Feasibility of Serodiagnosis of Ovarian Cancer by Mass Spectrometry

ARTICLE *in* ANALYTICAL CHEMISTRY · FEBRUARY 2009

Impact Factor: 5.64 · DOI: 10.1021/ac802293g · Source: PubMed

7 AUTHORS, INCLUDING:

Rasmus Bro
University of Copenhagen
272 PUBLICATIONS **11,169** CITATIONS
SEE PROFILE

Federico Marini
Sapienza University of Rome
104 PUBLICATIONS **1,179** CITATIONS
SEE PROFILE

Niels H H Heegaard
Statens Serum Institut
201 PUBLICATIONS **4,623** CITATIONS
SEE PROFILE

Anal. Chem. **2009,** *81,* 1907–1913

Feasibility of Serodiagnosis of Ovarian Cancer by Mass Spectrometry

Mikkel West-Nørager,*,† Rasmus Bro,‡ Federico Marini,§ Estrid V. Høgdall,|| Claus K. Høgdall,⊥ Lotte Nedergaard,⊗ and Niels H. H. Heegaard†

Department of Clinical Biochemistry and Immunology, Statens Serum Institut, DK-2300 Copenhagen S, Denmark, Department of Food Science, University of Copenhagen, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark, Dipartimento di Chimica, Università di Roma "La Sapienza", P.le Aldo Moro 5, 00185 Rome, Italy, Department of Virus, Hormones, and Cancer, Institute of Cancer Epidemiology, Danish Cancer Society, 2100 Copenhagen Ø, Denmark, The Gynecological Clinic, The Juliane Marie Center, Rigshospitalet, Copenhagen University Hospital, 2100 Copenhagen Ø, Denmark, and Department of Pathology, Centre of Diagnostic Investigations, Rigshospitalet, Copenhagen University Hospital, 2100 Copenhagen Ø, Denmark

The emergence of new biological disease markers from mass spectrometric studies of serum proteomes has been quite limited. There are challenges regarding the analytical and statistical procedures, preanalytical variability, and study designs. In this serological study of ovarian cancer, we apply classification methods in a strictly designed study with standardized sample collection procedures. A total of 265 sera from women admitted with symptoms of a pelvic mass were used for model building. We developed a rigorous approach for building classification models suitable for the highly multivariate data and illustrate how to evaluate and ensure data quality and optimize data preprocessing and data reduction. We document time dependent changes in peak profiles up to 15 months after sampling even when storing samples at −20 °C. The developed classification model was validated using completely independent samples, and a cross validation procedure which we call cross model validation was applied to get realistic performance values. The best models were able to classify with 79% specificity and 56% sensitivity, i.e., an analytical accuracy of 68%. However, the existing serum marker (CA-125) alone gave a better analytical accuracy (81%) in the same sample set. Also, the combination of mass spectrometric data and levels of CA-125 data did not improve the predictive performance of models. In conclusion, proteomic approaches to biomarker discovery are not necessarily yielding straightforward diagnostic leads but lay the foundation for more work.

The concept of discovering new diagnostic markers on the basis of mass spectrometric (MS) analyses of biofluids has been advocated strongly during the last 5−8 years.[1−6] However, despite many efforts, no new biomarkers originating from such studies have yet been approved for routine use in clinical settings.[7−12] Typical obstacles preventing the transfer of proteomics discoveries into useful diagnostic tools are deficiencies in study designs (including sample handling and sample types), incomplete or missing results validation, preanalytical and analytical variability, and difficulties with feature selection, improper statistical methods, as well as insufficient clinical data and uncertain clinical relevance. Compared with traditional clinical biochemistry, the sample complexity together with the enormous dynamic range (concentration differences) among the multiple analytes in a biological matrix such as blood is a challenge to study design, analytical methods, and data handling.[13]

Different solutions to the problem with the wide dynamic range in biological fluids have been proposed. One approach is to remove the most abundant proteins from the samples before analysis. In contrast, other approaches examine some of the most abundant proteins for their biomarker cargo by eluting and characterizing bound ligands.[14−16] Few of these sample preparation methods have undergone a systematic documentation of reproducibility and

* Corresponding authors. Statens Serum Institut, 81/544, Artillerivej 5, DK-2300 Copenhagen, Denmark. Phone: +45 32683268. Fax: +45 32683876. E-mail: mwn@ssi.dk (M.W.-N.); nhe@ssi.dk (N.H.H.H.).

† Statens Serum Institut.

‡ University of Copenhagen.

§ Università di Roma "La Sapienza".

|| Danish Cancer Society.

⊥ The Gynecological Clinic, The Juliane Marie Center, Rigshospitalet, Copenhagen University Hospital.

⊗ Department of Pathology, Centre of Diagnostic Investigations, Rigshospitalet, Copenhagen University Hospital.

