6

Metabolic Biomarker Identification with Few Samples

Pietro Franceschi, Urska Vrhovsek, Fulvio Mattivi and Ron Wehrens
IASMA Research and Innovation Centre
Via E. Mach, 1 38010 S. Michele all'Adige (TN)
Italy

1. Introduction

Biomarker selection represents a key step in bioinformatic data processing pipelines; examples range from DNA microarrays (Tusher et al., 2001; Yousef et al., 2009) to proteomics (Araki et al., 2010; Oh et al., 2011) to metabolomics (Chadeau-Hyam et al., 2010). Meaningful biological interpretation is greatly aided by identification of a "short-list" of features – biomarkers – characterizing the main differences between several states in a biological system. In a two-class setting the biomarkers are those variables (metabolites, proteins, genes ...) that allow discrimination between the classes. A class or group tag can be used to distinguish many situations: it can be used to discriminate between treated and non-treated samples, to mark different varieties of the same organism, etcetera. In the following, we will – for clarity – restrict the discussion to metabolomics, and the variables will constitute concentration levels of metabolites, but similar arguments hold *mutatis mutandis* for other -omics sciences, such as proteomics and transcriptomics, where the variables correspond to protein levels or expression levels, respectively.

There are several reasons why the selection of biomarker short-lists can be beneficial:

- Predictive purposes: using only a small number of biomarkers in predictive class modeling in general leads to better, i.e., more robust and more accurate predictions.
- Interpretative purposes: it makes sense to first concentrate on those metabolites that show clear differences in levels in the different classes, since our knowledge of metabolic networks in many cases is only scratching the surface.
- Discovery purposes: the complete characterization of unknown compounds identified in untargeted experiments is time- and resource-consuming. The primary focus should thus be placed on a carefully selected group of "unknowns" to be characterized at structural and functional level.

Two fundamentally different statistical approaches to biomarker selection are possible. With the first, experimental data can be used to construct multivariate statistical models of increasing complexity and predictive power – well-known examples are Partial Least Square Discriminant Analysis (PLS-DA) (Barker & Rayens, 2003; Kemsley, 1996; Szymanska et al., 2011) or Principal Component Linear Discriminant Analysis (PC-LDA) (Smit et al., 2007; Werf et al., 2006). Inspection of the model coefficients then should point to those variables that are important for class discrimination. As an alternative, univariate statistical tests can be

applied to individual variables, treating each one independent of the others and indicating which of them show significant differences between groups (see, e.g., Guo et al. (2007); Reiner et al. (2003); Zuber & Strimmer (2009)). Multivariate techniques are potentially more powerful in pin-pointing weak differences because they take into account correlation among the variables, but the models can be too much adapted to the experimental data, leading to poor generalization capacity. Univariate approaches, in contrast, both could miss important "weak" details and could overestimate the importance of certain variables, because correlation between variables is not taken into account.

As for many sciences with the "omics" suffix, in metabolomics the number of experimental variables usually greatly exceeds the number of objects, especially with the development of new mass-spectrometry-based technologies. In MS-based metabolomics, high resolution mass spectrometers are often coupled with high performance chromatographic techniques, like Ultra Performance Liquid Chromatography (UPLC). In these experiments, the variables, i.e., the metabolites, are represented by mass/retention-time combinations, and it is typical to have numbers of features varying from several hundreds to several thousands, depending on the experimental and analytical conditions. This increase in experimental possibilities, however, does not correspond to a proportional increase in the number of available samples, which can be limited by the availability of biological samples, by laboratory practice, in particular when complex protocols are required, and also by ethical issues, when, for example, experiments on animals have to be planned.

All these constraints produce *small sample sets*, presenting serious challenges for the statistical analysis, mainly because there is simply not enough information to model the natural biological variability. The situation is critical for multivariate approaches where the parameters of the statistical model need to be optimized (e.g., the number of components in a PLS-DA model). For this purpose, the classical approach is to use sub-sampling in combination with estimates of predictive power, like crossvalidation (Stone, 1974). In extreme conditions, i.e., really small sample sizes, this sub-sampling can give rise to inconsistent sub-models and tuning in the classical way becomes virtually impossible. In Hanczar et al. (2010), as an example, conclusions are focussed on ROC-based statistics (see below), but they are equally relevant for classical error estimates like the root-mean-square error of prediction, RMSEP) multivariate techniques can be still applied to the full data set, but it is not possible to assess the reliability of the biomarker selection pipeline, even if it is still reasonable to think that the biomarkers are strongly contributing to the statistical model. In these situations, univariate methods seem the best solution, also considering the presence of several strategies able to determine cut-off values in t-test based techniques (e.g., thresholding of p values subjected to some form of multiple testing correction (Benjamini & Hochberg, 1995; Noble, 2009; Reiner et al., 2003)). Regardless of the statistical strategy, for the "biomarkers" extracted in these conditions there is no obvious validation possible in the statistical sense; however, the results of the experiments are extremely important in the hypothesis generation phase to plan more informative investigations.

Interestingly, there is no literature on the effect of sample size on biomarker identification in the "omics" sciences, and the objective of this contribution is to fill this gap. We focus on a two-class problem, and in particular on small data sets. In our approach, real class differences have been introduced by spiking apple extracts with selected compounds, analyzing them using UPLC-TOF mass spectrometry, and comparing the feature lists to those of unspiked apple extracts. Using these data we are able to run a comparison between two multivariate

methods (PLS-DA and PC-LDA) and the univariate *t*-test, leading to at least a rough estimate of how consistent biomarker discovery can be when small sample sizes are considered. In particular, we compare the effect of sample size reduction on multivariate and univariate models on the basis of Receiver Operating Characteristics (ROC) (Brown & Davis, 2005).

