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 6

Optimal use of paired proteomics data†

We present a multilevel classification approach that uses pairing of samples in the data analysis. As an example we use a cervical cancer proteomics data set. Cervical cancer is one of the most common malignant diseases of women worldwide. Squamous cell carcinoma antigen (SCC-ag) is a serological tumour marker for patients with squamous cell carcinoma of the cervix. Concentration of SCC-ag in the serum correlates with the stage of disease, the presence or absence of risk factors, the effect of treatment, and the development of disease. SCC-ag level has, however, poor predictive value, especially at an early stage of the disease. It is therefore important to find additional (protein) markers.

To this end, blood samples were obtained from cervical cancer patients at the time of diagnosis (case samples) and again on average 6 to 12 months after treatment (control samples) and the serum proteome was analyzed by Liquid Chromatography - Mass Spectrometry after depletion of high-abundance proteins and trypsin digestion. Differences between cases and controls were extracted using multivariate classification techniques.

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 are paired and this can be exploited by employing paired data analysis schemes.

6.1 Introduction

Proteomics techniques are applied in clinical research to gain insight in the mechanism of a certain disease. One of the goals is developing tools for diagnosis in early stages of disease and prognosis of progression. Case and control samples are measured with complex analytical techniques, such as 2D gels, mass spectrometry or liquid chromatography coupled to mass spectrometry. The data sets produced by these techniques are large, and contain many isolated compounds (variables) and usually few samples. Differences in patterns between the case and control samples can be extracted using a large range of multivariate (classification) techniques.\textsuperscript{33,37,53} The biological variation between individuals can be large, obscuring smaller disease-related differences. Reducing biological variation increases the chance of finding patterns related to disease rather than individual differences. One way to decrease biological variation is by using different samples from the same individual, for example blood samples before and after treatment or samples of affected and unaffected tissue of the same organ.\textsuperscript{145}

Apart from the reduction in biological variation, measuring samples from the same individual has another benefit. The resulting data are paired and this can be exploited in their analysis. When only one variable is measured, the difference between two groups can be assessed with a t-test. In data with paired samples, using the pairing in the data in a paired t-test can show significant differences that might otherwise not be uncovered. In classification methods, paired analysis can also be implemented, in a way analogous to the multilevel analysis that Jansen et al. proposed for longitudinal data.\textsuperscript{146} In a study of rhesus monkeys, they used multilevel component analysis, expressing the urine NMR spectra of one animal with respect to its mean spectrum to investigate biorhythms in the metabolites. In this chapter we adopt a similar approach to improve classification performance in a cervical cancer proteomics study. Blood samples were obtained at the time of diagnosis with cervical cancer and again on average 6 to 12 months after treatment and serum was prepared. Proteins in serum, depleted of the six most abundant proteins, were measured with LC-MS after trypsin digestion.\textsuperscript{147}

Cervical cancer is one of the most common malignant diseases of women worldwide.\textsuperscript{148} It is much more common in developing countries, where 83% of cases occur and where cervical cancer accounts for 15% of female cancers, with a risk before age 65 of 1.5%.\textsuperscript{149} The highest incidence rates are observed
in sub-Saharan Africa, Melanesia, Latin America and the Caribbean, south-central Asia, and southeast Asia where facilities for prognosis and handling of the disease are inadequate. Squamous cell carcinoma antigen (SCC-ag) is a serological tumour marker for patients with squamous cell carcinoma of the cervix. Concentration of SCC-ag in serum correlates with the stage of disease, the presence or absence of risk factors, the effect of treatment, and the development of disease. The cut-off level at 1.9 ng/ml of SCC-ag is used to distinguish between diseased and healthy patients. Furthermore, cervical cancer patients with complete remission after treatment generally show low SCC-ag levels that remain low, while patients with recurrence continue to show elevated SCC-ag levels. In many cases, however, the measurement of the serum SCC-ag level by ELISA does not correlate with the presence of disease when it is still in an early stage and has thus poor predictive value. It does not contribute to better survival and is therefore not widely used for clinical follow-up of patients already treated for cervical cancer. Therefore, it is important to find alternative biomarkers for early diagnosis and for the clinical follow-up of cervical cancer patients after treatment. The patients included in our study were divided into four different groups based on the stage of cancer before treatment, the value of the SCC-ag marker, and the disease status 6 to 12 months after treatment. We compared classification performance with and without using pairing of the samples in the data analysis to show how classification can be improved upon. Two classification methods were used: Principal Component Discriminant Analysis (PCDA) and Support Vector Machines (SVM). As discussed in previous chapters, all results were statistically validated with (double) cross validation and permutation tests.

