Anal. Chem. 2005, 77, 5114–5123

Sample Handling for Mass Spectrometric Proteomic Investigations of Human Sera

Mikkel West-Nielsen,[†] Estrid V. Høgdall,[‡] Elena Marchiori,[§] Claus K. Høgdall,[∥] Christian Schou,[†] and Niels H. H. Heegaard*,[†]

Department of Autoimmunology, Statens Serum Institut, DK-2300 Copenhagen S, Denmark, Department of Virus, Hormones, and Cancer, Institute of Cancer Epidemiology, Danish Cancer Society, DK-2100 Copenhagen Ø, Denmark, Department of Computer Science, Vrije Universiteit Amsterdam, 1081 HV, Amsterdam, The Netherlands, and The Gynecological Clinic, The Juliane Marie Center, Rigshospitalet, DK-2100 Copenhagen Ø, Denmark

Proteomic investigations of sera are potentially of value for diagnosis, prognosis, choice of therapy, and disease activity assessment by virtue of discovering new biomarkers and biomarker patterns. Much debate focuses on the biological relevance and the need for identification of such biomarkers while less effort has been invested in devising standard procedures for sample preparation and storage in relation to model building based on complex sets of mass spectrometric (MS) data. Thus, development of standardized methods for collection and storage of patient samples together with standards for transportation and handling of samples are needed. This requires knowledge about how sample processing affects MS-based proteome analyses and thereby how nonbiological biased classification errors are avoided. In this study, we characterize the effects of sample handling, including clotting conditions, storage temperature, storage time, and freeze/thaw cycles, on MS-based proteomics of human serum by using principal components analysis, support vector machine learning, and clustering methods based on genetic algorithms as class modeling and prediction methods. Using spiking to artificially create differentiable sample groups, this integrated approach yields data that—even when working with sample groups that differ more than may be expected in biological studies—clearly demonstrate the need for comparable sampling conditions for samples used for modeling and for the samples that are going into the test set group. Also, the study emphasizes the difference between class prediction and class comparison studies as well as the advantages and disadvantages of different modeling methods.

In recent years, mass spectrometry (MS) has become the mainstay of efforts to exploit biofluid proteomic patterns for the discovery of novel biomarkers for diagnosis, prognosis, choice of therapy, and disease activity assessment, especially in the field of cancer.[1,2] Compared with the only other high-resolution method for biofluid proteome analysis—two-dimensional gel electrophoresis—MS has the obvious virtues of providing a digitized output, requiring less material, being easily automated, being less labor intensive, having higher throughput capabilities, and being capable of analyzing the peptide population below 5000–10 000 molecular weight. Also, molecular identification is, in principle, integral to the approach. Used on adequate patient and control materials and combined with proper data mining methods, quite large mass-to-charge intensity data sets may be readily pruned for peak patterns specific and predictive for disease. Accordingly, a number of publications reporting MS proteomic analysis-based biomarker discoveries in bladder, prostate, breast, and lung cancer in addition to ovarian cancer have now appeared.[3–11]

Concurrently, concerns have been voiced about the validity of the approach. This is mainly due to nonreproducible results and less obvious biological correlates of findings. One problem is, for example, that very low molecular weight (below 500) candidate biomarkers are reported in several studies.[4,12] Such components may originate from experimental bias and noise effects without biological provenience even though they appear to facilitate model discrimination between patients and normal.[13–17] Furthermore,

* Corresponding author. Phone: +45 32683378. Fax: +45 32683876. E-mail: nhe@ssi.dk.

[†] Statens Serum Institut.

[‡] Institute of Cancer Epidemiology.

[§] Vrije Universiteit Amsterdam.

[∥] The Gynecological Clinic.

[1] Rosenblatt, K. P.; Bryant-Greenwood, P.; Killian, J. K.; Mehta, A.; Geho, D.; Espina, V.; Petricoin, E. F., III; Liotta, L. A. *Annu. Rev. Med.* **2004**, *55*, 97–112.

[2] Diamandis, E. P. *Mol. Cell. Proteomics* **2004**, *3*, 367–78.

[3] Wellmann, A.; Wollscheid, V.; Lu, H.; Ma, Z. L.; Albers, P.; Schutze, K.; Rohde, V.; Behrens, P.; Dreschers, S.; Ko, Y.; Wernert, N. *Int. J. Mol. Med.* **2002**, *9*, 341–7.

[4] 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*, 572–7.

[5] Adam, B. L.; Vlahou, A.; Semmes, O. J.; Wright, G. L., Jr. *Proteomics* **2001**, *1*, 1264–70.

[6] Carrette, O.; Demalte, I.; Scherl, A.; Yalkinoglu, O.; Corthals, G.; Burkhard, P.; Hochstrasser, D. F.; Sanchez, J. C. *Proteomics* **2003**, *3*, 1486-94.

[7] Zhukov, T. A.; Johanson, R. A.; Cantor, A. B.; Clark, R. A.; Tockman, M. S. *Lung Cancer* **2003**, *40*, 267–79.

[8] Schaub, S.; Wilkins, J.; Weiler, T.; Sangster, K.; Rush, D.; Nickerson, P. *Kidney Int.* **2004**, *65*, 323–2.

[9] Paweletz, C. P.; Trock, B.; Pennanen, M.; Tsangaris, T.; Magnant, C.; Liotta, L. A.; Petricoin, E. F., III. *Dis. Markers* **2001**, *17*, 301–7.

[10] Coombes, K. R.; Fritsche, H. A., Jr.; Clarke, C.; Chen, J. N.; Baggerly, K. A.; Morris, J. S.; Xiao, L. C.; Hung, M. C.; Kuerer, H. M. *Clin. Chem.* **2003**, *49*, 1615–23.

[11] Villanueva, J.; Philip, J.; Entenberg, D.; Chaparro, C. A.; Tanwar, M. K.; Holland, E. C.; Tempst, P. *Anal. Chem.* **2004**, *76*, 1560–70.

[12] Alexe, G.; Alexe, S.; Liotta, L. A.; Petricoin, E.; Reiss, M.; Hammer, P. L. *Proteomics* **2004**, *4*, 766–83.

10.1021/ac050253g CCC: $30.25 © 2005 American Chemical Society
Published on Web 06/28/2005

analyses of the same sample sets in independent laboratories have revealed nonreproducible discriminatory patterns.[2] Differences in instruments, in data acquisition methods, and in data transformation and data mining software may account for many discrepancies, but the lack of standardization and the lack of systematic studies of the influence of sample handling and sample treatment on MS-based proteomic data probably also contribute significantly to these discrepancies. The performance of data mining methods on more or less deteriorated samples has not been examined systematically. Several initiatives are currently being undertaken to deal with sampling issues and with controversies of proteomic pattern recognition and biomarker discovery.[11,16,18–23] To investigate the stability of serum and plasma samples in relation to diagnostically valid, mass spectrometry-based proteomic classification models, we here address the issues of sampling and sample handling and the robustness of modeling as a function of sample state for human serum and plasma. The approach focuses on analytes below 10 000 molecular weight using sample extraction with C8-derivatized magnetic beads and standard MALDI-TOF MS. Data analysis was carried out using principal components analysis (PCA), support vector machines (SVMs), and clustering based on genetic algorithms. The parameters of clotting and handling and storage time, storage temperature, and number of freeze/thaw cycles were evaluated in unmodified samples and in paired sets of nonspiked and spiked samples to evaluate the capabilities of the integrated approach for group classification.

