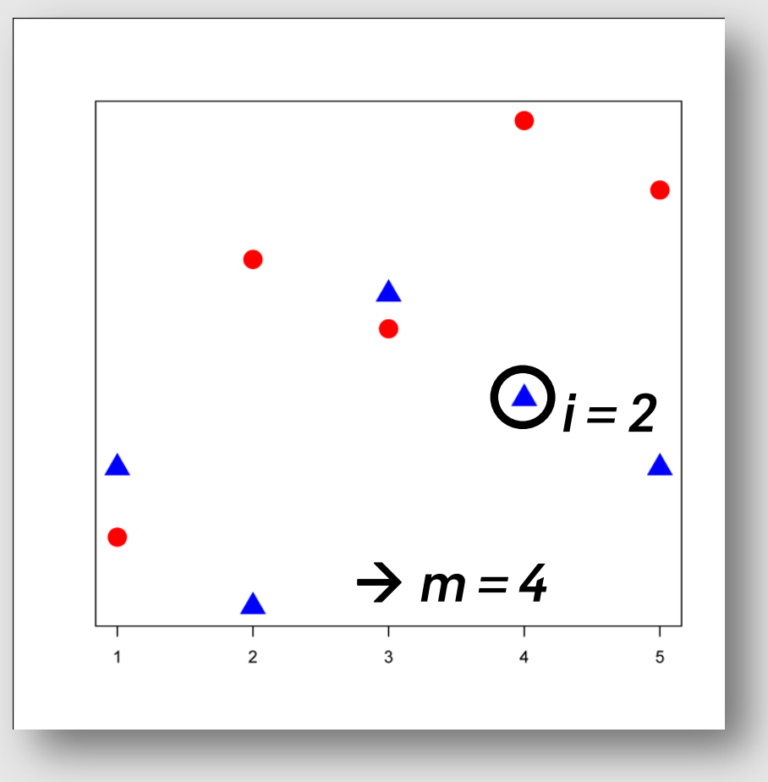


Fig. 4: The basic idea of the M Score for a given feature is to count its number of intensity values in one group ("*m*") ≥ the *i*-th largest value in the other group. Then, the probability of this scenario is computed. This is done for all *i*-th largest values and both group perspectives. Finally, the M Score is set to the smallest probability for this feature.

More precisely:

Step 1: To rate a feature, count its number of values in group j ≥ the i-th largest value in group (l-j):

$$M_j^{(i)} = \sum_{k=1}^{n_j} 1\left(x_{j,k} \geq x_{l-j,(i)}\right) = m,$$

$$\begin{aligned}
&\text{where:}\\
&j \in \{1,2\}: \text{the current group},\\
&l = |\{1,2\}| + 1 = 3,\\
&(l-j): \text{the "other group"},\\
&n_j: \text{number of values in group } j,\\
&n_{(l-j)}: \text{number of values in group } (l-j),\\
&i \in \{1,\dots,n_{l-j}\},\\
&x_{l-j,(i)}: i\text{-th largest value in } (l-j),\\
&1\left(x_{j,k} \geq x_{l-j,(i)}\right) = \begin{cases} 1, & \mathit{if}\ x_{j,k} \geq x_{l-j,(i)} \\ 0, & \mathit{otherwise} \end{cases}
\end{aligned}$$

Step 2: Then, compute the probability of having at least such a m-value:

$$P\left(M_j^{(i)} \geq m\right) = \sum_{k=m}^{n_j} P\left(M_j^{(i)} = k\right), \text{where}$$