6.2 Data set

Serum samples from cervical cancer patients were obtained from the Department of Gynecological Oncology (University Medical Centre Groningen, The Netherlands). The stage of cervical cancer was determined according to the International Federation of Gynecology and Obstetrics (FIGO) classification system. After diagnosis of cervical cancer the patients received the following therapy: primary operation, radiochemotherapy, radiotherapy, uterus operation. The patients were selected in order to form four groups with different stages of disease before treatment and different disease status after treatment (see Table 6.1) with the objective of finding discriminating proteins for early diagnosis of cervical cancer. To this end, cases in which the level of the
Table 6.1: Overview of patient groups. The groups are characterized by the stage of cancer, the level of SCC-ag at the time of diagnosis (SCC-ag A) and after the treatment when patients seem recovered (SCC-ag B) and the disease status long term after treatment (status time C).

<table>
<thead>
<tr>
<th>Group</th>
<th>stage</th>
<th>SCC-ag A</th>
<th>SCC-ag B</th>
<th>status time C</th>
<th>group size</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>early</td>
<td>high</td>
<td>low</td>
<td>recovered</td>
<td>10</td>
</tr>
<tr>
<td>II</td>
<td>early</td>
<td>low</td>
<td>low</td>
<td>recovered</td>
<td>10</td>
</tr>
<tr>
<td>III</td>
<td>advanced</td>
<td>high</td>
<td>low</td>
<td>recovered</td>
<td>10</td>
</tr>
<tr>
<td>IV</td>
<td>various</td>
<td>high</td>
<td>low</td>
<td>relapsed</td>
<td>12</td>
</tr>
</tbody>
</table>

Known marker SCC-ag corresponds with the presence of disease (higher than the threshold of 1.9 ng/ml) (Group I, 10 patients) and where not (Group II, 10 patients) were selected. All patients in Groups I and II were diagnosed with stage I cervical cancer. Group III (10 patients) was used as a positive control and consisted of advanced stage cervical cancer (stage III and IV) with corresponding high levels of SCC-ag before treatment. The patients in Groups I - III all recovered after treatment. Samples were taken before treatment (time point A) and 6 to 12 months after treatment when the patients had recovered (time point B). The 12 patients in Group IV were at stage I, II and III of cervical cancer, having SCC-ag levels higher than the threshold before treatment (time point A). After treatment, at time point B, the patients were symptom-free, but patients in this group relapsed long term after treatment (time point C). Patients in this group showed low SCC-ag levels at time point B but had elevated SCC-ag levels at time point C (recurrent disease). SCC-ag analysis at time point B had no predictive value whether patients were cured or were at risk for recurrence of cancer.\textsuperscript{151} The study protocol was in agreement with local ethical standards and the Declaration of Helsinki of 1964, as revised in 2004.

Samples were depleted of high abundant proteins on a Multiple Affinity Removal column (Agilent) according to the manufacturers instructions on a Beckman Gold HPLC system and digested with trypsin.\textsuperscript{147} The depleted, trypsin-digested serum samples were analyzed with an 1100 series HPLC system (Agilent) containing an Atlantis\textsuperscript{TM} dC18 in-line trap column (Waters) embedded in a Universal Sentry Guard Holder assembly (Waters) and an analytical capillary column of the same material (Waters) coupled on-line to an MSD-Trap-SL ion trap mass spectrometer (Agilent). The analysis is described
in detail previously. Within each patient group the samples were analyzed in randomized order. After data reduction in the m/z mode and time alignment (with correlation optimized warping), peaks were detected and matched between data files of one group. The generated peak matrix, created from the peak lists of the individual samples, consisted of a peak(row)-sample(column)-intensity(value) matrix. This peak matrix, $X$, was used for multivariate statistical analysis. LC-MS analysis, data processing (time alignment, peak picking) and statistical analysis were done separately for each of the four groups, resulting in different numbers of peak intensities per group.

### 6.3 Data analysis

**Paired samples**

The current multivariate data set has a paired nature: samples are taken from the same patient before and after treatment. We want to adapt the data analysis to the paired structure of the data in much the same way this is done in univariate data. As an example, Figure 6.1 A shows the univariate data for a group of cases and a group of controls. The values of the groups overlap largely and a t-test would show that they do not differ significantly. In Figure 6.1 B, the values are the same as in Figure 6.1 A, but now they concern patients before and after treatment. By considering the measurements per patient, we see that the values are increased after treatment. A paired t-test, where the link between the measurements is considered, shows that the group after treatment is in fact different from the group before treatment.

