File ID	105467
Filename	Chapter 9 Assessing the statistical validity of proteomics based biomarkers

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type	Dissertation
Title	Detection of biomarkers for lysosomal storage disorders using novel technologies
Author	M.J. van Breemen
Faculty	Faculty of Medicine
Year	2008
Pages	248

FULL BIBLIOGRAPHIC DETAILS:
http://dare.uva.nl/record/271774

Chapter Nine

Assessing the statistical validity of proteomics based biomarkers

Suzanne Smit[a], Mariëlle J. van Breemen[b], Huub C. J. Hoefsloot[a], Age K. Smilde[a], Johannes M. F. G. Aerts[b] and Chris G. de Koster[b,c]

[a] Swammerdam Institute for Life Sciences, University of Amsterdam, Amsterdam, The Netherlands
[b] Department of Medical Biochemistry, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands
[c] Clinical Proteomics Facility, Academic Medical Center, University of Amsterdam, Amsterdam, The Netherlands

Abstract

A strategy is presented for the statistical validation of discrimination models in proteomics studies. Several existing tools are combined to form a solid statistical basis for biomarker discovery that should precede a biochemical validation of any biomarker. These tools consist of permutation tests, single and double-cross validation. The cross-validation steps can simply be combined with a new variable selection method, called rank products. The strategy is especially suited for the low-samples-to-variables-ratio (undersampling) case, as is often encountered in proteomics and metabolomics studies. As a classification method, Principal Component Discriminant Analysis is used; however, the methodology can be used with any classifier. A data set containing serum samples from Gaucher patients and healthy controls serves as a test case. Double cross-validation shows that the sensitivity of the model is 89% and the specificity 90%. Potential putative biomarkers are identified using the novel variable selection method. Results from permutation tests support the choice of double cross-validation as the tool for determining error rates when the modelling procedure involves a tuneable parameter. This shows that even cross-validation does not guarantee unbiased results. The validation of discrimination models with a combination of permutation tests and double cross-validation helps to avoid erroneous results which may result from the undersampling.

Introduction

One area of interest in the study of disease is the proteomics based search for disease markers. Theoretically, proteomics considers all proteins in an organism, but usually only part of the proteome is measured. Surface-enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI-TOF MS) is a relatively new technique. It combines absorption of a subproteome on a chip with time-of-flight mass spectrometric detection. A subset of the protein complement of the sample is bound to the chip and measured. The advantage of SELDI-TOF MS over conventional techniques is the possibility of applying complex body fluids such as saliva, urine and blood directly to the chip. Mass spectra of samples of diseased and (healthy) control individuals are measured with the objective of distinguishing between the control and diseased groups. Data analysis methods are used to find differences, which can be single protein markers or different patterns in the protein profiles [1-5]. When these differences prove to be statistically valid, their biochemical meaning can be ascertained, so that they may be put to use in the clinic. The focus of this paper is on data analysis and statistical validation.

The data analysis may start by building a discrimination model that separates the groups as well as possible and that describes for which (combination of) variables they are most distinct. The large number of variables in the proteomics setup generates modelling and validation challenges commonly referred to as the curse of dimensionality [6] or undersampling. In short, the curse of dimensionality means that the number of samples needed to accurately describe a (discrimination) problem increases exponentially with the number of dimensions (variables) measured. Due to the limited availability and/or cost of measurement the number of samples is usually relatively small, in the tens or hundreds. The number of samples is then too small to accurately describe the groups. If that is the case, good discrimination results for the original control-diseased problem are possibly not significant. A permutation test can evaluate this possibility and can help to decide whether to look further into the biochemical validity of these differences between the control and diseased groups.

A permutation test gives information about the discrimination performance of the model, but the model should also be able to correctly classify new samples as diseased or control preferably using a low number of variables. Due to the limited number of samples, it is often not possible to test the ability of the model to classify new samples on a masked test set. The test data cannot be incorporated in the model and as a result the model would be trained on insufficient data. Additionally, the test set would contain very few samples, and the error in assigning only a few samples would not give a reliable indication of the prediction error. Cross-validation is often the validation method of choice, because it makes better use of the data. As Ambroise and McLachlan [7] and Simon et al. [8] have shown, cross-validation only gives a reliable error rate when the complete modelling procedure is cross validated. Failure to do so results in optimistic error rates. When the model requires the determination of a tuneable parameter (for example the number of components in principal component analysis) this has to be incorporated in the cross-validation.

