Statistical data processing in clinical proteomics
Berkenbos-Smit, S.

Citation for published version (APA):

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: http://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, Singel 425, 1012 WP Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (http://dare.uva.nl)
Chapter 3

Assessing the statistical validity of proteomics based biomarkers†

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 preceding a biochemical validation of any biomarker. These tools consist of permutation tests, single and double cross validation. The cross validation steps can conveniently be combined with a new variable selection method, called Rank Products. The strategy is especially suited for the undersampled 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 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 undersampling.

3.1 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 analytical 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 differing patterns in the protein profiles. 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 chapter 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 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 could then be 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 proceed with the biochemical validation of the 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 estimate 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 and Simon et al. have shown, cross validation only gives a reliable error rate when the complete modelling procedure is cross validated. Leaving out parts of the procedure during cross validation 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 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. 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. Also two specific Gaucher cell markers are known: chitotriosidase and CCL18. Chitotriosidase shows a thousandfold increased activity in serum of symptomatic Gaucher patients. Plasma CCL18 levels are elevated ten to fiftyfold in symptomatic Gaucher patients. 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. 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.

3.2 Data set

The objects of the data set consist of serum protein profiles of 19 Gaucher patients (10 males and 9 females; 15-65 years old at the initiation of therapy) and 20 controls (7 male and 13 female healthy volunteers). All patients with Gaucher disease (type I) studied were known to the Academic Medical Centre (Amsterdam, The Netherlands). All patients received either enzyme replacement or substrate reduction therapy. Serum samples were obtained before initiation of therapy. Approval was obtained from the local Ethics Committee. Informed consent was provided according to the Declaration of Helsinki. Serum samples were surveyed for basic proteins with SELDI-TOF-MS making use of the anionic surface of CM10 ProteinChip®. The resulting protein profiles are mass spectra composed of the mass to charge ratios (m/z) and the intensities of the desorbed (poly)peptide ions. The control and Gaucher samples were randomly assigned to different spots and different chips. All preprocessing (spot-to-spot calibration, baseline subtraction, peak detection) of the SELDI-TOF-MS data was performed using Ciphergen software. An example of the resulting spectra can be found in Figure 3.1.
3.3 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 the 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

$$R = \frac{d' Bd}{d' W d}$$

(3.1)

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 corresponding to the largest eigenvalue of $W^{-1} B$.\textsuperscript{111} Because there are more m/z values than samples, the matrix $W$ is singular. This means that $W^{-1}$ 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 meta-parameter the value of which can be decided upon using cross validation, which is described in section Cross validation. The combination of FLDA with PCA yields Principal Component Discriminant analysis (PCDA).\textsuperscript{30–32,112}

Permutation test

Once a PCDA model is found that discriminates between the healthy and diseased groups, what can then 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.113

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 size 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 data set 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 data set on which the model is built is only one subset from the entire control and diseased population, hence the model and corresponding prediction error are one possible outcome. Another subset would 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 tenfold 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 that is built 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.
3.3 Methods

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 that leads to the lowest cross validation error is selected. 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 be used to find a good estimate of the prediction error in a lightly altered procedure. 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. Double cross validation consists of two nested cross validation loops (Figure 3.2). 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
The statistical validity of proteomics biomarkers

**Figure 3.2:** Double cross validation. The original data set is split into a training set (outer train) and test set (outer test) ten times in the outer cross validation loop (Outer CV). In the inner loop the outer training set is split up nine times in a training set (inner train) and a test set (inner test). Every number of principal components (PCs) for the PCA step that is considered is used to build a model on the inner training set. This model then predicts the classes of the samples in the inner test set, 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 outer training set for the model in the outer loop. The data in the outer test set is predicted with this model to give an error. The errors made in the ten different outer test sets are combined in the prediction error.

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.

**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.
3.4 Results

In the tenfold 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. 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 estimated with double cross validation.

3.4 Results

Data

Serum samples of controls and Gaucher patients were measured with SELDI-TOF-MS. 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. This resulted in a model that discriminated perfectly between the Gaucher and control groups: all samples were assigned to the correct class. Hence, the resubstitution error, the 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 Figure 3.3 A. 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, 4 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 tenfold 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. Figure 3.3 B 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 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.

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 ninefold 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 tenfold 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 (Figure 3.3 C). 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 permuted 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 tenfold 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.\textsuperscript{108,110} 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 than 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 for this data set. With this number the tenfold 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 is
Table 3.1: Top ten best discriminating m/z values and their Rank Products (RP) according to the Rank Products method.

<table>
<thead>
<tr>
<th>m/z</th>
<th>RP</th>
</tr>
</thead>
<tbody>
<tr>
<td>4058.0</td>
<td>36</td>
</tr>
<tr>
<td>5852.6</td>
<td>288</td>
</tr>
<tr>
<td>3685.4</td>
<td>115 \cdot 10^3</td>
</tr>
<tr>
<td>4546.0</td>
<td>435 \cdot 10^3</td>
</tr>
<tr>
<td>2067.9</td>
<td>292 \cdot 10^5</td>
</tr>
<tr>
<td>4214.8</td>
<td>113 \cdot 10^6</td>
</tr>
<tr>
<td>3840.1</td>
<td>136 \cdot 10^7</td>
</tr>
<tr>
<td>1008.2</td>
<td>228 \cdot 10^7</td>
</tr>
<tr>
<td>4016.2</td>
<td>503 \cdot 10^8</td>
</tr>
<tr>
<td>8949.4</td>
<td>781 \cdot 10^8</td>
</tr>
</tbody>
</table>

\( \left( \frac{590}{2} \right)^{10} = 5 \cdot 10^{24} \). Table 3.1 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 focused on relatively small proteins (molecular masses below 10,000 Da). It is known that various proteases, particularly cathepsins, are elevated in Gaucher plasma.\(^{114}\) 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.

Another question is how many m/z values with low rank product would have to be selected for a good predictive model. Figure 3.4 shows how the classification error rate depends on the number of m/z values selected for the model. The error rates in Figure 3.4 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 Figure 3.4 shows, incorporating 10 m/z
Figure 3.4: Error rate vs. 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.

values or less resulted in error rates of 8 out of 39 and higher. The lowest prediction error was achieved when 210 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.

3.5 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 discrimi-
nation 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 was 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 a single cross validation then suffices.