2. Material and methods

2.1 Biomarker Identification

There are many strategies for identifying differentially expressed variables in two-class situations – a recent overview can be found in Saeys et al. (2007). A general approach is to construct a model with good predictive properties, and to see which variables are important in such a model. Given the low sample-to-variable ratio, however, one can not expect to be able to fit very complicated models, and in many cases a linear model is the best one can do (Hastie et al., 2001). The oldest, and most well-known technique is Linear Discriminant Analysis (LDA, McLachlan (2004)). One formulation of this technique, dating back to R.A. Fisher, is to find a linear combination of variables $\boldsymbol{a}$ that maximizes the ratio of the between-groups sums of squares, $\boldsymbol{B}$, and the within-groups sums of squares $\boldsymbol{W}$:

$$\boldsymbol{a}^T \boldsymbol{B} \boldsymbol{a} / \boldsymbol{a}^T \boldsymbol{W} \boldsymbol{a} \tag{1}$$

That is, $\boldsymbol{a}$ is the direction that maximizes the separation between the classes, both by having compact classes (a small within-groups variance) and by having the class centers far apart (a large between-groups variance). Large values in $\boldsymbol{a}$ indicate which variables are important in the discrimination. Another formulation is to calculate the Mahalanobis distance of a new sample $\boldsymbol{x}$ to the class centers $\boldsymbol{\mu}_i$:

$$d(\boldsymbol{x}, i) = (\boldsymbol{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu}_i) \tag{2}$$

The new sample is then assigned to the class of the closest center. This approach is equivalent to Fisher's criterion for two classes (but not for more than two classes). In this equation, $\boldsymbol{\Sigma}$ is the (estimated) pooled covariance matrix of the classes. If the Mahalanobis distance to each class center is calculated using the individual class covariance matrices, the result is Quadratic Discriminant Analysis (QDA), which as the name suggests, no longer leads to linear class boundaries. A final formulation is to use regression using indicator variables for the class. In a two-class situation one can use, e.g., the values of -1 and 1 for the two classes; positive predictions will be assigned to class one, and negative predictions to class -1. In many other cases, 0 and 1 are used, and the class threshold is put at 0.5. When there are more than two classes, one can use a separate column in the dependent variable for every class – if a sample belongs to that class the column should contain 1, else 0. Again, the size of the regression coefficients indicates which of the variables contribute most to the discrimination.

For most applications in the "omics" fields, even the most simple multivariate techniques such as Linear Discriminant Analysis (LDA) cannot be applied directly. From Equation 2 it is clear that an inverse of the the covariance matrix $\boldsymbol{\Sigma}$ needs to be calculated, which is impossible in cases where the number of variables exceeds the number of samples. In practice, the number of samples is nowhere near the number of variables. For QDA, the situation is even worse: to allow a stable matrix inversion, every single class should have at least as many samples as variables (and preferably quite a bit more). A common approach is to compress the information in the data into a low number of latent variables (LVs), either using PCA (leading

to PC-LDA, e.g. Smit et al. (2007); Werf et al. (2006)) or PLS (which gives PLS-DA; see Barker & Rayens (2003); Kemsley (1996)), and to perform the discriminant analysis on the resulting score matrices. These are not only of low dimension, but also orthogonal so that the matrix inversion, the calculation of $\mathbf{\Sigma}^{-1}$, can be performed very fast and reliably. Both for PC-LDA and PLS-DA, the problem is more often usually cast in a regression context, where again the response variable $\boldsymbol{Y}$ can take values of either 0 or 1. The model thus becomes:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \mathcal{E} \approx \boldsymbol{T}\boldsymbol{P}^T\boldsymbol{B} + \mathcal{E} \tag{3}$$

where $\mathcal{E}$ is the matrix of residuals. Matrix $\boldsymbol{X}$ is decomposed into a score matrix $\boldsymbol{T}$ and a loading matrix $\boldsymbol{P}$, both consisting of a very low number of latent variables, typically less than ten or twenty. The coefficients for the scores, $\boldsymbol{A} = \boldsymbol{P}^T\boldsymbol{B}$, can therefore be easily be calculated in the normal way of least-squares regression:

$$\boldsymbol{A} = (\boldsymbol{T}^T\boldsymbol{T})^{-1}\boldsymbol{T}^T\boldsymbol{Y} \tag{4}$$

which by premultiplication with $\boldsymbol{P}$ lead to estimates for the overall regression coefficients $\boldsymbol{B}$:

$$\boldsymbol{B} = \boldsymbol{P}\boldsymbol{A} \tag{5}$$

These equations are the same for both PLS-DA and PC-LDA. The difference lies in the decomposition of $\boldsymbol{X}$. In PC-LDA, $\boldsymbol{T}$ and $\boldsymbol{P}$ correspond to the scores and loadings, respectively, from PCA. That is, the class of the samples is completely ignored, and the only criterion is to capture as much variance as possible from $\boldsymbol{X}$. In PLS-DA, on the other hand, the scores and loadings are taken from a PLS model and the decomposition of $\boldsymbol{X}$ *does* take into account class information: the first PLS components by definition explain more, often much more, variance of $\boldsymbol{Y}$ than the first PCA components.

Both methods, PC-LDA as well as PLS-DA, are usually very sensitive to the choice of the number of LVs. Taking too few LVs will lead to bad predictions since important information is missed. Taking too many, the model will be too flexible and will show a phenomenon known as *overtraining*: it is more or less learning all the examples in the training set by heart but is not able to generalize and to make good predictions for new, unseen samples. As discussed, the assessment of the optimal number of LVs is neigh impossible with small sample sets. In the case under consideration, the extent of this effect is investigated by constructing several models with increasing numbers of LVs. Using real and simulated data sets (see below), models with 1–4, 6, and 8 LVs, respectively, are compared.

