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

Citation for published version (APA):

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

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