(1) Carrette, O.; Demalte, I.; Scherl, A.; Yalkinoglu, O.; Corthals, G.; Burkhard, P.; Hochstrasser, D. F.; Sanchez, J. C. *Proteomics* **2003**, *3*, 1486–1494.
(2) Conrads, T. P.; Zhou, M.; Petricoin, E. F., III.; Liotta, L.; Veenstra, T. D. *Expert Rev. Mol. Diagn.* **2003**, *3*, 411–420.
(3) Li, J.; Zhang, Z.; Rosenzweig, J.; Wang, Y. Y.; Chan, D. W. *Clin. Chem.* **2002**, *48*, 1296–1304.
(4) Paweletz, C. P.; Trock, B.; Pennanen, M.; Tsangaris, T.; Magnant, C.; Liotta, L. A.; Petricoin, E. F., III. *Dis. Markers* **2001**, *17*, 301–307.
(5) Petricoin, E. F.; Ardekani, A. M.; Hitt, B. A.; Levine, P. J.; Fusaro, V. A.; Steinberg, S. M.; Mills, G. B.; Simone, C.; Fishman, D. A.; Kohn, E. C.; Liotta, L. A. *Lancet* **2002**, *359*, 572–577.
(6) Petricoin, E. F., III.; Ornstein, D. K.; Paweletz, C. P.; Ardekani, A.; Hackett, P. S.; Hitt, B. A.; Velassco, A.; Trucco, C.; Wiegand, L.; Wood, K.; Simone, C. B.; Levine, P. J.; Linehan, W. M.; Emmert-Buck, M. R.; Steinberg, S. M.; Kohn, E. C.; Liotta, L. A. *J. Natl. Cancer Inst.* **2002**, *94*, 1576–1578.
(7) Diamandis, E. P. *Clin. Chem.* **2003**, *49*, 1272–1275.
(8) Diamandis, E. P. *Clin. Chem.* **2006**, *52*, 771–772.
(9) Sorace, J. M.; Zhan, M. *BMC Bioinf.* **2003**, *4*, 24.
(10) Xu, R.; Gamst, A. *J. Natl. Cancer Inst.* **2005**, *97*, 1226.
(11) Barker, P. E.; Wagner, P. D.; Stein, S. E.; Bunk, D. M.; Srivastava, S.; Omenn, G. S. *Clin. Chem.* **2006**, *52*, 1669–1674.
(12) Omenn, G. S. *Proteomics* **2006**, *6*, 5662–5673.
(13) Anderson, N. L.; Anderson, N. G. *Mol. Cell. Proteomics* **2002**, *1*, 845–867.
(14) Geho, D. H.; Jones, C. D.; Petricoin, E. F.; Liotta, L. A. *Curr. Opin. Chem. Biol.* **2006**, *10*, 56–61.

10.1021/ac802293g CCC: $40.75
Published on Web 01/29/2009

sensitivity toward in vitro processing artifacts.[17−19] Despite these reservations, MS is the only viable approach for comprehensive analysis of low-molecular weight proteins and peptides from biological fluids and thus in theory has a great potential for discovering biomarkers that are not accessible for analysis by other techniques.[20]

We have previously addressed issues related to sample handling and processing in an ongoing study searching for new ovarian cancer biomarkers in sera. Using normal blood samples, we applied a simple, robust, and reproducible prefractionation protocol based on chemically derivatized magnetic beads to retrieve and analyze subproteomes of sera. With specified experimental guidelines as well as standardized sample collection procedures we showed that it was possible to obtain quite reproducible peak intensities and peak positions in serum mass profiling using MALDI-TOF MS. To prune MS-data as much as possible for nonrelevant, nonbiological variation, the additional development of standardized protocols for analytical parameters and for handling the MS data was required and gave knowledge about how different preanalytical and analytical parameters contributed to the outcome of MS-based proteomic analyses.[17,18,21]

We here apply the methodological optimization reported in our previous studies to carefully collected serum samples from patients with ovarian cancer or benign pelvic conditions. A total of 265 sera are analyzed in the present study using (Cu^{2+})-immobilized metal affinity chromatography (IMAC)-magnetic beads for sample preparation. We show how MS data can be used to evaluate sample quality as a first step required for reliable pruning of the data for discriminatory peaks. Furthermore, we demonstrate how easy it is to be misled by classification models if possible confounding effects are not taken properly into account. We propose the use of a new cross validation method (cross model validation) which yields more realistic model performance measures in multivariate data analysis than simple cross validation. Our analyses gave a model based on peak patterns that discriminate malignant from benign conditions in about 68% of the cases at 56% sensitivity and 79% specificity levels. This performance was not better than when the same samples were analyzed using the standard ovarian cancer marker CA-125 alone, and combinations with MS-data were not capable of improving the performance. However, despite the inability to boost diagnostic performance, the data obtained here are encouraging because peak patterns with some specificity emerged and thus the results may be a useful starting point in the discovery of novel ovarian cancer biomarkers.

MATERIALS AND METHODS

The Pelvic Mass Study. The Pelvic Mass study is a Danish prospective study of ovarian cancer, covering biochemistry and molecular biology with the purpose of identifying prognostic factors as well as factors that differentiate benign and malignant conditions. Samples originate from women referred to an outpatient clinical because of symptoms of a pelvic mass. The study has been approved by the Science Ethics committees in the study area, KF01-227/03 and KF01-143/04.

Biological Material and Study Design. From October 2004, all women admitted to the Gynecological Clinic at Rigshospitalet, Copenhagen, with a pelvic mass or pelvic pain where exploratory surgery was planned were included in the prospective study. Blood samples were taken less than 15 days before surgery. All blood samples were collected and handled according to the conclusions of a previous study.[17] Handling and treatment of the blood samples were performed within one working day (i.e., less than 6 h). Thus, blood samples were left to clot at room temperature, and serum was isolated by centrifugation at 2000*g* for 10 min. All biological materials were stored at −80 °C in aliquots until analyses were performed.