A simplification of statistical modeling can be obtained by ignoring all possible correlations between variables and assuming a diagonal covariance matrix, which leads to diagonal discriminant analysis (DDA). It can be shown that using the latter for feature selection corresponds to examining regular t-statistics (Zuber & Strimmer, 2009), and this is the approach we will take in this paper. For each variable, the difference between the class means $\bar{x}_{1i}$ and $\bar{x}_{2i}$ is transformed into a z-score by dividing by the appropriate standard deviation estimate s_i:

$$z_i = |\bar{x}_{1i} - \bar{x}_{2i}|/s_i \tag{6}$$

Using the appropriate number of degrees of freedom, these z-scores can be transformed into p values, which have the usual interpretation of the probability under the null hypothesis of encountering an observation with a value that is at least as extreme. In biomarker identification, p values can be used to sort the variables in order of importance and it is also

possible to decide a cut-off value to identify variables which show "significant" differences from the null hypothesis.

Generally speaking, the absolute size of coefficients is taken as a measure for the likelihood of being a true marker: the variable with the largest coefficient, in a PLS-DA model for example, is the first biomarker candidate, the second largest the second candidate, and so on. Note that this approach assumes that all variables have been standardized, i.e., scaled to mean zero and unit variance. This is often done in metabolomics to prevent dominance of highly abundant metabolites. Statistics from a *t*-test can be treated in the same way.

2.2 Quality assessment

To evaluate the performance of biomarker selections one typically relies on quantities like the fraction of true positives, i.e., that fraction of the real biomarkers that is actually identified by the selection method, and the false positives – those variables that have been selected but do not correspond to real differences. Similarly, true and false negatives can be defined. These statistics can be summarized graphically in an ROC plot (Brown & Davis, 2005), where the fraction of true positives (y-axis) is plotted against the fraction of false positives (x-axis). These two characteristics are also known as the sensitivity and the (complement of) specificity. An ideal biomarker identification method would lead to a position in the top left corner: all true biomarkers would be found (the fraction of true positives would be one, or close to one) with no or only very few false positives. Gradually relaxing the selection criterion, allowing more and more variables to be considered as biomarkers, generally leads to an increase in the true positive fraction (upwards in the plot), but also to an increase in the false positive fraction (in the plot to the right). The best biomarker selection method is obviously the one that finds all biomarkers very quickly, leading to a very steep ROC curve at the beginning.

A quantitative measure of the efficiency of a method can be obtained by calculating the area under the ROC curve (AUC). A value of one (or close to one) indicates that the method does a very good job in identifying biomarkers – all true biomarkers are found almost immediately. A value of one half indicates a completely random selection (this corresponds to the diagonal in the ROC plot). Values significantly lower than one half should not occur. In many cases, the most important area in the ROC plot is the left side, which indicates the efficiency of the model in selecting the most important biomarkers. Consequently, it is common to calculate a partial area under the curve (pAUC), for instance up to twenty percent of false positives (pAUC.2). In a method with higher pAUC, the true biomarkers will be present in the first positions of the candidate biomarkers list, hence this is the quantity that will be considered in the current paper.

2.3 Apple data set

Twenty apples, variety Golden Delicious, were purchased at the local store. Extracts of every single fruit were prepared according to Vrhovsek et al. (Vrhovsek et al., 2004). The core of the fruit was removed with a corer and each apple was cut into equal slices. Three slices (cortex and skin) from the opposite side of each fruit were used for the preparation of aqueous acetone extracts. The samples were homogenized in a blender Osterizer model 847-86 at speed one in a mixture of acetone/water (70/30 w/w). Before the injection, acetone was evaporated by rotary evaporation, the samples were brought back to the original volume with ethanol and were filtered with a 0.22 μm filter (Millipore, Bedford, USA). UPLC-MS spectra were

HPLC	ACQUITY UPLC (Waters)
Column	BEH C18 1.7 μm, 2.1*50 mm
Column temperature	40°C
Injection volume	5μl
Eluent flux	0.8 mlmin^{-1}
Solvent A	0.1% formic acid in H_2O
Solvent B	0.1% formic acid in MeOH
Gradient	linear gradient
	from 0 to 100% of solvent B in 10 minutes
	100% of B for 2 minutes
	100% A within 0.1 minutes
	Equilibration for 2.9 minutes.
Mass Spectrometer	SYNAPT Q-TOF (Waters)
Mass range	50-3000 Da.
Capillary	3 kV
Sampling cone	25 V
Extraction cone	3 V
Source temperatures	150°C
Desolvation temperatures	500°C
Cone gas flow	50 Lh^{-1}
Desolvation gas flow	1000 Lh^{-1}

Table 1. Chromatographic and spectrometric conditions of the spiked-apple data set.

acquired on a ACQUITY - SYNAPT Q-TOF (Waters, Milford, USA) in positive and negative ion mode with the chromatographic conditions summarized in Table 1. No technical replicates were performed. Raw data were transformed to the open NetCDF format by the DataBridge built-in utility of the MassLynx software.

Class differences were introduced by spiking ten of the twenty extracts with a number of selected compounds, leaving the other ten as "untreated" controls. The majority of the spiked compounds are known to be commonly present in apples, while two of them (*trans*-resveratrol and cyanidin-3-galactoside) are not naturally present in the chosen matrix. The concentrations of the specific compounds in the pooled extract are presented in Table 2; markers were added in different concentrations to test the identification pipeline in conditions which mimic those found in a typical metabolomic experiment, where variation is usually present at different concentration levels. As an example of what the data look like, the first control sample, measured in positive mode, is shown in Figure 1. The horizontal axis shows the chromatographic dimension, and the vertical axis the mass-to-charge ratio. Circles indicate features that have been identified in this plane. In the remainder only the extracted triplets for the features, consisting of retention time, mass-to-charge ratio and intensity, will be used.