In the data analysis, we are looking for similar trends for different patients, rather than for differences between two groups. In Figure 6.1 C we centered the measurements per patient (subtract the patient’s mean from both measurements). Now, the controls will all be on comparable levels and so will the cases, making the differences between classes more pronounced. The trend then becomes obvious: before treatment all values are negative, after treatment all are positive.

The pairing of samples can also be explicitly exploited in the analysis of multivariate (proteomics) data. The samples taken before treatment will be viewed as cases and the samples after treatment as controls. In order to compare cases and controls, the peak intensities of a patient are centered (see Figure 6.2). The
matrix $X$ on the left hand side represents the data matrix of one of the Groups I-IV, arranged by patient. On the right hand side are two matrices: $X_{\text{between}}$ contains the mean peak intensities of each patient and $X_{\text{within}}$ contains the differences between the peak intensities and the mean of the patient. The vector $Y$ contains the class labels: $-1$ represents before treatment and $1$ represents after treatment.

Classifiers

Support Vector Machines
Support Vector Machines (SVM)\textsuperscript{154} are used to find a hyperplane to separate the cases from the controls in variable space. In this paper, we use a linear kernel. If the classes overlap, some objects will be on the wrong side of the hyperplane (misclassification). The extent to which objects can be on the wrong side of the hyperplane is bound by a penalty $\gamma$. A high value for $\gamma$ means it is very costly to cross the hyperplane; this may result in overfitting of the data. Small values for $\gamma$ may lead to hyperplanes that are not very effective in separating the classes. We use the misclassification rate in an inner cross validation loop to tune $\gamma$ to build a model that is generalizable to new objects.\textsuperscript{13,154}

Principal Component Discriminant Analysis
PCDA is a combination of Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA).\textsuperscript{30} The method was described in detail in Chapter 3. In short, PCA is first applied to the data to reduce the dimensionality. The PCA scores are then used in LDA to find a direction that discriminates
Figure 6.2: Splitting of the original data $X$ into levels of individual compounds per patient ($X_{\text{between}}$) and fluctuations of these levels around their mean ($X_{\text{within}}$).

between the two groups. The number of components in the PCA step is regulated by internal cross validation to minimize the number of misclassifications.$^{33}$

Validation

Prediction error
The comparison between using $X$ and $X_{\text{within}}$ is made based on the prediction error of the PCDA and SVM models. We calculate the prediction error as the misclassification rate in a tenfold double cross validation scheme (twelvefold for Group IV). In each fold, both samples of one patient form the test set, and the model is trained on the remaining samples. Training of the models includes tuning of the meta-parameters (the number of PCs in PCDA and $\gamma$ in SVM). Tuning of the meta-parameters is also done by cross validation, which gives an error associated with each chosen value for the meta-parameter. Because this cross validation first uses all of the data before choosing the meta-parameter, the cross validation error is not based on independent data. Hence, this cross validation error is optimistically biased and is therefore an inappropriate measure of the prediction error.$^{94,95}$ In double cross validation the tuning of the meta-parameters is done in an internal cross validation loop. An external cross validation loop, in which new samples are predicted using
the meta-parameter from the inner loop, is applied to find an independent estimate of the prediction error.\textsuperscript{94,95}

**Permutations**

The p-values of the prediction error are determined using permutation tests in which the labels of the data are randomized. The randomized data sets thus created are subjected to the same data analysis as the original data, including double cross validation to estimate the prediction error. In unpaired data the labelling of the data in the permutations is completely free, provided that the classes remain the same size as in the original data. Due to the paired structure of the data, the possibilities for permutation are limited. The spectra of one patient form a pair, so if in a permutation one of the spectra is given the label "case", then automatically the other is given the label "control".

**Normalization and scaling**

All data were normalized by dividing the intensities by their median intensities, making the measurements of the samples comparable. Thereafter, the $X$ matrix is auto scaled: all variables have zero mean and unit variance. In auto scaled data, the contribution of a variable to the classification model is not dependent on the intensity of the signal, but on the relative difference in signal intensity between the classes.