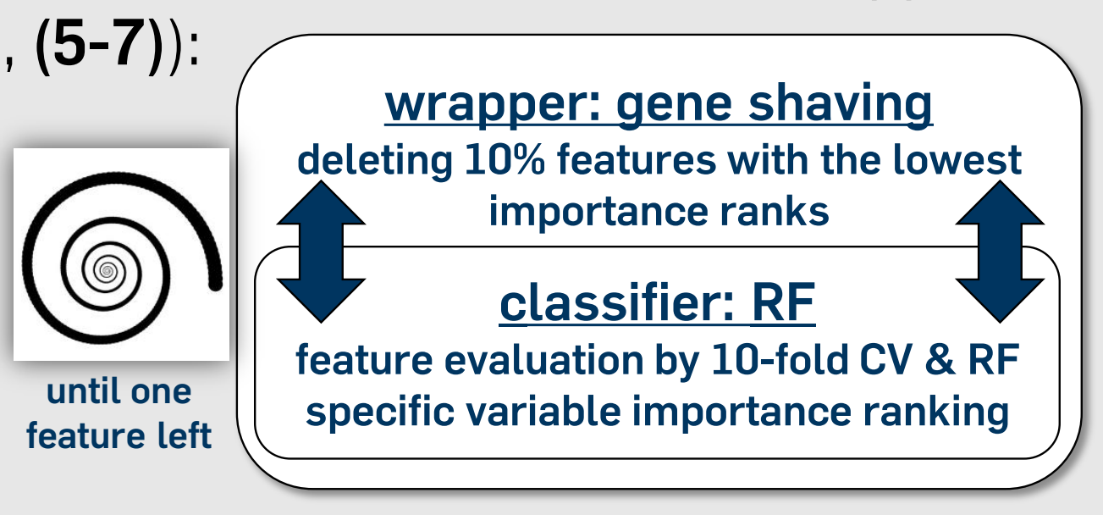
$$P\left(M_j^{(i)} = k\right) = \frac{\binom{n_j + n_{(l-j)} - k - i}{n_{(l-j)} - i}\binom{k+i-1}{i-1}}{\binom{n_j + n_{(l-j)}}{n_j}}$$

Step 3: The reported M score for the considered protein is computed by minimizing the latter probabilities with respect to both groups and all i-th largest values:

$$\text{M score} = \min_{j \in \{1,2\}}\left(\min_{i \in \{1,\dots,n_{l-j}\}} P\left(M_j^{(i)} \geq m\right)\right)$$

Gene Shaving

For multivariate feature selection PAA provides a random forest (RF)-based **(4)** backwards elimination wrapper method (gene shaving, GS, **(5-7)**):


until one feature left
wrapper: gene shaving
deleting 10% features with the lowest importance ranks
classifier: RF
feature evaluation by 10-fold CV & RF specific variable importance ranking

GS starts with the set containing all pre-selected protein features (***S*****₁**). In the following iterations RF models are trained and evaluated (10-fold cross validation) on variable set $\boldsymbol{S_i}$ to receive the variable ranking vector $\boldsymbol{imp_i}$ and the classification accuracy $\boldsymbol{acc_i}$ for that variable set. Then, the 10% weakest variables are discarded to receive $\boldsymbol{S_{i+1}}$. Finally, the variable set ($\boldsymbol{S_{best}}$) with the best accuracy value ($\boldsymbol{acc_{best}}$) is obtained and classification within the independent test set is performed for returning the overall accuracy.

References

(1) The package limma by Gordon Smyth et al. can be downloaded from Bioconductor http://www.bioconductor.org/).

(2) Sboner A, et al.: Robust-linear-model normalization to reduce technical variability in functional protein microarrays. J Proteome Res 2009, 8(12):5451-5464.

(3) Love B: The Analysis of Protein Arrays. In: *Functional Protein Microarrays in Drug Discovery.* CRC Press; 2007: 381-402.

(4) Liaw A WM: Classification and Regression by randomForest. R News 2002, 2(3):18-22.

(5) Hastie T, et al.: 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. Genome Biol 2000, 1(2):RESEARCH0003.

(6) Díaz-Uriarte R, Alvarez de Andrés S: Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006, 7:3.

(7) Jiang H, et al.: Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. BMC Bioinformatics 2004, 5:81.

(8) Michiels S, Koscielny S, Hill C: Prediction of cancer outcome with microarrays: a multiple random validation strategy. Lancet 2005, 365(9458):488-492.

(9) Baek S, Tsai CA, Chen JJ: Development of biomarker classifiers from high-dimensional data. Brief Bioinform 2009, 10(5):537-546.

(10) Nagele E, et al.: Diagnosis of Alzheimer's disease based on disease-specific autoantibody profiles in human sera. PLoS One 2011, 6: e23112.

(11) Han M, et al: Diagnosis of Parkinson's disease based on disease-specific autoantibody profiles in human sera. PLoS One 2012, 7: e32383.