Feature extraction is performed with XCMS (Smith et al., 2006) and all statistical analyses are carried out in R (R Development Core Team, 2011). The CentWave peak-picking algorithm (Tautenhahn et al., 2008) is applied, using the following parameter settings: ppm = 20, peakwidth = c(3,15), snthresh = 2, prefilter = c(3,5). The average numbers of detected features per chromatogram are 1179 and 610 for positive and negative ion mode, respectively.

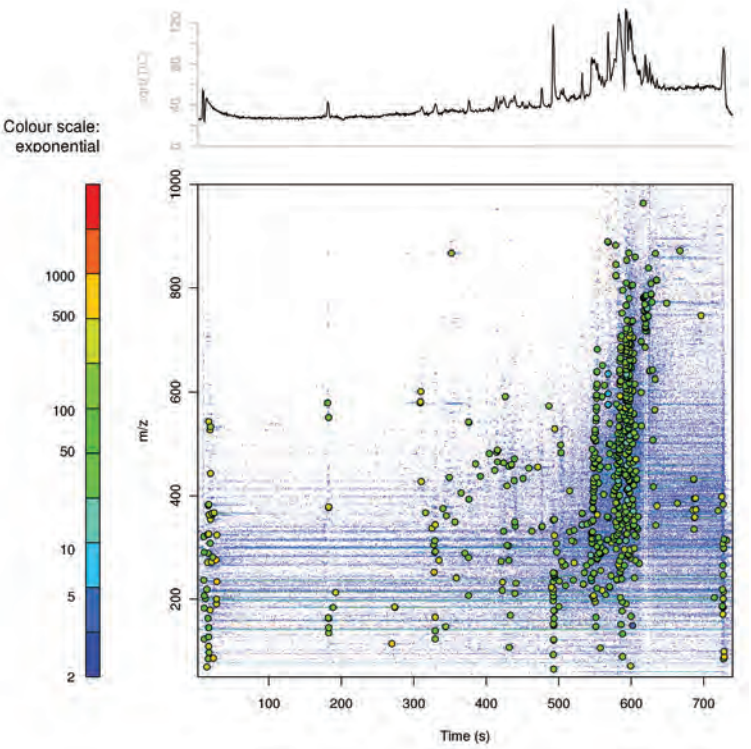

Fig. 1. Visualization of the data of the first control sample, measured in positive mode. The top of the figure shows the square root of the Total Ion Current (TIC); background color indicates the intensity of the signal in the plane formed by retention time and *m/z*axes. Circles indicate features found by the peak picking; the fill colour of these circles indicates the intensity of the features.

Compound	mgl^{-1} pool	Δ Conc. (mgl^{-1})
quercetin-3-galactoside (querc-3-gal)	5.69	1.48
quercetin	0.006	0.008
quercetin-3-glucoside (querc-3-glc)	1.05	0.3
quercetin-3-rhamnoside (querc-3rham)	3.64	3.55
phloridzin	2.92	2.3
cyanidin-3-galactoside (cy-3-gal)	n.d.	0.57
trans-resveratrol	n.d.	0.4

Table 2. Spiked compound summary. The difference in concentration is relative to the one measured in the pooled extract. Cyanidin-3-galactoside and *trans*-resveratrol are not normally found in Golden Delicious.

After grouping across samples, features are screened for isotopes, clusters and common adducts with in-house developed software.

Due to fragmentation occurring in the ionization source, it is common for a single neutral molecule to give rise to several ionic species. A single spiked compound can then generate several "biomarkers" in the MS peak table. Adducts, isotopes and common clusters are automatically screened, but fragments must be included in the biomarker list, as in real metabolomic experiments no a priori knowledge can be used to distinguish molecular from fragment ions. For the apple data set, the characteristic couples mass/retention time for all spiked metabolites were identified by manual inspection of the UPLC-MS profiles of standards. For negative ions the following numbers of features have been associated with the spike-in compounds: querc-3-gal/querc-3-glc (1 feature), phloridzin (2 features), *trans*-resveratrol (1), querc-3-rham (1). In the case of positive ion mode the numbers are cy-3-gal (1), *trans*-resveratrol (1), querc-3-rham (1), quercetin (1) and phloridzin (4). These feature are now taken to be the "true" biomarkers and they are used to construct ROC curves. The data set, as well as a more extended version including different concentrations of spiked-in compounds is publicly available in the R package BioMark (see `http://cran.r-project.org/web/packages/BioMark`, Wehrens & Franceschi (2011)) and has been used to evaluate a novel stability-based biomarker selection method (Wehrens et al., 2011).

In this application, the effects of decreasing sample size are investigated by subsampling the original set of twenty samples: sample sizes of 16, 12, 8 and 6 apples, respectively, are considered. In all cases, both classes (spiked and control) have equal sizes, which is the most easy case for detecting significant differences. Results are summarized by analysis of ROC curves – to prevent effects from accidentally easy or difficult subsets, the final ROC curves are obtained by averaging the results of 100 repeated re-samplings.

2.4 Simulated data sets

To assess the behaviour of biomarker selection for larger data sets, we resort to simulation. Simulated data sets have been constructed as multivariate normal distributions, using the means and covariance matrices of the experimental data: both classes (untreated and spiked) have been simulated separately. Simulations are performed for both positive and negative modes; in every simulation, one hundred data sets are created. The outcomes reported here are the averages of the results for the one hundred simulations. Data sets consisting of 10, 25, 50 and 200 biological samples per class have been synthesized.