In this paper, cross-validation is used for determination of a tuneable model parameter and for candidate biomarker selection in a proteomics example. The discrimination and classification performance of the model is assessed with (double-) cross-validation in combination with a permutation test [9,10]. The example of choice is Gaucher disease. Gaucher disease is a rare inherited enzyme deficiency disorder that results in enlarged spleen and liver and bone disease. Gaucher disease is chosen because previous studies have demonstrated that several proteins show elevated blood levels in Gaucher patients. Plasma levels of tartrate-resistant acid phosphatase 5b, β-hexosaminidase, angiotensin converting enzyme and lysozyme are increased in Gaucher patients [11]. Also two specific Gaucher cell markers are known: chitotriosidase and CCL18. Chitotriosidase shows a thousand-fold increased activity in serum of symptomatic Gaucher patients [12]. Plasma CCL18 levels are elevated ten to fifty-fold in symptomatic Gaucher patients [5]. SELDI-TOF MS is used to create protein profiles of the serum of 20 Gaucher patients and 20 controls. Due to the measuring conditions, the protein profiles do not contain proteins that are known to be differentially expressed in Gaucher patients. Nevertheless, the groups of serum protein profiles are expected to differ, due to the large clinical differences between the groups.

Principal component discriminant analysis (PCDA) is used to discriminate between the groups of protein profiles. The significance of the discrimination is evaluated in a permutation test. Double cross-validation is used to estimate the error of the model in classifying unknown samples. The cross-validation procedure generates several models. From these models discriminating proteins are selected using the rank products procedure as described by Breitling [13]. Combining PCDA, permutation tests, double cross-validation and variable selection with rank products results in a strategy for the discovery and rigorous statistical validation of candidate biomarkers.

Materials

Patients

All patients with Gaucher disease (type I) studied (10 males and 10 females; 15-65 years old, at the initiation of therapy) were known by referral to the Academic Medical Center. Of the 20 patients, 18 received enzyme replacement therapy (alglucerase, imiglucerase, [individualized dosing], Genzyme, Cambridge, MA) and 2 patients received substrate reduction therapy (chronic oral administration of an iminosugar inhibitor of glucosylceramidesynthesis, *N*-butyldeoxynojirimycin, Oxford Glycosciences, Abingdon, United Kingdom). The control group consisted of 7 male and 13 female healthy volunteers, 23-68 years old.

Serum samples

Blood samples were collected from patients (between 1991 and 2001) and healthy volunteers (between 1994 and 2002) in 7 mL, BD Vacutainer, ‘red-top’ tubes (BD # 367625), and sera were prepared. Collection protocols were the same for both groups.

Blood samples were allowed to clot at room temperature for 30 minutes. Subsequently, blood samples were centrifuged at 1300 RCF for 10 minutes at room temperature. All serum samples were stored at -20°C until required. Serum samples of Gaucher patients were obtained before initiation of therapy. Approval was obtained from the Ethic Committee. Informed consent was provided according to the Declaration of Helsinki.

SELDI-TOF MS

Serum samples were surveyed for basic proteins with SELDI-TOF MS making use of the anionic surface of CM10 ProteinChip® Arrays (Ciphergen Biosystems Inc., Fremont, CA, USA). Serum samples (10 µL) were first mixed with 90 µL of denaturation solution (9 M urea [Sigma Chemical Company, St Louis, MO, USA], 2% CHAPS [Fluka Biochemika, Buchs, Switzerland], and 1% DTT [Sigma Chemical Company, St Louis, MO, USA]) and incubated at room temperature for 1 hour. An aliquot (10 µL) of this solution was mixed with 90 µL binding buffer (50 mM Tris [Sigma Chemical Company, St Louis, MO, USA] + 0.1% Triton X-100 [BHD Laboratory Supplies, Poole, Dorset, UK], adjusted to pH 7 with hydrochloric acid [Merck, Darmstadt, Germany]). Before application of the sample to a CM10 ProteinChip® Array, all spots were equilibrated. To equilibrate the CM10 ProteinChip® Array, spots were washed with 200 µL of binding buffer (2 times, 5 minutes on a platform shaker) by using a Ciphergen Biosystems 96-well bioprocessor. After equilibration, buffer was removed and samples were added. Gaucher and control samples were applied in random order. The samples were allowed to bind to the anionic surface for 40 minutes at room temperature on a platform shaker. Subsequently the ProteinChip® Arrays were washed with 200 µL binding buffer (2 times, 5 minutes on a platform shaker). Next the ProteinChip® Arrays were washed with 200 µL binding buffer without Triton X-100 (2 times, 5 minutes on a platform shaker). After a brief wash with deionised water (to remove salts) ProteinChip® Arrays were dried on air. Prior to SELDI-TOF MS analysis, matrix was added to each spot (2 times 0.5 µL of sinapinic acid [Fluka Biochemika, Buchs, Switzerland] (10 mg/mL) in 50% aqueous acetonitrile [Merck, Darmstadt, Germany] containing 1% TFA [Fluka Biochemika, Buchs, Switzerland]). After co-crystallization of the (bound) proteins with the matrix molecules, a pulsed nitrogen laser was used for ionization of the samples. ProteinChip® Arrays were analyzed using a PBSIIc ProteinChip® Reader (Ciphergen Biosystems Inc., Fremont, CA, USA), a linear laser desorption/ionization time-of-flight mass spectrometer equipped with time-lag focussing. The result is a mass spectrum composed of the mass to charge ratios (*m/z* values) and intensities of the desorbed (poly)peptide ions. All spectra were acquired in the positive-ion mode.

