Serum protein profiling for diagnosis of breast cancer using SELDI-TOF MS

MARIE-CHRISTINE W. GAST[1], CARLA H. VAN GILS[2], LODEWIJK F.A. WESSELS[3,4], NATHAN HARRIS[5], JOHANNES M.G. BONFRER[6], EMIEL J.TH. RUTGERS[7], JAN H.M. SCHELLENS[8,9] and JOS H. BEIJNEN[1,9]

[1]Department of Pharmacy and Pharmacology, The Netherlands Cancer Institute/Slotervaart Hospital, P.O. Box 90440, 1006 BK Amsterdam; [2]University Medical Center Utrecht, Julius Center for Health Sciences and Primary Care, P.O. Box 85500, 3508 GA, Utrecht; [3]Department of Molecular Biology, Bioinformatics and Statistics, The Netherlands Cancer Institute, P.O. Box 90203, 1006 BE Amsterdam; [4]Faculty of EEMCS, Technical University of Delft, P.O. Box 5031, 2600 GA Delft, The Netherlands; [5]Vermillion Inc., 6611 Dumbarton Cirlce, Fremont, CA 94555-3603, USA; [6]General Clinical Laboratory, Antoni van Leeuwenhoek Hospital, P.O. Box 90203, 1006 BE Amsterdam; Departments of [7]Surgery, and [8]Clinical Oncology, Division of Medical Oncology, The Netherlands Cancer Institute/Antoni van Leeuwenhoek Hospital, P.O. Box 90203, 1006 BE Amsterdam; [9]Department of Pharmaceutical Sciences, Division of Biomedical Analysis, Utrecht University, Faculty of Science, P.O. Box 80082, 3508 TB Utrecht, The Netherlands

Received February 17, 2009; Accepted April 10, 2009

DOI: 10.3892/or_00000426

Abstract. In search for novel markers for breast cancer, we aimed to identify and validate novel serum protein profiles specific for breast cancer, and assess the influence of clinical (subjects age) and pre-analytical (sample storage duration) variables on the constructed classifiers. To this end, sera of breast cancer patients (n=152) and healthy controls (n=129), randomly divided into a training and test set, were analysed by surface-enhanced laser desorption/ionisation time-of-flight mass spectrometry (SELDI-TOF MS). In the training set, 14 peak clusters were found to differ significantly in expression between cases and controls. None of the peak clusters were influenced by subjects age and sample storage duration. Ten peak clusters were also found significantly discriminative in the test set. Peak clusters were structurally identified as C3a des-arginine anaphylatoxin, (tentative) inter-α-trypsin inhibitor heavy chain 4 fragments and a fibrinogen fragment. Logistic regression analyses on the training set yielded a classification model with a moderate performance on the test set, corresponding to those reported in previously performed validation studies. Most likely originating from the highly heterogeneous nature of breast cancer, selection of breast cancer subgroups for comparison with healthy controls is expected to improve results of future diagnostic SELDI-TOF MS studies.

Correspondence to: Marie-Christine W. Gast, Department of Pharmacy and Pharmacology, Slotervaart Hospital, P.O. Box 90440, 1006 BK Amsterdam, The Netherlands
E-mail: marie-christine.gast@slz.nl

Key words: breast cancer, diagnosis, protein profiling, serum, surface-enhanced laser desorption/ionisation time-of-flight mass spectrometry

Introduction

The American Cancer Society has estimated that breast cancer will be the most commonly diagnosed cancer among women in the USA in 2008, as it is expected to account for 26% of all new cancer cases among women (1). Following lung cancer, breast cancer currently is the second leading cause of cancer deaths in women (1). As the 5-year survival rate decreases from 98% for localised disease to 26% for distant stage disease (2), early detection is of paramount importance in reducing breast cancer related mortality. The diagnosis of breast cancer is, however, hampered by a lack of adequate detection methods, resulting in detection of only 63% of breast cancers at an early stage (1). Although mammography currently is the most widely applied imaging test today, its predictive value is lower in women with dense breast tissue and smaller lesions. Moreover, no molecular markers are recommended for the (early) detection of breast cancer hitherto. Currently used serum tumour markers in breast cancer, e.g., Cancer Antigen 15.3, lack adequate sensitivity and specificity to be applicable in early detection, and are therefore approved by the FDA only for monitoring therapy of advanced breast cancer or recurrence (3).

The application of a single biomarker in the detection of breast cancer may, however, not be feasible, as a single marker is unlikely to cover the high heterogeneity of breast cancer. Instead, a panel of markers is expected to better reflect breast cancer complexity, yielding improved sensitivity and

specificity. With cancer being, for a large part, a genetic disease, researchers initially searched for biomarkers by employing genomic and transcriptomic approaches. Although this has greatly expanded our insight into the genetic basis of cancer, it is currently understood that the functional 'end-units' of the genome, the proteins, cannot be predicted by genetic and transcriptomic data alone. Due to amongst other post-transcriptional mRNA modifications (e.g., alternative splicing) and post-translational protein modifications, one gene can encode multiple proteins, reflecting both the intrinsic genetic programme of the cell and the impact of its immediate environment (4). As such, the proteome provides a more realistic and detailed view of the biological status, offering a richer source of potential biomarkers.

