Metabolite profiles as a reflection of physiological status – a methodological validation

Matthias Steinfath [1] [*], Dirk Repsilber [1] [*], Manuela Hische [1], Nicolas Schauer [2], Alisdair R. Fernie [2], Joachim Selbig [1]

[1]Institute of Biochemistry and Biology, University Potsdam, c/o Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, D - 14476 Potsdam-Golm, Germany

[2]Max Plack Institute of Molecular Plant Physiology, Am Mühlenberg 1, D - 14476 Potsdam-Golm, Germany

Summary

Biological "omics" data comprise numerous variables (metabolites, gene expression, physiological quantities) and comparatively few samples. These samples represent either measurements for slightly different genotypes in identical environments, or for different environmental conditions affecting the same genotype. Given this kind of data, it is intriguing to ask for possible measurable associations between molecular variables and the phenotypical or physiological status.

To evaluate such correlations we need a model for the functional dependency of the physiological state on given molecular variables. Supervised machine learning methods such as neural networks, decision trees, or support vector machines may be used to reveal such correlations. The simplest model is certainly a linear approach. To investigate the association between molecular and phenotypical variables, we ask if the correlation between predictor and response is statistically significant, and how much of the phenotypical variance of the response can be explained by a given set of predictors. When confronted with a set of molecular data not all of them are generally relevant for each physiological trait. Given this fact the problem of feature selection arises.

Different regression methods have been developed to answer this question: Ordinary Least Squares (OLS) yields an unbiased solution, but normally has a high mean square error. In particular, there is no dimension reduction included in this method and, hence, overfitting is a critical problem. In contrast, Principle Component Regression (PCR) offers such a dimension reduction, however, the principle components are found without considering the response. Partial Least Squares Regression (PLSR) is utilised as an alternative method since it considers the variance within the predictors as well as between predictors and response, whilst Ridge Regression is a further alternative worthy of consideration.

In our study we applied these methods to data resulting from a tomato metabolite experimental series. Comparison of the results for this dataset with experimentally relevant correlation structure between variables and samples allows us to test the relative merits of the regression methods with respect to the questions raised above. Given certain prerequisite knowledge it also allows us to conjecture the true biological correlation. Our results show that under most circumstances OLS is worst with respect to prediction. However, the ranking of methods seems to change considerably if the question of feature selection is considered. Understanding and discussing these differences is a relevant contribution to the task of choice of suitable approach of correlation analysis for "omics" datasets with respect to the biological interpretation in question.

[*]contributed equally

1 Introduction

Uncovering the relation between molecular and phenotypical traits is a fundamental objective of molecular biology. Moreover, in the post-genomic era, as life-science researchers are increasingly confronted with high-throughput, complex and noisy data, this objective has become a bioinformatics issue, specifically a question of integrative bioinformatics. Finding correlations for data matrices where the number of variables (e.g. genes or metabolites) reaches or even exceeds sample size is a difficult task. Within this field, methodological advice is difficult to obtain, benchmarking has not yet been established, and methodological studies are rare. Here we aim to provide first recommendations regarding the interdependency of study objectives and suitability of certain multivariate analysis approaches for correlation analyses.

In our case, we are concerned with correlating metabolome data to phenotypical traits. A general hypothesis which is often adapted in this context regards variation at the molecular levels (gene expression, protein and metabolite levels) as underlying features, whereas phenotypical traits are regarded as the ultimate manifestation of variance at the molecular level. Hence, molecular data can be regarded as predictors, while phenotypical properties are conceived as responses. Most recently metabolism has been linked to macro-organismic events such as apoptosis [1] or whole organismal morphological properties [2]. In the latter example, as is commonly the case, pairwise correlations between metabolite concentrations and physiological properties were examined.

However, *combinations* of metabolites rather than individual metabolites are responsible for most phenotypes. This is partly reflected by the network structure of metabolism [1]. It is moreover known that more than two molecular entities frequently act in a combined way to accomplish certain molecular functions - a property of molecular organisation which has previously been defined as a hypergraph [3]. Such combinations of molecular entities can be represented in a first approximation by linear models. The coefficients, which appear in such models, are directly linked to the strength of the influence of the variable (e.g. the metabolite) on the phenotypical trait in question. This consideration leads directly towards a multivariate linear regression approach, in which solutions are found for the estimation of these coefficients.

Within these lines of multivariate approaches several different methods are available: The Ordinary Least Squares (OLS) approach yields an unbiased estimation of the coefficients. However, it is well-known that this property is often out-weighed by large mean square errors of prediction [4]. This is especially true in cases for which the number of variables is of the same order than the sample size. If the number of variables exceeds the sample size, OLS is not longer applicable. If, on the other hand, the sample size is much larger than the number of variables, all multivariate methods perform almost equally well. However, as we deal with the example of a metabolomics dataset, we are typically confronted with similar numbers of variables and samples.