Frequency-Based Validation

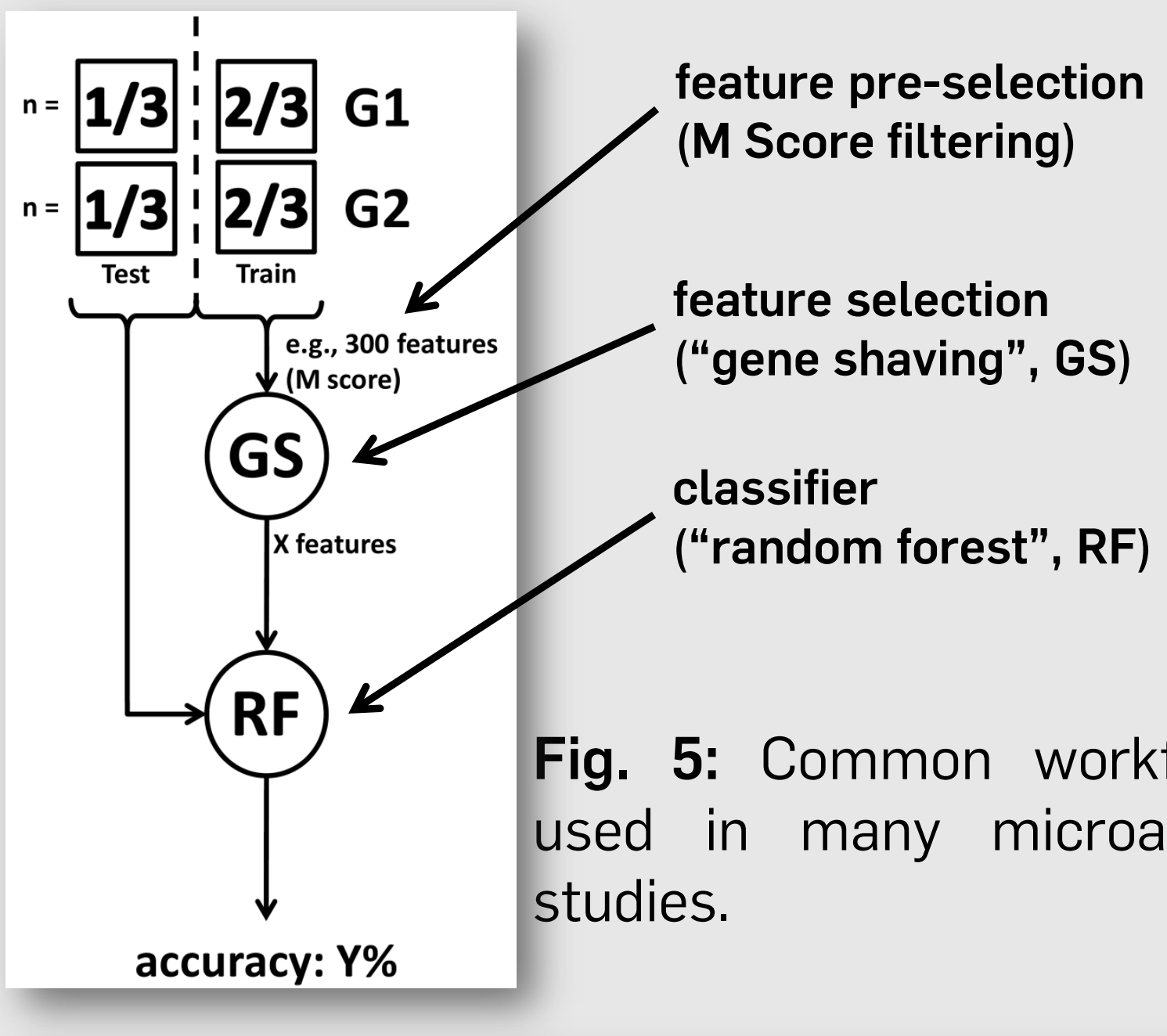

n =
1/3
2/3
G1
n =
1/3
2/3
G2
Test
Train
e.g., 300 features (M score)
GS
X features
RF
accuracy: Y%
feature pre-selection (M Score filtering)
feature selection ("gene shaving", GS)
classifier ("random forest", RF)

Fig. 5: Common workflow used in many microarray studies.