A total of 134 patients with ovarian cancers (all epithelial) (mean age 65), 39 with ovarian borderline tumors (mean age 54), and 396 benign pelvic mass patients (mean age 45) were enrolled as of April 2007. Preoperative blood samples as well as tumor tissue samples were obtained from all patients. The samples included in the present study consisted of the first 335 obtained sera. The final diagnosis was based on evaluation of tumor tissues by pathologists. All histopathological classifications of the ovarian tumors were based on the typing criteria of the WHO.[22] Clinical and pathologist-examined information was obtained from the Danish Gynecological Cancer Database. In this study we used the information about the benign or malignant status of the tumor.

Sample Logistics. Samples were collected over a period of 23 months as described above. Samples were received in three different batches (October 2005 (batch 1), March 2006 (batch 2), and August 2006 (batch 3)). In order to automate the sample preparation, the 335 samples were divided randomly and without knowledge of diagnoses into four different sample preparation runs using a 96 well-sample robot.

About 8% of the analyzed samples (25 samples) were rejected by the automated acquisition feature of the FlexControl software because of poor spectrum quality using parameters described in the section about MALDI-TOF MS. A total of 45 samples representing borderline tumors together with samples from other cancer groups were omitted in the further analysis presented in this paper (cf. Scheme 1 for an overview of the samples). This left a total of 265 samples of which 65 represented malignant cases to be included in the final data analysis. The group of malignant samples included samples from stage I−IV of ovarian cancer, distributed among various stages as follows: stage I, 10; stage II, 7; stage III, 34; and stage IV, 14.

Materials. The purification kit (IMAC-Cu magnetic beads) and matrix α-cyano-4-hydroxycinnamic acid (HCCA) were purchased from Bruker Daltonics GmbH (Leipzig, Germany).

Clinical Biochemistry. CA-125 values were determined for all the samples using standard procedures.[23]

(15) Lowenthal, M. S.; Mehta, A. I.; Frogale, K.; Bandle, R. W.; Araujo, R. P.; Hood, B. L.; Veenstra, T. D.; Conrads, T. P.; Goldsmith, P.; Fishman, D.; Petricoin, E. F., III.; Liotta, L. A. *Clin. Chem.* **2005**, *51*, 1933–1945.

(16) Mehta, A. I.; Ross, S.; Lowenthal, M. S.; Fusaro, V.; Fishman, D. A.; Petricoin, E. F., III.; Liotta, L. A. *Dis. Markers* **2003**, *19*, 1–10.

(17) West-Nielsen, M.; Hogdall, E. V.; Marchiori, E.; Hogdall, C. K.; Schou, C.; Heegaard, N. H. *Anal. Chem.* **2005**, *77*, 5114–5123.

(18) West-Norager, M.; Kelstrup, C. D.; Schou, C.; Hogdall, E. V.; Hogdall, C. K.; Heegaard, N. H. *J. Chromatogr., B: Anal. Technol. Biomed. Life Sci.* **2007**, *847*, 30–37.

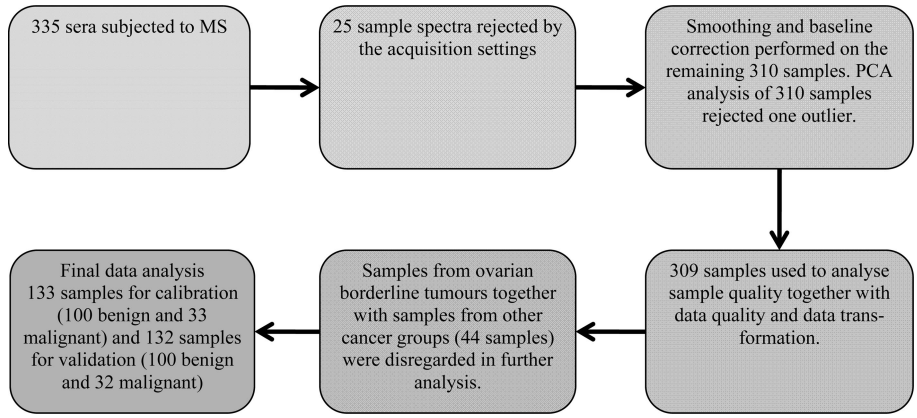
(19) Villanueva, J.; Lawlor, K.; Toledo-Crow, R.; Tempst, P. *Nat. Protoc.* **2006**, *1*, 880–891.

(20) Heegaard, N. H. H. *Expert Opin. Med. Diagn.* **2008**, 1–4.

(21) Villanueva, J.; Philip, J.; Entenberg, D.; Chaparro, C. A.; Tanwar, M. K.; Holland, E. C.; Tempst, P. *Anal. Chem.* **2004**, *76*, 1560–1570.

(22) Scully, R. E. *Natl. Cancer Inst. Monogr.* **1975**, *42*, 5–7.