EXPERIMENTAL PROCEDURES

Materials. Purification kit MB-HIC 8 and matrix α-cyano-4-hydroxycinnamic acid were purchased from Bruker Daltonik GmbH (Leipzig, Germany). Calibration standards containing seven peptides and four proteins were used as artificial markers (Bruker Daltonik) and consisted of the following molecules with average molecular masses ($z = 1$) given in parentheses: angiotensin II (1047.20), angiotensin I (1297.51), substance P (1348.66), bombesin (1620.88), ACTH clip 1–17 (2094.46), ACTH clip 18–39 (2466.73), somatostatin 28 (3149.61), insulin (5734.56), ubiquitin I (8565.89), and (with $z = 2$) cytochrome *c* (6181.05) and myoglobin (8476.77). Ethanol, HPLC grade water, acetonitrile, and acetone (Lichrosolv grade) were purchased from Merck.

Biological Samples. Clotting time experiments were conducted with samples from two healthy individuals. The freshly drawn blood samples were divided and left to clot for 2, 4, 8, and 24 h at 4 and 24 °C, respectively, before serum was separated by centrifugation at 3000*g* for 6 min at room temperature. We also analyzed eight sera from a 4-year-old sample set (prepared by centrifugation at 2000*g*, 10 min at room temperature), stored at −20 °C. These samples were from a study designed to investigate the stability of future biochemical markers in relation to clotting time and temperature.[24–26]

For the freeze/thaw experiments, sera from blood samples of three healthy volunteers were divided into two identical sample sets. One set was stored at −20 °C, and the other set was stored at −80 °C.

The temperature and storage time experiments were performed on fresh blood samples from eight healthy volunteers (five men, three women, age range 26–55). Paired serum and plasma samples were obtained. For serum, the blood was left to clot in standard blood collection tubes for 2 h at room temperature. Serum was then separated by centrifugation at 3000*g* for 6 min at room temperature. For plasma, the blood was collected in 5-mL tubes containing 100 μL of 0.45 mM EDTA and left for 2 h at room temperature. The plasma was then isolated by centrifugation at 3000*g* for 6 min at room temperature. Serum and plasma samples from the individual volunteers were divided into two equal fractions of 400 μL. One fraction of serum and of plasma was spiked with 20 μL (4 pmol/μL) peptide calibration standard and 4 μL (10 pmol/μL) protein calibration standard, i.e., a total of 11 markers. The control fraction was left unspiked. The spiked and nonspiked fractions were subsequently further divided into two equal (200 μL) fractions that were left at 4 and 24 °C, respectively, with collection of aliquots (5 μL) for MALDI-TOF MS analysis at 0, 1, 4, 8, 24, and, 48 h. (See also Figure 3).

Sample Preparation for Mass Spectrometry. Paramagnetic nonporous C8 particles, 0.8 μm in diameter, (MB-HIC 8, Bruker Daltonics, Bremen, Germany) were used for preparing subfractions of serum and plasma samples as outlined by the manufacturer. All operations were performed at room temperature. Briefly, 10 μL of MB-HIC binding solution and 5 μL of serum or plasma were transferred to a 0.2-mL thin-walled PCR tube (ABgene). A 5-μL aliquot of a homogeneous magnetic particle solution was added, mixed, and left for 1 min. The tubes were placed in a 2 × 8 well magnetic bead separator (MBS) (Bruker Daltonik) for 30 s for magnetic fixation of the MB-HIC 8 particles. The supernatant was aspirated, the tubes were removed from the MBS device, and 100 μL of wash solution was added and carefully mixed with the magnetic beads. The tube was then moved back and forth sequentially between adjacent wells on each side of the magnetic bar in the MBS device. This enhances washing of the magnetic particles as they are fixed to the tube wall and then move through the washing solution in succession. The sequential washing was done 20 times. After fixation of the magnetic beads for 30 s in the MBS device, the supernatant was aspirated. The entire washing procedure was repeated 3 times. After the final washing step, bound molecules were eluted by incubation with 10 μL of acetonitrile (50%) for 1 min before collecting the elution solution using the MBS device. One microliter of the eluent was mixed with 10 μL of matrix solution (0.5 g/L α-cyano-4-hydroxycinnamic acid in ethanol/acetone 2:1). One microliter was spotted onto a

[13] Sorace, J. M.; Zhan, M. *Bioinformatics* **2003**, *4*, 24.

[14] Baggerly, K. A.; Morris, J. S.; Coombes, K. R. *Bioinformatics* **2004**, *20*, 777–85.

[15] Li, L.; Umbach, D. M.; Terry, P.; Taylor, J. A. *Bioinformatics* **2004**, *20*, 1638–40.

[16] White, C. N.; Chan, D. W.; Zhang, Z. *Clin. Biochem.* **2004**, *37*, 636–41.

[17] Diamandis, E. P. *Clin. Chem.* **2003**, *49*, 1272–5.

[18] Cordingley, H. C.; Roberts, S. L.; Tooke, P.; Armitage, J. R.; Lane, P. W.; Wu, W.; Wildsmith, S. E. *Biotechniques* **2003**, *34*, 364–73.

[19] Petricoin, E., III; Liotta, L. A. *Clin. Chem.* **2003**, *49*, 1276–8.

[20] Hulmes, J. D.; Bethea, D.; Ho, K.; Huang, S.; Ricci, D. L.; Opiteck, G. J.; Hefta, S. A. *Clin. Proteomics* **2004**, *1*, 17–32.

[21] Pusch, W.; Flocco, M. T.; Leung, S. M.; Thiele, H.; Kostrzewa, M. *Pharmacogenomics.* **2003**, *4*, 463–76.

[22] Villanueva, J.; Tempst, P. *Nature* **2004**, *430*, 611.

[23] Diamandis, E. P. *J. Natl. Cancer Inst.* **2004**, *96*, 353–6.

[24] Hogdall, E. V.; Johansen, J. S.; Kjaer, S. K.; Price, P. A.; Blaakjaer, J.; Hogdall, C. K. *Scand. J. Clin. Lab. Invest.* **2000**, *60*, 247–51.

[25] Hogdall, E. V.; Hogdall, C. K.; Kjaer, S. K.; Xu, F.; Yu, Y.; Bast, R. C.; Blaakaer, J.; Jacobs, I. J. *Clin. Chem.* **1999**, *45*, 692–4.

[26] Riisbro, R.; Christensen, I. J.; Hogdall, C.; Brunner, N.; Hogdall, E. *Int. J. Biol. Markers* **2001**, *16*, 233–9.

600-μm-diameter spot size 384 AnchorChip target plate (Bruker Daltonik) and left to dry. The peptide and protein calibration standard (0.5 μL) in the same matrix was applied to target spots in proximity to the serum samples for external calibration of the instrument.

During the study, an automated sampling handling robot (ClinProt robot, Bruker Daltonik) became available. The work flow conducted by the robot is similar to the work flow performed manually.

Mass Spectrometry. For all MS experiments, an UltraFlex MALDI-TOF mass spectrometer (Bruker Daltonik) controlled by FlexControl software v. 2.0 (Bruker Daltonik) was used. 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. All acquitions were generated by an automated acquisition method included in the instrument software and based on averaging 150 randomized shots over 5 positions (30 shots/acquisition cycle). The acquisition laser power was set between 32 and 36%. Before each acquisition cycle, the target area was pretreated with 10 laser shots at 45% laser power to improve spectra quality.[11] Spectra were acquired in positive linear mode geometry under 20 kV of ion acceleration, and with ion selector deflection of mass ions of >900 m/z in the mass range 925–15 180 Da. This is the analyzed mass range throughout the study. Pulsed ion extraction was set to 320 ns to ensure appropriate time lag focusing. Data acquisition was automated through FlexControl using acceptance parameters as follows: peak resolution threshold ≥200 and S/N ratio equal to or greater than 15 in the m/z range 2000–6000.