3. Results and discussion

As a first step, the data are visualized using Principal Component Analysis (PCA). Since the intensities of the features can vary enormously, standardized data are used. The score plots for the positive and negative data sets are shown in Figure 2 for the positive ion mode, and in Figure 3 for the negative mode. In both cases, control and spiked data sets are not completely separated and the same is also true for the other PCs (not shown). This fact indicates that the "inherent" variability of the data set is not perturbed to a significant extent by spiking, as could be expected considering the small number of affected variables.

Even with this data structure, biomarker selection strategies can still perform efficiently. Figure 2 and Figure 3 also display the score plots of a PCA analysis performed considering only the top 10 variables selected by univariate *t*-testing. In these conditions, the separation

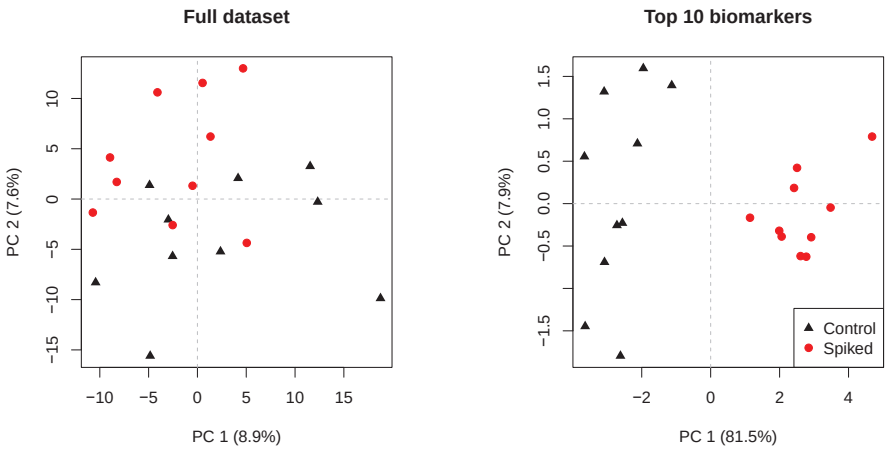

Fig. 2. PCA score plot (PC1 vs PC2) for the positive ion mode data set after standardization. In the left plot the principal components have been calculated on the full data set. In the right panel PCA analysis has been performed considering only the top 10 variables selected by a *t*-test.

between control and spiked samples is evident, thus indicating that this subset of the variables separates the two classes. Whether these ten variables contain the true biomarkers remains to be seen: especially in small data sets there may be chance correlations causing false positives, and seeing differences between the two groups in the score plots after *t*-testing in fact is trivial. The score plot is merely showing that the variables, selected on the bases of their discriminating power, are separating the two classes. As already discussed, small data sets will in general not capture all relevant biological variability, which implies that the predictive power of statistical models based on small data sets usually is very low. To illustrate this effect, the predictive power, i.e., the fraction of correct predictions for PC-LDA and PLS-DA models is presented in Figure 4. Four subsets of different sizes are considered as training sets, and the estimate of predictive power is based on predictions for the apples not in the training set. Again, the results are the average over 100 different subsamplings. Even if the control and spiked subsets are different, it can be seen that the predictive power of the multivariate methods is comparable to random selection, meaning that for every subset different variables will be important in the models and no consistency can be achieved. However, it is important to point out that this fact does not mean that some of the true biomarkers are not consistently selected upon subsetting, but rather that the more important variables are changing from a subset to another: even with models that are unpredictive it is possible to extract relatively good lists of putative biomarkers. Obviously, with very different characteristics for the two classes there *will* be predictive power, but for realistic data sets like the one used in this paper, where differences are small, it is unwise to focus solely on prediction.

To evaluate the efficiency of the different methods as far as biomarker selection is concerned, ROC curves for the *t*-test and two-component PLS-DA and PC-LDA models are presented in Figure 5, for 3, 4, 6 and 8 biological samples per class, respectively. The ROC curves indicate that all three variables selection methods perform significantly better than random selection.

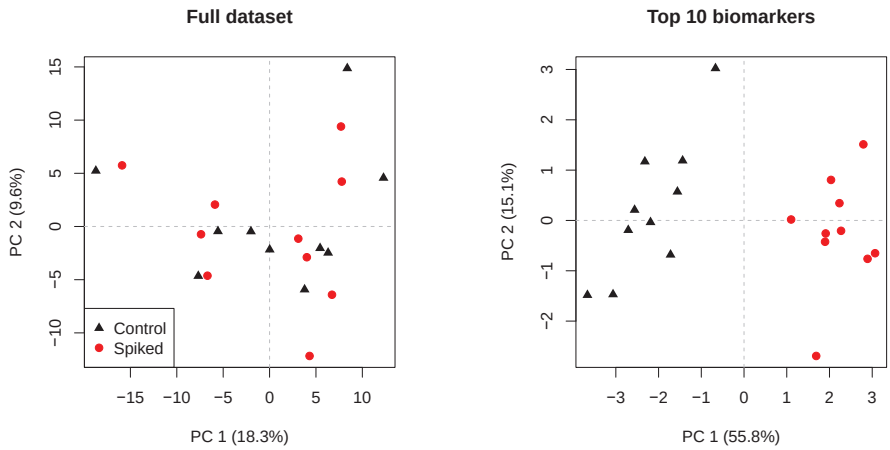

Fig. 3. PCA score plot (PC1 vs PC2) for the negative ion mode data set after standardization. In the left plot the Principal Components have been calculated on the full data set. In the right panel PCA analysis has been performed considering only the top 10 variables selected by a *t*-test.