Preprocessing of SELDI-TOF MS data for further analysis

External calibration

Spectra were externally calibrated against a mixture of known peptides (All-in-1 Peptide Standard, Ciphergen Biosystems Inc., Fremont, CA, USA). The pre-mixed peptide standard includes arg^8-vasopressin (1084 Da), somatostatin (1637 Da), porcine dynorphin

(2147 Da), human adrenocorticotropic hormone (1-24) (2933 Da), bovine insulin β-chain (3495 Da), human insulin (5807 Da), and hirudin BHVK (7033 Da).

Spot-to-spot calibration

Spot-to-spot calibration is a feature of the ProteinChip® Software that accounts for the spot to spot variation that can occur on an individual array. To determine the correction factors for the different positions on an array we used a set of peaks that is always present in our spectra. The correction factors for the different positions on an array are applied to the corresponding mass spectra and used in the recalculation of the masses.

Baseline subtraction

The ProteinChip baseline algorithm removes offsets in the spectra that are the result of how the signal is collected electronically and of chemical noise contributed from the energy absorbing molecules in the matrix. The algorithm is a modified piecewise convex-hull that attempts to find the bottom of the spectra and correct the peak height and area. Baseline subtraction is applied to all spectra.

Peak detection

Peaks were detected with Biomarker Wizard™. Biomarker Wizard™ is a feature of the ProteinChip® Software that is used for preparing data generated by ProteinChip® Software for further analysis. Biomarker Wizard™ performs the peak picking across the samples. The resolution of a peak at *m/z* 100 is 300, where the resolution R= m/Δm, m the mass of the ion and Δm is designated as the full width at half-maximum. An example of the obtained spectra can be found in Fig. 1.

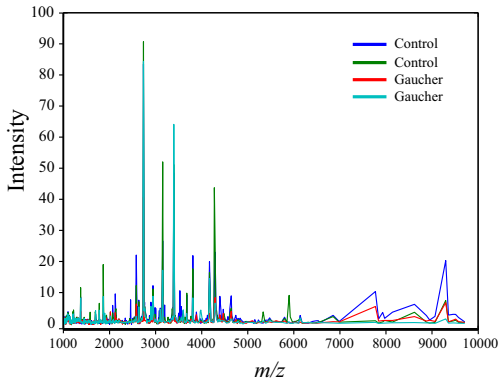


Figure 1. Four spectra, two controls and two patients, after the pre-processing steps.

Methods

Principal component discriminant analysis

Differences have to be found between the SELDI-TOF MS protein profiles of serum of controls and Gaucher patients to classify individuals as healthy or diseased. A simple method for discrimination between two groups is Fisher's linear discriminant analysis

(FLDA). Good discriminating directions are directions in *m/z* space in which the differences between the groups are large compared to the differences within the groups. In the two-group case, this direction is given by the vector d that maximizes the ratio R:

$$\mathbf{R} = \frac{\mathbf{dBd}}{\mathbf{dWd}}$$

where **W** is the pooled within class sample covariance matrix and **B** is the between class sample covariance matrix. The discriminating direction is the eigenvector of $\mathbf{W}^{-1}\mathbf{B}$ corresponding to its largest eigenvalue [14]. Because there are more *m/z* values than samples, the matrix W is singular. This means that $\mathbf{W}^{-1}$ does not exist and FLDA cannot be applied directly. This problem can be overcome by using principal component analysis (PCA), which finds new "variables" or principal components to describe the data. These components are linear combinations of the original *m/z* values. The first principal component (PC) describes as much of the variation in the data as possible, the second describes as much of the remaining variation as possible, etc. By keeping only a few of the principal components the dimensionality of the data can be reduced to a point where FLDA is applicable, while preserving most of the information in the data. The number of components in the model is a tuneable metaparameter the value of which can be decided upon using cross-validation, which will be described shortly. The combination of FLDA with PCA yields principal component discriminant analysis (PCDA) [15-18].