As alternatives for OLS, a number of regression methods originally developed in other scientific areas have become increasingly important in molecular biology, e.g. Partial Least Squares Regression (PLS) and Principal Component Regression (PCR). Frank et al. [4] have investigated the theoretical properties and the performance on artificial data of the afore mentioned methods. While OLS maximises the correlation between a linear combination of the predictors and the phenotypical response, PLS maximises the corresponding covariance and PCR the variance within the predictor data set. The purpose of the latter two methods is to find vari-

ables or combinations thereof which are important in determining the response. The remaining variables are then omitted for further analyses, thereby considerably reducing the mean square error. However, the success of this approach is strongly dependent on the correlation structure of the data [5]. The data are usually obtained as matrices, wherein the columns represent the predictor variables, e.g. metabolite levels, and the rows represent the samples, e.g. different genotypes or treatments under investigation. Such molecular "omics" data often show the following characteristics:

1. subsets of the variables (columns), and also subsets of the samples (rows) are correlated
2. both predictors and response are noisy, while the assumption of the linear model is that only the response has a random error [6]
3. number of independent variables are the same or greater than the sample size
4. the actual data do not follow a normal distribution. In addition, there is often very little or no repetitions at all .

The data in our study result from a tomato metabolite experimental series [2]. It is therefore possible to build our methodological validation approach on a *non-artificial* correlation structure.

Validity for analysing this kind of data can be summarised by three key words: Significance, prediction and interpretation. Here, *significance* means to determine if there is any linear relation between the molecular and physiological data in focus which is beyond a random finding. *Prediction* stands for our ability to successfully generalise the information gained from one pair of datasets to other datasets. *Interpretation* in our approach means the identification of metabolites or pathways as explaining a certain physiological trait, i.e. feature selection. Here we focus on the last two questions.

In our study we investigate the performance of the collection of standard analysis methods mentioned above under a realistic scenario regarding a correlation structure for both samples and variables as typical for experimental biological datasets. We aim at validating the prediction and feature selection properties of the methods under study. Results show the necessity for a careful choice of method, depending on the primary interest of the analysis in question. The performance ranking is changing between the objectives of *optimal prediction* and *optimal feature selection* properties. Further, we are able to hint the true correlation between phenotypical response and molecular predictor, given sample size and inherent technical variation in the study. As an example, and for illustration, we also apply all methods to two experimentally measured responses from the tomato dataset.

We discuss that the kind of preparative study, such as that we demonstrate it here, is necessary to assess the significance and relevance of an analysis involving the actually measured response variables.

2 Material and Methods

2.1 Datasets

Datasets used within our study are based upon the experimental dataset by Schauer *et al.*[2]. In the main part, we consider the metabolome part of this dataset together with a simulated phenotypical response, Y, to enable validation studies of the multivariate methods under question. As an illustrative example, we also analyse a series of cross-validations of two of the experimental phenotypical traits.

2.1.1 Experimental Data

Experimental data measured by [2] from tomato isogenic introgression lines served as a model dataset. For methodological details of metabolite profiling as well as genetics of the lines used in these experimental series, we refer back to the original publication. In summary, 77 different introgression lines from two seasons were harvested and skinned pericarp material from six independent plants per line was subjected to metabolite profiling. The metabolite profiles for 74 metabolites, represented by 115 analytes, were measured by GC-MS as detailed in [7] and [8]. Data are available as annotated heatmaps via
http://www.nature.com/nbt/journal/v24/n4/suppinfo/nbt1192_S1.html,
or if requested from the authors. Missing values in this data matrix were imputed using Bayesian Principal Component Analysis (BPCA) [9].

The original experimental data show marked correlation structures, on the one hand between the metabolite profiles (variables), refer Figure 1 left, but also between the introgression lines used (samples), see Figure 1 right. It has already been demonstrated that these correlations are based on molecular biological functional relations between the associated metabolites, as well as on genotypic similarities between the assessed introgression lines [2].

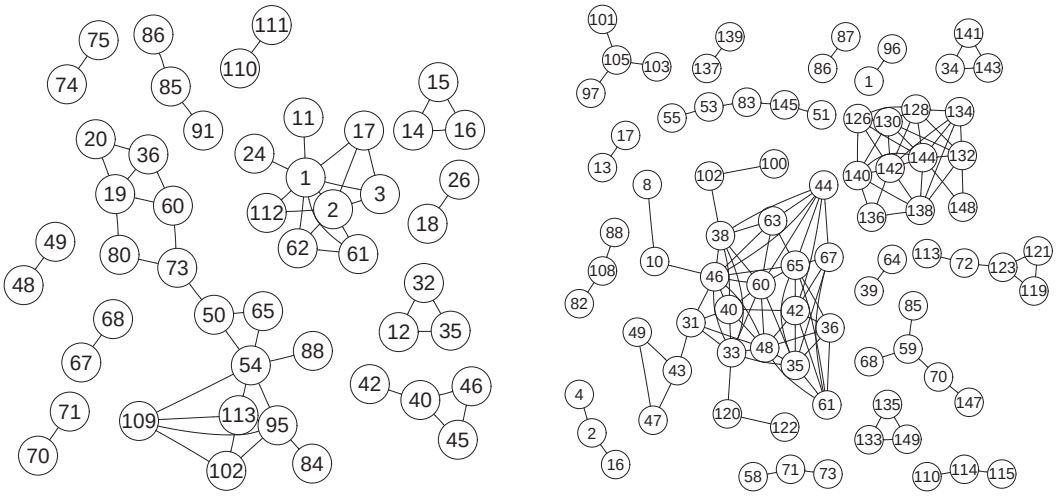

Figure 1: Correlation networks, using a threshold of $\theta = 0.7$, showing the complex correlation structure within the original experimental data metabolites (left) and samples (right).