Data Handling. Spectra were converted into ASCII file format using the FlexAnalysis software. Data were subsequently analyzed by using commercially available analysis programs, ClinProTools v. 1.0 (Bruker Daltonics) and Unscrambler v. 8.05 (Camo, Oslo, Norway) as well as by linear SVMs (available at, e.g.: http://www.kernel-machines.org/software.html). As our sample sets have small sizes, we used a leave-one-out cross-validation (LOOCV) procedure to analyze the performance of the techniques. In LOOCV, all spectra but one are used for generating a model, which is then applied to the spectrum left out. This process is repeated a number of times equal to the size of the data. The average error in applying the generated models to the left-out spectra provides an almost unbiased estimate of the true error of the classification model.[27] Accuracy, defined as the proportion of correctly classified spectra, is reported, as well as specificity and sensitivity, defined as the fraction of true negative (normal correctly classified) of the total negative (normal misclassified + normal correctly classified) and the fraction of true positive (spiked correctly classified) of the total positive (spiked misclassified + spiked correctly classified), respectively.

Predictive Model Building. *ClinProTools.* Complex protein profile comparison, potential biomarker detection, and differential model tests were carried out using ClinProTools. This software compares large data sets of different classes, builds predictive models from spectra profiles based on discriminators found in the data, and returns sensitivity and specificity of the model. The predictive models are developed on the basis of genetic algorithms that search for combinations of biomarkers (i.e., m/z values) that give the best discrimination of the classes under consideration;[4] i.e., class prediction is supervised.[28] Genetic algorithms are stochastic search procedures inspired by the rules of natural selection and natural genetics. Two model building techniques are available in this program (k-nearest neighbor and centroid clustering). Only centroid clustering was used in the present study and the Few Nearly Clean Cluster parameter was applied. The mass spectra are viewed as points in an N-dimensional space (where N is the number of peaks picked), and similarity between pairs of profiles is measured by means of Euclidean distances. The centroid clustering process is an iterative classification technique that uses the distances between the spectra and the centroids of already defined clusters. Spectra are assigned to clusters independently of their class membership. Every new spectrum is classified by considering the cluster with centroid nearest to the spectrum and assigning the class label of its elements, which occurs more often. The genetic algorithm employs four parameters: model size minimum, model size maximum (corresponding to the minimum and maximum number of biomarkers to be selected), together with mutation rate and crossover rate. The model size was set between 5 and 10, with a mutation rate of 0.01 and a crossover rate of 0.9. For tuning the model building the parameters used for data preparation are S/N ratio, minimum relative peak height, baseline correction, and normalization. Standard settings were used throughout except for the minimum relative peak height, which was empirically set to 0.2 to minimize nonpeak picking without losing any real peaks.

Principal Components Analysis. PCA, principal component regression (PCR), and partial least squares regression (PLS) were used for linear analysis of the data. Objectives of PCA are to discover or reduce the dimensionality of the data and to identify new meaningful underlying variables. PCR consists of PCA followed by multiple linear regression applied to the selected principal components. PLS finds components that have high variance and have high correlation with the class labels (normal and spiked) of the training data. PCA transforms a number of possibly correlated variables (e.g., m/z values) into a smaller number of uncorrelated variables called principal components. The first principal component accounts for as much variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible. The original data are described as linear combinations of loadings and scores. The loadings can be understood as the weights for each original variable when calculating the principal component, and thus, the m/z values can be retrieved. The scores describe the original data in a rotated coordinate system. As model building in the Unscrambler program takes every data point in the spectra into account, we reduced the number of data points by a factor of 4 resulting in 22 572 m/z values/spectrum. All spectra were normalized by a range normalization algorithm, which scales the samples to a common range, between 0 and +1.

Support Vector Machines. Linear SVMs[29,30] were used as adjunct methods to conduct data analysis related to spiked samples. Data

[27] Evgeniou, T.; Pontil, M.; Elisseeff, A *Machine Learn.* **2004**, *55*, 71–97.

[28] Simon, R.; Radmacher, M. D.; Dobbin, K.; McShane, L. M. *J. Natl. Cancer Inst.* **2003**, *95*, 14–8.

[29] Vapnik, V. N. *Statistical Learning Theory*; John Wiley & Sons: New York, 1998.

[30] Cristianini, N.; Shawe-Taylor, J. *Support Vector Machines*; Cambridge Press: New York, 2000.

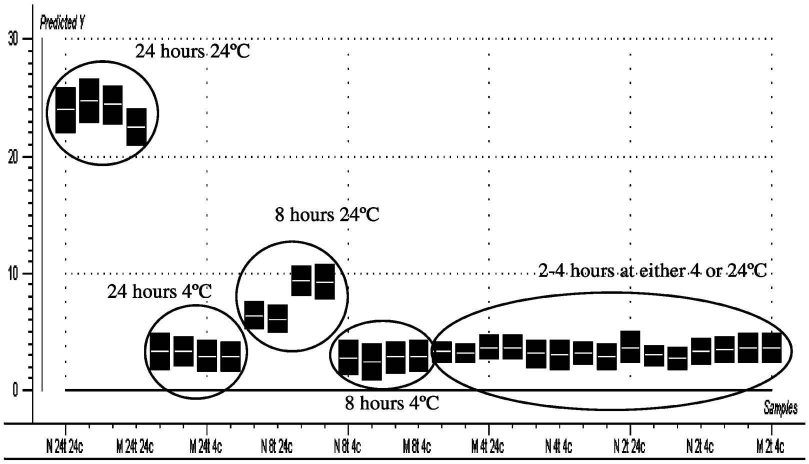

Figure 1. PCA-based prediction of spectra from serum left to clot at different temperatures and time. Serum left to clot at room temperature is much more subject to changes over time than serum left to clot at 4 °C. One sample was rejected during the automated MS data acquisition.

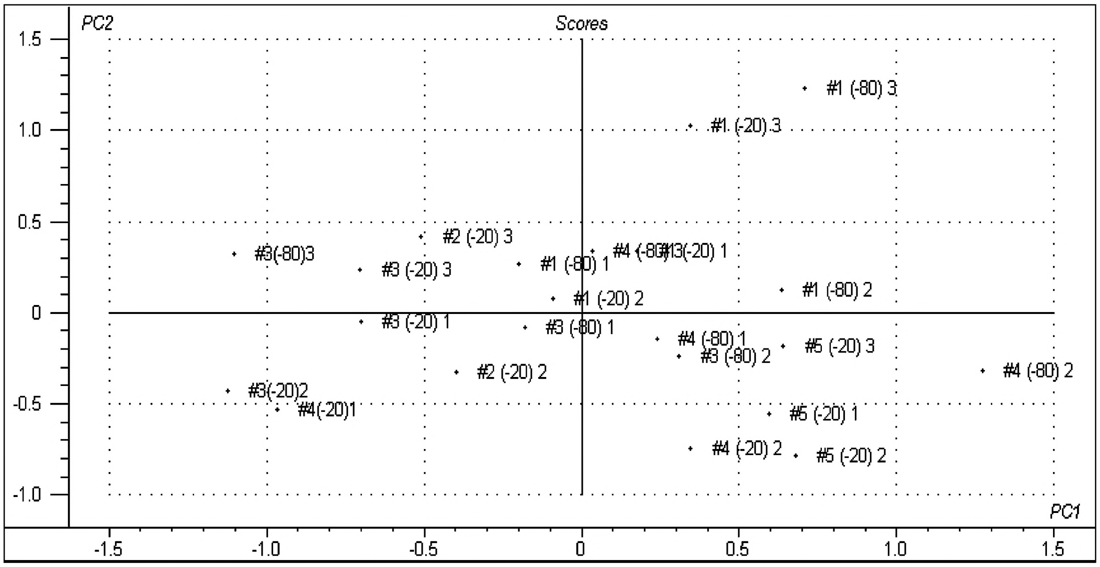

Figure 2. PCA of the different freeze/thaw cycles of sera stored at −20 and −80 °C. The scores show the locations of the samples along each principal component. Outlier data points were excluded as explained in the text.