Permutation test

Once a PCDA model is found that discriminates between the healthy and diseased groups, what can be said about the significance of the discrimination? Because of the size of the data set – there are many more *m/z* values than there are samples – it might be possible to find two arbitrary groups that can be well separated. In that case, a good discrimination in the original problem may very well be a coincidence and may not be very significant. A permutation test can evaluate this possibility. In a permutation test the class labels of the samples are randomly permuted: Every sample is randomly assigned a label while the number of control and diseased labels is the same as in the original problem. The permuted problem is treated in exactly the same way as the original problem. If the results are comparable to or better than the results of the original problem, the discrimination is probably a coincidence, or the result of confounded variables in poorly matched diseased and control samples. However, when a lot of permutations give groups for which the discrimination is worse, the result for the original problem may be significant [19].

Cross-validation

As mentioned before, the number of components in the PCDA model is determined with cross-validation. Cross-validation has two distinct applications. In the first place, it is a method that can give an estimate of the prediction error when the sample group is small. Cross-validation gives other information about the model than a permutation test because the latter does not assess the classification performance (i.e. it does not give a prediction error). When the dataset contains many samples, the predictions of one larger separate test set can also give an independent prediction error. This error differs in one important aspect

from the information obtained by cross-validation. The dataset on which the model is built is only one subset from the entire control and diseased population. The model and corresponding prediction error are one possible outcome. Another subset would surely result in a different model and error. Cross-validation evaluates the effects of using only one subset by splitting the available data several times into different test and training sets. In ten-fold cross-validation, for example, the modelling and subsequent prediction is repeated ten times. Every time, ten percent of the data is masked; the remaining ninety percent is the training set that is used for modelling. Although the training sets overlap partly, they are different subsets from the data and they result in different models. The ten different models from the cross-validation give insight in the variability of the model we build on the complete data set. In addition, a possible lucky subset that results in an optimistic prediction error is averaged out by the other subsets.

The second use of cross-validation is in estimating a tuneable parameter. For PCDA models the tuneable parameter is the number of components. For estimation of the parameter the complete cross-validation procedure is repeated for all possible parameters. The parameter is chosen that leads to the lowest cross-validation error. With this choice, information from the masked test sets is brought into the model. It makes the cross-validation error corresponding to the chosen number of components an optimistically biased estimate of the prediction error of the model.

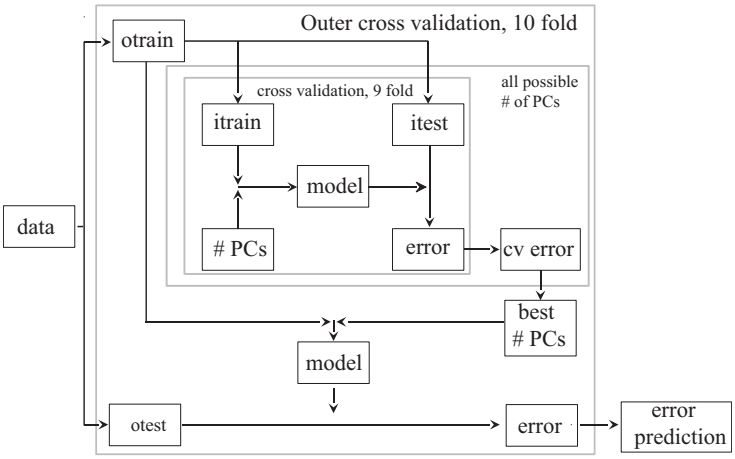
Taking many components conserves the original data best. Restricting the number of components reduces the amount of noise after the PCA step. Calculating the number of components in the PCA model with cross-validation is an appropriate way of obtaining a correct number of components. This number of components is capable of retaining the crucial information for the discrimination while discarding noise.

Double cross-validation

Cross-validation can still be used to find a good estimate of the prediction error, but it has to be used in a different way. Determining the tuneable parameter with cross-validation is part of the procedure to build a model. The entire modelling procedure has to be cross validated in order to obtain the prediction error. This can be done in a double cross-validation [20,21]. Double cross-validation consists of two nested cross-validation loops (Scheme 1). The modelling procedure, including the cross-validation that determines the tuneable parameter, forms the inner loop. The cross-validation for the error estimation takes place in the outer loop.

The outer loop starts by masking a few samples. The remainder of the data enters the inner loop. In the inner loop cross-validation estimates the tuneable parameter for the model as described above. The estimated parameter is used to build a model on all the data that entered the inner loop. This model is returned to the outer loop where it predicts the samples that were masked thus far. The masking, parameter estimation, model building and predicting of masked samples is repeated until each sample is masked exactly once in the outer loop. The double cross-validation error is a reliable estimate of the error of the modelling procedure, because the predicted samples are completely new to the model.

Double cross-validation also gives insight in the variability of the tuneable parameter and the model. Every outer loop generates a different subset on which the parameter is estimated and the model is built. Each different subset results in a different estimate for the parameter and in a different model.



