Statistical data processing in clinical proteomics

Suzanne Smit

Statistical data processing in clinical proteomics

Suzanne Smit

Statistical data processing in clinical proteomics

Suzanne Smit

Statistical data processing in clinical proteomics

ACADEMISCH PROEFSCHRIFT

ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. dr. D.C. van den Boom
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Agnietenkapel

op dinsdag 22 september 2009, te 10.00 uur

door

Suzanne Smit
geboren te Opperdoes

Promotiecommissie:

Promotores:

- prof. dr. A.K. Smilde
- prof. dr. C.G. de Koster

Copromotor:

- dr. ir. H.C.J. Hoefsloot

Overige leden:

- prof. dr. J.M.F.G. Aerts
- prof. dr. R.P.H. Bischoff
- prof. dr. A.H.C. van Kampen
- prof. dr. F.A. Wijburg
- prof. dr. P.H.C. Eilers

Faculteit der Natuurwetenschappen, Wiskunde en Informatica

The research reported in this thesis was carried out at the Swammerdam Institute for Life Sciences, Faculty of Science, Universiteit van Amsterdam. The publication of this thesis was made possible by the Netherlands Bioinformatics Centre (NBIC).

Voor Arjen en Fiona

Contents

1 Introduction **1**

2 Statistical data processing in clinical proteomics **3**
2.1 Introduction 5
2.2 Feature selection 7
2.3 Classification methods 10
2.4 Biomarker candidate selection 15
2.5 Comparison studies 16
2.6 Statistical validation 17
2.7 Proteomics data analysis: a framework 25
2.8 Black spots and open issues 26
2.9 Conclusions 27

3 Assessing the statistical validity of proteomics based biomarkers **29**
3.1 Introduction 30
3.2 Data set 32
3.3 Methods 33
3.4 Results 37
3.5 Conclusion 42

4 Limited value of serum protein profiling for discrimination of patients suffering from Fabry disease **45**
4.1 Introduction 46
4.2 Data set 47
4.3 Results 50
4.4 Discussion 52

5 Discriminating healthy from diseased. A classification strategy for clinical proteomics data **55**
5.1 Introduction 56
5.2 Materials and Methods 58
5.3 Results and discussion 63
5.4 Conclusions 68

6 Optimal use of paired proteomics data **71**
6.1 Introduction 72
6.2 Data set 73
6.3 Data analysis 75
6.4 Results 78
6.5 Conclusions 79

7 Enhancing classification performance: covariance matters **81**
7.1 Introduction 82
7.2 Data set 82
7.3 Classification methods: PCDA, SVM, SIMCA 83
7.4 Results 84
7.5 Conclusions 92

Outlook **93**

Bibliography **95**

Publications **105**

Summary **107**

Samenvatting **111**

Dankwoord **115**

Chapter 1

Introduction

Proteins play important roles in cells and organisms. As well as being part of the immune system proteins transport substances through the body and catalyse chemical reactions in the cell. The protein content of a cell depends on the function of the cell. It can change in response to (outside) influences, for example illness. On the other hand, changes in proteins can also cause disease. This means that if it is possible to measure such a change in a person with a certain disease we may learn something about the disease. We may also be able to use knowledge about the change in protein composition in diagnosing the disease. Often it is unknown which proteins might be involved. The research is then not aimed at a specific protein, but at many proteins at the same time. This is the domain of clinical proteomics.

Proteomics is the study of the proteome, which in its widest definition includes all proteins that are expressed in an organism. In practice it is not possible to measure all proteins, but with modern techniques it is possible to measure many proteins simultaneously. With for example mass spectrometry it is possible to analyse clinical samples (blood, urine, tissue) from patients and healthy controls. This results in intensities for many proteins for each sample, which is called the protein profile of the sample. The next step is to find differences between the protein profiles of groups of patients and controls. These differences are potential biomarker leads. Occasionally there may be an obvious difference: one protein that is present in patients but not in controls or one protein that is clearly underexpressed in patients. Often the differences are much more subtle and data analysis methods are needed to uncover them. The analysis of clinical proteomics data is the subject of this thesis.

Chapter 2 is an introduction to statistical analysis of clinical proteomics data.

In this chapter data analysis strategies for the discovery of biomarkers in clinical proteomics are reviewed. An overview of some widely used variable selection methods and classification methods is given. We present a framework in which most of the methods fall.

With the use of data mining methods comes the issue of statistical validation: How can we analyse the data in such a way that information of the statistical validity of the results is obtained? A strategy is put forward for a thorough statistical assessment of the entire data analysis procedure, combining permutation testing and cross validation. This strategy is tested in two case studies: the classification of SELDI-TOF-MS protein profiles of Gaucher patients and controls in Chapter 3 and of Fabry patients and controls in Chapter 4. We also use the validation protocol for assessing different statistical classification methods in Chapter 5.

The second part of the thesis gives two examples of how tailoring the data analysis to the structure of the data can enhance the performance. Proteomics studies are sometimes designed to compare samples from one patient, for example healthy and diseased tissue from the same organ or blood samples before and after treatment. This design results in a data set with a paired nature. When one variable per sample is measured, applying a paired test makes it easier to discover a difference. We considered whether applying a paired analysis to multivariate paired data would have the same effect. In Chapter 6 we present a classification approach that explicitly uses pairing of samples in a cervical cancer proteomics data set, obtaining a higher classification performance compared to ignoring the paired structure of the data.

Finally, we study the properties of some classification methods themselves, more specifically their behaviour with respect to covariances. In Chapter 7 we show an example of a data set that two common methods (Principal Component Analysis followed by Linear Discriminant Analysis (PCDA) and Support Vector Machines (SVM) perform poorly on, while Soft Independent Modelling of Class Analogy (SIMCA) performs much better. The data set consists of serum protein profiles of recovering and relapsing cervical cancer patients. The characteristics of this data set cause PCDA and SVM to fail where SIMCA can be successful, exemplifying that selecting a classification method that suits the data structure can improve results.

Chapter 2

Statistical data processing in clinical proteomics[†]

This chapter reviews data analysis strategies for the discovery of biomarkers in clinical proteomics. Proteomics studies produce large amounts of data, characterized by few samples of which many variables are measured. A wealth of classification methods exists for extracting information from the data. Feature selection plays an important role in reducing the dimensionality of the data prior to classification and in discovering biomarker leads. The question which classification strategy works best is yet unanswered.

Validation is a crucial step for biomarker leads towards clinical use. Here we only discuss statistical validation, recognizing that biological and clinical validation is of utmost importance. First, there is the need for validated model selection to develop a generalized classifier that predicts new samples correctly. A cross validation loop that is wrapped around the model development procedure assesses the performance using unseen data. The significance of the model should be tested; we use randomisations of the data for comparison with uninformative data. This procedure also tests the correctness of the performance validation. Preferably, a new set of samples is measured to test the classifier and rule out results specific for a machine, analyst, laboratory or the first set of samples. This is not yet standard practice.

We present a modular framework that combines feature selection, classification, biomarker discovery and statistical validation; these data analysis aspects are all discussed in this chapter. The feature selection, classification and biomarker discovery modules can be incorporated or omitted to suit the data

[†]This chapter is based on S. Smit, H.C.J. Hoefsloot, A.K. Smilde, *J. Chromatogr. B* **2008**, *866*, 77. DOI:10.1016/j.jchromb.2007.10.042

analysis problem and the preference of the researcher. The validation modules are an integral part of the data analysis that ensures its quality. In each module, the researcher can select from a wide range of methods, since there is not one unique way that leads to the correct model and proper validation. We discuss many possibilities for feature selection, classification and biomarker discovery. For validation we advice a combination of cross validation and permutation testing, a validation strategy supported in the literature.

2.1 Introduction

Modern developments in analytical techniques such as mass spectrometry (MS) make it possible to measure protein concentrations on a large scale; this area of research is called proteomics. The hope is that proteomics studies can contribute to healthcare. In clinical proteomics thousands of proteins or peptides can be measured in a single experiment. This chapter describes how information is obtained from preprocessed clinical proteomics data and how to validate the information using statistical procedures. The clinical proteomics experiments that we discuss in this paper can be seen as a discovery tool for biomarkers. A possible workflow for biomarker discovery is given in Figure 2.1. It starts with a biological question, which leads to a carefully designed experiment, sampling and measurements. Preprocessing of the data is necessary to remove instrumental noise and make the measurements comparable. The result is a data matrix consisting of N objects (samples) and m variables or features which is used in the subsequent data analysis. A preliminary answer to the biological question is obtained in the three blocks that are encircled in Figure 2.1: *Data processing*, *Biomarker pattern*, and *Statistical validation*. After the discovery of statistically valid biomarker leads, external testing and biological validation will show whether they truly answer the biological question.

Biomarkers can be used to predict the state of a patient, in diagnosis, to monitor the response to treatment, and to determine the stage of a disease. In the search for diagnostic markers, but not essentially different for the other goals, samples from cases and controls are measured. The measurements are usually stored in a data matrix and class labels are stored in a response vector. Data analysis tools try to find the differences in measurements that predict the state of a patient. This information is preferably in just a few proteins (biomarkers) that are indicative for the biological state. Alternatively, the interplay of multivariate data can provide the desired information. Results should be subjected to validation: statistical as well as biological. The statistical validation should investigate the performance of the biomarker, as well as the possibility of a chance result. The biological validation is concerned with the question whether the biomarkers are involved in processes that can be related to the disease. If the result of both validation processes is satisfactory a putative biomarker is established. Many more steps have to be taken before this leads to an established biomarker.[1]

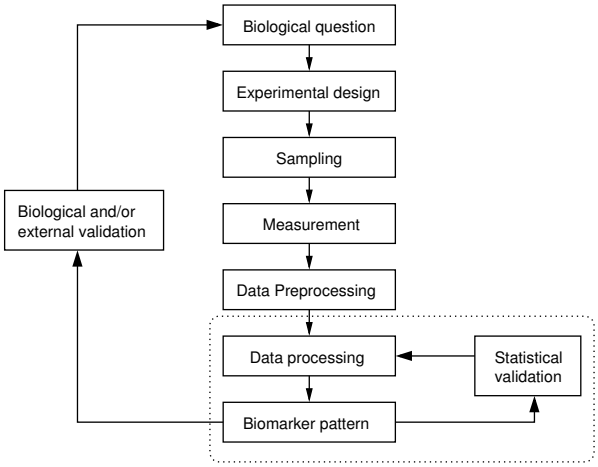

Figure 2.1: Biomarker discovery workflow. From biological question to biomarker leads. The blocks *Data processing*, *Biomarker pattern* and *Statistical validation* form the subject of this chapter.

MS is not the only technique used for proteomics investigations. Protein arrays and 2D gels also play an important role in the field.[2] However, most of the literature on data analysis in clinical proteomics discusses MS studies. Reviews on the application of MS in proteomics are available;[3,4] this chapter does not discuss the many types of MS experiments. We restrict ourselves mainly to data analysis in single MS experiments (such as liquid chromatography-MS, matrix assisted laser desorption/ionisation MS and surface enhanced laser desorption/ionisation) although our conclusions also hold for other types of (omics) experiments. In single MS experiments many different issues play a role. Among these are experimental design, selection of patients, sample handling, preprocessing of the spectra and biological validation.[5–12] We are not taking up these issues here but we focus on classification methods for proteomics studies and the statistical validation tools that are used in combination with the classification methods.

Classification methods applied in proteomics are developed in different sciences, such as machine learning, chemometrics, data mining and statistics. A wide range of methods is available, with many different characteristics. We try to give an overview of the methods that are popular in proteomics. The reason that validation in classification methods is an important and still open issue is mainly caused by the characteristics of a proteomics data set. Usu-

ally, a mass spectrum contains thousands of different mass/charge (m/z) ratios. The sample size, e.g. the number of patients, is relatively small. This results in a so-called high dimensionality small sample problem. This type of problem suffers from the curse of dimensionality,[13] which means that the number of samples needed to accurately describe a (discrimination) problem increases exponentially with the number of dimensions (variables). In proteomics studies, the number of samples is usually low compared to the number of variables, due to the limited availability or the cost of measurements. This undersampling leads to the possibility of discovering a discriminating pattern between two populations, even when these two populations are statistically not distinct. Working with high dimensional data can easily lead to overfitting: the derived model is specific for the training data and does not perform well on new samples.

Literature provides several approaches to overcome these problems. One approach is to reduce the dimensionality of the data. This can be done before a classification is performed or it can be combined with a classifier. Other techniques to cope with high dimensional data are statistical validation strategies, such as cross validation and permutation tests.

This chapter starts with an overview of the most encountered methods for classification and biomarker discovery in clinical proteomics. We present a framework in which most of the methods fall. And finally a strategy is put forward for a thorough statistical assessment of the entire data analysis procedure.

2.2 Feature selection

Feature selection plays an important role in clinical data analysis for three reasons. First, using all features in forming the classification rule in general does not give the best performance. Increasing the number of features from zero enhances performance to some point, after which adding more feature leads to a deteriorating performance, because many features are uninformative and they can conceal information in relevant features. This is called the peaking phenomenon.[14–16] The second reason is a technical one: some classification methods require the number of objects to be larger or equal to the number of features. Since proteomics data sets usually consist of far more features than samples, a selection has to be made before constructing the classification

rule. Third, one of the goals of a proteomics study is to find leads for potential markers for disease. Hence, the number of variables in the final model should be small to enhance the interpretability of the model. To this end, finding a good classifier is combined with selection of discriminating variables.

We distinguish different categories of feature selection methods. Filter methods and variable transformation reduce the number of features independent of a classification method (unsupervised), while wrappers select variables in concert with a classification method (supervised). Sometimes, feature selection is intrinsic to a classification method, for example in classification trees. Another category is variable selection after classification, where the information in the classification rule is used to find the most informative variables.

Filters, variable transformation and wrappers are discussed in this section, and section 2.4 describes variable selection intrinsic to classification and after classification. This division reflects that wrappers, filters and variable transformation are mostly used to deal with the peaking phenomenon and to solve the technical issues, while leads for biomarkers are often sought in the classification rule.

We realize that this is by no means a strict distinction. Wrapper[17] and filter[18] methods have also been used for biomarker selection, and vice versa: some intrinsic methods are used for preselection-selection to provide input for other classification methods.[19,20] We would like to point out that statistical validation is as important in variable selection as it is throughout the entire data analysis. In undersampled data sets, with fewer samples than variables, it may very well be possible to select a set of features that discriminate between cases and controls, which turns out to be uninformative when new samples are classified. Thorough statistical validation can prevent overfitting, and we discuss it in section 2.6.