were normalized and reduced as described above for the PCA data handling. Linear SVM constructs a linear classifier described by a hyperplane in the *m/z* values space. This hyperplane separates the two classes in such a way that its margin is maximized. The margin is the sum of the minimum distances of training points of each class from the hyperplane. Regularization is used for dealing with nonlinearly separable classes, where a parameter *C* (here set to 10) is used for penalizing misclassification. Biomarker detection based on linear SVM is performed using recursive feature elimination, (SVM-RFE).[31] SVM-RFE recursively eliminates chunks of *m/z* values that are considered less relevant than the remaining *m/z* values, according to a relevance criterion based on the weights associated to *m/z* values resulting from the SVM training. We terminate this process when 10 *m/z* values (i.e., the maximum number of *m/z* values used in the ClinProTools biomarker selection method) are left, thus choosing for a model dimension equal to 10.

RESULTS AND DISCUSSION

Clotting, Temperature, and Handling Time. The influences of blood clotting, handling time, and temperature on the characteristics of the MS data sets were assessed by PCA of spectra. Two groups appeared from this analysis. Mass spectra from blood left to clot for 8 and 24 h at 24 °C were separated by PCA from the rest (blood clotted for 2 and 4 h at 24 °C and 2−24 h at 4 °C) (data not shown). By PCR on spectra from samples incubated at 2, 4, 8, and 24 h at 24 °C, a linear model ($y = 0.98x + 0.16h$, $r = 0.99$) describes the observed data variation with respect to the length of clotting time. A full cross-validation (predicted) yields similar parameters ($y = 0.96x + 0.50h$, $r = 0.96$) showing the consistency of the model. When analyzing spectra from samples left to clot at 4 °C for any length of time, this model places them in the same group as the samples left to clot at 2 and 4 h at 24 °C (Figure 1). Thus, the model evaluates all MS profiles at 4 °C as belonging to the group of 24 °C samples clotted for a short time. No differences in the MS profiles of serum from blood left to clot from 2 to 24 h at 4 °C compared with blood left to clot at 2 and 4 h at 24 °C thus were discernible by PCA (Figure 1). Conversely,

[31] Guyon, I.; Weston, J.; Barnhill, S.; Vapnik, V *Machine Learn.* **2002**, *46*, 389−422.

the samples left to clot for longer times (8–24 h) at 24 °C displayed clearly changed m/z profiles (as the possibility of building a clotting time-based linear model from the 24 °C data indicates). The results indicate the need to avoid different clotting and handling times and temperatures when collecting material for clinical proteomics and that less than 4 h at 24 °C is preferable.

Analysis of archival sera (stored four years at −20 °C) representing blood samples from eight women clotting at different temperatures (4 °C or room temperature) and for different times (1, 3, and 72 h) was initially performed as above, but PCR did not yield a linear time dependency as with the fresh samples (data not shown). However—using PLS analysis of the mass spectra—it could be shown that the only sample group that was differentiated by PCA of m/z spectra in this sample set was samples left to clot for 72 h at room temperature (data not shown). This is in agreement with the results above using fresh samples and with the original results from Hogdall et al.,[24] who only found changes in a specific protein marker (YKL-40) in the samples left to clot at room temperature for 72 h.

Previous studies have indicated that freeze/thaw cycles have no effect on a number of specific protein markers in serum.[24−26] We analyzed three different serum samples divided into two groups that were stored at either −20 or −80 °C. The −20 °C samples were thawed and frozen 5 times over 5 days, and samples kept at −80 °C were thawed and frozen 4 times over 4 days. Data analysis of mass spectra by PCA indicated no gross differences in the m/z data distribution in the samples kept at one specific temperature and repeatedly sampled daily. Also, it was not possible to observe any differences between the samples stored at −20 and −80 °C, cf. Figure 2. One spectrum (−20 °C, second thaw) was not acquired because it was rejected by the resolution and S/N settings of the MS data acquisition program. Further, four data sets (three from the second freeze/thaw experiment of −80 °C samples and one −20 °C set from the fourth freeze/thaw experiment) became outliers because of an artifactual peak shoulder-induced m/z shift (that reverted to original values in analyses of the ensuing freeze/thaw cycles of the same samples). These data sets have been omitted from Figure 2. In larger sample sets in clinical proteomics, this type of confounder would be expected to be equalized by the other samples in the sample set. However, rigorous standardization and reproducibility with respect to crystallization conditions are underscored by these observations.

Spiked Samples. In the freeze/thaw and blood clotting experiments, we observed that small changes in sample handling and data acquisition heavily influenced the quality of the mass spectra. In an effort to devise sample handling and sample storage guidelines as well as optimal data handling procedures, we designed an artificial sample set of spiked and nonspiked serum and plasma samples from eight volunteers. The experimental approach is outlined in Figure 3 and is based on different storage conditions of paired samples that are nonspiked or peptide-spiked after initial preparation of serum from blood left to clot for 2 h at room temperature. A total of 24 mass spectra of each individual serum sample pair, i.e., a total of 192 mass spectra, were acquired in this experiment.

The changes taking place upon storage of sera were readily discernible in the mass spectra. Figure 4 exhibits a zoom-overlay plot of spectra of serum samples from one person as a function of incubation time at 4 and 24 °C. It is evident that the storage temperature is responsible for the largest changes in the mass spectra. The most striking changes are the diminishing amounts of species below $m/z = 2000$. Despite these changes, many of the spiked species were consistently recovered at varying signal strengths. Four of the spiked molecules, however, were never recovered. This may reflect losses during the sample processing procedure (i.e., the added material may not be eluted from the reverse-phase beads or be degraded during the sample preparation). The nonrecovered molecules were as follows: substance P (1348.66), ACTH clip 1−17 (2094.46), cytochrome *c* (6181.05), and myoglobin (8476.77).