Scheme 1. Double cross-validation. The original data set is split into a training set (otrain) and test set (otest) ten times in the outer cross-validation loop. In the inner loop otrain is split up nine times in a training set (itrain) and a test set (itest). Every number of components (PCs) for the PCA step that is considered is used to build a model on itrain. This model then predicts the classes of the samples in itest, leading to an error. The errors of all the inner cross-validation models that have the same number of components are combined in the cross-validation error (CV error). The number of components that leads to the lowest cross-validation error is selected and used together with the corresponding otrain for the model in the outer loop. The data in otest is predicted with this model to give an error. The errors made in the ten different outer test sets are combined in the cross-validation error.

Rank products

In the cross validating procedure several models are built. The rank products procedure seems to be a natural partner for cross-validation to evaluate the overall importance of a variable. The discriminant vector found with PCDA represents the differences between the control and the diseased groups. Since the largest peaks in this vector are most important for the discrimination, we can select *m/z* values based on their absolute value in the discriminant vector. In the ten-fold cross-validation ten different discriminant vectors are found in which the importance of the *m/z* values may differ. The information in the ten discriminant vectors can be combined using the rank products selection method [13]. For each of the discriminant vectors, the *m/z* values are ranked according to their absolute value. The *m/z* value with the largest absolute value gets rank 1, the next largest gets rank 2, etcetera. The ten ranks of each *m/z* value are multiplied to obtain the rank product, and the *m/z* values with the lowest rank product are the ones with the largest discriminative power. In this way, single cross-validation in combination with rank products can be used for variable selection. The prediction error associated with the selected variables is determined with double cross-validation.

Results

Data

Serum samples of 20 controls and 20 Gaucher patients were measured with SELDI-TOF MS. On visual inspection of the spectra, the mass spectrum of one Gaucher sample (a female receiving enzyme replacement therapy) appeared to be flawed. Consequently, this mass spectrum was excluded from the study. Preprocessing of the spectra was performed according to the descriptions given above. The resulting data set contained 20 control and 19 Gaucher spectra, each consisting of 590 *m/z* values between 1000 en 10.000. The protein profiles were normalized by dividing each profile by its median to arrive at comparable spectra. To prevent the largest peaks in the protein profiles from dominating the PCA part of the model, the data were auto-scaled. For (double) cross-validation, auto-scaling was always performed on the training data before modelling and then the test data was scaled prior to prediction with the scaling parameters of the training set. By doing this, it is ensured that the prediction of the test data is truly independent.

Discrimination

A discrimination model was built based on all data. A single cross-validation pointed at 15 principal components to be used (see further on). This resulted in a model that perfectly discriminated between the Gaucher and control groups: all samples were assigned to the correct class. Hence, the resubstitution error – error made in classifying samples used to model the data - was zero. With a permutation test, the significance of the discrimination was evaluated. The class labels of the samples were randomly permuted 10,000 times and PCDA models were made. A histogram of the resubstitution errors of the resulting models is shown in Fig. 2A. Although the average resubstitution error for the permutations (8.6) was larger than the resubstitution error for the original data (0), it was much smaller than the average error expected for randomly permuted problems (19.5: a flip of the coin result). Also, four of the permuted problems resulted in a resubstitution error of zero, like the original problem. This shows the well known overfitting phenomenon and a resubstitution error which is a severely optimistically biased prediction error.

The number of principal components for the discrimination model on all data was determined with cross-validation. The number of components was restricted between 2 and 20. For each possible number of components, a ten-fold single cross-validation was performed. In each fold, two samples were masked from both classes. Since there were 19 Gaucher samples, only one Gaucher sample was masked in the last fold. The single cross-validation error was lowest when 15 components were used; of the 39 samples 1 control and 2 Gaucher samples were misclassified. The same cross-validation strategy was applied to the 10.000 permuted problems. Fig. 2B shows a histogram of the number of misclassifications. None of the permutations gave a lower single cross-validation error of the original problem, but one permutation resulted in the same cross-validation error (three misclassifications). On average, the permutations resulted in 16.6 misclassifications in the single cross-validation. Like the average resubstitution error this number is still lower than the expected number of misclassifications in random permutations. This confirms and

illustrates that the single cross-validation error is also optimistically biased when it is used for tuneable parameter estimation and validation simultaneously.