Independent feature selection

Filter methods are applied to the preprocessed data before the construction of the classifier. Examples are significance tests such as the t-test, which compares differences in means between the case and the control groups. When the measurements for a variable differ significantly between the two groups, it is retained. The t-test assumes normality of the data. The Wilcoxon-Mann-Whitney test assesses differences between two groups without making this

assumption.

These significance tests are designed to deal with univariate data, and a variable is considered to differ significantly when its test statistic is smaller than some value, α (generally, $\alpha = 0.05$ or $\alpha = 0.01$). Since proteomic analysis involves testing many individual variables simultaneously, applying the same value for α leads to many false positives.[21] The Bonferroni correction sets an α-value for the entire set, so that the test statistic for each individual variable is compared to a value of α/(number of variables) and the false positive rate or family wise error rate (FWER) is controlled.

A less conservative correction for multiple testing is controlling the false discovery rate (FDR): the number of false positives among all positives.[22,23] Significance Analysis of Microarrays (SAM) uses a t-test with a threshold to select features. The false discovery rate is obtained by comparing the results with results in permutations.[24]

Like filtering methods, variable transformation is performed before classification. Projection methods reduce the dimensionality of the data in a multivariate approach. Principal Component Analysis (PCA) looks for linear combinations of the original variables that describe the largest amount of variation in the data.[25] The linear combinations (principal components) become new features that describe the data in a lower dimensional space.

Wrappers

Wrappers are feature selection methods that work in concert with a classification method. The classification method is used to test relevance of the variables. Variables that lead to good performance are selected. Forward selection starts with an empty set and selects the variable that gives the best classification result. Given this first variable, another variable is added that realizes the largest improvement of performance.[13] Variables are added until the performance does not improve or a set criterion is met. Backward elimination works similarly, starting with the full set of features and sequentially removing features from the set.[13] Genetic algorithms create many feature sets that are tested simultaneously for performance, given a classification method. The best sets are recombined to create a new generation of improved feature sets. The algorithm is stopped when the performance does not improve over several generations or when a preset performance measure is achieved.[26]

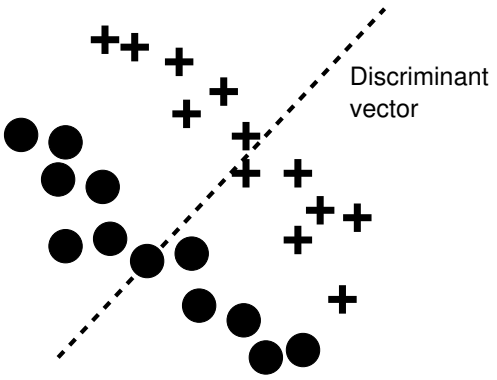

Figure 2.2: Linear discriminant analysis.

2.3 Classification methods

Discriminant Analysis

Discriminant analysis (DA) was first introduced by Fisher, who used it to discriminate between different Iris species.[27] In the feature space, a direction is sought that maximizes the differences between the classes with respect to the covariance within the control and case classes (Figure 2.2). This direction, the discriminant vector, can be used to classify new samples. DA uses the covariance matrix to find the discriminant vector. Linear Discriminant Analysis (LDA) assumes the within-class covariance matrices to be equal, which leads to linear decision boundaries. When the covariance matrices are unequal, Quadratic Discriminant Analysis (QDA) is applied. The decision boundary in QDA is quadratic.

Usually, in proteomics data, undersampling causes the within-class covariance matrix to be singular, which makes it impossible to find the discriminant vector. This can be solved by filtering features[19] or by selecting features with a wrapper method as described in the previous section. Other solutions lie in adjusting the DA algorithm to repair the singularity of the covariance matrix. Regularized Discriminant Analysis (RDA)[28] shrinks the covariance matrix towards a multiple of the identity matrix. In Diagonal Discriminant Analysis the covariance matrix is assumed to be diagonal, setting all off-diagonal elements to zero (see for example[29]).

A popular variant of DA in omics studies is Principal Component Discriminant Analysis (PCDA).[30] It solves the singularity by reducing the dimensionality of the data with PCA, after which DA is performed on the PCA scores. PCDA has been used for omics data analysis under a variety of names. As uncorrelated discriminant analysis, Ye *et al.* used it for the analysis of several publicly available gene expression data sets.[31] The maximum number of principal components is used in the classifier. In a proteomics study of SELDI-TOF-MS data concerning ovarian cancer and prostate cancer, Lilien *et al.* used the Q5 algorithm, also a combination of PCA and LDA to discriminate healthy from diseased.[32] Again, the maximum number of principal components is retained. The classification probability is calculated from the distance on the discriminant vector between the spectrum and the nearest class mean. Spectra with classification probabilities smaller than a threshold are not classified. Smit *et al.* applied PCDA to SELDI-TOF-MS measurements of serum to discriminate Gaucher patients from healthy controls.[33] The number of components was tuned with cross validation, showing that the maximum number of components does not always lead to the best model.

Partial Least Squares

Partial Least Squares (PLS)[34] is similar to PCA, but in extracting the new features, PLS also takes the covariance of the data with the response vector (vector of class labels) into account. PLS tries to find the relations between the data matrix and the vector of class labels, it is a latent variable approach to modelling the covariance structure of the data and the class labels. A PLS model will try to find the multidimensional direction in the space of the data matrix that explains the maximum variance in the class label space. When it is used for classification, it is referred to as partial least squares discriminant analysis (PLSDA).[35]

PLSDA is a much used method in metabolomics studies. It has for example been applied in a human metabolomics study into obesity to differentiate between obese and lean individuals.[36] In a proteomics dementia data set, Gottfries *et al.* employed PLSDA for discrimination between different classes of dementia and healthy individuals.[37] More examples of PLSDA applications in clinical metabolomics studies can be found in an overview by Trygg *et al.*[38]

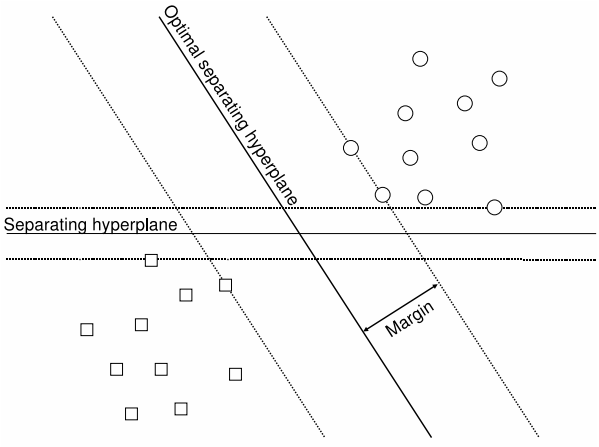

Figure 2.3: The optimal separating hyperplane separates the classes with the widest margin.

Support Vector Machines

The support vector classifier constructs a hyperplane that separates two classes. When the classes are linearly separable, the optimal hyperplane maximizes the distance from the closest objects to the hyperplane, as is shown in Figure 2.3. This distance is called the margin. The class assignment of new samples depends on which side of the hyperplane they are. In the case that the classes are not perfectly separable, some objects will be on the wrong side of the hyperplane (misclassification). The amount to which objects are allowed to be on the wrong side of the hyperplane is bound by a penalty. A high value for the penalty means it is very costly to cross the hyperplane. Consequently, in the original feature space the boundary will be wiggly to accommodate all samples; this may result in overfit. Small values can lead to hyperplanes that are not very effective in separating the classes.[13,39]

In Support Vector Machines (SVM), the data are transformed to a larger feature space. This makes it possible to accommodate discrimination problems for which a linear decision boundary is inappropriate. A nonlinear transformation of the data can be chosen in such a way that the classes are (almost) separable by a hyperplane in the higher dimensional feature space. The linear separation in the high dimensional feature space translates to a nonlinear decision boundary in the original feature space. The new, higher dimensional

feature space does not have to be considered explicitly, the hyperplane can be computed using a kernel function. There are many possibilities for transforming the data, which makes SVM a versatile method.[39] The same data transformations could also be coupled to other classifiers, such as PCDA and PLSDA.

The SVM methodology is a popular method for classification in clinical proteomics. Among recent applications are studies of tuberculosis,[40] ovarian and prostate cancer,[41] response to therapy in rectal cancer patients,[42] heart failure,[43] and breast cancer.[44]

Logistic Regression

The odds is defined as the ratio of the probability of a sample being a member of one class to the probability that the sample is outside that class. Logistic Regression models use linear regression to fit the data to the natural logarithm of the odds. It ensures that the probabilities are between zero and one and that they sum to one. Logistic Regression is similar to LDA, but it makes fewer assumptions about the underlying distributions. Like in DA, the large number of variables in proteomics data constitutes a problem, which can be tackled in several ways. Variable selection prior to modelling was used by Bhattacharyya *et al.* in a proteomics study of pancreatic cancer[45] and by Zhu *et al.* on microarray data in three cancer diagnosis data sets.[46] Others have combined PLS with logistic regression.[47,48] In Penalized Logistic Regression, a penalty is set on the regression coefficients. As a result, some coefficients become zero, which effectively reduces the number of features.[49,50]

Nearest Shrunken Centroids

In nearest centroid classification, a sample is assigned to the class with the nearest class mean. To accommodate classification of gene expression data, Tibshirani *et al.* developed the Nearest Shrunken Centroids (NSC) method.[51] It shrinks the class centroids towards the overall centroid, thereby selecting genes. NSC, like Diagonal Discriminant Analysis assumes a diagonal within-class covariance matrix. Tibshirani employed NSC for the discrimination of different cancer types. To predict the tissue of origin of 60 cancer cell lines, Shankavaram applied NSC to gene expression profiles.[52] In a proteomics

study of kidney patients with and without proteinuria, Kemperman *et al.* selected discriminating proteins using NSC.[53]

Artificial Neural Networks

Artificial Neural Networks (ANN) refers to a class of nonlinear modelling methods. Three parts can be discerned in an ANN: the neurons in the input layer (data), neurons in one or more hidden layers, and the output layer neurons (predicted responses). The neurons in the hidden layer are formed by basis transformations of the input. The parameters of the basis transformations are learnt from the data, as are the weights assigned to the hidden neurons to create the output.[13] Bloom applied ANN for the detection of the tissue of origin of adenocarcinomas, which were analyzed by 2D gel electrophoresis.[54] Other applications are prediction in breast cancer[55] and kidney disease.[56]

Classification Trees

A Classification Tree algorithm recursively splits the data in a parent node into two subsets called child nodes. The decision for the split is based on the value for one protein. The aim is to maximize homogeneity in the child nodes and the protein that gives the largest decrease in heterogeneity is chosen. The child nodes then become parent nodes and new variables are selected to split these nodes in turn. This process continues until all variables have been used or all terminal nodes are homogeneous. The last step is pruning of the tree to avoid overfitting. Several measures of heterogeneity are employed in different tree algorithms.[13] Some applications of decision trees in proteomics are clinical studies of pancreatic cancer,[45] clinical behaviour after treatment in leukaemia patients,[57] and ectopic pregnancy.[58] In this last study, Gerton *et al.* first built two trees to optimize separately for sensitivity and specificity, which they then combined to form one classification model.

Ensemble classifiers

Ensemble classifiers are formed by combining several single classification rules (base classifiers), with the goal to construct a predictor with superior

performance. A new sample is classified by all individual classifiers and the ensemble prediction can be made by majority voting. The ensemble method is successful when each individual rule makes correct prediction for more than half of the samples and if the rules are diverse (give independent predictions).[59]

Different types of ensemble methods exist. Using several different classification methods to construct the base classifiers is one way to create diverse rules.[60] Alternatively, the rules can all be constructed with the same classification method, for example ANN.[61] Diversity of the rules can then be introduced by resampling the subjects with cross validation,[62] bootstrapping,[61,63–65] and boosting.[66,67] A combination of bagging and boosting is used by Dettling in BagBoosting, where in each boosting step a bagged classifier is constructed.[68] Alternatively, resampling of the variables also leads to diverse base classifiers.[69–72] After construction of the base classifiers, their diversity can be evaluated by comparing their predictions[60,69] or the structure of the individual classifiers.[63] The final step is the combination of the base classifiers to arrive at one prediction for a sample. Several fusion methods exist,[73] of which weighted voting and majority voting are most applied.[60,62,64]

A well known ensemble classifier is the Classification Forest. The Classification Forest is an extension of the Classification Tree, where multiple trees are constructed and used in an ensemble to predict new samples. Examples of forest classifiers are the Random Forest (RF),[74–76] and the Decision Forest.[77,78]

2.4 Biomarker candidate selection

With biomarker candidate selection we refer to feature selection with the aim to discover which proteins are promising leads for biomarkers. We place this module after the classification methods, because the classification rules contain information about the contribution of each variable to the classification. This information reveals the proteins of interest, which may prove to be biomarkers. Two methods that determine the interesting variables directly are the classification tree,[13] which classifies samples based on their values for a small number of proteins and the NSC algorithm, which as a by-product of constructing a classification rule selects variables.[51]

Other classification methods carry relevant information about the variables in

the form of weights and regression coefficients (linear SVM, DA). This information is used in many applications to select relevant sets of proteins. Guyon developed Recursive Feature Elimination (RFE), a backward feature selection method, which eliminates the feature with the smallest weight in a linear SVM rule.[17] Rank products was initially designed for gene selection using gene expression differences between two groups directly,[79] but it has also been employed for selection of proteins using a PCDA classification rule.[33]

Bijlsma used a threshold on the regression coefficients in PLSDA to select potential metabolite biomarkers.[36] Another feature extraction method for PLSDA is Variable Importance in the Projection (VIP). The VIP value of a variable reflects its importance in the model with respect to the response vector as well as to the projected data.[80] It has been used in the selection of metabolites in studies of liver function in Hepatitis B[81] and intestinal fistulas.[82] Variable selection in ensemble methods is perhaps less straightforward, due to the amount of information that comes from using multiple classification rules. The random forest algorithm estimates the importance of a variable by permuting the measurements for that variable, leaving the rest of the data intact and classifying new samples.[74] It is also possible to use the information from significance tests (t-test, Wilcoxon-Mann-Whitney test) to select disease markers, without running a classification algorithm.[18]

2.5 Comparison studies

Many more classification algorithms are available; the list of classifiers and variable selection methods we discuss is not exhaustive. The question arises which method is best suited for classification of proteomics data. It is hard to compare results from different studies because conditions vary. This is due to the fact that preprocessing, reporting of performance and validation schemes are not the same. There are some studies that describe performance of several classification methods applied to the same data set, with the aim to compare classifiers.