One of the techniques currently applied in proteomics research of breast cancer is surface-enhanced laser desorption/ionisation time-of-flight mass spectrometry (SELDI-TOF MS). Until now, eleven studies have been published in which the SELDI-TOF MS platform was applied with varying success in the identification and validation of serum markers for diagnosis (5-12), prognosis (13), or monitoring of therapy-efficacy (14) or -toxicity (15) in breast cancer. However, issues have been raised concerning the robustness and validity of alleged markers discovered by SELDI-TOF MS. A potential drawback of analysing high-dimensional proteomic (SELDI-TOF MS) data for disease associated biomarkers is the propensity to discover patterns among variables that are the result of pre-analytical artefacts in a given sample set, rather than of the pathology of interest (16). Several lines of evidence indicate that pre-analytical variables, e.g., sample collection, processing and storage, can exert profound effects on protein profiles, regardless of true biological variation. In addition, clinical characteristics, such as patients age, could also introduce bias (17). Despite these concerns, only few studies investigating the serum proteome for discovery of breast cancer specific biomarkers investigate the possible influence of pre-analytical and patient-related variables on the expression of potential biomarkers.

The raised issues on the validity and robustness of alleged biomarkers can, however, also be addressed by validation and structural identification (16). Nonetheless, thus far, in breast cancer, only two panels of biomarkers discovered by SELDI-TOF MS have been validated by analysis of independent sample sets, resulting in partial (10,11) or no validation (18). Moreover, only few biomarkers discovered by SELDI-TOF MS breast cancer research have been structurally identified.

In the current study, we aimed to discover and validate novel serum protein profiles specific for breast cancer. To this end, archival sera of breast cancer patients and healthy controls were analysed using SELDI-TOF MS. Spectral data were merged in one file, after which they were randomly and evenly split into a training and test set. In the training set, we detected 14 discriminating peak clusters, one cluster of which was structurally identified. Furthermore, the relationship between the intensity of the classifier peak clusters and breast cancer status was adjusted for demographic and pre-analytical variables (i.e., subjects age and sample storage duration). Finally, the samples in the test set were applied for validation purposes.

Materials and methods

Study population. Archival sera of 152 breast cancer patients (BC) and 129 female healthy controls (HC), collected between January 2003 and July 2005, were analysed on different occasions in our laboratory using standardised analytical procedures. All sera were collected prior to any therapy, with individuals informed consent after approval by the institutional review boards. All sera originate from the Netherlands Cancer Institute serum bank, where they had been collected and stored for 3-50 months at -30˚C according to standard procedures.

Chemicals. All used chemicals were obtained from Sigma, St. Louis, MO, USA, unless stated otherwise.

SELDI-TOF MS protein profiling. Serum protein profiling was performed using the ProteinChip SELDI (PBSIIc) reader (Bio-Rad Labs, Hercules, CA, USA). Various chip chemistries, binding and washing procedures and sample pretreatments were initially evaluated to determine which affinity chemistry and sample pretreatment procedure provided the best serum profiles in terms of number and resolution of proteins. Immobilised metal affinity capture (IMAC30) arrays were selected for further analysis. Samples were analysed in three batches (Batch 1: BC: n=40, HC: n=40; Batch 2: BC: n=43, HC: n=46; Batch 3: BC: n=69, HC: n=43). The samples in Batch 1 were analysed in singlicate, while the samples in Batch 2 and 3 were analysed in duplicate. Throughout the assay, arrays were assembled in a 96-well bioprocessor, which was shaken on a platform shaker at 300 rpm.

Arrays were charged twice with 50 μl of 100 mM nickel sulphate (Merck, Darmstadt, Germany) for 15 min, followed by three rinses with deionised water (Braun, Emmenbrücke, Germany) and two equilibrations with 200 μl phosphate-buffered saline (PBS; 0.01 M) pH 7.4/0.5 M sodium chloride/0.1% TritonX-100 (binding buffer; sodium chloride from Merck) for 5 min. Unfractionated serum samples were thawed on ice and denatured by 1:10 dilution in 9 M urea/2% 3-[(3-cholamidopropyl)dimethylammonio-]-1-propanesulfonic acid (CHAPS). Pretreated samples were diluted 1:10 in binding buffer and randomly applied to the arrays. After a 30 min incubation, the arrays were washed twice with binding buffer and twice with PBS pH 7.4/0.5 M sodium chloride for 5 min. Following a quick rinse with deionised water, arrays were air-dried. A 50% sinapinic acid (Bio-Rad Labs) solution in 50% acetonitrile (Biosolve, Valkenswaard, The Netherlands)/0.5% trifluoroacetic acid (Merck) was applied twice (1.0 μl) to the arrays as the matrix. Following air-drying, the arrays were analysed using the ProteinChip SELDI (PBS IIc) reader. As the three batches were analysed on different occasions (with PBS IIc reader maintenance in between), data acquisition was optimised for each sample set separately (data not shown), to obtain similar spectra. For mass accuracy, the instrument was calibrated on each day of measurements with All-in-One peptide standard (Bio-Rad Labs).