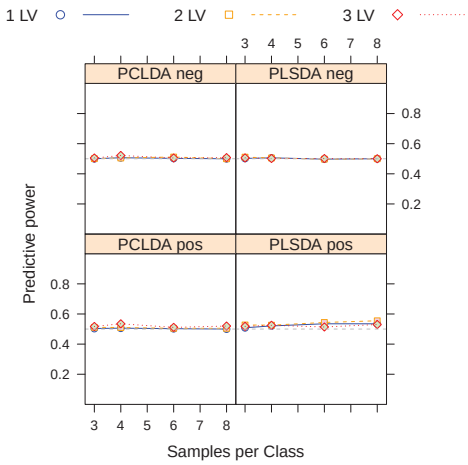

Fig. 4. Predictive power of multivariate PLS-DA and PC-LDA on a subset of the initial data set for positive and negative ion mode. Different lines are relative to models constructed with an increasing number of LVs. The horizontal dashed line indicates random selection.

Of the three, PC-LDA is always the least efficient, while PLS-DA and the *t*-test have a very similar performance. In absolute terms, the efficiency of the three methods increases with the number of biological samples. ROC curves for all possible conditions were constructed and the results are summarized in terms of early AUC (pAUC.2) in Figure 6, for positive and negative ion mode, respectively. From these figures it is possible to extract some clear trends:

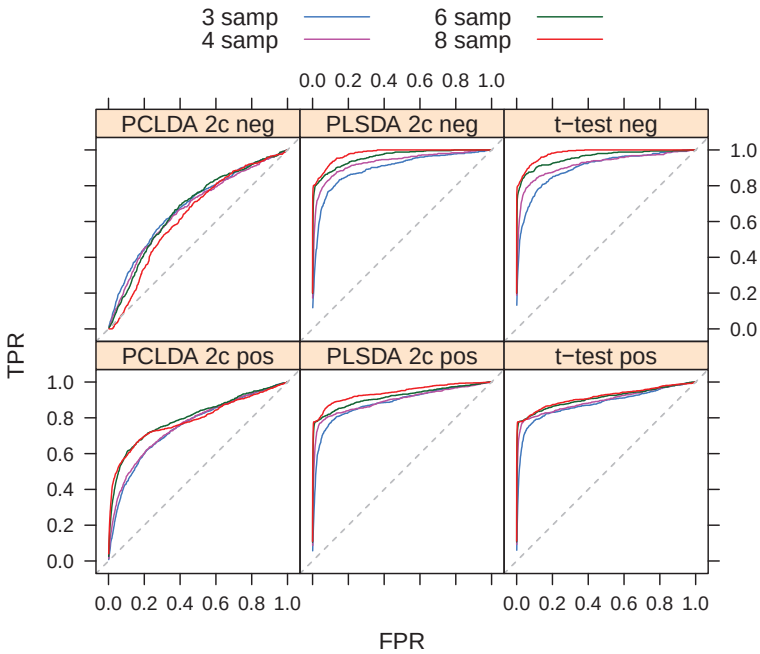

Fig. 5. ROC curves for the *t*-test and two component PLS-DA and PC-LDA as a function of the number of samples per class.

1. The performance of the methods improves by increasing the number samples per class.
2. The performance of PLS-DA is not particularly sensitive to the number of components.
3. PC-LDA does not show top class performance in any of the conditions considered.
4. The performance of PC-LDA is very much dependent on the number of components.
5. Multivariate approaches do not show a definitive advantage over univariate *t*-testing.

As expected, the performances of all the methods in terms of biomarker identification decrease with a reduction of the data set size. However, it is important to point out that even in the worst possible case (3 samples per class) early AUC for PLS-DA and the *t*-test are significantly greater than that obtained for completely random selection. This indicates that both methods can be used effectively in the biomarker selection phase, even with a low number of samples. In other words, features related to spiked compounds are consistently present in the top positions of the ordered list of experimental variables, which implies that also models constructed with very few samples can be relied upon to recognize these features.

The performance of PC-LDA is very much dependent on the number of components taken into account. This behavior can be explained by considering that in PC-LDA the variable reduction step is performed without any knowledge of class labels, only selecting the directions of greater variance. If these directions show little discriminating power, their supervised linear combination leads to poor modeling. However, performance improves with

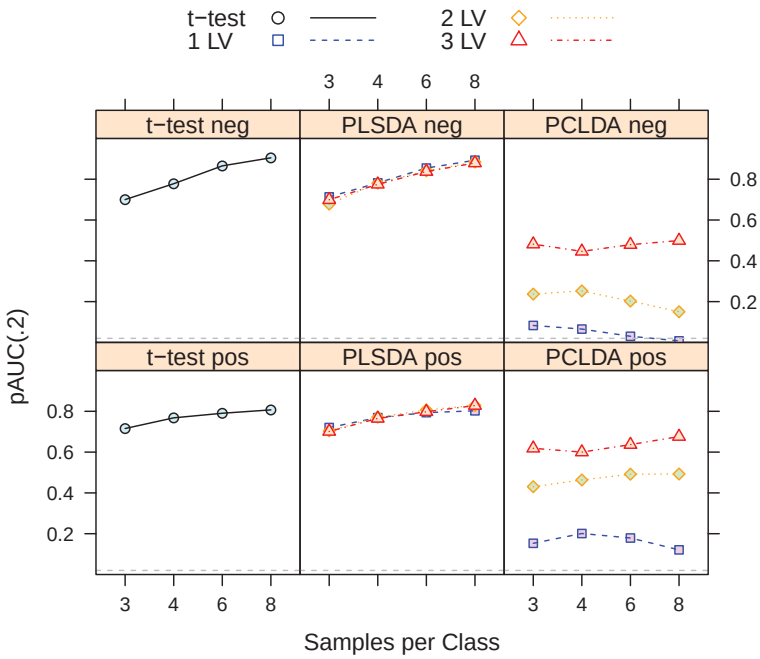

Fig. 6. pAUC.2 for PLS-DA, PC-LDA and *t*-test as a function of the number of samples per class and the number of LVs. The gray dashed line indicates the pAUC.2 of random selection.

the number of components, as an increase of the number of LVs leads to a better "coverage" of the data space. These limitations do not affect PLS-DA, as the variable reduction step is already performed in a supervised framework, where discriminating power is the main request. This means that the first PLS components are by definition more relevant than the first PCA components in biomarker identification. The other side of the coin is the danger of overfitting, very real in the application of PLS-DA (Westerhuis et al., 2008) – we will come back to this point later.