Liu *et al.* investigated six feature selection methods on leukaemia gene expression data and on ovarian cancer MS data.[83] After feature selection, four classifiers were applied to the reduced data. For the gene expression set Entropy Feature Selection, which selects the features based on their discriminatory power, came out first. A correlation based feature selection (this method

selects a subset of features that correlate with response but not with one another) led to the best performance in the ovarian cancer data. A special issue of Proteomics in 2003 covered the data analysis efforts of several research groups on one lung cancer data set.[84] Many strategies are applied in this issue to obtain a classifier. Due to the use of different validation schemes and different preprocessing it is very difficult to compare the performance. In a comparison study of simple DA classifiers with aggregated Classification Trees (as representative for more sophisticated machine learning approaches) on three gene expression data sets, Dudoit *et al.* found that the DA methods performed very well.[29] Wagner compared several linear and nonlinear DA methods and a linear SVM for classification of prostate cancer MS data.[85] Although the performances of the methods were comparable, the linear DA and linear SVM performed slightly better than nonlinear DA methods. Wu *et al.* combined two feature selection methods and several classification algorithms to classify ovarian cancer MS data.[19] They concluded that RF outperformed the other methods (among which SVM, DA, bagged and boosted classification trees), but their conclusion was mainly based on the results after feature selection with RF. Feature selection based on the t-statistic resulted in superior performance of SVM and LDA, closely followed by RF. For classification of MS data of Gaucher disease, Hendriks *et al.* applied six classification methods.[67] The most successful were SVM, Penalized Logistic Regression and PCDA.

These comparison studies show there is no consensus about the best classifier. This is due to the fact that different data sets have different characteristics and therefore no classifier will have a high performance for all data sets. The performance not only depends on the data but also on the feature selection step and on the individual experience of the data analyst.[86] Experience with a method is likely to give better results. We have found no set of guidelines for selecting a classifier.

2.6 Statistical validation

The next step towards clinical utility is validation. First, the results of a preliminary clinical proteomics study should be subjected to thorough statistical assessment. Next, a new set of samples should be measured independently in time and/or place from the first data set to test the classifier. If the preliminary results warrant the investment, the following step would be identification of the relevant proteins to determine biological validity. In this section

we describe two tools, permutation tests and cross validation, to assess the statistical validity of the classifier, based on the preliminary data set only. An overview of validation strategies in proteomics literature is given. We start by discussing different performance measures that are used in clinical proteomics.

Performance measures

The performance of a classifier in clinical applications is usually given in two measures. The sensitivity is the fraction of cases that are classified as cases. The specificity is the fraction of controls that is correctly identified. The sensitivity and specificity can take values between zero and 1, where zero means all samples in that class are misclassified and 1 means that they are all correctly identified. They are both reported, because they each show a different characteristic of the classifier and can be very different.[87] The sensitivity and specificity can be altered by shifting the threshold for assignment to the case or control class. This may lead to a classifier with more desirable characteristics, such as a higher sensitivity, usually at the cost of specificity. The sensitivity and specificity can be plotted together in a receiver operating characteristic (ROC) curve. An example of an ROC curve is given in Figure 2.4. The sensitivity is plotted on the y-axis and the x-axis represents the false positive fraction ($1 -$ specificity). The lower left corner represents the case where all controls are correctly classified (specificity equals 1), but all the cases are classified as controls (sensitivity is zero). The opposite case occurs in the upper right corner, where the sensitivity is 1 and the specificity is zero. Both corners are always part of the ROC curve. In between, the sensitivity and false positive fractions for different values of the threshold are plotted. Ideally, the resulting curve would go from the lower left corner to the upper left corner and then to the upper right corner. This represents a classifier that is able to distinguish perfectly between cases and controls for some value for the threshold. The information in an ROC plot is summarized by the area under the curve (AUC). The AUC of a perfect classifier is 1, whereas an uninformative classifier has an AUC of 0.5.[21,87]

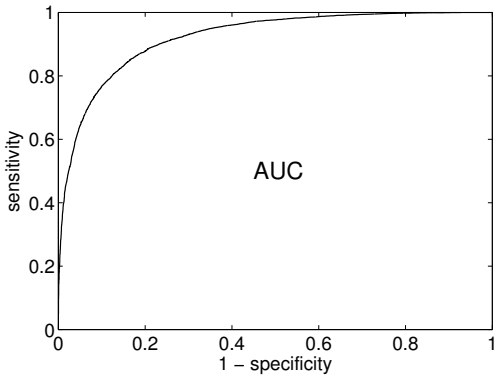

Figure 2.4: An ROC curve shows how the sensitivity and specificity of a certain classifier are connected. Changing the decision boundary influences the sensitivity and specificity, improving one of these at the expense of the other.

Cross validation for performance estimation

A classifier is trained on a limited data set at some point in time with the objective to correctly classify samples that will be measured in the future. At the time of construction, it is not possible to foresee how well a classifier will perform on newly acquired samples, because the samples are not yet available. Therefore, the performance is estimated on data that is available. Nevertheless, the performance estimate should be based on an unseen set of samples, which are not in any way used in creating the classifier. If the performance is estimated using samples that have somehow been used in the modelling procedure, the estimate will be overly optimistic.[13] A second requirement of the performance estimate is that it should take into account the variability of the classifier. The data set from which the parameters of the classifier are estimated is a sample from the entire population and therefore this classifier is one possible realization. Other samples from the same population would result in different parameter estimates. The variability of the classifier should be reflected in the performance estimator.

Both requirements are met in cross validation. Cross validation makes efficient use of the available data, which is especially helpful in small data sets. The general idea is to split the data into several approximately equal-size parts. Each part is masked in turn (test set), while the remaining parts combined are used to train the classifier (training set). The classifier is then

applied to the masked set for prediction. This is repeated until all parts have been masked once, and then the error made in the blinded test sets is combined to give an independent estimate of the performance of the classifier. Because the training sets are different in each repetition, the cross validated performance estimate incorporates the variability of the classifier.

There exist different variants of cross validation. When the test set is made up of one sample it is called leave-one-out (LOO) cross validation. In k-fold cross validation, the data are divided in k parts. If k equals the number of samples it is leave-one-out cross validation. A variant of k-fold cross validation is leave-multiple-out cross validation, where repetitions are allowed in the test sets.[88] Often, the ratio of the class sizes is preserved in the training and test sets, making them accurate representations of the original data. This is called stratified cross validation.[33,89,90]

Cross validation for meta-parameters and feature selection

Many of the classification methods described in the previous section require the optimization of model tuning parameters. For example, in PCDA and PLSDA, the number of retained latent variables should not be too low, because valuable information would be discarded. On the other hand, incorporating too many latent variables means uninformative noise is incorporated in the model. Care has to be taken to avoid overfitting of the model to the available data, as the data are typically highly undersampled. The choice of the tuning parameters should be such that the generalization error of the resulting model (the error made in new samples) is low. This is also true for the selection of (a subset of) proteins for prediction. The selection should not only give good predictions for the available data, but also on newly acquired data. The tuning parameters and protein subset selection are called meta-parameters.

Cross validation is a much employed method to tune meta-parameters in proteomics, as well as in other 'omics' studies, chemometrics, and Quantitative Structure-Activity Relationship research. In this section we will borrow from research on cross validation in these fields and transfer relevant findings to clinical proteomics. For meta-parameter tuning, the cross validation procedure is repeated for different choices of the meta-parameter. The performances of classifiers with different values for the meta-parameters are compared to choose the parameter with the lowest cross validation error. Because

the test sets are not used in training the classifiers, overfitting of the model is prevented.

In the previous section we mentioned that cross validation reflects the variability of the classifier that is due to the data being a sample from a population. This is also of importance for the selection of a meta-parameter, since the goal is to construct a representative classifier. In LOO cross validation, the training sets are very similar to the full data set and to each other. This means that the classifiers constructed on the training sets will not vary much and there is still a risk of overfitting. K-fold cross validation introduces more variability, because the training sets are smaller and less similar.[13] This forces the selection procedure to recognize general patterns, rather than individual data points.[88] A good value for k depends on the data: with smaller values for k, the test sets are larger and the training sets in undersampled data sets may become too small for building meaningful models. Moreover, the bias inherent to cross validation increases with smaller values for k. This inherent bias results from the training sets being smaller than the full data.[13] Generally, 5 or tenfold cross validation is used.[91] There are many ways to split the data into different parts in k-fold cross validation. The estimate of the performance may depend on the choice of split.[89] Therefore, it is recommended to repeat the cross validation several times with different splits of the data. Kohavi and John let the number of repeats depend on the standard deviation of the performance estimate.[92] They repeat until the standard deviation becomes sufficiently small. This way, large data sets are cross validated fewer times than small ones, in which the variance will be higher. It saves computing time and it gives a criterion for the number of repeats of cross validation necessary.

Cross validation can be performed with restrictions. Baumann restricts the number of variables (proteins) or latent variables to be selected.[88] However, this requires a priori knowledge of the data. Kohavi and John implement a complexity penalty in their evaluation to favour smaller subsets of variables.[92]

Double cross validation for meta-parameter selection and performance estimation

When selecting a model with cross validation, the corresponding cross validation error is an inappropriate estimate of the prediction error of the model.

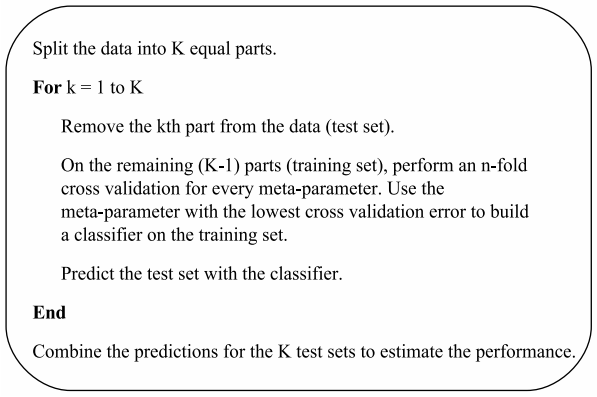

Figure 2.5: Pseudocode for double cross validation.

In that case the cross validation error is not based on an independent test set, because with the choice for a certain model, all of the data - the test samples as well as the training samples - is used. To solve this, Stone introduced the cross validatory paradigm: the cross validated choice of parameters requires cross validatory assessment to avoid overly optimistic performance estimates.[93] This means a nested cross validation scheme is needed to estimate the prediction error, where the parameter optimization is executed in an internal loop and the prediction error is estimated in an external loop on a completely independent set of samples. Pseudocode for this cross validation scheme is given in Figure 2.5. It is often called cross-model validation or double cross validation. For modelling procedures in which parameters are tuned in another way than with cross validation, for example by bootstrapping, all these training steps have to be taken into account in the validation of the performance.

Several researchers have investigated the extent of the bias of the cross validation error when not all model training steps are evaluated within the cross validation. Taking two microarray data sets as an example (using SVM with RFE), Ambroise and McLachlan showed that, while single cross validation suggests that the error rate was negligible, the test error was far from that.[94] Double cross validation error is a much better estimate of the performance. In addition, they calculated the single and double cross validation error rates for 20 permutations of the data. Although no information is present in the permuted data sets, the cross validation error that is obtained with the se-

lection of genes was almost zero. In contrast, double cross validation error estimates were much more realistic, between 40% and 45%. Similar results were reported by Simon *et al.*,[95] Varma *et al.*[96] and Smit *et al.*[33] The bias that is introduced in the performance estimate by ignoring the meta-parameter selection in the validation process is called the parameter selection bias. Double cross validation removes the parameter selection bias, but it does have the slight bias inherent to cross validation that is the result of the lower number of samples in the training set than in the full data set.[96]

It may seem a bit unclear what model is validated with double cross validation, because the internal loop returns different meta-parameters for different training sets.[97] This is very much the same as what we described for cross validation in the previous section. The variability of the classifier – in this case, the variability of the meta-parameters as well as the estimated parameters – is taken into account in estimating the performance with double cross validation.[33] Consequently, in double cross validation the entire model optimization procedure is validated.[96] The final classifier can be constructed in several ways. Stone chooses the tuning parameter with a cross validation and uses this parameter to build a model on the full data set.[93,97] Other possibilities are retaining all k classification rules from the double cross validation and use them together as an ensemble classifier for new samples or using the most frequently selected parameter in the internal loop on the full data set.[98]

Permutation test

In a permutation test the class labels are repeatedly removed and randomly reassigned to samples to create an uninformative data set of the same size as the data under study. One application of permutation tests is determining the relevance of a model. Building and testing a classifier on many permutations of the data gives a distribution of the performance found by chance, to which the performance of the classifier on the original data can be compared. The same classifier building protocol that is applied to the data is applied to the permutations, including any filtering or other selection of variables and parameter tuning.[88]

Permutation testing was already mentioned in the previous section where it appeared as a tool to investigate the bias of different cross validation methods.[94,98] The rationale behind the use of the permutation test in this manner

is that with uninformative data that are divided into two groups, a classifier would on average assign 50% to the wrong class. A validation method that returns an error rate that is on average much deviating from the expected 50% error rate is biased. Permutation tests thus answer two questions: whether the information in the data is truly relevant and whether the performance estimation is carried out properly.

In the literature, the number of executed permutations varies substantially. Ambroise *et al.* use 20 permutations to investigate the bias of incomplete cross validation,[94] while Bijlsma *et al.* and Smit *et al.* use 10,000 permutations to determine the significance of the performance of a classifier.[33,36] So how many permutations are needed? For very small data sets it may be feasible to perform an exhaustive permutation test in which all possible permutations are considered. The number of possible permutations quickly rises, even for moderate class sizes. As an alternative, a test can be performed with only a subset of all permutations. The number of permutations determines accuracy and the lower bound of the p-value; with 100 permutations the lowest possible p-value is 0.01. Since the variance of the performance in permutations can be very large, a large number of permutations are needed to obtain a reliable result.

Strategies and applications

In this section we provide some examples of validation strategies applied in transcriptomics, metabolomics and proteomics literature.