2.1.2 Simulated phenotypical responses

For our simulation approach three quantities must be generated: The predictor matrix X, the response Y and the vector of true coefficients β.

As far as the predictor matrix X is concerned, the following settings were used: The original, experimental correlation structure can either be preserved, as for most our analyses, or destroyed by permutation of the columns. In the case that part of the true coefficients are zero, we can distinguish between true and pseudo-predictors. For true predictors the corresponding coefficients are greater than zero, while all other variables are referred to as pseudo-predictors. In the feature selection part of the results we will look for these true predictors.

The correlations between true predictors and non-predictive variables have always been destroyed by permutations, while those within these two sets were conserved.

In the simulation we can also distinguish between the true and the measured value of the predictors. The difference between these two is given by experimental noise and can be described by

$$X_\delta = X + \delta \tag{1}$$

In our approach, the experimental error δ is always chosen to be normally distributed with zero mean and standard deviation σ_p. σ_p was chosen for each variable individually as ratio of its experimentally measured variation. Its standard deviation varies with variance of the true predictor values. The true values are used to generate the response, while the measured values are used in the evaluation.

The response variable is produced from the predictor variables assuming the following linear model:

$$Y = X\beta + \varepsilon \tag{2}$$

The error is again normally distributed with zero mean and standard deviation σ_ε. The correlation between Y and $X\beta$ is then given by:

$$corr(Y, X\beta)^2 = \frac{var(X\beta)}{\sigma_\varepsilon^2 + var(X\beta)} \tag{3}$$

This allows us to choose a predefined canonical correlation between predictor set and response, referred to as *true correlation* C_{true} in the sequel. Subsequently we calculate σ_ε.

The error term consists of a biological and an experimental error: $\varepsilon = \varepsilon_e + \varepsilon_b$. In the absence of repetitions we can not discriminate between these two, so the difference is only used in the discussion, where we estimate the effect of different noise sources given a certain biological correlation.

In summary, the following parameters were varied:

- correlation between predictor variables
- β the vector of coefficients

- $\varepsilon \sim N(0, \sigma_r)$ the error in the response
- $\delta \sim N(0, \sigma_p)$ the error in the predictors

2.1.3 Output of the simulation

The principal output is an estimated vector of coefficients $\hat{\beta}$ along with the predicted response vector, $\hat{Y}$. These quantities were estimated from the training set and compared to the true β and Y in the test set, compare section 2.2.5.

2.2 Regression methods

Here we describe 4 different regression methods: Ordinary Least Squares, Partial Least Squares, Principal Components Regression and Ridge Regression.

2.2.1 Ordinary Least Squares and Canonical Correlation Analysis

Ordinary Least Squares (OLS) is the oldest and probably most obvious method to predict one variable, the response, from another, the predictor, if there is a linear relationship between both. In the sequel $\mathbf{X}$ denotes the set of predictor variables and $\mathbf{Y}$ the set of response variables. In the case of a multvariate response, the corresponding method is called Canonical Correlation Analysis (CCA) developed by [10]. For the following part of this study $\mathbf{Y}$ is treated as univariate.

OLS and CCA solve the following maximisation problem

$$argmax_{\beta,\alpha} corr(\mathbf{X}\beta, \mathbf{Y}\alpha). \tag{4}$$

Where $corr$ stands for the Pearson correlation. The vectors β and α, have the same direction as the vectors which minimise the squared errors. The dimensionality of these four vectors is determined by the number of predictor and response variables. For univariate response α is a scalar, in this case β is the vector of regression coefficients without intercept. This correlation is called the canonical correlation. The β, α are called canonical vectors, and $\mathbf{X}\beta, \mathbf{Y}\alpha$ canonical variates.

The advantage of the OLS method is that the estimated regression coefficients are unbiased. These estimates are, however, not very precise. Typically, high mean square errors render OLS improper regarding the prediction objective. A particular problem is, that the predictor variables are often highly correlated and multi-collinear. This problem is severe for OLS, but better tackled by the methods described in the sequel.

2.2.2 Principal Component Regression

Here the regression is not on the predictor variables $\mathbf{X}$ directly but on a linear combination of them. The new variable set is given by

$$\mathbf{Z} = \mathbf{X}\mathbf{v}_i, i = 1, .., M. \tag{5}$$

where M depicts the number of predictor variables used. The vectors $\mathbf{v}_i$ are the eigenvectors of $\mathbf{X^T X}$, i.e. of the covariance matrix of $\mathbf{X}$. This corresponds to the following maximum principle

$$argmax_{\mathbf{v}} var(\mathbf{Xv}). \tag{6}$$

The new variables are ordered in a natural way by their eigenvalues. An OLS regression is then performed on those so-called *latent* variables.