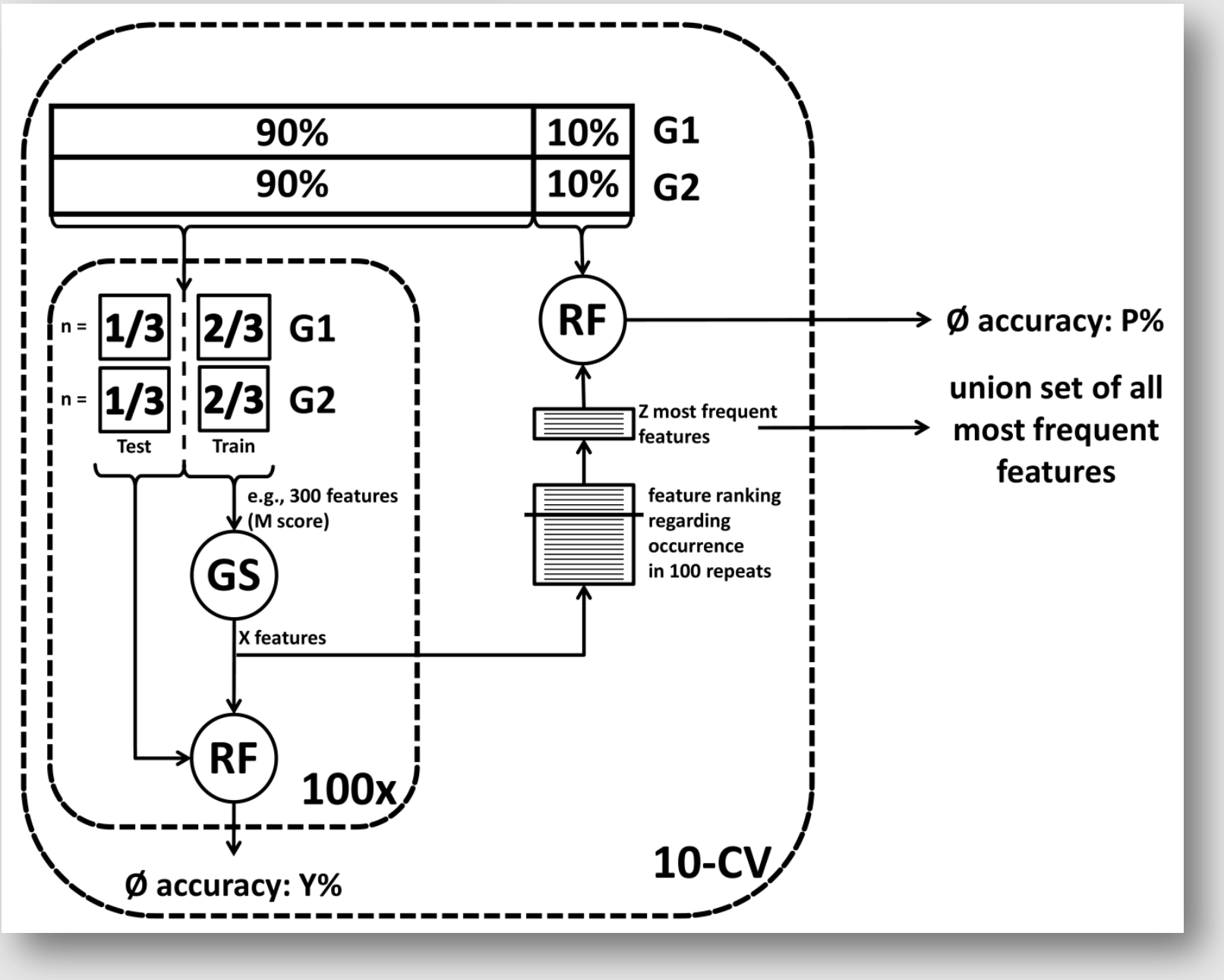


Fig. 6: The frequency-based **(8, 9)** PAA workflow selects most frequent features from 10-fold cross-validated 100 GS repeats and reports the average accuracy.

Exemplary Data

data set	G1	G2	N1	N2	P	Ø accuracy
"Alzheimer"*	AD	NDC	20	20	24	86.075%
"Amyotrophic Lateral Sclerosis"**	ALS	NDC	20	20	30	90.725%
"Parkinson"*** (biased!)	PD	NDC	29	20	18	100%

Table 1: Analysis results of three exemplary two-group data sets are shown. Column "G1" shows the label of the respective diseased group and "G2" the label of the non diseased controls ("NDC"). Columns "N1" and "N2" show the number of samples in G1 and G2. Finally, column "P" shows the number of selected features and column "Ø accuracy" the average classification accuracy of the 10-fold cross validation.

