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

Citation for published version (APA):
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 confirms 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 compared 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.