In this small-sample set, the *t*-test does as well as the best multivariate methods. This shows that modeling the correlation structure is not necessarily an advantage if the number of samples is low, or, alternatively, that the true correlation structure has not been captured well enough from the few samples that are available to allow meaningful inference. A definite advantage of the *t*-test is that it has no tunable parameters and can be applied without further optimization. It should be noted that we do not need to apply multiple-testing corrections in this context since we only use the order of the absolute size of the *t*-statistics to construct the ROC curves, and not a specific cut-off level α. In other applications, however, this aspect should be taken into account.

To extend the comparison between different models beyond the limits imposed by the apple experiment, ROC curves and early AUC were calculated for the simulated data using larger sample sizes (10, 25, 50, 200), both for positive and negative ion modes. The dependence of

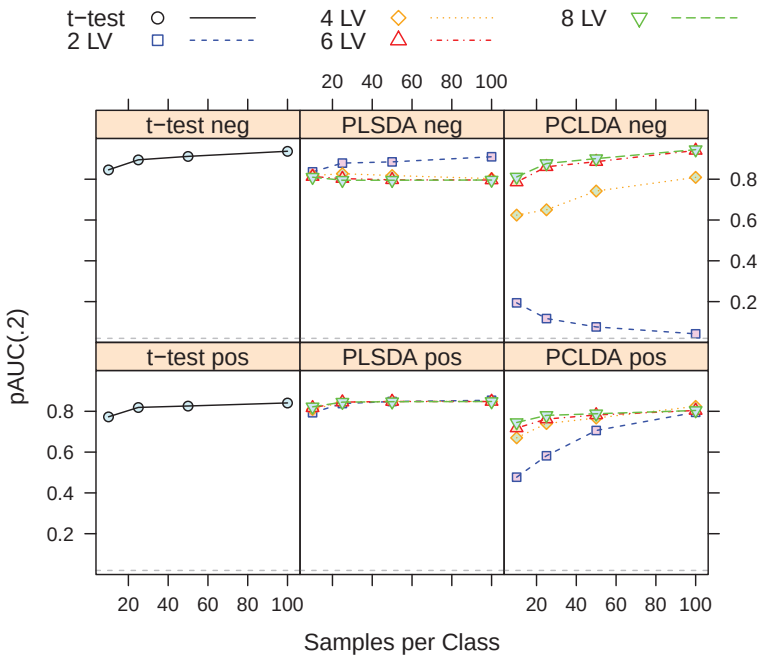

Fig. 7. pAUC.2 for PLS-DA, PC-LDA and *t*-test as a function of the number of samples per class and the number of LVs. Simulated data set. Gray dashed line indicates the pAUC.2 of random selection

pAUC.2 on the number of replicates and of components is presented in Figure 7, comparing the multivariate methods to the *t*-test and to "random" selection.

This analysis shows that PC-LDA only becomes effective if a large number of LVs is considered: the true biomarkers should have appreciable weight in the latent variables and it is by no means certain that this is the case for the first couple of LVs. Is it worth noting that for negative ion mode, the model with 2 LVs is comparable to random selection. In the case of PLS-DA, this dependence on the number of LVs is less evident and shows an opposite trend: the best performance is obtained with the smallest number of LVs. This is in agreement with the explanation given earlier: the relevant variables are captured in the very first PLS components, and the effect of overtraining leads to deterioration if more components are added. If anything, it is surprising that the overtraining effect is relatively small for these data.

The results on the simulated data sets are in agreement with the conclusions from the apple data. Differences between the methods decrease with increasing sample sizes, but even with the largest number of objects (200 in each group) the *t*-test still performs as well as PLS-DA. Multivariate testing is slightly more effective for the positive ion mode, while the *t*-test shows a slight advantage for the negative ion mode. This behaviour is probably due to the different characteristics of both ionization modes, leading to different levels of correlation

between biomarkers. Indeed, in positive ion mode, the ionization shows a more pronounced fragmentation (phloridzine, for example, gives rise to four different biomarkers).

4. Conclusions

In this paper we have investigated the effects of sample set size on the performance of some popular strategies for biomarker identification (PLS-DA, PC-LDA and the *t*-test). The experiments are performed on a spiked metabolomic data set measured in apple extracts by UPLC-QTOF. The efficiency of the different statistical approaches is compared in terms of ROC curves, and in order to assess general trends, simulated data have been used to extend the data set. The experimental results clearly show that Linear Discriminant Analysis carried out on the Principal Components (PC-LDA) is the least efficient strategy for biomarker identification among the ones we considered. PLS-DA and the *t*-test show comparable performances in all the considered conditions. These results, and the observation that PLS-DA based selection is relatively consistent for different numbers of components, indicate that multivariate and univariate approaches are equally efficient for the apple data set. It is perhaps surprising that relatively good results in terms of biomarker selection are obtained, even for models that have very poor predictive performance. One should realise, however, that this is not a paradox at all: it merely is the result from the low sample-to-variable ratio, leading to chance correlations of intensities of metabolite signals with class. The true biomarkers are often present among the most significant variables in, e.g., a PLS-DA model, but many other false positives are, too, destroying the predictive power. One recently published approach actually utilizes this variability by focusing only on those variables that are *consistently* present in the most important variables upon disturbance of the data by jackkifing or bootstrapping (Wehrens et al., 2011).