2.2.3 Partial Least Squares

Partial Least Squares (PLS) is a method developed in the field of chemometrics [11, 4]. The procedure maximises the covariance between the variable sets as described by the following equation:

$$argmax_{\beta,\alpha} cov(\mathbf{X}\beta, \mathbf{Y}\alpha) = argmax_{\beta,\alpha} var(\mathbf{X}\beta) corr(\mathbf{X}\beta, \mathbf{Y}\alpha) var(\mathbf{Y}\alpha) \tag{7}$$

Thus PLS can be described as CCA penalised with principle components [12].

The three maximum principles explained above can be described by a single maximisation criterion with a continuous parameter γ [13]

$$argmax_{\beta} var(\mathbf{X}\beta)^{\gamma/(\mathbf{1}-\gamma)} \mathbf{corr}(\mathbf{X}\beta, \mathbf{Y}) \tag{8}$$

Then $\gamma = 0$ corresponds to OLS, $\gamma = 0.5$ to PLS and $\gamma = 1$ to PCR.

2.2.4 Shrinkage methods: Ridge Regression

Shrinkage methods do not select a discrete number of variables but subject instead the coefficient vector to a restriction. The ridge regression for instance estimates the coefficient vector by:

$$\hat{\beta}^{ridge} = argmin \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} x_{ij}\beta_j \tag{9}$$

Here, N denotes the sample size. The parameter λ can be linked to continuous number df of independent variables by:

$$df(\lambda) = trace[X(X^T X + \lambda I)^{-1} X^T] = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda} \tag{10}$$

Here d_j stands for the jth singular value of X[14].

For the regression with the last three methods a number of $M < p$ latent variables was taken for regression. The difficulty remains to find the correct M.

2.2.5 Model selection for PLS, PCR and Ridge Regression

Validation of the performance regarding the prediction task is based on a cross-validation approach. This is straightforward in the case of OLS. However, for the three alternative methods also the optimal number of latent variables has to be trained.

For this purpose, the dataset is divided into three subsets. With the first two parts the parameters of the model are determined, specifically the number of latent variables, M, used for PLS, PCR and Ridge Regression. The third part serves as test set.

The typical training/test-procedure runs as follows: The coefficients calculated in the first set, the trainings set, are used to predict the response in the second set. The optimal number of coefficients thus determined is then applied to the combined first and second set to calculate a new coefficient vector, $\hat{\beta}$, which is used to predict the response,$\hat{Y}$, in the third set, the test-set. From this last step the correlation between the prediction, $\hat{Y}$, and the true trait vector, Y, are scored.

3 Results

3.1 Prediction performance

The following set of parameters was used throughout for generation of the datasets:

True coefficients: $\beta = (1, ...1, 0, ...0)^T$.

Here always the 10 first coefficients were set to 1, the rest to zero.

True correlations $C_{true} = 0.7, 0.8, 0.9, 1.0$.

Noise in the predictor $\sigma_{p,j}/\sqrt{Var(X_j)} = 0.0, 1.0, 2.0$ (j=1,..,p).

Sample size $N = 100$.

Number of predictor variables $p = 25, 50, 70$ from the original 115.

Using this approach are 36 different combinations are possible. These combinations were applied first to an input matrix, which conserves the original correlation structure of the experimental data. A second investigation was then carried out for the permuted experimental data matrix, where the original correlation structure was destroyed.

To investigate a broader variety of different datasets the columns of the original data matrix were permuted. For each permutation the procedure (36 combinations) as described above was applied. Thus different sets of predictors were obtained. The following results represent an average over 10 permutations, with the exception of the results of the feature selection which represent the results of 100 permutations.

3.1.1 Overall predictive performance

To approximately compare the performance of the regression methods, the results of all 36 combinations were summed. The squared difference between the correlation of the estimated

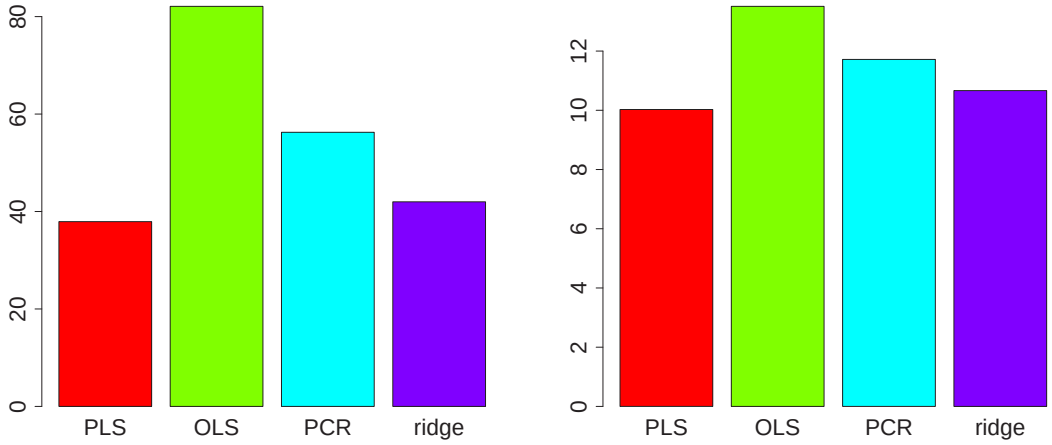

Figure 2: Overall performance of the regression methods: Sum of the squared differences between the true and the predicted correlation: for the original correlation structure (left), and for destroyed correlations (right).