A microarray data analysis workflow is suggested by Wessels *et al.*[89] Their validation protocol consists of 100 repeats of a stratified double cross validation, where the outer loop is a threefold cross validation and the inner loop is tenfold. They report the average of the sensitivity and the specificity. For a metabolomics obesity study, Bijlsma *et al.* developed a strategy for data preprocessing, processing and validation.[36] The PLSDA classifier performance is evaluated with single cross validation and 10,000 permutations. Potential biomarkers are selected that have regression coefficients above a certain threshold. The information carried in the selection is tested by building models with only the selected variables. Additionally, non-informative models are build on the data without the selected variables to test if all relevant information is captured in the selected variables.

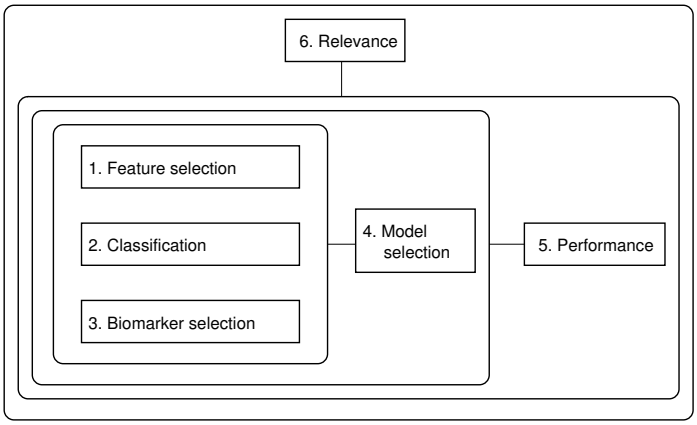

Figure 2.6: Modular view of proteomics data analysis.

In proteomics research there are also several examples of statistical validation strategies. Lee validated PLSDA results on MS data with double cross validation and by comparing the performance with 20 permutations of the original data.[99] Similar statistical strategies in clinical proteomics studies are used by Tong *et al.*[78] and Smit *et al.*[33]

2.7 Proteomics data analysis: a framework

Data analysis methods extract information from the data to predict the class. As shown, there are many methods for feature selection, classification, biomarker candidate selection and statistical validation. It is possible to combine methods in different ways, leading to many data analysis approaches. We propose a modular data analysis framework (Figure 2.6), in which most data analysis strategies fit. While it is possible to make a selection from the feature selection, classification and biomarker discovery modules to form a good working classifier, the validation modules form an integral part of the strategy which should not be left out. For each module the researcher can use his or her method of choice. In the remainder of this section we will discus the modules and their interactions.

Module 1 is the feature selection. This module is optional, but for high dimensional data the choice of classification method sometimes demands feature se-

lection, for example when discriminant analysis or logistic regression is used. Module 2 is the classification method, this module is only necessary if one of the aims is to obtain a classification rule. Module 3 represents the biomarker selection, it is to be used if biomarker discovery is the purpose of the study and the biomarker selection is not intrinsic to the classification method.

The next three modules are statistical validation methods that are all discussed in section 2.6. From a statistical point of view it is recommendable to use these modules if possible since they give generalizable models (module 4), performance estimates (module 5) and insight in the relevance of the model and the data (module 6). Invoking these validation tools enhances the trustworthiness of the model and the biomarkers.

2.8 Black spots and open issues

External test set

If there is only one data set available a cross validation approach makes efficient use of the data.[98] However, an external test set is always of added value.[95] An external data set obtained in a different way can show whether the model is not too specific for the data set that is used to construct the classification rule. For example the measurement could be performed on another instrument, by a different person, and the samples could have been obtained from a different population of patients. In the omics literature several examples of the use of external test set can be found.[76,100]

Power calculations

An issue that we have not yet addressed in this chapter is power calculations. A power calculation determines the sample size necessary to observe a known effect. Such calculations are standard in clinical trials,[101] but are not yet developed for clinical proteomics. There are two problems involved in power calculations for clinical proteomics: i) unknown effect size, ii) highly multivariate data. For power calculations the expected effect size (or the minimal wanted effect size) has to be known a priori. This is problematic in clinical proteomics. Moreover, power calculations are well developed for univariate

analysis, but the results for multivariate analysis are very limited.[102]

Obviously, the larger the sample sets, the more accurate the result. Unfortunately, the number of measurements is usually limited due to the cost of measurements or the limited availability of suitable samples. Validation strategies help overcome some problems. However, Rubingh shows that statistical tests become unreliable for data sets with small sample size.[103]

Increasing complexity of data sets

The technology of mass spectrometry is improving, see for example the developments in hyphenated techniques, such as the combination of liquid chromatography and mass spectrometry (LC-MS). This implicates that the data sets, which are already complex, will be even more complex in the future. We observe a tendency in the literature to analyze combinations of different types of omics data.[3]

2.9 Conclusions

Proteomics research, despite the large effort in recent years, knows many issues that are still subject to debate. This chapter discussed some issues related to the analysis of proteomics data. Due to the complex nature and high dimensionality of the data it is easy to find differences between groups. But these differences are possibly just chance results. The goal is to develop classifiers and/or biomarkers that perform well on new data. Furthermore, a proper estimate of the performance is desirable for forming realistic expectations for the prediction of future samples. Additionally, the relevance of the model should be investigated.

In this chapter we have shown that there are some good examples of performing statistical validation. We urge to set some standards in reporting results from models derived from proteomics data. Such a standard could include that sensitivity and specificity are only to be reported on test sets that have not been used during model building. Furthermore, also a p-value, possibly obtained from a permutation test, should be reported in order to assess the probability of a chance result.

A statistically valid biomarker should always be subjected to biological validation. This answers the question whether the biomarkers are specific for the disease. A statistical valid biomarker can be biologically irrelevant, for example: if the experiment is on a healthy control group and a group with cancer, the biomarker might be indicative for a secondary effect like inflammation that is not specific for cancer. Even the most thorough statistical procedure can not safeguard against this type of findings.

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.

[†]This chapter is based on S. Smit, M.J. van Breemen, H.C.J. Hoefsloot, A.K. Smilde, J.M.F.G. Aerts, C.G. de Koster, *Anal. Chim. Acta.* **2007**, *592*, 210. DOI:10.1016/j.aca.2007.04.043

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.[104–108] 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[13] 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[94] and Simon *et al.*[95] 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.[78,99] 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.[109] Also two specific Gaucher cell markers are known: chitotriosidase and CCL18. Chitotriosidase shows a thousandfold increased activity in serum of symptomatic Gaucher patients.[110] Plasma CCL18 levels are elevated ten to fiftyfold in symptomatic Gaucher patients.[108] 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.[79] Combining PCDA, permutation tests, double cross validation and

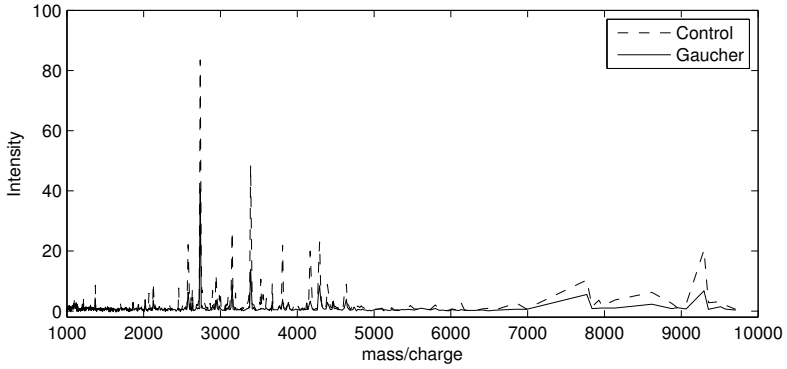

Figure 3.1: Examples of a SELDI-TOF-MS spectrum of a control subject and a Gaucher patient after preprocessing.

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

$$R = \frac{\mathbf{d}'\mathbf{B}\mathbf{d}}{\mathbf{d}'\mathbf{W}\mathbf{d}} \tag{3.1}$$

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 $\mathbf{W}^{-1}\mathbf{B}$.[111] 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 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).[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.

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.[93] 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

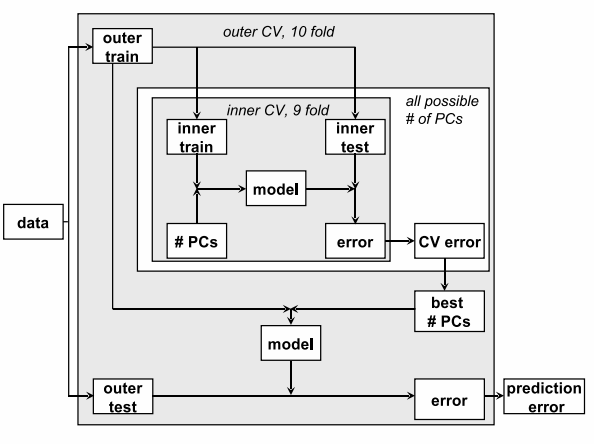

Figure 3.2: Double cross validation. The original data set is split into a training set (outer train) and test set (outer test) ten times in the outer cross validation loop (Outer CV). In the inner loop the outer training set is split up nine times in a training set (inner train) and a test set (inner 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.

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.[79] 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

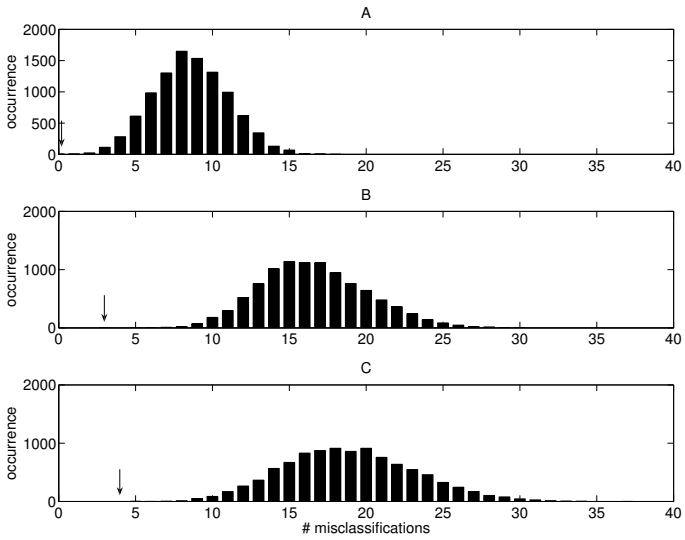

Figure 3.3: Permutation test. Histogram of the number of misclassifications in 10,000 permutations. A: resubstitution error; B: cross validation error; C: double cross validation error. The arrows indicate the number of misclassifications in the original problem.

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 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. 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 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 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.[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 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 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.

m/z	RP
4058.0	36
5852.6	288
3685.4	$115{\cdot}10^3$
4546.0	$435{\cdot}10^3$
2067.9	$292{\cdot}10^5$
4214.8	$113{\cdot}10^6$
3840.1	$136{\cdot}10^7$
1008.2	$228{\cdot}10^7$
4016.2	$503{\cdot}10^8$
8949.4	$781{\cdot}10^8$

$\left(\frac{590}{2}\right)^{10} = 5 \cdot 10^{24}$. 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

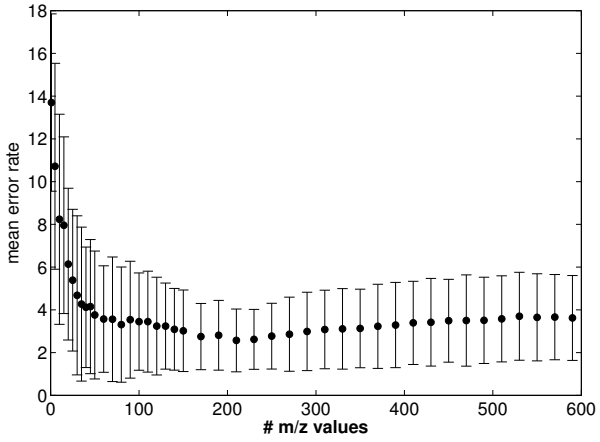

Figure 3.4: Error rate vs. number of variables. For 1, 5 and 10 variables LDA was used to build the model, for larger number of variables we used PCDA with 10 PCs. The reported error rates are averages of 100 different double cross validations.

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.

Outlook

We have discussed some aspects of the analysis of clinical proteomics data. By tailoring the data analysis method (Chapters 6 and 7) it is possible to find effects in the data that would otherwise remain hidden. The combination of cross validation and permutation testing forms a thorough statistical validation which creates a solid foundation to continue developing differences between patient groups into clinically valuable biomarkers. Nevertheless, there remain many open issues regarding the analysis of proteomics data and we discuss some of those here. These issues are the subject of ongoing and future research. We briefly touched upon the issues of power calculations and increasingly complex data sets at the end of Chapter 2. In this chapter we elaborate on these issues.

Power calculations

Power calculations provide the relationship between sample size, effect size and the power of a statistical test. When the effect size is known or estimated, the sample size can be calculated given the power desired. An appropriate sample size, not too many or too few, gives rise to effective experimental designs at controlled costs. For clinical proteomics and other omics disciplines power calculations are not standard procedure. The reason for this is twofold. First, in clinical proteomics studies the effect size is usually unknown. The search for differentially expressed proteins is performed in a shotgun approach. Whether differentially expressed proteins will be measured and, if so, how large an effect can be expected is not known beforehand. Estimates for these could probably be obtained from pilot studies with 5-10 observations per class.[159] The second problem stems from the high-dimensionality of the data. While power calculations are well developed in univariate analysis, results for multivariate data are very limited. Recently some results have been obtained for multiple testing problems[102,159,160] using the (local) false discovery rate.[22] However, the issue is still open for high-dimensional data. Computer simulations using biological knowledge might be a good approach.

Increasing complexity of data sets

The improvement in mass spectrometry technology and the development of hyphenated techniques, for example liquid chromatography coupled to mass spectrometry (LC-MS, see for example Chapters 6 and 7) leads to ever more complex data sets. Different platforms and different measuring parameters, e.g. different columns, allow for measuring different parts of the proteome. Integration of the resulting data can be achieved at several levels. The data sets may be combined to form one larger set in which they are analyzed together. Alternatively, each set is analyzed separately and the results are combined to give an expanded view. Another form of increasingly complex data sets results from the integration of different types of 'omics' data, for example gene expression and proteomics data. The findings in one data set can be used to confirm findings in the other, or together they can bring to light new discoveries.[3] The best method for fusing data sets remains a topic for future research.