Statistics and bioinformatics. Spectra were processed per batch by the ProteinChip Software v3.1 (Bio-Rad Labs). Spectra were baseline subtracted, followed by normalisation to the

Table I. Patient and sample characteristics of the study population.

Parameter	Breast cancer	Healthy control
N	152	129
Age (years), median [IQR]	61.1 [50.3-67.0]	52.0 [42.0-57.7]
Stage[a]		NA
0	7	
1	30	
2A/2B	68/28	
3A/3C	13/6	
Diagnosis[a]		NA
DCIS	6	
IDC	116	
ILC	16	
IDC and ILC	5	
Other	9	
Sample storage time (months), median [IQR]	26.0 [14.1-36.7]	20.1 [12.6-31.9]
Sample collection interval	Apr '03-Jul '05	Jan '03-Jul '05

DCIS, ductal carcinoma *in situ*; IDC, invasive ductal carcinoma; ILC, invasive lobular carcinoma; IQR, interquartile range; NA, not applicable. [a]Pathologically determined stage and diagnosis.

total ion current. Spectra with normalisation factors >2 or <0.5 were excluded from further analysis. The Biomarker Wizard (BMW) software package was applied for peak detection. BMW settings were optimised for each batch separately (data not shown), to ascertain correct detection of real peaks (instead of peaks that merely represent noise). Peak information was subsequently exported as spreadsheet files, and peak intensities from the duplicate analyses in Batch 2 and 3 were averaged. The three batches were analysed on three separate occasions, a parameter known to influence spectral data (19,20). As such, merging peak intensity data of the three batches could lead to spurious results. To this end, first, the intensities of peaks occurring across all three batches were log transformed to obtain normal distributions. Next, the log transformed peak intensities were converted to standard Z-values per batch, by subtracting the mean and dividing by the standard deviation. The log-Z transformed data of the three batches were merged in one file. After this, cases and controls were randomly divided over a training (BC: n=76, HC: n=65) and test (BC: n=76, HC: n=64) set. In the training set, the parametric T-test was applied for the comparison of the mean log-Z transformed peak intensities between cases and controls. Resulting p-values were corrected for multiple testing by the Bonferroni method, by multiplying p-values with the number of peak clusters detected and tested.

To estimate the influence of subjects age and storage duration on the relationship between the 14 discriminating peak clusters and breast cancer status, logistic regression analyses were performed on the training set. We calculated a crude odds ratio per peak cluster, using a univariate model (i.e., by inclusion of only one peak cluster as continuous variable). Next, multivariate odds ratios adjusted for subject's age (categorized according to tertiles: ≤51.3 years, 51.3-61.4 years, >61.4 years), and storage duration (categorized according to tertiles: ≤14.5 months, 14.5-31.7 months, >31.7 months) were calculated. Both parameters were considered confounders if the adjusted odds ratio was 10% different from the crude odds ratio.

To investigate the relationship between a combination of the log-Z transformed peak intensities and the presence of breast cancer, crude odds ratios for each of the peak intensities (as continuous variables) were estimated in a logistic regression model with the inclusion of all peak clusters detected based on forward entry ($p<0.05$). Again, to investigate whether the relationship between peak intensities and the presence of breast cancer could be explained by the age of the subjects and/or sample storage duration, the odds ratios were adjusted for these parameters.

The classification performance of the logistic regression model was evaluated by estimation of the area under the receiver operating characteristic (ROC) curve (AUC) and accompanying 95% confidence interval. The model was subsequently applied to the test set for validation purposes. All statistical analyses were performed using SPSS statistical software, version 13.0 (SPSS Inc., Chicago, IL, USA).

Peptide purification and identification. Structural identification of potential biomarkers was performed previously (Gast *et al*, unpublished data). Briefly, potential markers were purified from serum using anion-exchange chromatographic, size exclusion, and gel-electrophoresis techniques, followed by trypsin digestion. The peptide map of the digest, acquired on the ProteinChip SELDI (PBS IIc) Reader, was investigated with the NCBI database using the ProFound search engine at http://prowl.rockefeller.edu/prowl-cgi/profound.exe. Confirmation of protein identity was provided by sequencing tryptic digest peptides by quadrupole-TOF (Q-TOF) MS (Applied Biosystems/MSD Sciex, Foster City, CA, USA) fitted with a ProteinChip Interface (PCI-1000). Fragment ion spectra were taken to search the SwissProt 44.2 database (Homo Sapiens: 11072 sequences) using the MASCOT search engine at www.matrixscience.com (Matrix Science Ltd., London, UK). Protein identity was further confirmed by immunoassay on ProteinA beads (Gast *et al*, unpublished data).