*GEO record "GSE29676", (10)

**ProtoArray study conducted in our lab

***GEO record "GSE29654", (11)

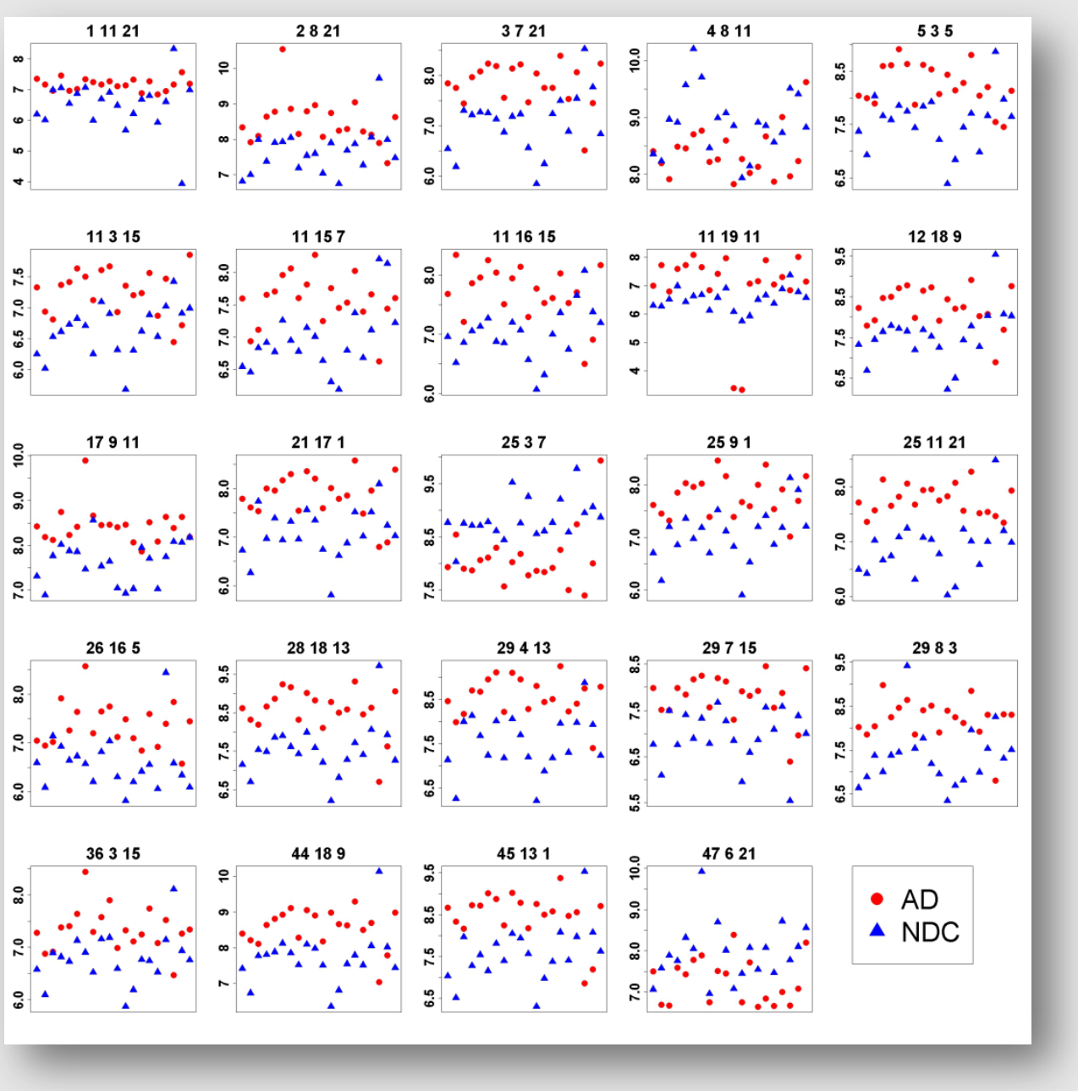


Fig. 7: PAA plots the intensities of all selected features (one sub-plot per feature) in group-specific colors. These plots can be used to check whether the selected features are differential.

Availability

A beta version of PAA can be downloaded from *http://www.medizinisches-proteom-center.de/PAA* or requested via *michael.turewicz@rub.de*