response and the known true correlation was used as measure of the performance of the methods. Figure 2, left panel, shows the result for the original case where the correlation structure was conserved. This can be compared with the corresponding results for the case, when the correlations were destroyed by permutations, as displayed by Figure 2, right panel.

Another measure is the correlation between the true and the estimated coefficients. This measure yields the same ranking of performance.

In the sequel we consider only the simulation results, for which the original experimental correlation structure was conserved.

3.1.2 Predictive performance as dependent on noise in predictor or response

The predictive power of the regression methods in dependence of the true correlation, i.e. the noise in the response, is shown in Figure 3, left. When examining the influence of noise in the predictor variables, this kind of random error leads to the effect that even OLS does not yield unbiased estimations of the coefficients. Figure 3, right, shows the results for a true correlation of $C_{true} = 0.8$. When 50 predictors were taken, the same tendency is observed, OLS is generally worse.

3.1.3 Predictive performance as dependent on the number of predictive variables

The impact of the number of predictors was simulated as shown in Figure 4 by setting the true correlation to $C_{true} = 0.8$, while leaving the noise in the predictors zero.

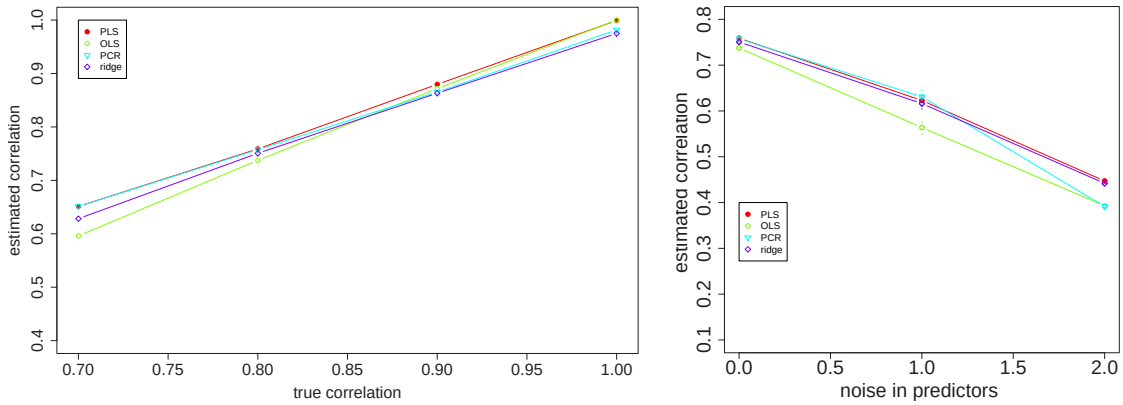

Figure 3: The predictive power of the regression methods as dependent on noise in the response (left) and noise in the predictors (right). Parameters to assess the relevance of noise in the response were $\sigma_p = 0.0$ and $p = 25$, for the right panel $\sigma_p/var(X)$ is displayed on the first axis. In both panels, the median estimated correlation over 30 different test sets is shown, the error bars depict the corresponding standard error of the mean.

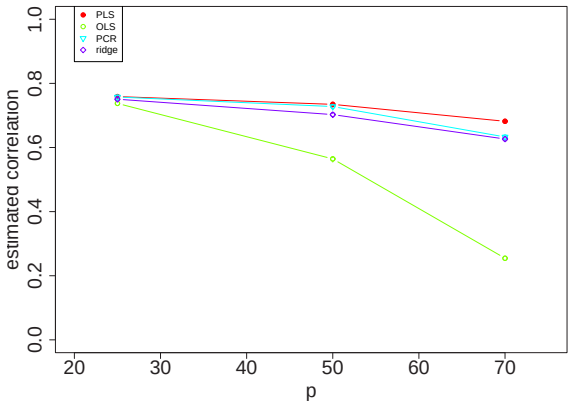

Figure 4: The predictive power of the regression methods as dependent on the number of predictors, for the original correlation structure. The points show the median estimated correlation over 30 different test sets, the error bars the corresponding standard error of the mean.