Results

Study population. Patient and sample characteristics are summarized in Table I. The healthy controls were significantly younger than the breast cancer patients at time of sample procurement [Mann-Whitney U test (MWU); $p<0.001$]. The majority of breast cancer patients had invasive ductal carcinoma (76%) and was diagnosed with Stage 2 (63%) disease. The median sample storage duration was slightly longer for breast cancer sera (median: 26.0 months) than for the healthy control sera (median: 20.1 months) (MWU; $p=0.018$).

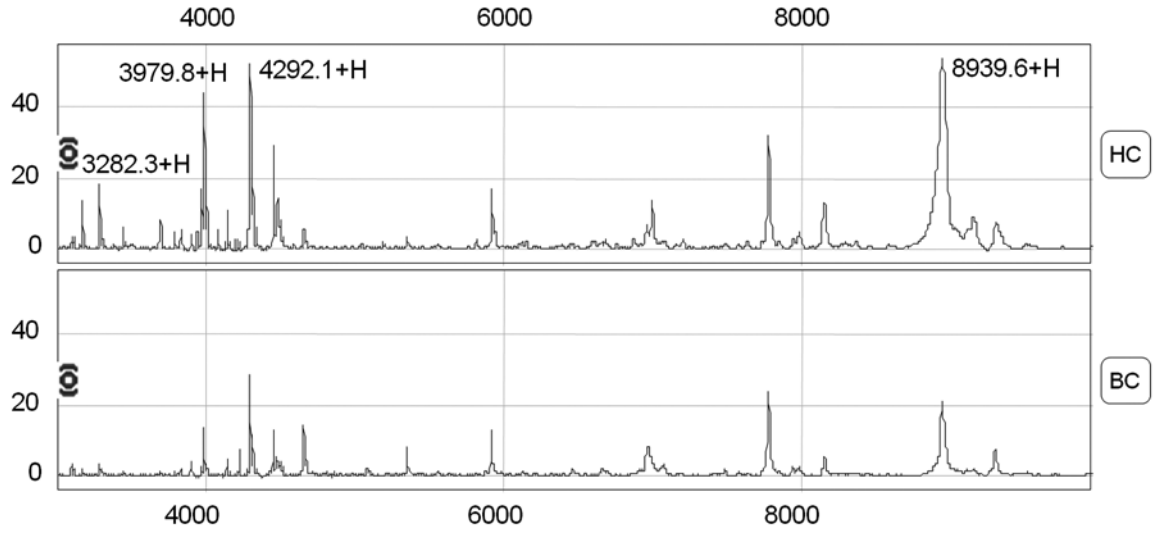

Figure 1. Representative example of protein profiles obtained from a healthy control (HC) and a breast cancer patient (BC).

Table II. Characteristics of the 14 clusters that differ significantly in expression between breast cancer and healthy control in the training set.

Cluster	T-test		Logistic regression analyses					
	Training set	Test set	Training set[a]		Training set[b] (adjusted)		Test set[a]	
(m/z)	p-value[c]	p-value[c]	OR	(95% CI)	OR	(95% CI)	OR	(95% CI)
2733	0.011	0.047	0.45	(0.28-0.70)	0.50	(0.31-0.81)	0.52	(0.34-0.79)
3166	<0.001	0.013	0.40	(0.26-0.62)	0.41	(0.25-0.67)	0.50	(0.34-0.74)
3282	<0.001	0.005	0.41	(0.26-0.64)	0.44	(0.28-0.70)	0.49	(0.34-0.72)
3299	<0.001	<0.001	0.42	(0.27-0.64)	0.41	(0.26-0.66)	0.40	(0.26-0.61)
3691	<0.001	NS	0.37	(0.23-0.59)	0.37	(0.22-0.63)	0.58	(0.40-0.85)
3782	0.004	0.004	0.44	(0.28-0.68)	0.48	(0.30-0.75)	0.49	(0.33-0.71)
3965	0.005	0.004	0.45	(0.29-0.69)	0.52	(0.33-0.81)	0.49	(0.33-0.71)
3980	0.007	NS	0.47	(0.31-0.71)	0.49	(0.31-0.77)	0.65	(0.45-0.94)
3997	0.003	NS	0.44	(0.29-0.67)	0.44	(0.27-0.70)	0.62	(0.44-0.87)
4219	0.046	NS	1.86	(1.27-2.72)	2.40	(1.51-3.80)	1.80	(1.23-2.64)
4292	0.002	0.028	0.39	(0.24-0.65)	0.42	(0.25-0.71)	0.50	(0.33-0.76)
4309	<0.001	0.007	0.32	(0.20-0.54)	0.32	(0.18-0.56)	0.48	(0.32-0.72)
8940	<0.001	0.004	0.37	(0.24-0.57)	0.35	(0.22-0.58)	0.48	(0.33-0.71)
11745	0.003	0.008	2.21	(1.46-3.33)	2.22	(1.39-3.56)	2.18	(1.43-3.34)