Towards clinical use

The goal in clinical proteomics research is to find protein markers that are of clinical use, for example in population screening programs. Finding a protein that is differentially expressed in one experiment does not necessarily translate to a clinical application. Pepe identifies several phases of development for markers intended for population screening.[1] The work presented in this thesis could be considered first phase studies where many leads are discovered and prioritized. Between this phase and actual use as a screening tool lie the phases of clinical assay development and evaluation. A challenge in these phases is setting acceptable thresholds for type I and type II errors (false positives and false negatives). A type I error means unnecessary psychological burden for the person tested falsely positive. In population screening, a test with a high type I error results in many costly follow-up procedures that would not have been performed without the screening programme. On the other hand, a high type II error leads to many people being falsely reassured. A good screening tool strikes an acceptable balance between the two.

Bibliography

[1] M. S. Pepe, R. Etzioni, Z. D. Feng, J. D. Potter, M. L. Thompson, M. Thornquist, M. Winget and Y. Yasui, Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* 2001, *93*, 1054-1061.

[2] S. Hanash, Disease proteomics. *Nature* 2003, *422*, 226-232.

[3] R. Aebersold and M. Mann, Mass spectrometry-based proteomics. *Nature* 2003, *422*, 198-207.

[4] B. Domon and R. Aebersold, Mass spectrometry and protein analysis. *Science* 2006, *312*, 212-217.

[5] G. S. Omenn, Strategies for plasma proteomic profiling of cancers. *Proteomics* 2006, *6*, 5662-5673.

[6] J. Villanueva, J. Philip, C. A. Chaparro, Y. B. Li, R. Toledo-Crow, L. DeNoyer, M. Fleisher, R. J. Robbins and P. Tempst, Correcting common errors in identifying cancer-specific serum peptide signatures. *Journal of Proteome Research* 2005, *4*, 1060-1072.

[7] R. A. R. Bowen, Y. Chan, J. Cohen, N. N. Rehak, G. L. Hortin, G. Csako and A. T. Remaley, Effect of blood collection tubes on total triiodothyronine and other laboratory assays. *Clinical Chemistry* 2005, *51*, 424-433.

[8] A. J. Rai and F. Vitzthum, Effects of preanalytical variables on peptide and protein measurements in human serum and plasma: Implications for clinical proteomics. *Expert Review of Proteomics* 2006, *3*, 409-426.

[9] M. Dijkstra, R. J. Vonk and R. C. Jansen, SELDI-TOF mass spectra: A view on sources of variation. *Journal of Chromatography B* 2007, *847*, 12-23.

[10] M. West-Nielsen, E. V. Hogdall, E. Marchiori, C. K. Hogdall, C. Schou and N. H. H. Heegaard, Sample handling for mass spectrometric proteomic investigations of human sera. *Analytical Chemistry* 2005, *77*, 5114-5123.

[11] A. E. Pelzer, I. Feuerstein, C. Fuchsberger, S. Ongarello, J. Bektic, C. Schwentner, H. Klocker, G. Bartsch and G. K. Bonn, Influence of blood sampling on protein profiling and pattern analysis using matrix-assisted laser desorption/ionisation mass spectrometry. *BJU International* 2007, *99*, 658-662.

[12] M. Hilario, A. Kalousis, C. Pellegrini and M. Muller, Processing and classification of protein mass spectra. *Mass Spectrometry Reviews* 2006, *25*, 409-449.

[13] T. Hastie, J. Friedman and R. Tibshiranie, *The elements of statistical learning. Data mining, inference and prediction*, Springer, New York, 2001.

[14] L. Kanal and B. Chandrasekaran, On dimensionality and sample size in statistical pattern classification. *Pattern Recognition* 1971, *3*, 225-234.

[15] A. Choudhary, M. Brun, J. P. Hua, J. Lowey, E. Suh and E. R. Dougherty, Genetic test bed for feature selection. *Bioinformatics* 2006, *22*, 837-842.

[16] E. R. Dougherty, J. P. Hua and M. L. Bittner, Validation of computational methods in genomics. *Current Genomics* 2007, *8*, 1-19.

[17] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, Gene selection for cancer classification using support vector machines. *Machine Learning* 2002, *46*, 389-422.

[18] M. K. Titulaer, I. Siccama, L. J. Dekker, A. L. C. T. van Rijswilk, R. M. A. Heeren, P. A. S. Smitt and T. M. Luider, A database application for pre-processing, storage and comparison of mass spectra derived from patients and controls. *BMC Bioinformatics* 2006, *7*, 403.

[19] B. L. Wu, T. Abbott, D. Fishman, W. McMurray, G. Mor, K. Stone, D. Ward, K. Williams and H. Y. Zhao, Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data. *Bioinformatics* 2003, *19*, 1636-1643.

[20] I. Levner, Feature selection and nearest centroid classification for protein mass spectrometry. *BMC Bioinformatics* 2005, *6*, 68.

[21] D. I. Broadhurst and D. B. Kell, Statistical strategies for avoiding false discoveries in metabolomics and related experiments. *Metabolomics* 2006, *2*, 171-196.

[22] Y. Benjamini and Y. Hochberg, Controlling the false discovery rate - a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society B* 1995, *57*, 289-300.

[23] J. D. Storey, A direct approach to false discovery rates. *Journal of the Royal Statistical Society B* 2002, *64*, 479-498.

[24] V. G. Tusher, R. Tibshirani and G. Chu, Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America* 2001, *98*, 5116-5121.

[25] I. T. Joliffe, *Principal component analysis*, 2nd ed., Springer, New York, 2002.

[26] R. Wehrens and L. M. C. Buydens, Evolutionary optimisation: A tutorial. *Trends in Analytical Chemistry* 1998, *17*, 193-203.

[27] R. A. Fisher, The use of multiple measurements in taxonomic problems. *Annals of Eugenics* 1936, *7*, 179-188.

[28] J. H. Friedman, Regularized discriminant analysis. *Journal of the American Statistical Association* 1989, *84*, 165-175.

[29] S. Dudoit, J. Fridlyand and T. P. Speed, Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 2002, *97*, 77-87.

[30] R. Hoogerbrugge, S. J. Willig and P. G. Kistemaker, Discriminant analysis by double stage principal component analysis. *Analytical chemistry* 1983, *55*, 1710-1712.

[31] J. Ye, T. Li, T. Xiong and R. Janardan, Using uncorrelated discriminant analysis for tissue classification with gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2004, *1*, 181-190.

[32] R. H. Lilien, H. Farid and B. R. Donald, Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *Journal of Computational Biology* 2003, *10*, 925-946.

[33] S. Smit, M. J. van Breemen, H. C. J. Hoefsloot, A. K. Smilde, J. M. F. G. Aerts and C. G. de Koster, Assessing the statistical validity of proteomics based biomarkers. *Analytica Chimica Acta* 2007, *592*, 210-217.

[34] S. Wold, A. Ruhe, H. Wold and W. J. Dunn III, The collinearity problem in linear regression. The partial least squares (PLS) approach to generalized inverses. *SIAM Journal on Scientific Computing* 1984, *5*, 735-743.

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

[36] S. Bijlsma, L. Bobeldijk, E. R. Verheij, R. Ramaker, S. Kochhar, I. A. Macdonald, B. van Ommen and A. K. Smilde, Large-scale human metabolomics studies: A strategy for data (pre-) processing and validation. *Analytical Chemistry* 2006, *78*, 567-574.

[37] J. Gottfries, M. Sjogren, B. Holmberg, L. Rosengren, P. Davidsson and K. Blennow, Proteomics for drug target discovery. *Chemometrics and Intelligent Laboratory Systems* 2004, *73*, 47-53.

[38] J. Trygg, E. Holmes and T. Lundstedt, Chemometrics in metabonomics. *Journal of Proteome Research* 2007, *6*, 469-479.

[39] V. Vapnik, *The nature of statistical learning theory*, 2nd ed., Springer-Verlag, New York, 2000.

[40] D. Agranoff, D. Fernandez-Reyes, M. C. Papadopoulos, S. A. Rojas, M. Herbster, A. Loosemore, E. Tarelli, J. Sheldon, A. Schwenk, R. Pollak, C. F. J. Rayner and S. Krishna, Identification of diagnostic markers for tuberculosis by proteomic fingerprinting of serum. *Lancet* 2006, *368*, 1012-1021.

[41] K. Jong, E. Marchiori and A. van der Vaart, Analysis of proteomic pattern data for cancer detection. *Applications of Evolutionary Computing* 2004, *3005*, 41-51.

[42] F. M. Smith, W. M. Gallagher, E. Fox, R. B. Stephens, E. Rexhepaj, E. F. Petricoin, L. Liotta, M. J. Kennedy and J. V. Reynolds, Combination of SELDI-TOF-MS and data mining provides early-stage response prediction for rectal tumors undergoing multimodal neoadjuvant therapy. *Annals of Surgery* 2007, *245*, 259-266.

[43] R. Willingale, D. J. L. Jones, J. H. Lamb, P. Quinn, P. B. Farmer and L. L. Ng, Searching for biomarkers of heart failure in the mass spectra of blood plasma. *Proteomics* 2006, *6*, 5903-5914.

[44] X. G. Zhang, X. Lu, Q. Shi, X. Q. Xu, H. C. E. Leung, L. N. Harris, J. D Iglehart, A. Miron, J. S. Liu and W. H. Wong, Recursive SVM feature selection and sample classification for mass-spectrometry and microarray data. *BMC Bioinformatics* 2006, *7*, 197.

[45] S. Bhattacharyya, E. R. Siegel, G. M. Petersen, S. T. Chari, L. J. Suva and R. S. Haun, Diagnosis of pancreatic cancer using serum proteomic profiling. *Neoplasia* 2004, *6*, 674-686.

[46] J. Zhu and T. Hastie, Classification of gene microarrays by penalized logistic regression. *Biostatistics* 2004, *5*, 427-443.

[47] A. Goncalves, B. Esterni, F. Bertucci, R. Sauvan, C. Chabannon, M. Cubizolles, V. J. Bardou, G. Houvenaegel, J. Jacquemier, S. Granjeaud, X. Y. Meng, E. T. Fung, D. Birnbaum, D. Maraninchi, P. Viens and J. P. Borg, Postoperative serum proteomic profiles may predict metastatic relapse in high-risk primary breast cancer patients receiving adjuvant chemotherapy. *Oncogene* 2006, *25*, 981-989.

[48] L. Shen and E. C. Tan, Dimension reduction-based penalized logistic regression for cancer classification using microarray data. *IEEE/ACM Transactions on Computational Biology and Bioninformatics* 2005, *2*, 166-175.

[49] P. H. C. Eilers, J. Boer, G. J. B. van Ommen and J. C. van Houwelingen, Classification of microarray data with penalized logistic regression. *Progress in Biomedical Optics and Imaging* 2001, *2*, 187-198.

[50] M. Dettling and P. Buhlmann, Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis* 2004, *90*, 106-131.

[51] R. Tibshirani, T. Hastie, B. Narasimhan and G. Chu, Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 2002, *99*, 6567-6572.

[52] U. T. Shankavaram, W. C. Reinhold, S. Nishizuka, S. Major, D. Morita, K. K. Chary, M. A. Reimers, U. Scherf, A. Kahn, D. Dolginow, J. Cossman, E. P. Kaldjian, D. A. Scudiero, E. Petricoin, L. Liotta, J. K. Lee and J. N. Weinstein, Transcript and protein expression profiles of the nci-60 cancer cell panel: An integromic microarray study. *Molecular Cancer Therapeutics* 2007, *6*, 820-832.

[53] R. F. J. Kemperman, P. L. Horvatovich, B. Hoekman, T. H. Reijmers, F. A. J. Muskiet and R. Bischoff, Comparative urine analysis by liquid chromatography-mass spectrometry and multivariate statistics: Method development, evaluation, and application to proteinuria. *Journal of Proteome Research* 2007, *6*, 194-206.

[54] G. C. Bloom, S. Eschrich, J. X. Zhou, D. Coppola and T. J. Yeatman, Elucidation of a protein signature discriminating six common types of adenocarcinoma. *International Journal of Cancer* 2006, *120*, 769-775.

[55] Q. H. C. Ru, L. W. A. Zhu, J. Silberman and C. D. Shriver, Label-free semiquantitative peptide feature profiling of human breast cancer and breast disease sera via two-dimensional

liquid chromatography-mass spectrometry. *Molecular & Cellular Proteomics* 2006, *5*, 1095-1104.

[56] J. C. Oates, S. Varghese, A. M. Bland, T. P. Taylor, S. E. Self, R. Stanislaus, J. S. Almeida and J. M. Arthur, Prediction of urinary protein markers in lupus nephritis. *Kidney International* 2005, *68*, 2588-2592.

[57] M. Albitar, S. J. Potts, F. J. Giles, S. O'Brien, M. Keating, D. Thomas, C. Clarke, I. Jilani, C. Aguilar, E. Estey and H. Kantarjian, Proteomic-based prediction of clinical behavior in adult acute lymphoblastic leukemia. *Cancer* 2006, *106*, 1587-1594.

[58] G. L. Gerton, X. J. Fan, J. Chittams, M. Sammel, A. Hummel, J. F. Strauss and K. Barnhart, A serum proteomics approach to the diagnosis of ectopic pregnancy. *Annals of the New York Academy of Sciences* 2004, *1022*, 306-316.

[59] L. K. Hansen and P. Salamon, Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 1990, *12*, 993-1001.

[60] G. Bhanot, G. Alexe, B. Venkataraghavan and A. J. Levine, A robust meta-classification strategy for cancer detection from MS data. *Proteomics* 2006, *6*, 592-604.

[61] B. Liu, Q. H. Cui, T. Z. Jiang and S. D. Ma, A combinational feature selection and ensemble neural network method for classification of gene expression data. *BMC Bioinformatics* 2004, *5*, 136.

[62] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson and P. S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature Medicine* 2001, *7*, 673-679.

[63] J. H. Hong and S. B. Cho, The classification of cancer based on DNA microarray data that uses diverse ensemble genetic programming. *Artificial Intelligence in Medicine* 2006, *36*, 43-58.

[64] M. P. A. Ebert, J. Meuer, J. C. Wiemer, H. U. Schulz, M. A. Reymond, U. Traugott, P. Malfertheiner and C. Rocken, Identification of gastric cancer patients by serum protein profiling. *Journal of Proteome Research* 2004, *3*, 1261-1266.