3.2 Feature selection

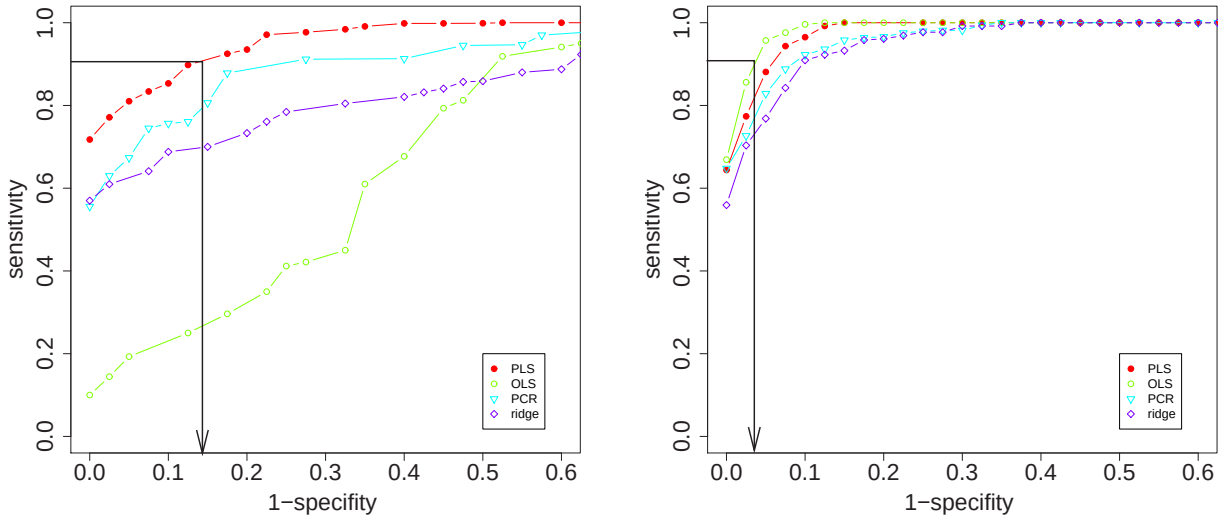

Figure 5: ROC-curves for results comparison of the *coefficients method* (left) and the *correlation method* (right). Parameters: $p = 50$, $C_{true} = 0.8$. The arrows indicate the readout for specificities if sensitivities are fixed at $sens = .0.9$.

Feature selection was assessed using two different approaches, the *coefficients method* and the *correlation method*. Methods' performances regarding feature selection were tested without noise in the predictors, the noise in the response was fixed such that a true correlation of $C_{true} = 0.8$ resulted, $p = 50$ predictors were used.

For the *coefficients method* we determined the regression coefficients, which are significantly different from zero [15]. Only in the case of OLS these are identical with the single variables' regression coefficients.

For PLS, we used high values of the *variable importance score*, VIP, instead [16]. To detect the true predictors permutation tests were carried out. To evaluate the results we calculated a receiver operating characteristic (ROC) for all methods.

Regarding the *correlation method* we looked for the correlation of each independent variable with the response estimated by the different methods [17]. Here the results depend strongly on the noise in the response i.e. the true correlation.

Figure 5, right, shows the result for a true correlation of 0.8 and 50 predictors. We see that already at false positive rate of 0.0 a sensitivity of 0.8 is reached for OLS and 0.05 for PLS.

Summarising Figure 5, PLS gives best results for the *coefficients method*, while OLS gives best results for the *correlation method*. OLS/correlation leads to seemingly small improvements. However, to really judge the *practical relevance* of different performances in feature selection, it is important to take into account, that it is not sensitivity or specificity of a feature selection which counts for the objective, but the false discovery proportion. This is because of the necessary validation experiments which have to follow each screening approach or high-throughput data analysis.

These validation experiments are relatively expensive, and, hence, it would be desirable to perform them as efficient as possible, *i.e.* with as few false candidates as possible. Thus, the

proportion of false positives within the candidate list of correlated features is a relevant measure to minimise.

The false discovery proportion, FDP, is defined as follows:

$$FDP = \frac{FP}{FP + TP} = 1 - \frac{p \cdot sens}{p \cdot sens + (1-p)(1-spec)}$$

where FP depicts the false positive features, TP the truly correlated features (true positives), sensitivity $sens$, specificity $spec$, and p the proportion of truly correlated features within all features (prevalence).

Concerning the comparison of coefficients/PLS versa correlation/OLS, assume the sensitivity fixed at a typically high value of e.g. $sens = 0.90$, and the prevalence of informative features at $p = 0.2$, as in our simulation studies. The belonging specificities for coefficients/PLS and correlation/OLS can be taken from figure 5 as indicated, resulting in $spec_{\text{coefficients/PLS}} = 0.86$ and $spec_{\text{correlation/OLS}} = 0.96$, respectively.

For this typical scenario, the false discovery proportions would result as

$$FDP_{\text{coefficients/PLS}} = 0.38$$

and

$$FDP_{\text{correlation/OLS}} = 0.15.$$

This reflects the fact that, a 10% increase in specificity can result in a 2.5-fold decrease for the proportion of false positives, which is a highly relevant improvement. Hence, correlation/OLS seems to be the recommendable method in our scenario.