95% CI, 95% confidence interval; NS, not significant; OR, odds ratio. [a]Crude logistic regression analyses, by inclusion of one peak cluster (continuous), [b]adjusted logistic regression analyses (training set only), by inclusion of one peak cluster (continuous), subject's age (categorical), and sample storage duration (categorical), [c]Bonferroni corrected p-values.

SELDI-TOF MS protein profiling. Representative SELDI-TOF MS spectra are presented in Fig. 1. Following spectrum pre-processing and normalisation, 73 (BC: n=36; HC: n=37), 89 (BC: n=43; HC: n=46), and 111 samples (BC: n=68; HC: n=43) were left for analysis in Batch 1, 2, and 3, respectively. The Biomarker Wizard detected 57 peak clusters across all three batches. In the training set, 14 peak clusters were found significantly different in expression between breast cancer and control (T-test; Bonferroni corrected $p<0.05$, Table II). Except for the m/z 4219 and m/z 11745 peak clusters, intensities were found decreased in breast cancer compared to control (Table II: logistic regression, odds ratio <1). Following correction for subject's age and sample storage duration, the adjusted odds ratios of three peak clusters (m/z 2733, 3965, and 4219) differed by >10% from the crude odds ratios. All three peaks remain, however,

Table III. Multivariate logistic regression analyses in the training set, by forward entry inclusion of all peak clusters detected, before and after adjustment for subjects age and sample storage duration.

	Multivariate model			Multivariate model, adjusted		
Variable	OR	(95% CI)	p-value	OR	(95% CI)	p-value
m/z 4219	1.94	(1.24-3.04)	0.004	2.78	(1.59-4.86)	<0.001
m/z 4309	0.26	(0.14-0.48)	<0.001	0.26	(0.13-0.52)	<0.001
m/z 5350	0.62	(0.39-0.97)	0.035	0.60	(0.36-1.01)	0.054
m/z 28183	0.53	(0.33-0.83)	0.006	0.49	(0.29-0.85)	0.011

Performance	Multivariate model
ROC AUC	
Training set	0.813 (0.742-0.884)
Test set	0.713 (0.626-0.800)
Training set	
Sensitivity	74.3%
Specificity	71.9%
Test set	
Sensitivity	72.6%
Specificity	61.3%

AUC, area under the receiver operating characteristic (ROC) curve; 95% CI, 95% confidence interval; OR, odds ratio; ROC, receiver operating characteristics curve.

significantly related to breast cancer status. Ten of these 14 peak clusters were found significantly different in peak expression between breast cancer and control in the test set as well.

Next, multivariate logistic regression analyses were performed on the training set. Following forward entry inclusion of all peak clusters detected spectrum-wide, four peak clusters (m/z 4219, 4309, 5350, and 29183) were incorporated in the model, resulting in a ROC AUC of 0.813 (85% CI: 0.742-0.884) (Table III). Two peak clusters (m/z 4219 and 4309) were already found significantly different in peak expression between breast cancer and healthy control. Of the four peak clusters included in this model, only m/z 4219 had an adjusted odds ratio that differed >10% from the crude odds ratio. Similar to the univariate analyses, however, after adjustment this peak cluster was even more strongly related to breast cancer status. The multivariate model classified the samples in the training set with a sensitivity and specificity of 74.3 and 71.9%, respectively. Model performance was lower following validation on the test set [ROC AUC: 0.713 (95% CI: 0.626-0.800); sensitivity: 72.6%, specificity: 61.3%].

Peptide purification and identification. One of the 14 peak clusters found significantly different between breast cancer and control was m/z 8940, which we previously identified as complement component 3 precursor by peptide mapping (ProFound; estimated Z-score 1.57, 4% sequence coverage) (Gast *et al*, unpublished data). Amino acid sequencing of 6 peptides in the tryptic digest by tandem MS on a Q-TOF identified the marker as C3a des-arginine anaphylatoxin ($C3a_{desArg}$, 61% sequence coverage), a 76 amino acid protein with theoretical mass 8939.46 Da and pI 9.54. This identity was confirmed by an immunoassay, for which ProteinA beads were loaded with a C3a polyclonal antibody (Abcam Ltd., Cambridge, UK) (Gast *et al*, unpublished data).