Scheme 1



Serum Prefractionation. Paramagnetic immobilized metal(Cu) affinity chromatography nonporous particles (MB-IMAC-Cu) were used for preparing subfractions of serum samples. The manual chromatographic fractionation procedure described by the manufacturer adapted to a ClinProt sample preparation robot was used. Briefly, 10 μL of MB-IMAC-Cu binding solution and 5 μL of serum were transferred to a 0.2 mL thin-walled PCR-tube (ABgene, U.K.). A volume of 5 μL of a homogeneous magnetic particle solution were added, mixed, and left for 1 min. In preliminary experiments, we did not observe better performance in terms of number of peaks using up to 10 μL of serum with 5 μL of IMAC-Cu beads. To preserve as much serum as possible, we decided to use only 5 μL throughout the study. After incubation, the tubes were placed in a 2 × 8 well magnetic bead separator (MBS) (Bruker Daltonik, Germany) for 30 s for magnetic fixation of the Cu-IMAC particles. The supernatant was discarded, and the tubes were removed from the MBS device. A volume of 100 μL of wash solution (unknown composition) (Bruker Daltonics, Germany) was added and carefully mixed with the magnetic beads. The tube was then replaced in the MBS device and moved back and forth sequentially between adjacent wells on each side of the magnetic bar in the MBS device for a total of 20 times. After fixation of the magnetic beads for 30 s in the MBS device, the supernatant was removed. The total washing procedure was repeated three times. After the final washing step, bound molecules were eluted by incubation with 10 μL of elution solution (TFA based solution) for 1 min before collecting the elution solution using the MBS device to remove the beads.

The eluent (1 μL) was then mixed with 10 μL of matrix solution (0.3 g/L HCCA in ethanol/acetone 2:1), and 1 μL of this solution was spotted onto a 600 μm diameter spot size 384 AnchorChip target plate (Bruker Daltonics, Germany) and left to dry. A volume of 0.5 μL of the peptide calibration standard was applied to target spots in close proximity to the serum samples for external calibration of the instrument. All operations were carried out in an automated sampling handling robot (ClinProt robot, Bruker Daltonics, Germany). Air humidty and temperature were controlled with temperature set to 22 ± 1 °C and humidity set to 35 ± 3%. Samples were analyzed randomly and blinded.

MALDI-TOF MS. The AnchorChip target plate was placed in an UltraFlex MALDI-TOF mass spectrometer controlled by FlexControl software v. 2.0. The instrument is equipped with a 337 nm nitrogen laser, delayed-extraction electronics, and a 2 GHz digitizer. The instrument was initially externally calibrated by standard procedures. Data were generated by an automated acquisition method included in the instrument software based on averaging of 150 randomized shots over 5 positions (30 shots/position). The acquisition laser power was set between 25−35%. Before each acquisition cycle the target position was pretreated with 10 laser shots at 40% laser power to improve spectra quality. The automated acquisitions were given the following spectral evaluation parameters: cutoff value for the signal-to-noise ratio, S/N = 15, and spectral resolution set to >200. Spectra that did not fulfill these specifications were rejected. Spectra were acquired in positive linear mode geometry below 20 kV of ion acceleration and with ion selector deflection of mass ions >900 m/z in the mass range 1000−10 000 Da. Pulsed ion extraction was set to 320 ns to ensure appropriate time lag focusing. Samples were randomized and blinded throughout the entire experiment, from collection point to the mass spectrometric serum profiling, and were analyzed in four batch runs as described above.

Chemometrics. Spectra were baseline-corrected and converted into ASCII file format using the FlexAnalysis software (Bruker, Daltonics). The converted data were subsequently analyzed by using both commercially available analysis programs and freely downloadable toolboxes running under Matlab (The MathWorks, Inc.) environment. Specifically, Unscrambler v. 8.05 (Camo, Oslo, Norway) was used for principal component analysis (PCA) and partial least-squares regression (PLS). Additionally, partial least squares discriminant analysis, PLS-DA, was used for classification through PLS_Toolbox 4.1.[24] To find the optimal number of PLS components to be used, cross-validation using randomly selected subsets was used with number of data splits set to 10 and averaging results over 15 repeated cross-validations.

Variable Reduction and Selection. The MALDI-TOF MS spectra contain roughly 33 000 data points (variables) but these 33 000 variables do not represent 33 000 independent chemical variations. Because of the correlated nature of the measured spectra, there is a high degree of correlation between neighbor variables. A sort of data reduction that seeks to remove this redundancy is very beneficial because it reduces the risk of

(23) Kenemans, P.; van Kamp, G. J.; Oehr, P.; Verstraeten, R. A. *Clin. Chem.* **1993**, *39*, 2509–2513.

(24) Wise, B. M.; Gallagher, N. B.; Bro, R.; Shaver, J. M.; Windig, W.; Koch, R. S. *PLS_Toolbox*, version 4.1; Eigenvector Research Inc.: Wenatchee, WA, 2006.

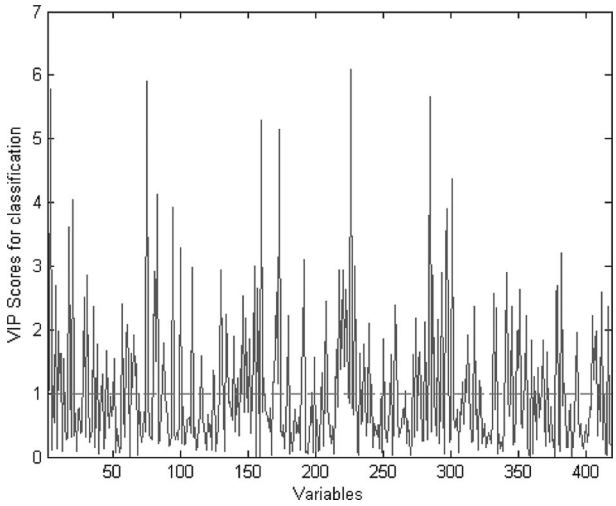


Figure 1. PLS-DA variable picking. The variable importance in projection (VIP) scores are shown for every variable used in the model (calibration sample set). Variables with scores above the dotted line are regarded as important variables, while variables below the line can be disregarded.