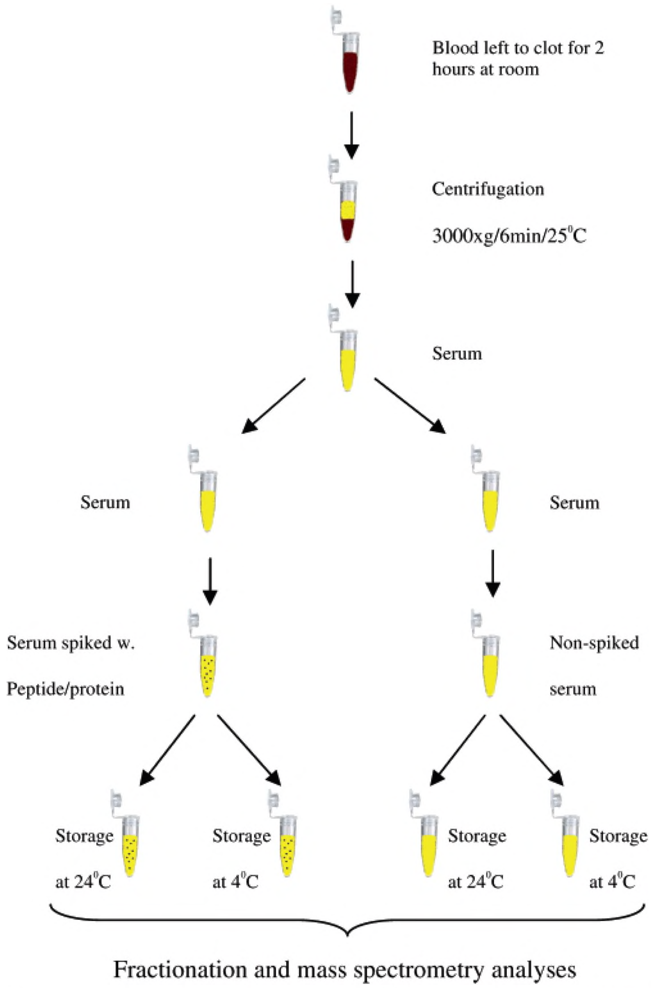

Figure 3. Design of sample preparation experiment. Samples (serum as well as plasma) from each individual were split into two equal parts. One part was spiked with a peptide and protein standard mixture (see Experimental Procedures); the other was not spiked. The spiked and nonspiked tubes were further divided into two parts and stored at increasing time intervals (0, 1, 4, 8, 24, and 48 h) at 4 and 24 °C before fractionation with C8-activated magnetic beads and mass spectrometric analyses.

With respect to plasma, we analyzed a number of paired serum and plasma samples and found that the two types of material required different sample treatment procedures, and thus, substitution of serum for plasma in future trials may not be recommended without optimizing sample preparation. The effect was not due to EDTA, and high-speed centrifugation did not change

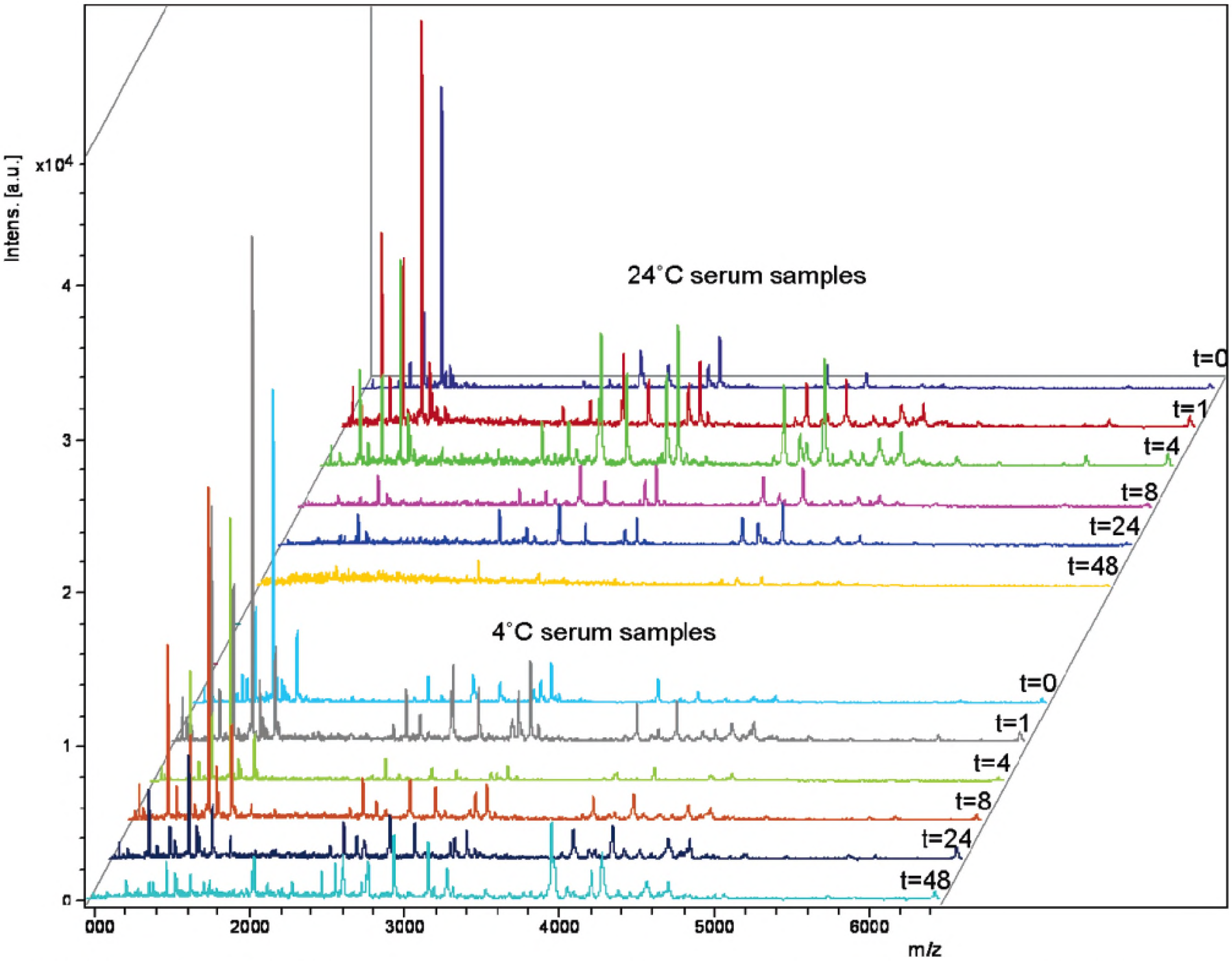

Figure 4. 3D overlay (zoom) plot of mass spectra of a serum sample from one person. Spectra of samples incubated at 0, 1, 4, 8, 24, and 48 h are shown grouped according to the serum incubation temperature at 4 and 24 °C, respectively. The major changes in serum are seen in the low *m*/*z* region (900−2000 Da.).

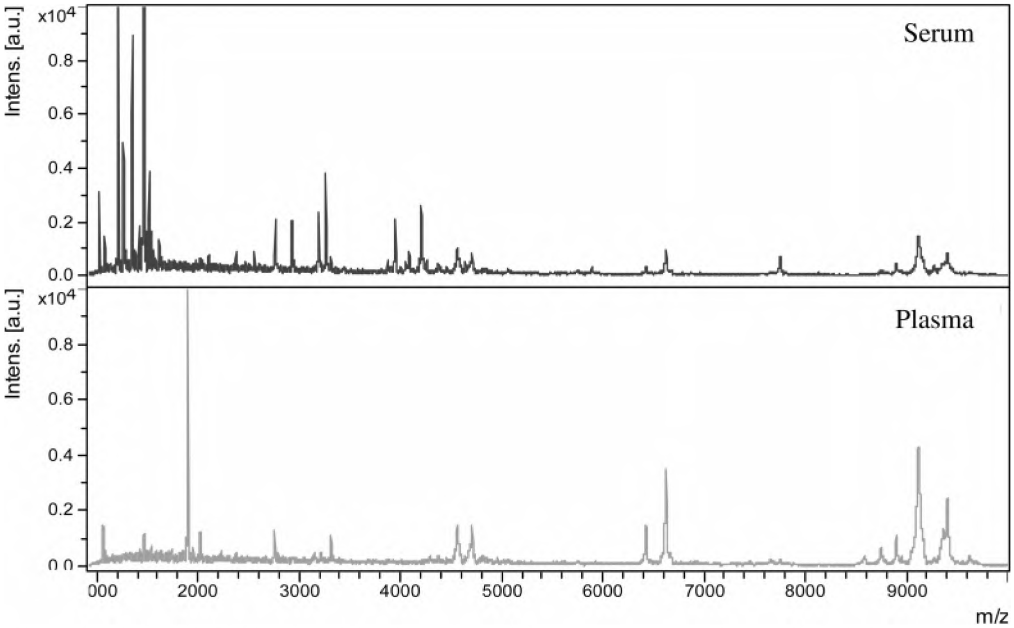

Figure 5. Comparison of MS analysis of serum and plasma samples. Two samples from same individual (serum, upper trace; plasma, lower trace) were processed for MS using three times the serum volumes for the plasma analysis. Samples were analyzed directly after thawing from storage at −20 °C.