3.3 Examples for correlation of experimental phenotypes with metabolite profiles

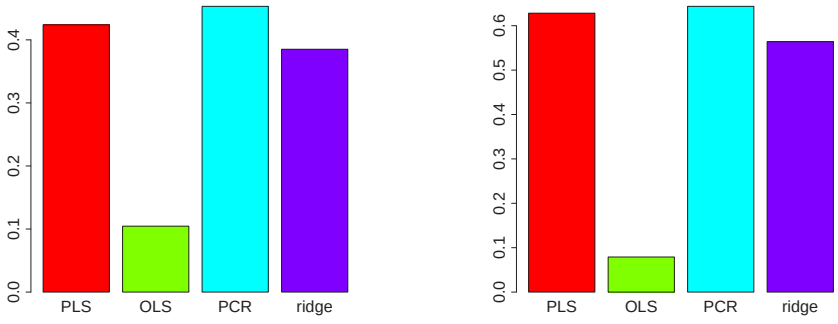

Figure 6: Comparisons of cross-validation correlations of predicted $\hat{Y}$ and the true Y for the four methods in question, for two phenotypical traits of the tomato dataset. Left: fruit length, right: harvest index. OLS performs worst in both cases, the three alternative methods show higher correlations of predicted and true responses.

As an illustrative example we also include the analysis of two phenotypical traits of the tomato dataset by Schauer *et al.* [2], *fruit length* and *harvest index*. Figure 6 shows mean prediction

errors of a cross-validation approach, using a 20% random testset of all measured samples, over 100 runs of analysis.

For both phenotypical traits results show the inferiority of OLS if compared to the other three multivariate methods regarding the *prediction* objective, as already studied within the simulation approach. Moreover, differences between these two phenotypical are also visible: levels of cross-validation correlations are considerably higher in the case of *harvest index* if compared to *fruit length*.

4 Discussion

Integrative bioinformatics seeks to combine data and analysis results spanning different methodologies and organisational levels of molecular biology. There are different levels approaching the ultimate aim of a comprehensive model integrating experimental results from different data types and/or levels of molecular organisation. Frequently, the first step towards this aim is to find out about the features governing the most important relationships among data from two different sources using large-scale correlation analyses [18]. This task becomes exceptionally challenging as the number of dimensions in the data grows. Often, we are therefore concerned with high-throughput transcriptional, protein or metabolomics data which are sought to be combined with each other or with phenotypical trait information available for the samples under investigation.

A metabolomics experimental series has been the basis for our methodological study of predictive and feature selection performances involving four standard regression methods frequently applied in bioinformatics' integrative data analysis (further examples e.g. [19, 20]). First, we were able to show that it is especially the typical correlation structure of the metabolic profiles which determines the investigated performances of analysis methods. Our approach is, hence, to compare methods on simulated datasets where this typical experimental correlation structure is preserved. Second, it turned out that different methods are to be preferred depending on the focus of analysis, *prediction* or *feature selection*. This choice also determines the relative importance of different error sources.

Interpretation of multiple regression results is essentially a vaguely defined problem. However, we approached it in our study having two objectives. On the one hand, we are attempting to define the true unknown predictors among many variables, the bulk of which do not contribute to a the considered phenotypical trait. This objective is known as the problem of *feature selection*. On the other, we are also interested in how well the selected features are in turn capable to predict the response variable. This we call the *prediction task*.

Regarding the *prediction task*, as shown in Figure 2 the failure of the OLS procedure is drastic in the case when the original correlation structure is conserved, while the other methods perform approximately equally well. However, when the correlations are removed by permutations, the relative performance of PCR declines, as expected. The difference between all methods is much smaller than in the case of strong correlations. In the case of 25 predictor variables the PLS, PCR and Ridge Regression reach nearly the maximum possible correlation (Figure 3, left panel) and are also robust against a rising number of predictors (Figure 4). All methods are also quite stable against noise in the predictors as long as the experimental variance is not greater than the biological. OLS is inferior even in favourable cases. Ridge Regression and PLS are

better or equally good as the other methods. In contrast, PCR performs well in the case of correlated data but fails in the case of weak correlations.

Regarding *feature selection*, there are markable differences between the *coefficients method* and the *correlation method*. Finally, *correlation/OLS* shows the most favourable ROC-curve, translating as the smallest false discovery proportion. A possible reason for the observed differences between the *coefficients* and the *correlation* approach may be, that the *coefficients method* possibly selects a subset of the variables which is optimal for regression. However, a lot of true predictors are not found with this approach. In the case of metabolites this could mean that, only few metabolites of a given pathway would be found, even if all metabolites of this pathway are responsible for a certain trait. The *correlation method*, looking for correlations with the estimated trait, seems to be the favourable alternative. In contrast to the results for the *prediction task*, OLS is never worse than the other methods and in many cases better than Ridge Regression or PLS. This may be due to the fact that PLS, PCR and Ridge Regression suppress the variables, which do not spread strongly in the predictor matrix, but are nevertheless important for the trait.