[65] G. Valentini, M. Muselli and F. Ruffino, Cancer recognition with bagged ensembles of support vector machines. *Neurocomputing* 2004, *56*, 461-466.

[66] E. Tamoto, M. Tada, K. Murakawa, M. Takada, G. Shindo, K. Teramoto, A. Matsunaga, K. Komuro, M. Kanai, A. Kawakami, Y. Fujiwara, N. Kobayashi, K. Shirata, N. Nishimura, S. I. Okushiba, S. Kondo, J. Hamada, T. Yoshiki, T. Moriuchi and H. Katoh, Gene-expression profile changes correlated with tumor progression and lymph node metastasis in esophageal cancer. *Clinical Cancer Research* 2004, *10*, 3629-3638.

[67] M. M. W. B. Hendriks, S. Smit, L. M. W. Akkermans, T. H. Reijmers, P. H. C. Eilers, H. C. J. Hoefsloot, C. M. Rubingh, C. G. de Koster, J. M. Aerts and A. K. Smilde, How to distinguish healthy from diseased? Classification strategy for mass spectrometry based clinical proteomics. *Proteomics* 2007, *7*, 3672-3680.

[68] M. Dettling, Bagboosting for tumor classification with gene expression data. *Bioinformatics* 2004, *20*, 3583-3593.

[69] Y. H. Peng, Integration of gene functional diversity for effective cancer detection. *International Journal of Systems Science* 2006, *37*, 931-938.

[70] A. Bertoni, R. Folgieri and G. Valentini, Bio-molecular cancer prediction with random subspace ensembles of support vector machines. *Neurocomputing* 2005, *63*, 535-539.

[71] K. J. Kim and S. B. Cho, Ensemble classifiers based on correlation analysis for DNA microarray classification. *Neurocomputing* 2006, *70*, 187-199.

[72] H. H. Won and S. B. Cho, Neural network ensemble with negatively correlated features for cancer classification. *Artificial Neural Networks and Neural Information Processing - Ican/Iconip 2003* 2003, *2714*, 1143-1150.

[73] L. I. Kuncheva, A theoretical study on six classifier fusion strategies. *IEEE Transactions On Pattern Analysis And Machine Intelligence* 2002, *24*, 281-286.

[74] L. Breiman, Random forests. *Machine Learning* 2001, *45*, 5-32.

[75] E. C. Gunther, D. J. Stone, R. W. Gerwien, P. Bento and M. P. Heyes, Prediction of clinical drug efficacy by classification of drug-induced genomic expression profiles in vitro. *Proceedings of the National Academy of Sciences of the United States of America* 2003, *100*, 9608-9613.

[76] K. Hoffmann, M. J. Firth, A. H. Beesley, N. H. de Klerk and U. R. Kees, Translating microarray data for diagnostic testing in childhood leukaemia. *BMC Cancer* 2006, *6*, 229.

[77] W. D. Tong, H. X. Hong, H. Fang, Q. Xie and R. Perkins, Decision forest: Combining the predictions of multiple independent decision tree models. *Journal of Chemical Information and Computer Sciences* 2003, *43*, 525-531.

[78] W. D. Tong, W. Xie, H. X. Hong, H. Fang, L. M. Shi, R. Perkins and 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. *Environmental Health Perspectives* 2004, *112*, 1622.

[79] R. Breitling, P. Armengaud, A. Amtmann and P. Herzyk, Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Letters* 2004, *573*, 83.

[80] L. Eriksson, E. Johansson, N. Nettaneh-Word and S. Wold, *Introduction to multi- and megavariate data analysis using projection methods (PCA & PLS)*, Umetrics, Umea, Sweden, 1999.

[81] J. Yang, X. J. Zhao, X. L. Liu, C. Wang, P. Gao, J. S. Wang, L. J. Li, J. R. Gu, S. L. Yang and G. W. Xu, High performance liquid chromatography-mass spectrometry for metabonomics: Potential biomarkers for acute deterioration of liver function in chronic Hepatitis B. *Journal of Proteome Research* 2006, *5*, 554-561.

[82] P. Y. Yin, X. J. Zhao, Q. R. Li, J. S. Wang, J. S. Li and G. W. Xu, Metabonomics study of intestinal fistulas based on ultraperformance liquid chromatography coupled with Q-TOF mass spectrometry (UPLC/Q-TOF MS). *Journal of Proteome Research* 2006, *5*, 2135-2143.

[83] H. Liu, J. Li and L. Wong, A comparative study on feature selection and classification methods using gene expression profiles and proteomic patterns. *Genome Informatics* 2002, *13*, 51-60.

[84] Special issue: Mining MALDI-TOF data. *Proteomics* 2003, *3*, 1667-1724.

[85] M. Wagner, D. N. Naik, A. Pothen, S. Kasukurti, R. R. Devineni, B. L. Adam, O. J. Semmes and G. L. Wright, Computational protein biomarker prediction: A case study for prostate cancer. *BMC Bioinformatics* 2004, *5*, 26.

[86] D. J. Hand, Breast cancer diagnosis from proteomic mass spectrometry data: A comparative evaluation. *Statistical Applications in Genetics and Molecular Biology* 2008, *7*, 15.

[87] M. S. Pepe, Evaluating technologies for classification and prediction in medicine. *Statistics in Medicine* 2005, *24*, 3687-3696.

[88] K. Baumann, Cross-validation as the objective function for variable-selection techniques. *Trends in Analytical Chemistry* 2003, *22*, 395-406.

[89] L. F. A. Wessels, M. J. T. Reinders, A. A. M. Hart, C. J. Veenman, H. Dai, Y. D. He and L. J. van't Veer, A protocol for building and evaluating predictors of disease state based on microarray data. *Bioinformatics* 2005, *21*, 3755-3762.

[90] R. Kohavi, A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* 1995, 1137-1145.

[91] P. Zhang, Model selection via multifold cross-validation. *Annals of Statistics* 1993, *21*, 299-313.

[92] R. Kohavi and G. H. John, Wrappers for feature subset selection. *Artificial Intelligence* 1997, *97*, 273-324.

[93] M. Stone, Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society B* 1974, *36*, 111-147.

[94] C. Ambroise and G. J. McLachlan, Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America* 2002, *99*, 6562-6566.

[95] R. Simon, M. D. Radmacher, K. Dobbin and L. M. McShane, Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. *Journal of the National Cancer Institute* 2003, *95*, 14-18.

[96] S. Varma and R. Simon, Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 2006, *7*, 91.

[97] R. G. Brereton, Consequences of sample size, variable selection, and model validation and optimisation, for predicting classification ability from analytical data. *Trends in Analytical Chemistry* 2006, *25*, 1103-1111.

[98] B. J. A. Mertens, M. E. De Noo, R. A. E. M. Tollenaar and A. M. Deelder, Mass spectrometry proteomic diagnosis: Enacting the double cross-validatory paradigm. *Journal of Computational Biology* 2006, *13*, 1591-1605.

[99] K. R. Lee, X. W. Lin, D. C. Park and S. Eslava, Megavariate data analysis of mass spectrometric proteomics data using latent variable projection method. *Proteomics* 2003, *3*, 1680-1686.

[100] N. P. Munro, D. A. Cairns, P. Clarke, M. Rogers, A. J. Stanley, J. H. Barrett, P. Harnden, D. Thompson, I. Eardley, R. E. Banks and M. A. Knowles, Urinary biomarker profiling in transitional cell carcinoma. *International Journal of Cancer* 2006, *119*, 2642-2650.

[101] W. J. Dixon and F. J. Massey, *Introduction to statistical analysis*, 4th ed., McGraw-Hill, New York, 1983.

[102] J. A. Ferreira and A. Zwinderman, Approximate sample size calculations with microarray data: An illustration. *Statistical Applications in Genetics and Molecular Biology* 2006, *5*, 25.

[103] C. M. Rubingh, S. Bijlsma, E. P. P. A. Derks, I. Bobeldijk, E. R. Verheij, S. Kochhar and A. K. Smilde, Assessing the performance of statistical validation tools for megavariate metabolomics data. *Metabolomics* 2006, *2*, 53-61.

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

[105] 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 and J. Wright, Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research* 2002, *62*, 3609-3614.

[106] A. Vlahou, A. Giannopoulos, B. W. Gregory, T. Manousakas, F. I. Kondylis, L. L. Wilson, P. F. Schellhammer, G. L. Wright and O. J. Semmes, Protein profiling in urine for the diagnosis of bladder cancer. *Clinical Chemistry* 2004, *50*, 1438-1441.

[107] S. G. Soltys, Q. T. Le, G. Y. Shi, R. Tibshirani, A. J. Giaccia and 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. *Clinical Cancer Research* 2004, *10*, 4806-4812.

[108] 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 and J. M. F. G. Aerts, Marked elevation of the chemokine CCL18/PARC in Gaucher disease: A novel surrogate marker for assessing therapeutic intervention. *Blood* 2004, *103*, 33.

[109] E. Beutler and G. A. Gabrowski, *Gaucher disease*, in *The metabolic and molecular bases of inherited disease*, edited by C. R. Scriver, A. L. Beadet, W. S. Sly and D. Valle, McGraw-Hill, New York 2001.

[110] C. E. M. Hollak, S. van Weely, M. H. J. van Oers and J. M. F. G. Aerts, Marked elevation of plasma chitotriosidase activity. A novel hallmark of Gaucher disease. *Journal of Clinical Investigation* 1994, *93*, 1288-1292.

[111] B. G. M. Vandeginste, D. L. Massart, L. M. C. Buydens, S. De Jong, P. J. Lewi and J. Smeyers-Verbeke, *Handbook of chemometrics and qualimetrics: Part B*, Elsevier, Amsterdam, 1998.

[112] P. Howland and H. Park, Generalizing Discriminant Analysis Using the Generalized Singular Value Decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2004, *26*, 995-1006.

[113] P. W. Mielke Jr and K. J. Berry, *Permutation methods: A distance function approach*, Springer, New York, 2001.

[114] M. T. Moran, J. P. Schofield, A. R. Hayman, G. P. Shi, E. Young and T. M. Cox, Pathologic gene expression in Gaucher disease: Up-regulation of cysteine proteinases including osteoclastic cathepsin K. *Blood* 2000, *96*, 1969-1978.

[115] R. O. Brady, A. E. Gal, R. M. Bradley, E. Martensson, A. L. Warshaw and L. Laster, Enzymatic defect in Fabry's disease. Ceramidetrihexosidase deficiency. *New England Journal of Medicine* 1967, *276*, 1163-1167.

[116] J. A. Kint, Fabry's disease: α-Galactosidase deficiency. *Science* 1970, *167*, 1268-1269.

[117] R. J. Desnick, Y. A. Ioannou and M. E. Eng, *α-Galactosidase A deficiency: Fabry disease*, in *The metabolic and molecular bases of inherited disease*, edited by C. R. Scriver, A. L. Beaudet, W. S. Sly and D. Valle, McGraw-Hill, New York 2001.

[118] K. Utsumi, N. Yamamoto, R. Kase, T. Takata, H. Okumiya, T. Suzuki and E. Uyama, High incidence of thrombosis in Fabry's disease. *Internal Medicine* 1997, *36*, 327-329.

[119] Y. Shen, P. F. Bodary, F. B. Vargas, J. W. Homeister, D. Gordon, K. A. Ostenso, J. A. Ostenso and D. T. Eitzman, α-Galactosidase A deficiency leads to increased tissue fibrin deposition and thrombosis in mice homozygous for the factor V Leiden mutation. *Stroke* 2006, *37*, 1106-1108.

[120] D. T. Eitzman, P. F. Bodary, Y. Shen, C. G. Khairallah, S. R. Wild, A. Abe, J. Shaffer-Hartman and J. A. Shayman, Fabry disease in mice is associated with age-dependent susceptibility to vascular thrombosis. *Journal of the American Society of Nephrology* 2003, *14*, 298-302.

[121] E. A. Diamantopoulos, C. V. Vassilopoulos and G. E. Marakomichelakis, Intermittent claudication unmasking underlying Fabry's disease. *International Angiology* 2002, *21*, 201-203.

[122] R. Schiffmann, A. Rapkiewicz, M. Abu-Asab, M. Ries, H. Askari, M. Tsokos and M. Quezado, Pathological findings in a patient with Fabry disease who died after 2.5 years of enzyme replacement. *Virchows Archiv* 2006, *448*, 337-343.

[123] P. F. Bodary, Y. Shen, F. B. Vargas, X. Bi, K. A. Ostenso, S. Gu, J. A. Shayman and D. T. Eitzman, α-Galactosidase A deficiency accelerates atherosclerosis in mice with apolipoprotein E deficiency. *Circulation* 2005, *111*, 629-632.

[124] F. Barbey, N. Brakch, A. Linhart, X. Jeanrenaud, T. Palecek, J. Bultas and D. Burnier, Increased carotid intima-media thickness in the absence of atherosclerotic plaques in an adult population with Fabry disease. *Acta Paediatica Supplement* 2006, *95*, 63-68.

[125] T. DeGraba, S. Azhar, F. Dignat-George, E. Brown, B. Boutiere, G. Altarescu, R. McCarron and R. Schiffmann, Profile of endothelial and leukocyte activation in Fabry patients. *Annals of Neurology* 2000, *47*, 229-233.

[126] K. Demuth and D. P. Germain, Endothelial markers and homocysteine in patients with classic Fabry disease. *Acta Paediatrica Supplement* 2002, *91*, 57-61.

[127] A. C. Vedder, E. Biro, J. M. F. G. Aerts, R. Nieuwland, A. Sturk and C. E. M. Hollak, *unpublished data*.

[128] C. M. Eng, N. N. Guffon, W. R. Wilcox, D. P. Germain, P. Lee, S. Waldek, L. Caplan, G. E. Linthorst and R. J. Desnick, Safety and efficacy of recombinant human α-Galactosidase A replacement therapy in Fabry's disease. *New England Journal of Medicine* 2001, *345*, 9-16.

[129] R. Schiffmann, J. B. Kopp, H. A. Austin III, S. Sabnis, D. F. Moore, T. Weibel, J. E. Balow and B. R. O., Enzyme replacement therapy in Fabry disease: a randomized controlled trial. *JAMA* 2001, *285*, 2743.