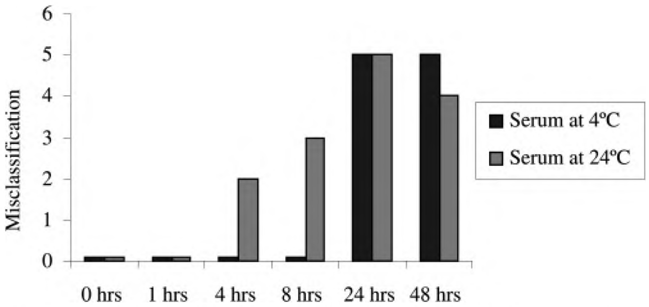

Figure 6. Misclassifications by ClinProTools of sample spectra from sera stored at the indicated time and temperature. Samples from both 4 and 24 °C, stored in the time interval 0−1 h are used for model building. Dark gray (4 °C) and light gray (24 °C) sample data are shown. Sixteen samples are in each time group, which gives values of up to 33% misclassification for samples stored at 24 and 48 h.

the quality of the spectra. Only after using three times the volume optimal for serum processing, did plasma samples under otherwise identical handling and analysis conditions yield spectra of the comparable quality (Figure 5). This may reflect protein degradation in plasma samples, but further characterization must be carried out to address this issue.

To specify a set of minimum requirements for handling of sera for clinical proteome analyses, we subsequently used three different biostatistical approaches for grouping the artificially created binary sample sets. The aim was to assess the influence of temperature and incubation time on the predictive performance of the classification models without and with (potential) biomarker detection. Thus, this part of the study concerns class prediction while the above stability studies represent class comparison issues.[28] Of the 96 potential m/z data sets at each temperature,

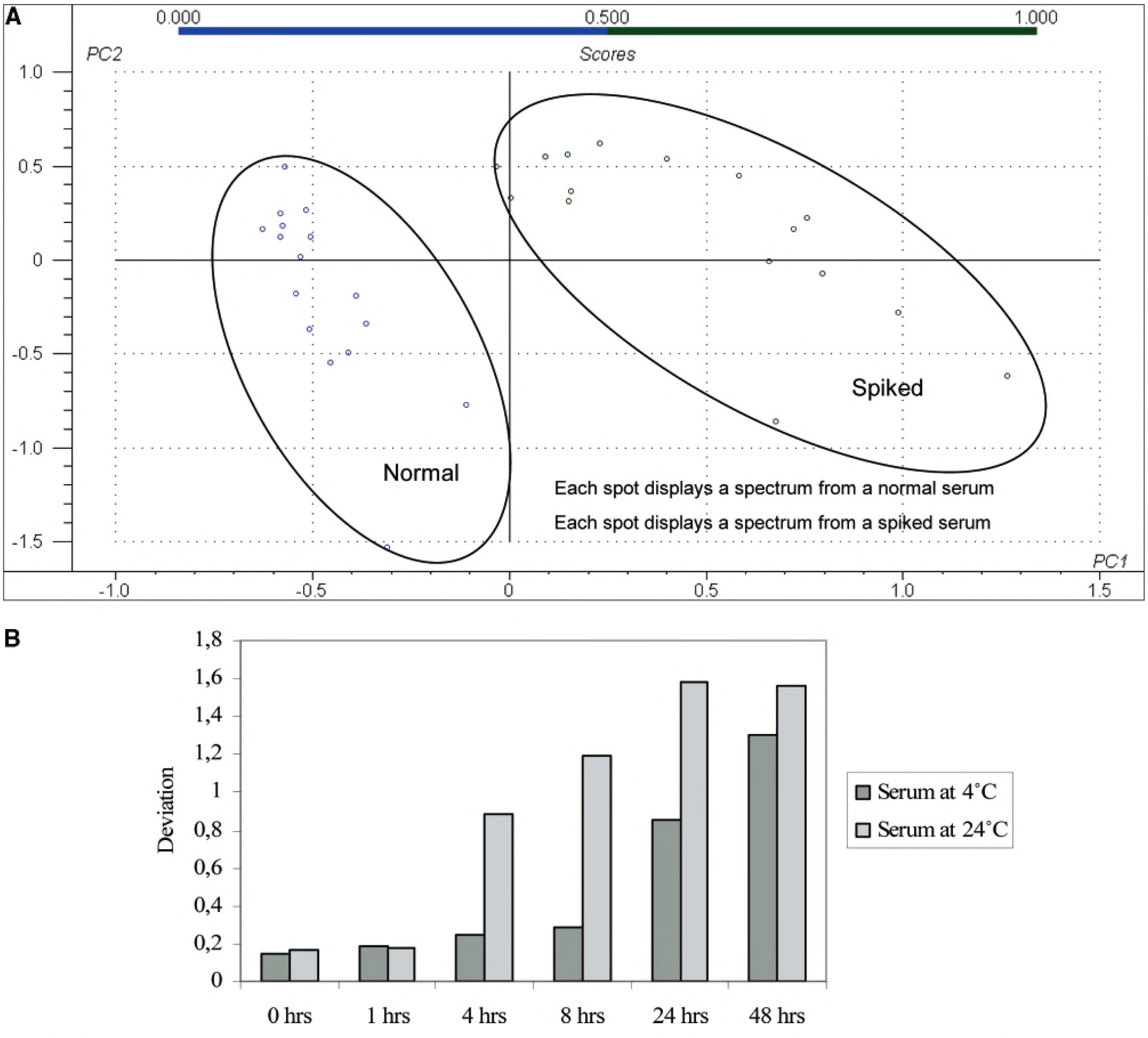

Figure 7. PCA performance. (A) Scores plot from a linear regression classification model (PLS1) of mass spectra from normal and spiked sera. Samples incubated for up to 1 h at 24 °C are included in the model. The sample data divide into two clearly distinguished groups represented by spiked (black circles) and nonspiked (blue circles) samples. (B) Deviation of sample prediction. A model created from ≤1-h samples incubated at 4 and 24 °C. All 0−48-h samples incubated at 4 (light gray) and 24 °C (dark gray) are used to validate the model. From the classification of each spectrum, a standard deviation value is given telling how well the each spectrum at a given incubation time is classified. Mean standard deviations are depicted as a function of incubation time and temperature. The standard deviation of the validation spectra increases dramatically between 1 and 4 h of incubation at 24 °C and after 8−24 h of incubation at 4 °C.