Fig. 2 depicts the correlation matrix presenting the (absolute) Pearson's correlation coefficients calculated between the peak intensities of the 14 peaks found significantly different in expression between breast cancer and healthy control. To preclude bias by group, all Pearson's correlation analyses were performed in the healthy controls of the total study population. As 11 peak clusters were found highly correlated to each other (Pearson's R>0.63, Fig. 2), we hypothesize these clusters to represent multiple fragments of one founder protein. Using data from previous publications, we suggest this founder protein to be inter-α-trypsin inhibitor heavy chain 4 (ITIH4). Eight of the alleged ITIH4 peak clusters had an observed mass corresponding to the theoretical mass of the different ITIH4 fragments described in the literature (Table IV). The peak clusters at m/z 4219 and 11745 were not correlated to any of the significantly different peak clusters. The m/z 4219 and m/z 5350 peak clusters, selected in the multivariable logistic regression analysis, were previously identified as (putative) fibrinogen fragments by our group.

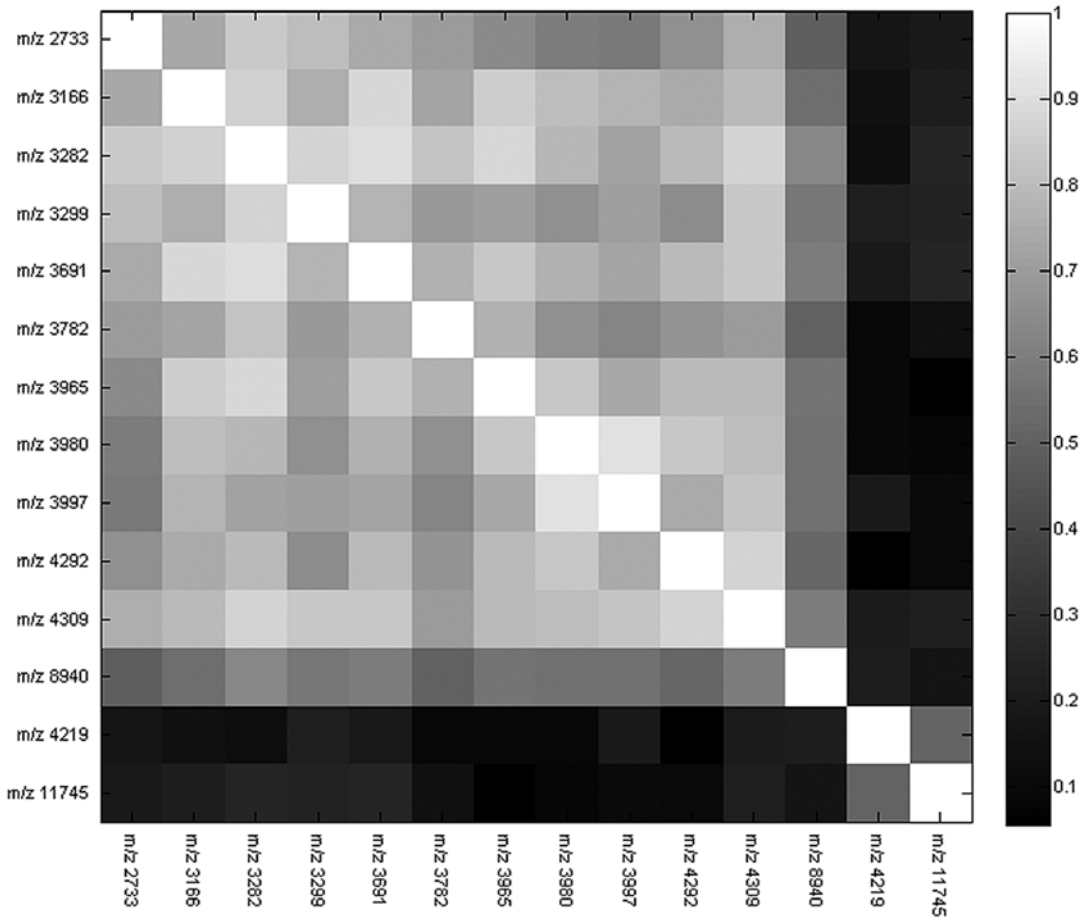

Figure 2. Peak intensity correlation matrix for the 14 peaks found significantly different in expression between breast cancer and healthy control in the training set (for clarity, Pearson's correlation coefficients were converted into absolute values).

Table IV. Stuctural identities of eight alleged ITIH4 peak clusters that significantly differ in expression between breast cancer and healthy control in the training set.