[130] C. Whybra, C. Kampmann, F. Krummenauer, M. Ries, E. Mengel, E. Miebach, F. Baehner, K. Kim, M. Bajbouj, A. Schwarting, A. Gal and M. Beck, The Mainz Severity Score Index: a new instrument for quantifying the Anderson-Fabry disease phenotype, and the response of patients to enzyme replacement therapy. *Clinical Genetics* 2004, *65*, 299-307.

[131] D. F. Moore, O. V. Krokhin, R. C. Beavis, M. Ries, C. Robinson, E. Goldin, R. O. Brady, J. A. Wilkins and R. Schiffmann, Proteomics of specific treatment-related alterations in Fabry disease: a strategy to identify biological abnormalities. *Proceedings of the National Academy of Sciences of the United States of America* 2007, *104*, 2873-2878.

[132] K. A. Baggerly, J. S. Morris and K. R. Coombes, Reproducibility of seldi-tof protein patterns in serum: Comparing datasets from different experiments. *Bioinformatics* 2004, *20*, 777-785.

[133] S. Michiels, S. Koscielny and C. Hill, Prediction of cancer outcome with microarrays: A multiple random validation strategy. *Lancet* 2005, *365*, 488-492.

[134] P. B. Deegan, M. T. Moran, I. McFarlane, J. P. Schofield, R. G. Boot, J. M. F. G. Aerts and T. M. Cox, Clinical evaluation of chemokine and enzymatic biomarkers of Gaucher disease. *Blood Cells Molecules and Diseases* 2005, *35*, 259-267.

[135] J. M. F. G. Aerts, C. E. M. Hollak, M. van Breemen, M. Maas, J. E. M. Groener and R. G. Boot, Identification and use of biomarkers in Gaucher disease and other lysosomal storage diseases. *Acta Paediatrica* 2005, *94*, 43-46.

[136] J.-H. Jiang, R. Tsenkova and Y. Ozaki, Principal discriminant variate method for classification of multicollinear data: Principle and applications. *Analytical Sciences* 2001, *17(ICAS2001)*, i471-i477.

[137] J.-H. Jiang, R. Tsenkova, Y. Wu, R.-Q. Yu and Y. Ozaki, Principal discriminant variate method for classification of multicollinear data: Applications to near-infrared spectra of cow blood samples. *Applied Spectroscopy* 2002, *56*, 488-501.

[138] P. Armitage, G. Berry and J. Matthews, *Statistical methods in medical research*, Blackwell Science, Malden, MA, 2002.

[139] J. Friedman, T. Hastie and R. Tibshirani, Additive logistic regression: A statistical view of boosting. *Annals of Statistics* 2000, *28*, 337-407.

[140] M. Dettling and P. Buhlmann, Boosting for tumor classification with gene expression data. *Bioinformatics* 2003, *19*, 1061-1069.

[141] G. C. Cawley, Matlab support vector machine toolbox (v0.55β), University of East Anglia, School of Information Systems, 2000.

[142] J. C. Platt, *Fast training of support vetor machines using sequential minimal optimization*, in *Advances in kernel methods: Support vector learning*, edited by B. Scholkopf, C. Burges and A. Smola, The MIT Press, Cambridge, MA 1999.

[143] M. de Fost, C. E. M. Hollak, J. E. M. Greener, J. M. F. G. Aerts, M. Maas, L. W. Poll, M. G. Wiersma, D. Haussinger, S. Brett, N. Brill and S. vom Dahl, Superior effects of high-dose enzyme replacement therapy in type 1 Gaucher disease on bone marrow involvement and chitotriosidase levels: A 2-center retrospective analysis. *Blood* 2006, *108*, 830-835.

[144] R. Mcgill, J. W. Tukey and W. A. Larsen, Variations of box plots. *American Statistician* 1978, *32*, 12-16.

[145] C. W. Lee, M. Y. Lin, W. C. Lee, M. H. Chou, C. S. Hsieh, S. Y. Lee and J. H. Chuang, Characterization of plasma proteome in biliary atresia. *Clinica Chimica Acta* 2007, *375*, 104-109.

[146] J. J. Jansen, H. C. J. Hoefsloot, J. van der Greef, M. E. Timmerman and A. K. Smilde, Multilevel component analysis of time-resolved metabolic fingerprinting data. *Analytica Chimica Acta* 2005, *530*, 173-183.

[147] N. I. Govorukhina, T. H. Reijmers, S. O. Nyangoma, A. G. J. van der Zee, R. C. Jansen and R. Bischoff, Analysis of human serum by liquid chromatography-mass spectrometry: Improved sample preparation and data analysis. *Journal of Chromatography A* 2006, *1120*, 142-150.

[148] E. L. Franco, N. F. Schlecht and D. Saslow, The epidemiology of cervical cancer. *Cancer Journal* 2003, *9*, 348-359.

[149] D. M. Parkin, F. Bray, J. Ferlay and P. Pisani, Global cancer statistics, 2002. *CA – A Cancer Journal for Clinicians* 2005, *55*, 74-108.

[150] J. Monsonego, HPV infections and cervical cancer prevention. Priorities and new directions. Highlights of EUROGIN 2004 International Expert Meeting, Nice, France, October 21-23, 2004. *Gynecologic Oncology* 2005, *96*, 830-839.

[151] M. D. Esajas, J. M. Duk, H. W. A. de Bruijn, J. G. Aalders, P. H. B. Willemse, W. Sluiter, B. Pras, K. ten Hoor, H. Hollema and A. G. J. van der Zee, Clinical value of routine serum squamous cell carcinoma antigen in follow-up of patients with early-stage cervical cancer. *Journal of Clinical Oncology* 2001, *19*, 3960-3966.

[152] J. L. Benedet, H. Bender, H. Jones, H. Y. S. Ngan and S. Pecorelli, Figo staging classifications and clinical practice guidelines in the management of gynecologic cancers. *International Journal of Gynecology & Obstetrics* 2000, *70*, 209-262.

[153] G. Tomasi, F. van den Berg and C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data. *Journal of Chemometrics* 2004, *18*, 231-241.

[154] J. A. K. Suykens and J. Vandewalle, Least squares support vector machine classifiers. *Neural Processing Letters* 1999, *9*, 293-300.

[155] Special issue: Competition on Clinical Mass Spectrometry Based Proteomic Diagnosis. *Statistical Applications in Genetics and Molecular Biology* 2008, *7*, 1-15.

[156] S. Wold, Pattern-recognition by means of disjoint principal components models. *Pattern Recognition* 1976, *8*, 127-139.

[157] H. H. Yue and S. J. Qin, Reconstruction-based fault identification using a combined index. *Industrial & Engineering Chemistry Research* 2001, *40*, 4403-4414.

[158] B. Mertens, M. Thompson and T. Fearn, Principal component outlier detection and SIMCA - a synthesis. *Analyst* 1994, *119*, 2777-2784.

[159] D. A. Cairns, J. H. Barrett, L. J. Billingham, A. J. Stanley, G. Xinarianos, J. K. Field, P. J. Johnson, P. J. Selby and R. E. Banks, Sample size determination in clinical proteomic profiling experiments using mass spectrometry for class comparison. *Proteomics* 2009, *9*, 74-86.

[160] B. Efron, Size, power and false discovery rates. *The Annals of Statistics* 2007, *35*, 1351-1377.

Publications

- E.J.J. van Velzen, J.A. Westerhuis, J.P. M. van Duynhoven, F.A. van Dorsten, H.C.J. Hoefsloot, D.M. Jacobs, S. Smit, R. Draijer, C.I. Kroner, and A.K. Smilde, Multilevel Data Analysis of a Crossover Designed Human Nutritional Intervention Study. *Journal of Proteome Research* 2008, *7*, 4483-4491. DOI:10.1021/pr800145j
- J.A. Westerhuis, H.C.J. Hoefsloot, S. Smit, D.J. Vis, A.K. Smilde, E.J.J. van Velzen, J.P.M. van Duijnhoven and F.A. van Dorsten, Assessment of PLSDA cross validation. *Metabolomics* 2008, *4*, 81-89. DOI:10.1007/s11306-007-0099-6
- H.C.J. Hoefsloot, S. Smit and A.K. Smilde, A Classification Model for the Leiden Proteomics Competition. *Statistical Applications in Genetics and Molecular Biology* 2008, *7*, 8. DOI:10.2202/1544-6115.1351
- J.M.F.G. Aerts, M.J. van Breemen, A.P. Bussink, K. Ghauharali, R. Sprenger, R.G. Boot, J.E. Groener, C.E. Hollak, M. Maas, S. Smit, H.C. Hoefsloot, A.K. Smilde, J.P. Vissers, S. de Jong, D. Speijer, C.G. de Koster, Biomarkers for lysosomal storage disorders: identification and application as exemplified by chitotriosidase in Gaucher disease. *Acta Paediatrica* 2008, *Suppl. 457*, 7-14. DOI:10.1111/j.1651-2227.2007.00641.x
- S. Smit, H.C.J. Hoefsloot, A.K. Smilde, Statistical data processing in clinical proteomics. *Journal of Chromatography B* 2008, *866*, 77-88. DOI:10.1016/j.jchromb.2007.10.042
- M.M.W.B. Hendriks, S. Smit, W.L.M.W. Akkermans, T.H. Reijmers, P.H.C. Eilers, H.C.J. Hoefsloot, C.M. Rubingh, C.G. de Koster, J.M.F.G. Aerts, A.K. Smilde, How to distinguish healthy from diseased? Classification strategy for mass spectrometry-based clinical proteomics. *Proteomics* 2007, *7*, 3672-3680. DOI:10.1002/pmic.200700046
- S. Smit, M.J. van Breemen, H.C.J. Hoefsloot, A.K. Smilde, J.M.F.G. Aerts and C.G. de Koster, Assessing the statistical validity of proteomics based biomarkers. *Analytica Chimica Acta* 2007, *592*, 210-217. DOI:10.1016/j.aca.2007.04.043

Summary

The subject of this thesis is the analysis of data in clinical proteomics studies aimed at the discovery of biomarkers. The data sets produced in proteomics studies are huge, characterized by a small number of samples in which many proteins and peptides are measured. The studies described in this thesis compare different patient groups (recovering vs. relapsing patients) or a group of patients with a group of healthy controls. The size of the data and the size of the differences between the groups call for special data analysis strategies.

Chapter 2 is a review of data analysis strategies for the discovery of biomarkers in clinical proteomics. A wealth of classification and feature extraction methods exists and in this chapter the most commonly applied methods are discussed. Due to the complex nature of the data and the high dimensionality it is easy to find differences between groups. However, these differences are possibly just chance results. The goal is to develop classifiers and/or biomarkers that can be used to classify new samples. Therefore, methods to test the validity of the results are part of a good data analysis strategy. A modular framework that fits most of the strategies described in the literature is presented. In this framework feature selection, classification, biomarker discovery and statistical validation are regarded as separate modules in the analysis of proteomics data. A strategy can be built from a combination of these modules in many ways, to suit the data analysis problem at hand. While it is possible to choose from the feature selection, classification and biomarker discovery modules to form a good working classifier, the validation modules are an integral part of the strategy. Which methods are used to execute a specific module is a matter of choice which depends in part on the structure of the data and in part on the preferences and expertise of the data analyst.

In Chapter 3 we present a strategy for the statistical validation of discrimination models in proteomics studies. It is illustrated on data from a proteomics study of Gaucher disease, a lysosomal storage disorder. Gaucher disease is chosen as a case study because it is known to cause dramatic changes in the

blood of patients. Samples from patients and healthy controls are measured with mass spectrometry and compared with Principal Component Discriminant Analysis (PCDA). The strategy combines permutation tests, single and double cross validation. The permutation test is part of the strategy to rule out the possibility of a chance result, by testing the classification method on randomized data. From the permutation test a p-value is obtained by comparing the performance of the classifier to the performance on randomized data. In the single cross validation the best PCDA model is selected, based on its generalizability towards new samples. In some studies the reported selectivity and specificity of a method is based on the single cross validation error. This error is biased, since the cross validation error is also the criterion that drives the model selection; Model construction and model evaluation are interwoven. In a permutation test this bias is uncovered because the average cross validation error of many permutations will be very different from the expected 50% (for two classes of equal size). An unbiased prediction error is obtained by validating the entire model selection procedure, which in our strategy leads to double cross validation. The permutation test confirmes that the double cross validation is an independent estimation of the performance. The double cross validated sensitivity in the Gaucher vs. control problem is 89% and the specificity is 90%.

Fabry disease is a lysosomal storage disorder for which currently no blood biomarker is available. In Chapter 4 we compare serum protein profiles of controls and Fabry patients, an approach that allowed classification of patients suffering from Gaucher disease in Chapter 3. Classification of Fabry patients and controls using PCDA results in high error rates, also after variable selection. With Support Vector Machines (SVM), the prediction error is lower. The permutation test shows that the classification result is significant, but the misclassification rate is still 16%. It might be argued that the procedure used for protein profiling is not sensitive enough to detect early manifestations of Fabry disease. However, concomitant with misclassification of Fabry patients as being normal, some control subjects are classified as diseased Fabry patients. Strikingly, all three unaffected relatives of Fabry patients (R1, R2 and R3) that were tested were classified as being patient, either using SVM or PCDA. This suggests that the discrimination may not be primarily based on the underlying disorder but rather on other characteristics shared by families. This illustrates the importance to use very closely matched control subjects in these types of studies.

In Chapter 2 we discussed many classification methods. One of the choices

to be made in a proteomics study comparing two classes of patients is the choice for a classification method. In Chapter 5 we apply several classification methods to one clinical proteomics data set, the Gaucher disease data from Chapter 3. The strategy developed in Chapter 3 is now used as a protocol which can be used for choosing among different statistical classification methods and obtaining figures of merit of their performance. The methods considered are PCDA, Penalized Logistic Regression (PLR), LogitBoost (LB), Principal Discriminant Variates (PDV), Nearest Shrunken Centroids (NSC), and SVM. In the extended cross validation study PCDA, PLR and SVM, performed equally well and PDV was almost as good. LB and NSC perform worse than the other four methods. Using a proper classification method, $82-90\%$ of the subjects were correctly classified.