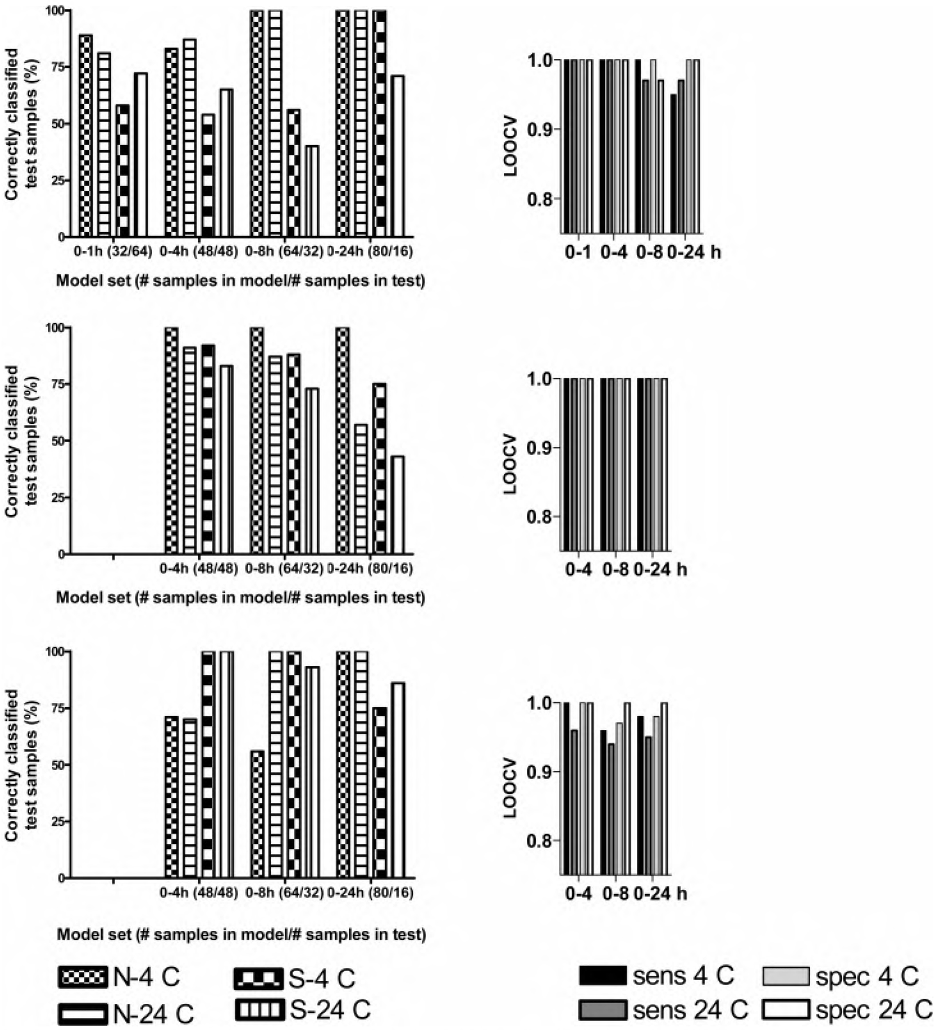

Figure 8. Modeling performance using sample sets of spiked (S) and nonspiked (N) sera (cf. Figure 3). Results are shown from models based on the following: centroid clustering using ClinProTools (upper row), SVM (middle row), and SVM-RFE (lower row). Classification error graphs (left panel) show the percentage of correctly classified test samples (i.e., in the subset of samples not used for training) as a function of the training sample sets (increasing incubation times at two different temperatures as indicated, numbers of samples in model set and in the test set are indicated in parentheses—a total of 96 data set at 4 °C and 94 (2 from the 48-h group did not yield spectra) at 24 °C were used). The 0−1-h sample sets were not used alone for training in the SVM and SVM-RFE analyses. Graphs in right column shows the outcome (average sensitivity (sens) and specificity (spec)) of LOOCV for each of the models based on different modeling sample sets as indicated.

two from the 24 °C group were rejected because their overall peak threshold was below that specified by the acquisition settings. These two samples belonged to the 48-h storage group.

While the PCA (Unscrambler) and SVM programs include all input data points, the ClinProTools software reduces the amount of data by peak picking routines. This process reduces data complexity from 90 288 data points to 150−200 discrete peaks in our spectra. The clustering model approach was used because initial comparison indicated that the k-nearest neighbor analyses had an overfitting bias (data not shown).[28] Each final model relies on a reduced number of marker peaks for correct classification of unknown sample spectra. We specified a minimum of 5 and a maximum of 10 markers for the model building in cluster analyses. Since modeling depends on maximizing sample group differences, a few of the added markers often sufficed in achieving a complete separation of spiked and nonspiked sera. The data show that model building is optimal when based on samples treated similarly to the test samples. Thus, the fragility of models based on sera stored for a short period of time for the predictive analysis of sera stored for ≥24 h is even more pronounced when only using the data of sera stored at ≤1 h (Figure 6). These results indicate that long-term (≥24 h) storage at any (24 or 4 °C) temperature of test samples is harmful for their classification but also that inclusion of data sets from samples stored at increasing time and temperature into the model building will yield better classification specificity and sensitivity of all the sample sets. This is not surprising as the mass spectra from sample data sets with increasing incubation time and temperature force the model to take into account the decreasing quality of the mass spectra. Since the trends in conditions that are interfering with proper model building are here obtained with artificially changed sera, it is very likely that the suggested detrimental effects of incompatible model and test sets in real biological samples will be much more pronounced.

When the same sample sets were analyzed with PCA/PLS (Unscrambler), spiked sera were discriminated from nonspiked sera when a model based on the entire data sets of samples stored up to 1 h at both 4 and 24 °C was used (Figure 7A, samples