Mr (obs)	Mr (calc)	Putative structural identity		
(m/z)	(Da)	Start - End	Amino acid sequence	Ref.
2733	2725.06	662-688	R.PGVLSSRQLGLPGPPDVPDHAAYHPF.R	(43-45)
3166	3157.58	617-644	R.NVHSGSTFFKYYLQGAKIPKPEASFSPR.R	(44-46)
3282	3273.72	658-688	R.MNFRPGVLSSRQLGLPGPPDVPDHAAYHPF.R	(43-45)
3299	3289.72	658-688	R.MNFRPGVLSSRQLGLPGPPDVPDHAAYHPF.R, Met-Ox	
3965	3957.46	654-690	A.AGSRMNFRPGVLSSRQLGLPGPPDVPDHAAYHPFRR.L	(43,46)
3980	3973.46	654-690	A.AGSRMNFRPGVLSSRQLGLPGPPDVPDHAAYHPFRR.L, Met-Ox	(43,46)
4292	4284.83	650-690	R.QAGAAGSRMNFRPGVLSSRQLGLPGPPDVPDHAAYHPFRR.L	(46)
4309	4300.83	650-690	R.QAGAAGSRMNFRPGVLSSRQLGLPGPPDVPDHAAYHPFRR.L, Met-Ox	(46)

ITIH4, inter-α-trypsin inhibitor heavy chain 4; Mr (obs), observed mass-to-charge ratio; Mr (calc), calculated mass from the matched peptide sequence.

Discussion

In the current study, sera of breast cancer patients (n=152) and healthy controls (n=129) were analysed using the SELDI-TOF MS technology. Spectra were divided into a training and test set, and 14 peak clusters were found to differ significantly in peak expression between breast cancer and healthy control in the training set. Ten of these 14 peak

clusters could also be validated in the test set. We previously identified one peak cluster as $C3a_{desArg}$, while 12 other peak clusters were tentatively identified as ITIH4 fragments and a fibrinogen fragment, respectively. A classification model was subsequently generated by multivariate logistic regression analysis on the training set. Its performance on the test set was similar to those reported by previously performed independent validation studies (10,11,18,21). Hence, our split-sample approach yielded reliable estimates of performance. Nonetheless, the diagnostic performances reported thus far are moderate. The identification of a general diagnostic biomarker is, however, seriously challenged by the molecular characteristics of breast cancer, which are highly heterogeneous (22-24). As such, selection of breast cancer subgroups for comparison with healthy controls is expected to improve results of future diagnostic SELDI-TOF MS studies.

Complement $C3a_{desArg}$. We discovered the expression of the serum m/z 8940 $C3a_{desArg}$ peak to be significantly decreased in breast cancer compared to controls in both the training and test set. Complement C3 is the most abundant (1.2 mg/ml) complement protein in serum (25), supporting the activation of all three pathways of complement activation (the classic, alternative, and lectin pathway) (26,27). Produced mainly in the liver and adipocytes, C3a is formed by cleavage of C3 (185 kDa) by C3-convertases into C3b (176 kDa) and C3a (8.9 kDa) (28). The anaphylatoxin C3a is only short lived in serum as carboxypeptidases cleave the C-terminal arginine residue, creating the more stable, but biologically inactive $C3a_{desArg}$ (8.9 kDa) (27-29).

As C3 is a positive acute phase reactant (30), elevated serum levels of C3 [and hence, $C3a_{(desArg)}$] in cancer compared to control are anticipated. Indeed, elevated serum C3 levels have been described in various cancer types, including neuroblastoma (31), lung cancer (32), and cancer of the digestive tract (33). Likewise, increased serum $C3a_{desArg}$ levels, determined by SELDI-TOF MS, have been reported in breast (9,10), hepatocellular (34), and colorectal cancer (35,36), and chronic lymphoid malignancies (37). We, on the other hand, observed decreased $C3a_{desArg}$ levels in breast cancer in the current study population, as well as in a subset thereof, which we analysed for validation of the 8.9 kDa marker reported by Li *et al* (21). Other studies have described decreased 8.9 kDa peak intensities in breast (7,38), and lung cancer (39). Moreover, Li *et al* observed decreased SELDI-TOF MS $C3a_{desArg}$ peak intensities in sera of metastatic breast cancer patients (10). Their finding is corroborated by the decreased serum C3 levels reported in patients with metastatic breast, gastric, and colorectal cancer (33) and brain tumours (31). Hence, complement activation seems an early event during tumourigenesis. This, however, can not explain the results of the current study, as we included only sera of patients with locally invasive breast cancer.

An other possible explanation for the observed inconsistensies in 8.9 kDa $C3a_{desArg}$ regulation can be the *in vitro* complement activation, caused by coagulation induced platelet activation (40). Banks *et al* (41) reported the intensity of an IMAC3 m/z 8939 peak (not structurally identified though) to significantly increase with prolonged coagulation times. Coagulation time is, however, an unlikely confounder, as studies generally apply standardised collection protocols for both cancer and control samples. $C3a_{desArg}$ levels can also be affected by sample storage time. In a previous study, we found the m/z 8939 $C3a_{desArg}$ peak intensity positively correlated to sample storage time during the first three years of storage, after which intensities remained stable. Although in the current study, the breast cancer sera were stored for a slightly longer period than the control sera, storage time of both sample groups was less than three years. Moreover, as the m/z 8939 peak performance was not influenced by adjustment for sample storage duration, this parameter is unlikely to have confounded results of the current study.