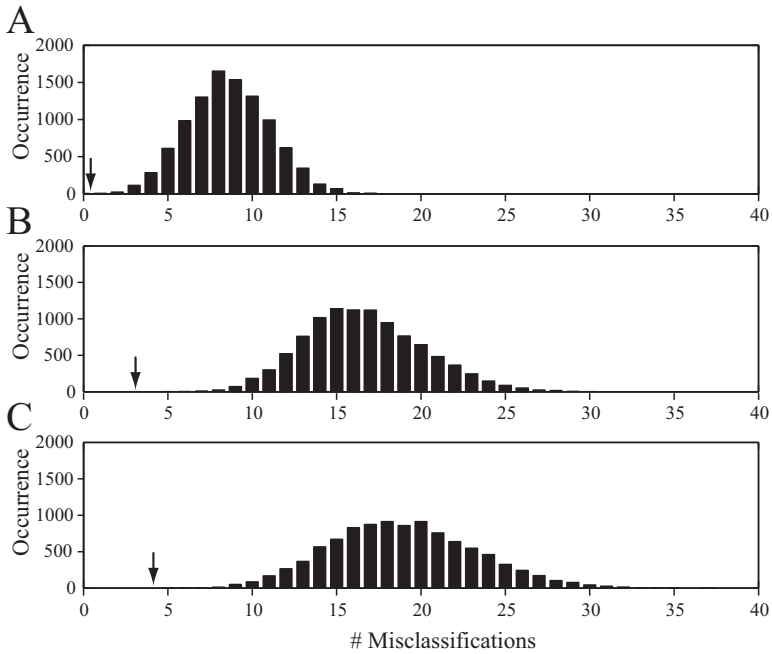


Figure 2. Permutation test. Histogram of the number of misclassifications in 10,000 permutations. A: resubstitution error; B: cross-validation error; C: double cross-validation error. The arrows indicate the number of misclassifications in the original problem.

Classification

The prediction error of the model in classifying unknown samples was established by double cross-validation. In the inner loop, the number of components for the model was determined by using nine-fold cross-validation. As in the single cross-validation, between two and twenty components were used in a model. The models from the inner loop were tested in the outer loop with ten-fold cross-validation. In the end, out of a total of 39 samples, two control and two Gaucher samples were misclassified. Thus, the sensitivity of the model was 89% and the specificity 90%.

These classification results are again compared to the double cross-validation results of 10.000 permutations (Fig. 2C). The double cross-validation errors of all the permuted problems were larger than the double cross-validation error of the original problem. The average prediction error was 19.9 misclassifications, which is approximately half of the 39 samples. This is what would be expected for random data: the model is not able to classify truly new samples. The best it can do is 'guess' at the class label, which leads to this flip-of-the-coin result. It illustrates the statement that the double cross-validation error is an independent estimate of the prediction error.

All three methods, re-substitution, single cross-validation and double cross-validation yield statistical significance in the permutation test. A P-value from each test could be calculated as the ratio of the number of equal or better performances with the permutated data and the total number of permutations. The significance of the double cross-validation is highest. Because the mean of the distribution of the double cross-validation is furthest away from zero the power of this test is also better than in the case of the other two methods.

Double cross-validation not only resulted in a prediction error for the model, it also gave information about the variability. The ten-fold outer loop resulted in ten different discriminant vectors at the end of the double cross-validation. The number of components in the PCA step of these models ranged from seven to twenty. However, the resulting ten discriminant vectors were very similar, which implies that PCDA is a robust method.
The combination of samples to form test sets in the outer loop was one possible order. The double cross-validation was repeated 100 times, each time with different combinations of samples in the test and training sets. This was done to exclude the possibility that a specific order of left-out objects would influence the results. The average number of misclassifications of those 100 runs was 4, which is the same as the number of misclassifications found in the double cross-validation discussed above. Hence, this is a stable result.

The validity of the sensitivity and specificity which was found depends on the matching of the Gaucher and control samples. In this study, the matching was not perfect: There is a difference in the distribution of sexes between the two groups. Also, the age of the patients and controls are not matched perfectly, but the groups do have the same large age range. Similar cohorts of patients and controls were used in studies that revealed the now well established Gaucher markers chitotriosidase and CCL18 [5,12]. The permutation test also gives information on (poor) matching of cases and controls. In a random permutation the (poor) matching is broken. In the 10,000 permutations there are many where for example the male/female matching is much poorer then in the original data. Still all the classification results turn out to be worse. From this it can be concluded that the matching was sufficient and that the difference due to Gaucher disease is the dominant effect.

Rank products

In the previous section it was determined that 15 components is the optimal number. With this number the ten fold cross-validation is performed. The ten discriminant vectors were used for variable selection using rank products. All *m/z* values per model were ranked and multiplied to obtain its rank product. The average rank product for a given *m/z* value $(590/2)^{10} = 5 \cdot 10^{24}$. Table 1 shows the ten *m/z* values with the lowest rank products, so the largest contributions to the discrimination. Surprisingly, all the top ten proteins are up-regulated in the group of Gaucher patients. It should be kept in mind that the analysis was focussed on relatively small proteins (molecular masses below 10.000 Da). It is known that various proteases, particularly cathepsins, are elevated in Gaucher plasma [22]. This may conceivably lead to unique low molecular mass degradation products. Alternatively, the

top ten ranking *m/z* values may also represent only one or a few proteins. Due to the action of proteases and singly and doubly charged states one protein could give rise to multiple peaks. The proteins with the lowest rank products are candidate biomarkers. A biochemical validation is the next step to assert the relevance of the putative markers before they can be viewed as true biomarkers, but this is beyond the scope of this paper.