**6.4 Results**

For each of the Groups I to IV the cases and controls are predicted with PCDA and SVM with a linear kernel in a tenfold double cross validation (twelvefold for Group IV) in two ways: 1) without taking the paired structure of the data into account, using $X$, and 2) using the paired nature of the samples, using $X_{\text{within}}$. In each fold of the double cross validation scheme, both samples from one patient (case and control) form the test set. The results are given in Table 6.2. For Groups I, II and III, using $X_{\text{within}}$ leads to fewer misclassifications than using $X$. When using $X_{\text{within}}$, if the case sample of a patient is misclassified then the control sample will also be misclassified, and vice versa. This is caused by the construction of $X_{\text{within}}$: by centering the data per patient, the peak intensities in the case sample of a patient become the negative of the peak intensities of the control sample.
In Groups I to III the medical treatment was successful and the patients were cured. The misclassification rate on $X_{\text{within}}$ decreases from 40% for Group I (early stage, low SCC-ag before treatment) to 10% for Group III (advanced stage, high SCC-ag before treatment) with PCDA and with SVM. The improved classification performance corresponds to expected larger differences in the samples taken before and after treatment due to a more advanced stage cancer in Group III patients.

We performed exhaustive permutation tests for all groups. For Groups I, II and III 512 possible permutations exist and for Group IV 2048 permutations. The p-values obtained in the permutation tests using $X$ and $X_{\text{within}}$ are given in Table 6.2. The p-value for the results in Group III using $X_{\text{within}}$ shows that classification is significant (p=0.04 with PCDA and p=0.03 with SVM) due to the paired data analysis.

Group IV is different from Groups I, II and III. The criteria for inclusion in one of the Groups I, II or III are very well defined (SCC-ag level, stage of cancer), resulting in homogeneous groups of patients. Group IV, however, consists of patients who relapse some time after treatment, regardless of the stage of cancer or the concentration of SCC-ag and is consequently much more diverse. The misclassification rate is 42% for $X$ and $X_{\text{within}}$ with PCDA and 25% with SVM. Apparently, the differences between the samples before and after treatment are very small and classification is difficult. This is in accordance with the fact that the patients did not truly recover, despite the fact that the SCC-ag concentration was low after treatment. The p-values for $X$ are much lower than for $X_{\text{within}}$, with SVM the classification result is significant. We can partly explain this by realizing that the number of misclassifications in $X$ can be even or odd, while in $X_{\text{within}}$ it can be only an even number. This also holds in permutations of $X$ and $X_{\text{within}}$. Consequently, the misclassification distribution for $X_{\text{within}}$ is coarser and the p-value larger.

### 6.5 Conclusions

Even though the data sets contain few samples, we can draw some conclusions with respect to the use of $X_{\text{within}}$ based on the results with the PCDA and SVM classifiers. With both classifiers, the prediction error is smaller in Group III than in Groups I and II. This is in line with the expectation that the differences between patient sera before and after treatment become larger.
Table 6.2: Classification results and corresponding p-values.

<table>
<thead>
<tr>
<th></th>
<th>Group I</th>
<th>Group II</th>
<th>Group III</th>
<th>Group IV</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PCDA</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>error</td>
<td>45%</td>
<td>40%</td>
<td>40%</td>
<td>42%</td>
</tr>
<tr>
<td>p</td>
<td>0.40</td>
<td>0.22</td>
<td>0.33</td>
<td>0.28</td>
</tr>
<tr>
<td>error</td>
<td>40%</td>
<td>20%</td>
<td>10%</td>
<td>42%</td>
</tr>
<tr>
<td>p</td>
<td>0.40</td>
<td>0.11</td>
<td>0.04</td>
<td>0.42</td>
</tr>
<tr>
<td><strong>SVM</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>error</td>
<td>45%</td>
<td>35%</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>p</td>
<td>0.43</td>
<td>0.17</td>
<td>0.09</td>
<td>0.03</td>
</tr>
<tr>
<td>error</td>
<td>40%</td>
<td>20%</td>
<td>10%</td>
<td>25%</td>
</tr>
<tr>
<td>p</td>
<td>0.38</td>
<td>0.11</td>
<td>0.04</td>
<td>0.12</td>
</tr>
</tbody>
</table>

based on the stage of the cancer and the level of SCC-ag before treatment. Discrimination of pre- and post-treatment samples in Group IV is difficult, likely because the patients did not really recover. We show that taking advantage of paired sampling by classifying using \( X_{\text{within}} \) results in lower prediction errors than assuming that all samples are independent (X). An exception are the results for Group IV, where the prediction error for X and \( X_{\text{within}} \) is the same.