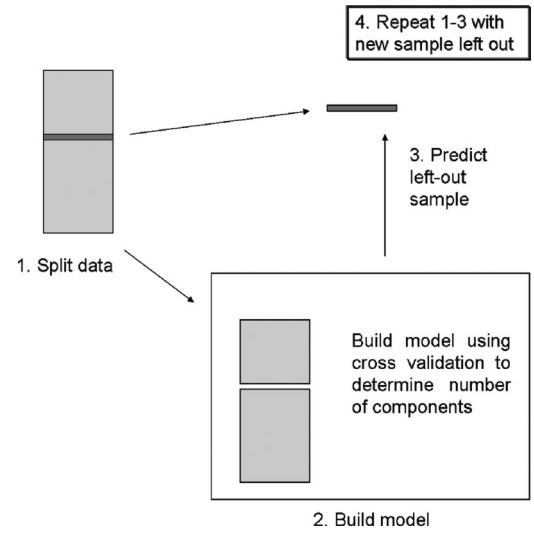


Figure 2. Cross model validation (CMV). Schematic overview of the different steps in CMV. The first step is splitting the data into a calibration and test set. The second step is building a model using normal cross validation for optimizing the model. The third step is testing the model using the left out test set and finally repeating steps 1−3 until every sample (step 1) has been left out once.

overfitting, i.e., conferring significance to minor, spurious correlations. This is particularly important when performing variable selection. Data reduction was achieved manually from an overlay of all 265 spectra, choosing a start and end point for each observed peak thereby removing any part of the spectra not containing peaks. A peak was defined as having a Gaussian shape with the maximum peak height well above baseline level with a signal-to-noise ratio above 3. This procedure reduced the amount of data points by 50%, and in total a master list of 117 peaks defined by m/z intervals were obtained. For each interval, a principal component analysis (PCA) was then calculated. If a peak represents only one underlying feature, it is expected that one principal component can explain the variation, which means that the initial many variables representing the peak can be reduced to a single number which is basically the peak area. If the peak contains information from several chemical compounds, then more than one component may be needed, but the number of components will always be dramatically lower than the number of initial variables. Hence representing each peak by principal components further substantially reduces the degree of redundancy in the data. The entire data reduction procedure reduces the number of variables from more than 30 000 to about 500 per sample. It is likely that part of the variables (peaks) does not have classifying power. To test if further variable selection would improve model prediction, PLS-DA models were made using stepwise variable selection. Variable selection was based on using variable importance in projection (VIP)[25] scores that estimate the importance of each variable used in the PLS-DA model. VIP scores, calculated by the PLS_Toolbox 4.1[24] close to or greater than 1 can be considered important and variables with VIP scores significantly less than 1 are less important and may be good candidates for exclusion from the model, see Figure 1 and the Results and Discussion.

Cross Model Validation. With a limited number of samples to be used for both the model calibration and model test set, the normal strategy in many cases will be cross-validation (CV). However, only using CV will often lead to too optimistic classification results especially when model building involves several steps. For example, when specific variables are selected to improve classification results, relying only on cross-validation will lead to overfitted models. Still, the result of the overfitting will not usually be observed because the variables of importance are estimated using the CV and hence improving both the model and the cross validation errors, i.e., validation is not independent of model building. We therefore propose to estimate model performance using a cross model validation (CMV)[26] strategy (Figure 2). This approach is able to give an accurate estimate of what would be obtained using true independent validation. In CMV, one data set (one sample) is put aside and the remaining data are subjected to the complete modeling procedure. This means that, for example, the following steps are performed: the number of components is found based on an "internal" cross-validation, variables are selected, etc. The classification model is then used to predict the left out samples, and the errors are saved. This is repeated until every sample has been left out once. It has been shown that cross model validation is able to overcome the highly biased results obtained when only ordinary cross-validation is used.[26]

Validation. After data reduction, the data matrix was divided into a calibration and a validation set of equal size. All modeling steps (variable selection and classification models) were calculated using the calibration data only. Subsequently, the PLS-DA model was validated using the independent validation data set yielding errors of prediction (ErrP). Output numbers are given as fractional misclassifications for each class (benign vs malignant).

RESULTS AND DISCUSSION

Assessing Sampling-Induced Variation. From previous work on sample stability and sample preparation, we found that

(25) Chong, I.-G.; Jun, C.-H. *Chemom. Intell. Lab. Syst.* **2005**, *78*, 103–112.

(26) Westerhuis, J. A.; Hoefsloot, H. C. J.; Smit, S.; Vis, D. J.; Smilde, A. K.; Van Velzen, E. J. J.; van Duijnhoven, J. P. M.; van Dorsten, F. A. *Metabolomics* **2008**, *4*, 81–89.