Table 1. Top 10 best discriminating *m/z* values according to the rank products method and their rank products

***m/z* value**	**Rank product**
4058	36
5852.6	288
3685.4	1.15E+05
4546	4.35E+05
2067.9	2.92E+07
4214.8	1.13E+08
3840.1	1.36E+09
1008.2	2.28E+09
4016.2	5.03E+10
8949.4	7.81E+10

Note: All 10 proteins are up-regulation in the Gaucher patients.

Another question is how many *m/z* values with low rank product would have to be selected for a good predictive model. Fig. 3 shows how the classification error rate depends on the number of *m/z* values selected for the model. The error rates in Fig. 3 are double cross-validation errors. The rank products were calculated in an inner cross-validation and models based on different numbers of *m/z* values were tested in the outer cross-validation. In this way, the performance of the selected *m/z* values in classifying unknown samples was tested. As Fig. 3 shows, incorporating ten *m/z* values or less resulted in error rates of 8 out of 39 and higher. The lowest prediction error was achieved when 200 *m/z* values were incorporated in the model. Selecting 50 or more *m/z* values leads to a performance that is comparable to the performance of the model without selection. Apparently, not all *m/z* values are needed in the model to achieve good prediction. In fact, the best predictions were obtained with less than half of the *m/z* values. On the other hand, it is not possible to reduce the number of *m/z* values to just a few without significant loss of performance.

Conclusion

A strategy is presented for the discovery of candidate disease markers and statistical validation thereof. It consists of building a discrimination model with PCDA and subsequent validation of its discriminative ability with a permutation test and of its predictive ability by double cross-validation. It was shown that it is possible to select candidate biomarkers by combining cross-validation with rank products. The strategy was applied to SELDI-TOF MS spectra of serum samples of Gaucher patients and healthy controls. Double cross-validation showed that the PCDA model has a sensitivity of 89% and a specificity of 90%. In addition, the permutation test proved that the discrimination was significant. The results of the resubstitution, cross-validation and double cross-validation permutations tests supported the use of double cross-validation. All three tests

indicated that the result obtained for the original problem were not a coincidence. However, the test with double cross-validation was the only test that gave the flip-a-coin result that can be expected for randomly permuted labels in the two group case. These results illustrate the need for a thorough validation of discriminant models in proteomics. In this study, PCDA was chosen to build a discriminant model on SELDI-TOF MS data, but the conclusions regarding the validation with permutation tests and double cross-validation also hold for other discrimination methods and other types of omics data. For a procedure in which no meta-parameter has to be estimated the same procedure as described in this paper can be used, but with a single cross-validation instead of a double cross-validation.

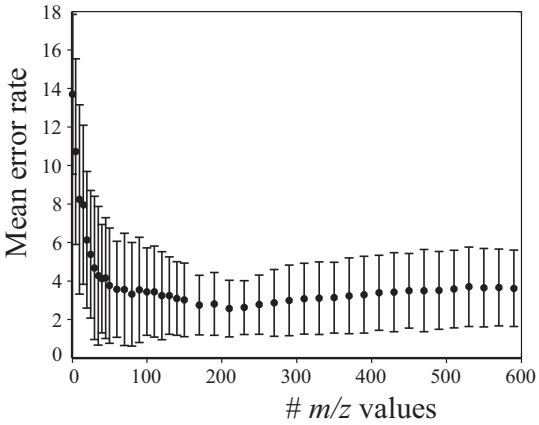


Figure 3. Error rate versus number of variables. For 1, 5 and 10 variables LDA was used to build the model, for larger number of variables we used PCDA with 10 PCs. The reported error rates are averages of 100 different double cross-validations.

Acknowledgments

The authors thank TNO Quality of Life (Zeist, NL) for providing PCDA m-files for use with Matlab®.

References

[1] E.F. Petricoin III, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, Use of proteomic patterns in serum to identify ovarian cancer, Lancet 359 (2002) 572-577.

[2] B.L. Adam, Y. Qu, J.W. Davis, M.D. Ward, M.A. Clements, L.H. Cazares, O.J. Semmes, P.F. Schellhammer, Y. Yasui, Z. Feng, G.L. Wright Jr., Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men, Cancer Res. 62 (2002) 3609-3614.