Chapter 6 introduces an approach tailored to classify paired data. The approach is demonstrated in a cervical cancer proteomics data set. Squamous cell carcinoma antigen (SCC-ag) concentration in serum correlates with the stage of disease, the effect of treatment, and the development of disease, but it has poor predictive value. This study was initiated to find additional cervical cancer markers. Samples were obtained from cervical cancer patients at the time of diagnosis (case samples) and again on average 6 to 12 months after treatment when all patients appear to have recovered (control samples). Measuring the same patients after treatment as controls has an advantage over measuring a separate set of healthy individuals, since the biological variation in the data is reduced, increasing the chance of finding patterns related to disease rather than differences between individuals. The resulting data has a paired structure and a strategy for analysing paired data is proposed. This strategy is compaired to an unpaired strategy in four patient groups, one group of patients that relapse some time after the control sample is taken and three groups of recovering patients. In the relapsing patient group the performance is the same for both methods, while in the three groups with recovering patients classification performance improves using the paired analysis approach.

In Chapter 7 we revisit the question of selecting a suitable classification method. The four patient groups from the cervical cancer study in Chapter 6 are considered together, with the objective to find differences between recovering and relapsing patients. SVM and PCDA – two methods that in the previous chapters proved to be good classifiers of clinical proteomics data – are unable to obtain a good classification in this problem. The reason for this is the position of the classes: they are not disjoint (they overlap). Because

the within-class covariances are very different, Soft Independent Modelling of Class Analogy (SIMCA) is able to distinguish between the classes, using the residuals from the classes' PCA models. The difference between PCDA and SIMCA, two seemingly similar methods, can be seen in the metrics they use. Although they can be expressed in a similar fashion, different aspects of the data are stressed, resulting in very different performances. This example shows how choosing an appropriate classification method can improve classification performance.

Samenvatting

Dit proefschrift behandelt de analyse van data in klinische proteomics studies gericht op het ontdekken van biomarkers. Proteomics datasets zijn heel groot en bestaan vaak uit weinig monsters waarin veel eiwitten en peptiden zijn gemeten. De studies in dit proefschrift vergelijken verschillende patiëntengroepen (herstelde en recidiverende patiënten) of een groep patiënten met een groep gezonde controles. Vanwege de omvang van de data en de te verwachten kleine verschillen tussen de groepen zijn speciale strategieën voor de data-analyse noodzakelijk.

Hoofdstuk 2 geeft een overzicht van data-analyse strategieën die toegepast worden in de zoektocht naar biomarkers. Er bestaan veel methoden voor classificatie en variabelenselectie en de meest toegepaste worden besproken. Doordat de data complex en hoog-dimensioneel is, is het vaak gemakkelijk verschillen tussen twee groepen te vinden. Toch kan zo'n verschil op toeval berusten in plaats van op biologische verschillen. Het doel is een classificator of biomarker te ontwikkelen die ook nieuwe data goed kan voorspellen. Om dit te testen behoren validatiemethoden deel uit te maken van een goede data-analysestrategie. Veel strategieën die te vinden zijn in de literatuur passen binnen het modulaire raamwerk dat in dit hoofdstuk wordt gepresenteerd. Voorselectie van variabelen, classificatie, de biomarkerselectie en statistische validatie zijn losse modules in de analyse van proteomics data. De modules kunnen op verschillende manieren gecombineerd worden tot een strategie die past bij het data-analyse probleem. De modules voor variabelenselectie, classificatie en biomarkerselectie kunnen naar keuze gebruikt worden; het doel is dat ze tezamen een goede classificator vormen. Een goede data-analysestrategie bevat wel altijd de validatiemodules. De invulling van elke module varieert en is deels afhankelijk van de structuur van de data en deels van de expertise en voorkeuren van de data-analist.

In het tweede hoofdstuk presenteren we een strategie voor de statistische validatie van discriminatiemodellen in proteomics studies. Een proteomics studie

naar de ziekte van Gaucher, een lysosomale stapelingsziekte, dient als voorbeeld. De keuze voor dit voorbeeld is gemaakt, omdat bekend is dat de ziekte van Gaucher zeer grote veranderingen in het bloed teweegbrengt. Massaspectra van serum van patiënten en gezonde controles worden gemeten en vergeleken met behulp van Principal Component Discriminant Analysis (PCDA). De strategie is een combinatie van een permutatietoets, enkelvoudige en dubbele kruisvalidatie. Met de permutatietoets kan de mogelijkheid van een toevallig resultaat getoetst worden door de classificatiemethode toe te passen op gerandomiseerde data. Door de resultaten van classificator te vergelijken met de resultaten op gerandomiseerde data wordt een p-waarde verkregen. Met behulp van enkelvoudige kruisvalidatie wordt een PCDA model gekozen dat generaliseerbaar is naar nieuwe metingen. Soms is de gerapporteerde voorspelfout van een methode gebaseerd op de fout die in deze kruisvalidatie gemeten wordt. Deze fout is echter onzuiver, doordat de fout in de kruisvalidatie het criterium voor de modelselectie is. Op deze manier raken modelselectie en modelevaluatie vermengd. Met de permutatietoets kan de onzuiverheid in de fout zichtbaar gemaakt worden. Gemiddeld zal de kruisvalidatiefout in randomisaties namelijk afwijken van de verwachte 50% (voor twee even grote klassen). Door de gehele procedure van het modelbouwen te valideren wordt een zuivere schatting van de voorspelfout verkregen. In onze aanpak betekent dit, dat er een dubbele kruisvalidatie nodig is. De permutatietoets bevestigt inderdaad dat de dubbele kruisvalidatie een onafhankelijke schatting van de voorspelfout maakt. De sensitiviteit van de classificator is 89% en de specificiteit 90%.

De ziekte van Fabry is een andere lysosomale stapelingsziekte, waarvoor nog geen biomarker in het bloed bekend is. In Hoofdstuk 4 vergelijken we eiwitprofielen in het serum van patiënten en gezonde controles op dezelfde manier waarmee we in Hoofdstuk 3 onderscheid tussen Gaucher patiënten en controles konden maken. Het PCDA-model voor de classificatie van Fabry en controles geeft een hoge voorspelfout, ook na variabelenselectie. De voorspelfout die met Support Vector Machines (SVM) gemaakt wordt is weliswaar kleiner, maar nog wordt 16% van alle monsters fout geclassificeerd. Een mogelijke verklaring zou kunnen zijn dat de procedure voor het meten van de eiwitprofielen niet gevoelig genoeg is om veranderingen in het beginstadium van de ziekte van Fabry te meten. Echter, niet alleen worden sommige Fabry-patiënten worden als controle geclassificeerd, sommige gezonde controles worden tot de groep patiënten gerekend. Het is opvallend dat de drie gezonde familieleden van Fabry-patiënten alle zowel met PCDA als SVM als patiënt aangemerkt worden. Dit doet vermoeden dat de classificatie niet zo-

zeer gebaseerd is op kenmerken die samenhangen met de ziekte van Fabry, maar wellicht met kenmerken binnen de families. Dit laat zien hoe belangrijk het is dat de controlepersonen goed overeenkomen met de patiënten in dit soort studies.

In Hoofdstuk 2 worden veel classificatiemethoden besproken. In proteomics studies zoals beschreven in dit proefschrift is de keuze voor een bepaalde classificatiemethode van groot belang. In het vierde hoofdstuk passen we een aantal classificatiemethoden toe op de dataset uit Hoofdstuk 3 (ziekte van Gaucher). De validatiestrategie uit Hoofdstuk 3 wordt nu toegepast als een protocol waarmee kwaliteitsparameters verkregen worden op basis waarvan een geschikte methode gekozen kan worden. Naast PCDA en SVM worden ook Penalized Logistic Regression (PLR), LogitBoost (LB), Principal Discriminant Variates (PDV) en Nearest Shrunken Centroids (NSC) beschouwd. PCDA, PLR en SVM presteren allemaal even goed en PDV was ongeveer zo goed als voornoemde methoden. LB en NSC presteren opvallend slechter. Met de betere methoden blijkt het mogelijk om 82 tot 90% van alle monsters correct te classificeren.

Hoofdstuk 6 laat zien hoe classificatiemethoden aangepast kunnen worden voor gepaarde data. Deze aanpak wordt gedemonstreerd op een baarmoederhalskanker dataset. Voor baarmoederhalskanker bestaat reeds een marker, Squamous Cell Carcinoma Antigen (SCC-ag). SCC-ag correleert met het stadium waarin de kanker zich bevindt, het effect van de behadeling en de ontwikkeling van de ziekte, maar de voorspellende waarde is gering. De studie in dit hoofdstuk was opgezet om aanvullende markers te vinden. Daartoe werden plasmamonsters afgenomen bij baarmoederhalskankerpatiënten ten tijde van de diagnose en nogmaals 6 tot 12 maanden na behandeling wanneer alle patiënt genezen lijken. Door monsters van dezelfde personen na behandeling te gebruiken als controlemonster is de biologische variatie tussen de groepen klein, wat de kans op het vinden van ziektegerelateerde verschillen vergroot. Het resultaat is een datset met een gepaarde structuur. Er wordt een strategie voorgesteld om gepaarde data te analyseren. Vier groepen patiënten, één groep patiënten bij wie de kanker enige tijd na het afnemen van het controlemonster terugkeert en drie groepen herstelde patiënten, werden geclassificeerd in een gepaarde en een ongepaarde aanpak. In de groep met recidiverende kanker maakt het niet uit of in de analyse rekening gehouden wordt met de gepaarde structuur. In de drie groepen met herstelde patiënten presteert de gepaarde aanpak beter.

In Hoofdstuk 7 wordt nogmaals ingegaan op het selecteren van een geschikte classificatiemethode. In de baarmoederhalskanker dataset uit Hoofdstuk 6 wordt gezocht naar verschillen tussen herstellende patiënten en patiënten met recidiverende kanker. SVM en PCDA werkten goed in Hoofdstuk 6, maar hun prestatie op het herstel-recidive probleem is niet erg goed. Een reden hiervoor is dat de klassen overlappen. Soft Independent Modelling of Class Analogy (SIMCA) beschouwt onder meer de residuen van de PCA modellen van beide klassen om te classificeren en kan op basis daarvan wel onderscheid maken. PCDA en SIMCA lijken op elkaar, maar de metriek die ze gebruiken verschilt. Weliswaar kan de metriek voor beide methoden op eenzelfde manier uitgedrukt worden, maar PCDA en SIMCA benadrukken andere aspecten van de data. Hierdoor presteren ze zeer verschillend. Dit hoofdstuk illustreert hoe de keuze voor een geschikte methode de classificatie kan verbeteren.

Dankwoord

Het is af! De afgelopen jaren hebben mijn collega's, vrienden en familie mij direct en indirect geholpen en daar ben ik ze heel dankbaar voor. Een aantal van hen wil ik hier in het bijzonder noemen.

Age en Huub, bedankt voor de gelegenheid om in de BDA-groep onderzoek te doen. Tijdens onze regelmatige besprekingen kreeg ik altijd weer goede moed om het werk voort te zetten, danwel het allemaal net iets anders te gaan doen. Chris bracht ons in contact met Hans en Mariëlle van het AMC met wie we de data-analyse van SELDI-TOF-MS data hebben opgezet. Chris, Hans en Mariëlle, hartelijk bedankt voor de prettige samenwerking.

Toen ik nog maar net begonnen was, bracht Age een aantal mensen bij elkaar die ieder hun eigen methode mochten gebruiken om de Gaucher data zo goed mogelijk te classificeren. Margriet, Theo, Wies, Carina en Paul wil ik graag bedanken voor hun inzet en enthousiasme voor deze 'shootout'. Voor de beste classificatie was een fles wijn in het vooruitzicht gesteld, maar die heeft volgens mij nog niemand in ontvangst mogen nemen...

I thank the Analytical Biochemistry group of the University of Groningen for the work on Chapters 6 and 7, in particular Rainer and Peter for their help with preprocessing and writing.

In het dagelijks leven op de UvA word ik omringd door veel fijne collega's. Ten eerste is er natuurlijk de BDA-groep: Age, Andrew, Antoine, Chengjian, Diana, Edoardo, Eric, Eva, Ewoud, Gooitzen, Hans, Henk-Jan, Huub, Janko, Jeroen, Johan J., Johan W., Jos, Maikel, Marcel, Kilian, Olja, Robert, Serge, Susana, Tunahan, mijn promotie werd mede door jullie een tijd om met veel plezier op terug te kijken. Daniel, ik ben blij dat je mijn paranimf wilt zijn, terwijl je zelf druk bent om je onderzoek af te ronden. De andere collega's van SILS op het Roeterseiland wil ik bedanken voor de gezamenlijke werkbesprekingen en de gezellige koffiepauzes op de achtste verdieping. Dan nog mijn ICU-collega's, jullie aanmoedigingen als ik beneden kwam buurten deden me

altijd goed.

Lieve Rinke, Sara en Willemijn, ik ben heel blij met onze vriendschap en kijk altijd uit naar onze afspraken en etentjes, meestal heel lang omdat we er tegenwoordig agenda's voor moeten trekken. Met Martijn en Mattijs heb ik leuke 'Klussen met Hout' projecten volbracht, ondanks het hoge Buurman-en-Buurman gehalte. Bedankt voor de zaterdagen tekenen, zagen, meten, en nog eens zagen – in die volgorde. Judith, je zult af en toe wel gek zijn geworden van ons, bedankt dat je dat allemaal kunt hebben en zelfs nog iets lekkers haalt voor bij de koffie. Floor en Mattijs, bedankt voor veel gezellige uurtjes samen.

Fred en Pineke, Nicolien, bedankt voor de leuke tijd die we samen hebben, thuis en tijdens onze vakanties. Pap en mam, bedankt voor alles: jullie steun tijdens de studie, jullie hulp zelfs wanneer ik er niet om vraag, jullie belangstelling voor wat mij bezighoudt en bovenal jullie liefde. Inge, Johan, Gideon en David, Jos, Emma, Linde, Nathan, Yvonne, Simon, Melvin, Steffanie, bedankt voor jullie gezelligheid, belangstelling en liefde. Ik ben gek op jullie!

Een dagelijks hoogtepunt is aan het begin van de avond, als ik net met Fiona thuis ben. We kunnen Arjen zien aankomen op de fiets, lach van oor tot oor, omdat hij ons op de uitkijk ziet staan. Lieve Fiona, ik kan niet zeggen dat je het voltooien van dit proefschrift bespoedigd hebt, maar met je heerlijke kusjes en je uitgelaten blijdschap maak je dat meer dan goed. Arjen, mijn lief, zonder jou was ik wellicht niet aan dit project begonnen en zonder jou had ik het zeker niet af kunnen maken. Ik ben blij jou naast me te weten.