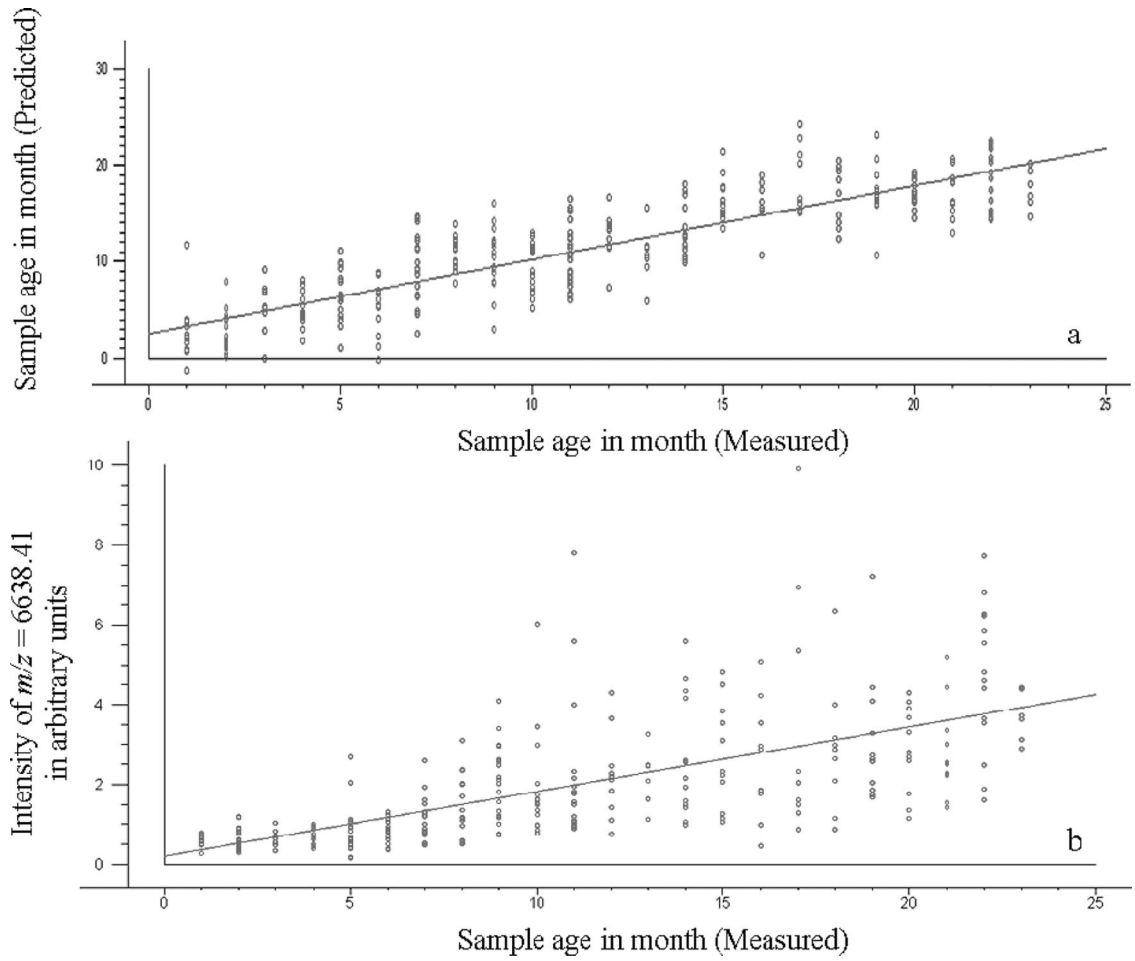


Figure 3. Long-term storage effects at −20 °C. (a) Modeling sample age using PLS (validated). Samples were collected over a period of 23 months. The samples are grouped according to the month in which the sample was taken, 09-2004 is number 23 and 07-2006 is number 1 on the abscissa. The slope is 0.78 with correlation $r = 0.86$. (b) The figure shows a variable with m/z 6638.41 with intensity increasing as a function of storage time. The variance between samples becomes larger with increased storage time. For other variables, e.g., at 4205.71 Da, the opposite trend was observed (not shown), i.e., diminished intensity as a function of storage time.

specific guidelines must be followed to minimize the introduction of sampling-induced data variation and to ensure reproducibility, such as day-to-day variation and target variation.[18] Guidelines for blood collection and storage have been implemented into the protocol of the present study. Initially, we evaluated the impact of these sampling and sample handling guidelines on the apparent quality of data. First, an exploratory analysis was performed where the main variation in the data was modeled and visualized to detect unforeseen systematic variation. Afterward, specific observations were confirmed by building regression models predicting observed systematic patterns. Although the mean age of the group of women with benign tumors (45 years) and women with malignant tumors (65 years) is different and could lead to a bias, we found no patient-age dependent correlation in the data using a supervised regression model (PLS).

Samples were analyzed in four batches (cf. Materials and Methods). A principal component analysis model (unsupervised modeling) of all samples was calculated on the raw untransformed data and showed that samples from the four different batch analyses were evenly scattered indicating no presence of severe outliers (data not shown) except for one sample that was discarded in all further analyses. It was not possible through the different principal components to find any sort of trend (grouping) in the data relating to the four different sample preparation dates. Thus, in this respect, the sample quality was satisfactory. Furthermore, the PCA did not indicate any obvious grouping according to the sample origin that could derive, e.g., from benign and malignant cases of ovarian cancer (data not shown). However, a minor trend could be seen according to sample age, and this was further analyzed using supervised modeling.