[3] A. Vlahou, A. Giannopoulos, B.W. Gregory, T. Manousakas, F.I. Kondylis, L.L. Wilson, P.F. Schellhammer, G.L. Wright, O.J. Semmes, Protein profiling in urine for the diagnosis of bladder cancer, Clin. Chem. 50 (2004) 1438-1441.

[4] S.G. Soltys, Q.T. Le, G.Y. Shi, R. Tibshirani, A.J. Giaccia, A.C. Koong, The use of plasma surface-enhanced

laser desorption/ionization time-of-flight mass spectrometry proteomic patterns for detection of head and neck squamous cell cancers, Clin. Cancer Res. 10 (2004) 4806-4812.
[5] R.G. Boot, M. Verhoek, M. de Fost, C.E.M. Hollak, M. Maas, B. Bleijlevens, M.J. van Breemen, M. van Meurs, L.A. Boven, J.D. Laman, M.T. Moran, T.M. Cox, J.M.F.G. Aerts, Marked elevation of the chemokine CCL18/PARC in Gaucher disease: a novel surrogate marker for assessing therapeutic intervention, Blood 103 (2004) 33-39.
[6] T. Hastie, J. Friedman, R. Tibshiranie, in: P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg, I. Olkin, N. Wermuth, S. Zeger (Series Advisors), The elements of statistical learning: data mining, inference and prediction, Springer, New York, 2001, pp. 22-27.
[7] C. Ambroise, G.J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data, Proc. Natl. Acad. Sci. U. S. A. 99 (2002) 6562-6566.
[8] R. Simon, M.D. Radmacher, K. Dobbin, L.M. McShane, Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification, J. Natl. Cancer Inst. 95 (2003) 14-18.
[9] K.R. Lee, X.W. Lin, D.C. Park, S. Eslava, Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method, Proteomics 3 (2003) 1680-1686.
[10] W.D. Tong, W. Xie, H.X. Hong, H. Fang, L.M. Shi, R. Perkins, E.F. Petricoin, Using decision forest to classify prostate cancer samples on the basis of SELDI-TOF MS data: assessing chance correlation and prediction confidence, Environ. Health Perspect. 112 (2004) 1622-1627.
[11] E. Beutler, G.A. Grabowski, Gaucher disease, in: C.R. Scriver, A.L. Beaudet, W.S. Sly, D. Valle (Eds.), The metabolic and molecular bases of inherited disease, McGraw-Hill, New York, 2001, pp. 3635-3668.
[12] C.E.M. Hollak, S. van Weely, M.H.J. van Oers, J.M.F.G. Aerts, Marked elevation of plasma chitotriosidase activity: a novel hallmark of Gaucher disease, J. Clin. Invest. 93 (1994) 1288-1292.
[13] R. Breitling, P. Armengaud, A. Amtmann, P. Herzyk, Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments, FEBS Lett. 573 (2004) 83-92.
[14] B.G.M. Vandeginste, D.L. Massart, L.M.C. Buydens, S. De Jong, P.J. Lewi, J. Smeyers-Verbeke, in: Handbook of Chemometrics and Qualimetrics: Part B, Elsevier, Amsterdam, 1998, pp. 213-219.
[15] R. Hoogerbrugge, S.J. Willig, P.G. Kistemaker, Discriminant analysis by double stage principal component analysis, Anal. Chem. 55 (1983) 1710-1712.
[16] R.H. Lilien, H. Farid, B.R. Donald, Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum, J. Comput. Biol. 10 (2003) 925-946.
[17] J. Ye, T. Li, T. Xiong, R. Janardan, Using uncorrelated discriminant analysis for tissue classification with gene expression data, IEEE/ACM Trans. Comput. Biol. Bioinformatics 1 (2004) 181-190.
[18] P. Howland, H. Park, Generalizing discriminant analysis using the generalized singular value decomposition, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2004) 995-1006.
[19] P.W. Mielke Jr., K.J. Berry, in: Permutation methods: a distance function approach, Springer, New York, 2001.
[20] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. R. Statist. Soc. B, 36 (1974) 111-147.
[21] L. Nørgaard, R. Bro, PLS regression in the food industry: a study of N-PLS regression and variable selection for improving prediction errors and interpretation, in: M. Tenenhaus, A. Morineau (Eds.), Les Methode PLS. Symposium International PLS'99, Cisia-Ceresta, 1999, pp. 187-202.
[22] M.T. Moran, J.P. Schofield, A.R. Hayman, G.P. Shi, E. Young, T.M. Cox, Pathologic gene expression in Gaucher disease: up-regulation of cysteine proteinases including osteoclastic cathepsin K, Blood 96 (2000) 1969-1978.