The main point of this contribution, however, is the relation between data set size and reliability of biomarker identification. As expected, all the methods become less efficient as the number of biological replicates decreases, but even in these conditions the use of PLS-DA and the *t*-test offer effective biomarker identification strategies. This observation is fundamentally important in all studies where it is impossible to acquire more samples, and suggests that small sample sizes can still allow reliable selection of biomarkers.

5. References

Araki, Y., Yoshikawa, K., Okamoto, S., Sumitomo, M., Maruwaka, M. & Wakabayashi, T. (2010). Identification of novel biomarker candidates by proteomic analysis of cerebrospinal fluid from patients with moyamoya disease using SELDI-TOF-MS, *BMC Neurology* 10: 112.

Barker, M. & Rayens, W. (2003). Partial least squares for discrimination, *J. Chemom.* 17: 166–173.

Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing, *J. Royal. Stat. Soc. B* 57: 289–300.

Brown, C. D. & Davis, H. T. (2005). Receiver operating characteristics curves and related decision measures: A tutorial, *Chemom. Intell. Lab. Syst.* 80: 24–38.

Chadeau-Hyam, M., Ebbels, T., Brown, I., Chan, Q., Stamler, J., Huang, C., Daviglus, M., Ueshima, H., Zhao, L., Holmes, E., Nicholson, J., Elliott, P. & Iorio, M. D. (2010). Metabolic profiling and the metabolome-wide association study: significance level for biomarker identification, *J. Proteome Res.* 9(9): 4620–4627.

Guo, Y., Hastie, T. & Tibshirani, R. (2007). Regularized discriminant analysis and its application in microarrays, *Biostatistics* 8: 86–100.

Hanczar, B., Hua, J., Sima, C., Weinstein, J., Bittner, M. & Dougherty, E. (2010). Small-sample precision of ROC-related estimates, *Bioinformatics* 28: 822–830.

Hastie, T., Tibshirani, R. & Friedman, J. (2001). *The Elements of Statistical Learning*, Springer Series in Statistics, Springer, New York.

Kemsley, E. K. (1996). Discriminant analysis of high-dimensional data: a comparison of principal components analysis and partial least squares data reduction methods, *Chemom. Intell. Lab. Syst.* 33: 47–61.

McLachlan, G. (2004). *Discriminant Analysis and Statistical Pattern Recognition*, Wiley-Interscience.

Noble, W. S. (2009). How does multiple testing correction work?, *Nat. Biotechnol.* 27: 1135–1137.

Oh, J., Craft, J., Townsend, R., Deasy, J., Bradley, J. & Naqa, I. E. (2011). A bioinformatics approach for biomarker identification in radiation-induced lung inflammation from limited proteomics data, *J. Proteome Res.* 10(3): 1406–1415.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0. URL: *http://www.R-project.org*

Reiner, A., Yekutieli, D. & Benjamini, Y. (2003). Identifying differentially expressed genes using false discovery rate controlling procedures, *Bioinformatics* 19(3): 368–375.

Saeys, Y., Inza, I. & Larranaga, P. (2007). A review of feature selection techniques in bioinformatics, *Bioinformatics* 23: 2507–2517.

Smit, S., Breemen, M. J. v., Hoefsloot, H. C. J., Aerts, J. M. F. G., Koster, C. G. d. & Smilde, A. K. (2007). Assessing the statistical validity of proteomics based biomarkers, *Anal. Chim. Acta* 592: 210–217.

Smith, C. A., Want, E. J., Tong, G. C., Abagyan, R. & Siuzdak, G. (2006). XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification, *Anal. Chem.* 78: 779–787.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions, *J. R. Statist. Soc. B* 36: 111–147. Including discussion.

Szymanska, E., Saccenti, E., Smilde, A. & Westerhuis, J. (2011). Double-check: validation of diagnostic statistics for PLS-DA models in metabolomics studies, *Metabolomics* .

Tautenhahn, R., Bottcher, C. & Neumann, S. (2008). Highly sensitive feature detection for high resolution LC/MS, *BMC Bioinformatics* 9: 504.

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

Vrhovsek, U., Rigo, A., Tonon, D. & Mattivi, F. (2004). Quantitation of polyphenols in different apple varieties, *J. Agr. Food. Chem.* 52(21): 6532–6538.

Wehrens, R. & Franceschi, P. (2011). *BioMark: finding biomarkers in two-class discrimination problems*. R package version 0.3.0.

Wehrens, R., Franceschi, P., Vrhovsek, U. & Mattivi, F. (2011). Stability-based biomarker selection, *Anal. Chim. Acta* 705: 15–23.

Werf, M. J. v. d., Pieterse, B., Luijk, N. v., Schuren, F., Vat, B. v. d. W.-v. d., Overkamp, K. & Jellema, R. H. (2006). Multivariate analysis of microarray data by principal component discriminant analysis: prioritizing relevant transcripts linked to the

degradation of different carbohydrates in Pseudomonas putida S12, *Microbiology* 152: 257–272.

Westerhuis, J., Hoefsloot, H., Smit, S., D.J., V., Smilde, A. K., van Velzen, E., van Duijnhoven, J. & van Dorsten, F. A. (2008). Assessment of PLSDA cross validation, *Metabolomics* 4: 81–89.

Yousef, M., Ketany, M., Manevitz, L., Showe, L. & Showe, M. (2009). Classification and biomarker identification using gene network modules and support vector machines, *BMC Bioinformatics* 10: 337.

Zuber, V. & Strimmer, K. (2009). Gene ranking and biomarker discovery under correlation, *Bioinformatics* 25: 2700–2707.