Each sample was grouped corresponding to the month of collection with the newest samples starting with number 1 and the oldest samples having number 23. Then, the analysis of sample variance as a function of different collection time points was carried out. A partial least-squares (PLS) regression model was applied trying to predict sample age (time from sampling to measurement) according to the month in which the sample was obtained. The predictions of sample age are seen in Figure 3a showing the cross-validated predictions. The average prediction error given as the root-mean-square error of prediction (RMSEP) is 3.3 months. Compared to the span of sample ages (22 months), it thus appears that it is possible to predict sample age from peak patterns albeit with some uncertainty. In other words, reproducible changes are seen in samples when they are stored for a longer period at −20 °C. Although the trend is linear, it seems that the curve flattens

out after 15 months (Figure 3a) indicating that there are less (systematic) changes after 15 months. Several m/z values are responsible for the sample age variability observed in Figure 3a. One clear example of an m/z variable (6638.41 Da) with this behavior is shown in Figure 3b. Here the peak height and variance increases over time. A few other peaks show the opposite trend, although less apparent. The previously described sample handling-related peptide markers, including complement component C3f, fibrinopeptides, and kallikreins, do not contribute to the predictive model extracted from the analyses of the long-term storage changes in the present samples.[17,18]

The data illustrate that it is important to keep sample age variation to a minimum in proteome investigations. Additionally, it is extremely important to ensure that any possible variation in sample age is not confounded with the response of interest as any observed diagnostic alteration would thus be prone to be a simple effect of sample age (see Confounding Effects of Sample Storage Time). In principle, sampling-induced variation will lead to a poorer signal-to-noise ratio for detecting the clinically relevant responses but will otherwise not affect the analysis as long as the irrelevant variation is not thought to represent relevant (diagnostic) variation as illustrated in the next section.

Confounding Effects of Sample Storage Time. The storage time-dependent alterations in sample composition as, e.g., shown above (Figure 3) have a significant potential for confounding proteomics studies. If, for example, all the malignant samples were collected at one time point and all the benign samples at a different time point, the validity of any observed correlation would be extremely questionable. This we can illustrate with the data of the present study by choosing 36 malignant samples from the first 10 months and comparing data of these samples with 43 benign samples from the last 6 months of the study. After randomly selecting 50 samples (27 benign vs 23 malignant) from this set for calibration of the PLS-DA model and using the rest (16 benign vs 13 malignant) samples for validation, a model is obtained that almost perfectly classifies the independent validation set of 29 benign vs malignant samples. In this case, 100% specificity (no false positives) and 92% sensitivity (one false negative) is obtained (Figure 4). However, rather than classifying benign and malignant samples, we are actually modeling differences in sample collection time points. This is verified when using a validation sample set with benign and malignant samples from the opposite time collection points, ending with a total misclassification of 84%. Thus, it is absolutely imperative to perform careful sampling and randomization in order to discover true biological relations.

CMV and Model Validation. When the number of samples is limited compared with the number of variables associated with each sample, there will be a risk of overfitting even when using cross validation; especially using supervised methods such as PLS-DA in conjunction with variable selection. The use of an independent validation set as in the present study is strongly encouraged but is in reality not always possible. To assess model robustness independent of a validation set, cross model validations (CMV) were made according to the procedure described in Materials and Methods (Figure 2). VIP scores are used in the inner cross validation (CV) removing less important variables (Figure 1). The final PLS-DA model is obtained when the error of prediction has

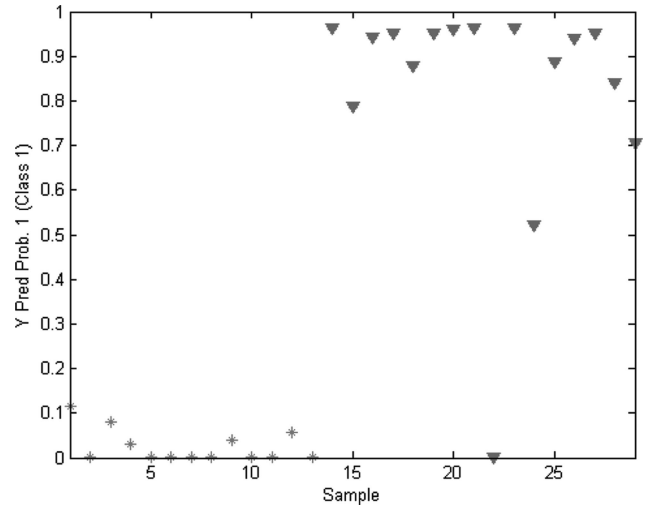


Figure 4. Example of spurious classification performance caused by storage artifacts. Probability scores for the prediction of 16 benign (*) and 13 malignant (▼) randomly selected samples from a model based on "wrong" sampling. Benign samples with a score less than 0.5 are correctly predicted, and malignant samples are correctly predicted when having scores greater than 0.5.

Table 1. Error of Prediction of the Different Cross Validation Methods[a]

validation methods	error of prediction (%)	min/max
cross validation (CV)	20	
cross model validation (CMV)	35	10%/66%

[a] CV includes all samples and therefore only one model and no min/max values are given, whereas CMV derives from many different models all adding up to a final prediction error with the minimum and maximum error of prediction given.

reached a minimum or when variables with VIP scores below a defined threshold (0.8) are absent, resulting in models with typically 70−100 variables. In Table 1, the mean error of prediction from the CMV procedure is listed. A large variation between the different submodels is observed with a minimum at 10% and a maximum at 66%. The mean error of prediction value from all submodels in the CMV procedure is 35%. This strongly indicates that little discriminating power can be found in these data. Furthermore, the discrepancy in performance from the different submodels in CMV may be ascribed to the indirect correlations between cancer and biomarkers found through mass spectrometry due to the large dynamic range in serum protein levels and the reduced dynamic range of both the chromatographic fractionation step and MALDI-TOF mass spectrometry,[14,15,27] i.e., the expectation is that the amount and hence concentration of secreted cancer biomarker is very small compared to the total protein level in the biological matrix. Thus, without a specific capture of these proteins they will not be measurable by mass spectrometry. In addition, this study includes a total of 65 malignant samples representing different stages in cancer development (stages I−IV) and hence the limited number of malignant samples from each stage, potentially characterized by unique sets of diagnostic patterns, may further dilute model robustness.