Our results also allow an estimation of error effects on measured correlations: We are able to estimate the biological correlation between a set of predictors, e.g. metabolites, with a response, e.g. a trait like growth, given an estimated correlation in cross validation. For this purpose, the experimental conditions - sample size, number of variables, experimental error in response and predictors must be known.
Consider the following example calculation. Let us assume that the biological variation increases linearly with the mean value of the trait, i.e. the coefficient of variation (CV) is constant. In this case a CV of 0.05 results in a correlation of ≈ 0.9. If the measurement error has a CV of 0.1, the correlation resulting from the combined effect of biological variation and experimental error is ≈ 0.80. Our simulation shows that even in the case of low errors in the independent variables the predicted correlation is 0.72.

As an application example, two of the experimental phenotypes measured by Schauer et al [2] were chosen as input for feature selection and cross-validation comparisons using all of our four multivariate analysis methods. First, also these experimental results reflect our findings from the simulation study, that OLS is inferior for solving the *prediction* objective. Moreover, the differences in achievable correlations of metabolite profiles with these two phenotypical traits might lead to the interpretation that the *harvest index* is rather more reflected in the metabolite profiles than this is the case for the trait *fruit length*.

As an outlook, we want to stress that also multivariate responses are worthy of investigation, as also *combinations* of phenotypical traits may be more adequately mirroring a phenotypical status which might be explained by molecular variables. Also, for *non-linear* correlations the appropriate methods have to be methodologically validated in a similar approach.

A systematic assessment of how a given experimental correlation structure limits the achievable performances in both prediction and feature selection should be a standard first step of data analysis – prior to taking into account the actually measured response variables and estimating the correlations of interest. This step is necessary for a valid interpretation of the actual results as well as for comparability with similar datasets and analyses.

References

[1] Janes K.A., Albeck J.G., and Yaffe M.B. A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science*, 310:1646–1653, 2005.

[2] Schauer N., Semel Y., Roessner U., Gur A., Balbo I., Carrari F., Pleban T., Perez-Melis A., Bruedigam C., Kopka J., Willmitzer L., Zamir D., and Fernie A.R. Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. *Nature Biotechnology*, 24:447 – 454, 2006.

[3] Olken F. Graph data management for molecular biology. *OMICS A Journal of Integrative Biology*, 7(1):75–78, 2003.

[4] Frank I. and Friedman J.H. A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135, 1993.

[5] Butler N.A. and Denham M. The peculiar shrinkage properties of partial least squares regression. *Journal of the Royal Statistical Society*, 62(3):585–593, 2000.

[6] Madansky A. The fitting of straight lines when both variables are subject to error. *JASA*, 54:173–205, 1959.

[7] Roessner U., Luedemann A., Brust D., Fiehn O., Linke T., Willmitzer L., and Fernie A.R. Metabolic profiling allows comprehensive phenotyping of genetically or environmentally modified plant systems. *Plant Cell*, 13:11–29, 2001.

[8] Roessner-Tunali U., Hegemann B., Lytovchenko A., Carrari F., Bruedigam C., Granot D., and Fernie A.R. Metabolic profiling of transgenic tomato plants overexpressing hexokinase reveals that the influence of hexose phosphorylation diminishes during fruit development. *Plant Physiology*, 133:84–99, 2003.

[9] Oba S., Sato M.A., Takemasa I., Monden M., Matsubara K., and Ishii S. A bayesian missing value estimation method for gene expression profile data. *Bioinformatics*, 19(16):2088–2096, 2003.

[10] Hotelling H. The most predictable criterion. *The Journal of Educational Psychology*, 26(2):139–143, 1935.

[11] Wold H. *Soft modelling by latent variables*. Academic Press. London, 1975.

[12] Barker M. and Rayens W. Partial least squares for discrimination. *Journal of Chemometrics*, 17:166–173, 2003.

[13] Stone M. and Brooks R. Continuum regression: Cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression. *Journal of the Royal Statistical Society*, 52(2):237–269, 1990.

[14] Hastie T., Tibshirani R., and Jerome J.F. *The elements of statistical learning*. Springer, 2001.

[15] Eriksson L., Johansson E., Kettaneh-Wold N., and Wold S. *Introduction to Multi- and Megavariate Data Analysis using Projection Methods (PCA & PLS)*. Umetrics AB, Umeå, Sweden, 1999.

[16] Chong I.G. and Jun C.H. Performance of some variable selection methods when multicollinearity. *Chemometrics and Intelligent Laboratory Systems*, 78:103–112, 2005.

[17] Razavi A.R., Gill H., Stahl O., Shahsavar N., and South-east Swedish breast cancer study group. Exploring cancer register data to find risk factors for recurrence of breast cancer – application of Canonical Correlation Analysis. *BMC Medical Informatics and Decision Making*, 2005.

[18] Aitchison J.D. and Galitski T. Inventories to insights. *Journal of Cell Biology*, 161(3):465–469, 2003.

[19] Nie L., Wu G., Brockman F.J., and Zhang W. Integrated analysis of transcriptomic and proteomic data of desulfovibrio vulgaris: zero-inflated poisson regression models to predict abundance of undetected proteins. *Bioinformatics*, 22(13):1641–1647, 2006.

[20] Gao F., Foat B.C., and Bussemaker H.J. Defining transcriptional networks through integrative modeling of mRNA expression and transcription factor binding data. *BMC Bioinformatics*, 5:31–40, 2003.