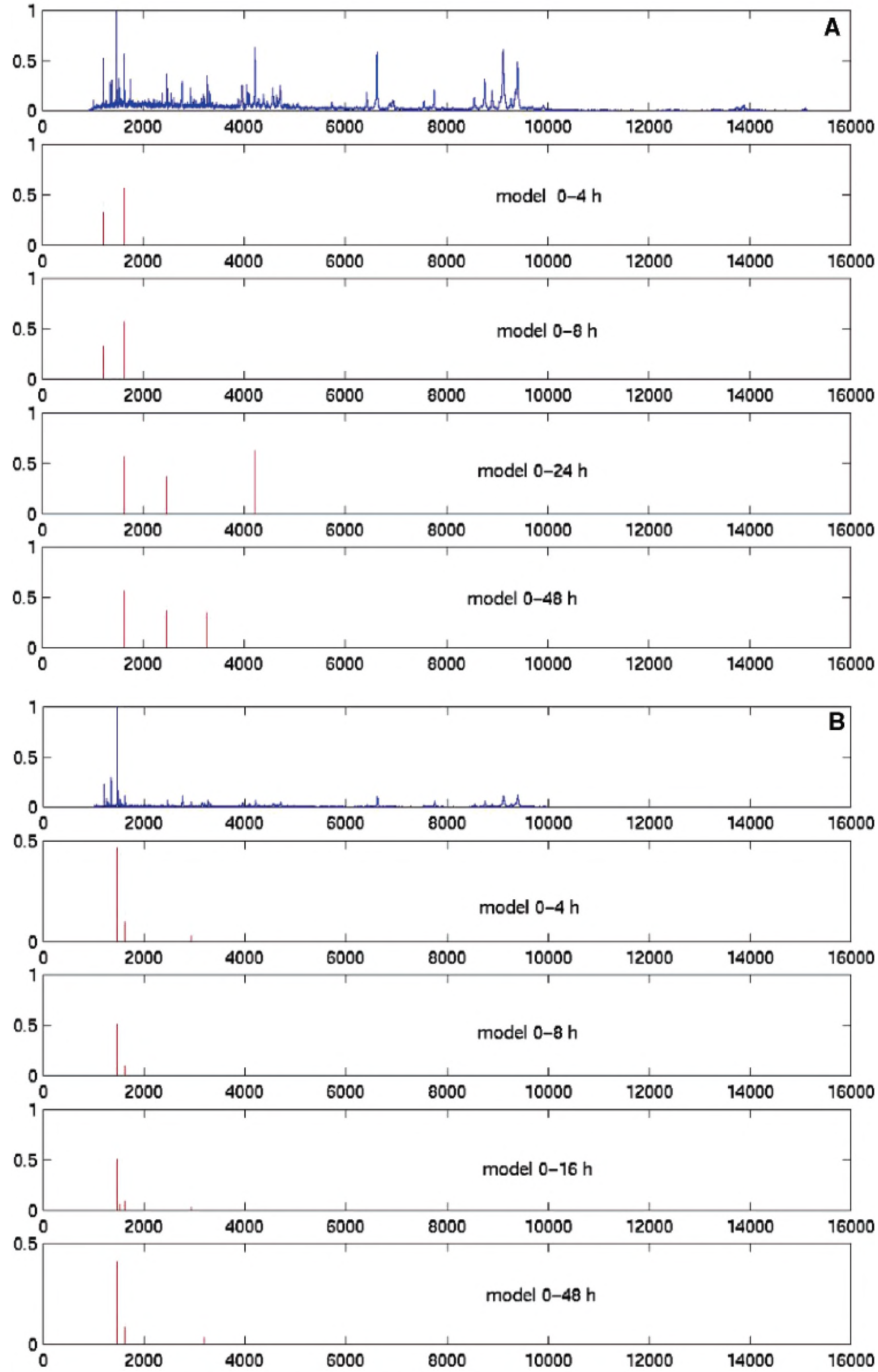

Figure 9. Frequency plots of models selected by SVM. Shown are the *m/z* values selected by each of the four models. (A) 4 °C; (B) 24 °C. Upper traces show an example of a spiked spectrum from each temperature.

incubated at 24 °C are shown). The actual m/z values are derived from loadings plots that show how much each variable (m/z value) contributes to the meaningful variation in the data (sample). A linear regression model of these data had a high predictive score ($r > 0.95$) and a high cross-validation score (LOOCV, $r > 0.93$) indicating that the model is robust. When using data from sera incubated more than 1 h for the model building, the linear regression model gets increasingly worse in discriminating the spiked from the nonspiked serum samples. This tendency is most profound for the models based on serum samples stored at 24 °C. Figure 7B depicts this in plots of the mean mass spectrum standard deviation against the serum incubation time at 4 and 24 °C. Data were calculated by a linear regression model based on the cumulated data sets from sera stored up to 1 h at 4 and 24

°C. For sera stored at 4 °C, the mean mass spectrum deviation starts to increase dramatically somewhere between 8 and 24 h of incubation whereas the mean mass spectra deviation in serum samples stored at 24 °C has a similar increase already between 1 and 4 h. Overall, the PCA/PLS analyses again stress the need for treating all samples in the same way to get a meaningful basis for developing models.

Data were further analyzed using SVM and SVM-RFE. Models were generated using four training data sets of increasing size and wider range of incubation time values and tested on the remaining data. Figure 8 summarizes the results of the SVM and SVM-RFE as well as the clustering modeling mentioned above together with predictive classifications of test sample sets. Model stability was assessed by applying LOOCV to each of the three training data sets (four in the case of centroid clustering) (Figure 8, right panel). SVM and SVM-RFE were applied to data sets containing 22 572 *m*/*z* values, i.e. the same data sets that were subjected to PCA (cf. above), while the cluster analysis as mentioned was based on peak picking (150–200 peaks). Results of SVM indicate a better performance when data at 4 °C are used instead of 24 °C. Not only does the SVM analysis become more stable with respect to small changes of the training set (LOOCV accuracy equal to 1), but its performance on the test set also improves, with two spiked misclassified spectra, both from the group of long incubation time (48 h) compared with six to seven misclassifications in the 24 °C group (all 48 h).

This difference in performance does not occur when recursive feature elimination is used (Figure 8, lower panel). However, the performance of SVM-RFE (average accuracy 0.97–1) is slightly worse than that of SVM (average accuracy 1), indicating that little information relevant for SVM classification is discarded when selecting only 10 out of 22 572 *m*/*z* values. Figure 9 shows frequency plots of *m*/*z* values selected by SVM at 4 (Figure 9A) and 24 °C (Figure 9B) in the four successive models generated from the accumulated data of the four time ranges as indicated. Observe that no *m*/*z* value higher than 4500 is selected in any of the LOOCV runs. At each of the LOOCV runs, most of the 10 selected *m*/*z* values were concentrated in the neighborhood of one or two true spikes, most of them near spikes of *m*/*z* values of 1620 (bombesin, *m*/*z* = 1620.88) and 2465 (ACTH clip 18–39, *m*/*z* = 2466.73). Thus, as observed in the other modeling approaches as well, one or two of the spikes yield *m*/*z* values of enough discriminatory power for optimal classification despite the presence of additional markers.

The SVM-based analyses, in agreement with PCA and the cluster analyses, support the importance of using model building samples and test samples that are treated similarly to avoid classification problems. SVM offers 10% improvement in accuracy on the test set over the cluster analyses while LOOCV does not readily support the choice of one computational approach over the other. It is important to note that modeling for discriminatory analysis does not yield a description of all the significantly changed analytes but stops at the number of analytes that attain maximum class prediction capabilities, i.e., is different from class comparison analyses. In real cases, the trends in the LOOCV specificity and sensitivity and in the accuracy error variations observed here with artificially enhanced differences between samples must be expected to be much more significant.

CONCLUSIONS

In the use of mass spectrometric proteome analysis for clinical diagnostic purposes, it is mandatory that samples are not compromised due to suboptimal handling, transport, and processing. Our results show that clotting time and temperature are important factors for serum quality and hence for model building, which is compromised in all the computational approaches tested here if blood is left to clot at room temperature for more than 3 h or more than 24 h at 4 °C. Likewise, storage of sera is detrimental if exceeding 4 h at room temperature or more than 24 h at 4 °C. There are no indications in the relatively short time span (7–14 days) of these experiments that it is advantageous to store samples at −80 °C rather than at −20 °C. Additionally, our findings suggest that thaw/freeze cycles have no or very little effect on the stability of serum and hence on model building and biomarker findings. Using plasma samples, our preliminary results indicate that more material must be applied for the sample extraction to yield measurable data. Whether this is due to changes in proteolysis in plasma as compared to serum remains to be established. If archival samples are used for development of MS-based proteomic diagnostics, our study emphasizes the need for similar treatment of all samples. Finally, the purpose of specific studies should be clear before embarking on data analysis; i.e., the analyses presented here aim at modeling the most efficient discrimination between two groups (supervised) and do not aim to characterize all significant differences as would be the case in class comparison studies.

ACKNOWLEDGMENT

This work has been supported by Sygeforsikringen danmark (to N.H.H.H.).

Received for review February 10, 2005. Accepted June 3, 2005.

AC050253G