A randomly selected calibration and validation data set was used for completely independent assessment of model perfor-

(27) Liotta, L. A.; Petricoin, E. F. *J. Clin. Invest.* **2006**, *116*, 26–30.

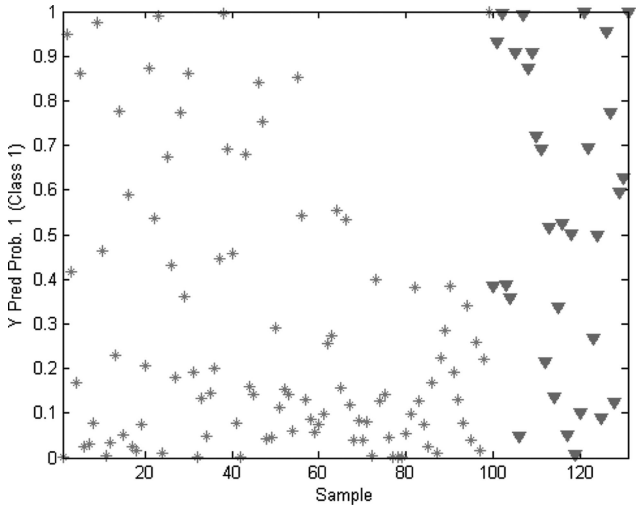


Figure 5. Overview of the performance of the final model in the independent validation sample set of 99 benign (*) and 32 malignant (▼) from the PLS-DA model obtained from a calibration sample set. Probability scores from 0 to 1 with probabilities less than 0.5 are classified as benign samples and scores greater than 0.5 classified as malignant samples.

Table 2. Sensitivity and Specificity of the Final Diagnostic Model When Assessed by Simple Cross Validation and by Model Prediction Using the Independent Validation Sample Set[a]

model/correctly predicted	benign	malignant
cross validation (CV)	86%	73%
model prediction (validation)	79%	56%
Simple CV		
test positive	14	24
test negative	86	9
Independent Validation		
test positive	21	18
test negative	78	14

[a] True positive (TP), false positive (FP), true negative (TN), and false negative (FN) given for both simple cross validation (CV) and model prediction using the independent validation sample set.

mance. The prediction of the validation data set is shown in Figure 5 and comes from a model consisting of 29 variables covering 26 peaks provided by the calibration data set. Stars represent benign samples, and these samples are correctly predicted when giving probability scores below 0.5. The triangles are samples from patients with malignant disease, and these are correctly predicted when having scores greater than 0.5. The classification power of both the cross-model-validation, cross-validation, and truly validated model is seen in Tables 1 and 2, and true positive, true negative, false positive and false negative values are given as well. As can be seen in Table 2, the cross-validation is overly optimistic and the true validation is closer to the cross model validation seen above (Table 1). The final model discriminates malignant from benign conditions in 68% of the cases with 56% sensitivity and 79% specificity. The differences in the predictive abilities for the two classes could be, at least in part, ascribed to the malignant case being less well represented than the benign ones, so that the analysis of a much larger number of samples (especially malignant) may be necessary. However, the ovarian cancer marker CA-125, which is in routine use in clinical biochemistry performs definitively better in the same sample set. Thus, CA-125 alone gives a specificity of 97% and a sensitivity of 68% and there is no diagnostic improvement upon combining the two biochemical approaches. Thus, MS profiling in this study do not contribute to or boost the diagnostic performance of the clinically well-established biomarker CA-125.

CONCLUSIONS

The use of principal component or other multivariate analyses to ascertain sample quality is to be recommended. Also, it may be recommended to use a standardized sample, e.g., a sample pool in all analyses to further help ascertaining sample analysis quality. Furthermore, samples should be analyzed so as to keep the storage age-induced variation to a minimum either by analyzing samples shortly after receiving them or storing samples at −80 °C. Storage effects do not seem to increase (systematically) after about 1.5 years at −20 °C, and we do not see any of the previously described sample-handling induced peptide (complement C3f, fibrinopeptides, and kallikreins) indicators for time/temperature induced variation and variables in the long term storage analysis. The time effect seen in this study affects the number of variables needed for model building and the robustness of such models. It is crucial to keep samples randomized throughout the entire experiment in order to avoid wrong conclusions. In fact, the time effect in the present data is bigger than the disease specific differences we aim to identify.

Simple cross validation may overfit data, and we recommend the use of model cross validation in further studies working with multivariate analysis of samples, as in mass spectrometry and most other spectroscopy methods. Variable selection slightly improves the classification power of the PLS-DA-model. Approximately 20−30 peaks were required in the present study. The performance of the model of the present study is not yet satisfactory for clinical use.

Despite the inability to perform better than the established biochemical marker, CA-125, the data obtained here are encouraging because peak patterns with some specificity emerged. Further analyses of correlations with other biochemical parameters, tumor staging, and additional MS analyses with alternative sample preparation matrixes as well as increased numbers of samples will all increase the likelihood of finding more robust differentiating patterns of diagnostic value for ovarian cancer.

ACKNOWLEDGMENT

The authors acknowledge the Pelvic mass group at Rigshospitalet, Denmark, and Julia Tanas Tanassi for helping with sample preparation.

Received for review October 31, 2008. Accepted December 31, 2008.

AC802293G