ITIH4 fragments. Of the 14 peaks we found significantly different in expression between breast cancer and healthy controls, 11 were identified as putative ITIH4 fragments. The peak intensities of all putative ITIH4 fragments were decreased in breast cancer compared to control. ITIH4, a 120 kDa plasma glycoprotein expressed mainly in the liver, acts as a positive acute phase reactant and is extensively proteolytically processed (42). Plasma kallikrein readily cleaves ITIH4 into an N-terminal 85 kDa and C-terminal 35 kDa fragment, after which the 85 kDa fragment is further cleaved into an N-terminal 57 kDa and a putative 28 kDa fragment. The latter fragment has not been detected in its entirety hitherto, as it is rapidly cleaved into subsequent smaller fragments (42). Changes in the abundance of different fragments have been found associated with various types of cancer (e.g., prostate, breast, ovarian, colorectal and pancreatic cancer) (42-44), indicating cancer-type specific proteolytic processing of ITIH4. Three of the 11 putative ITIH4 fragments (i.e., m/z 2733, m/z 3282, and m/z 4292) have been reported as potential markers for breast cancer (42). Unlike our results, however, this study found increased peak intensities of the three fragments in cancer compared to control (42).

The m/z 4292 ITIH4 fragment has also been described by Li *et al* (9,10). They initially observed a 4.3 kDa ITIH4 fragment to be downregulated in breast cancer (9), but found this peak upregulated upon validation (10). In their original discovery study, the cancer sera were collected during a (non-specified) longer time interval than the control sera, whereas in the validation study, sera of both cases and controls were collected within a two-year time interval. Combined with the postulated instability of the ITIH4 fragment (causing further truncation during prolonged storage), this could indeed explain their discrepant results. Nonetheless, following analysis of prospectively collected sera, Mathelin *et al* (11) also observed a decreased expression of the m/z 4292 ITIH4 peak intensity in breast cancer. This decrease was also observed following analysis of a subset of the current study population for validation of the markers reported by Li *et al* (21). However, the decrease of m/z 4292 observed in the breast cancer cases of the current study could not be explained by the difference in storage duration between the cancer and control sera, as correction for this parameter by logistic regression analyses did not affect the performance of the m/z 4292 peak. In addition, evidence for the alleged 4.3 kDa ITIH4 fragment instability is only limited. Peak intensities of

this fragment were found to be both increased (45) and decreased by different (pre-) analytical parameters (42,43), though the fragmentation pattern was not altered (42,43). Perhaps the discrepant results of the various studies are caused by differences between the patient populations investigated in the various studies.

Other markers. Of the 14 peak clusters found significantly different in expression between breast cancer and healthy controls, both m/z 4219 and m/z 11745 were not correlated to any of the other peak clusters. While we previously identified m/z 4219 as a putative fibrinogen fragment, the identity of m/z 11745 peak is yet unknown. The m/z 5350 peak cluster, included in the logistic regression model, was identified earlier as a fibrinogen fragment as well (i.e., fibrinogen α-E fragment $FGA_{576-625}$). The multivariate classification model furthermore designated the m/z 28183 peak cluster as a potential marker, in combination with m/z 4219, 4309, and 5350. Although peak intensities of both m/z 5350 and m/z 28183 were not significantly different in expression between breast cancer and healthy control, combination with other markers evidently improved their diagnostic performance. The m/z 5350 peak cluster, though not structurally identified, has been reported earlier as significantly increased in sera of patients with lung cancer (39) and hypopharyngeal squamous cell carcinoma (46). Based on the observed mass, we hypothesise the m/z 28183 peak cluster to represent apolipoprotein A-I. This protein was previously identified by our group as a potential marker for colorectal cancer by serum SELDI-TOF MS analyses (47). Synthesised both in the liver and small intestine, apolipoprotein A-I constitutes the major component of high-density lipoproteins (48). It is a negative acute phase reactant (49), explaining the decreased expression we observed in cancer vs. healthy control (Table II, crude odds ratio <1). Its decreased expression in cancer is confirmed by other studies investigating breast (48), ovarian (50), colorectal (47), and hepatocellular cancer (51).

In conclusion, using SELDI-TOF MS, we discovered and validated 10 peak clusters that significantly differ in expression between sera of breast cancer patients and healthy controls. These peak clusters were structurally identified as the high abundant $C3a_{desArg}$ anaphylatoxin, and putative ITIH4 and fibrinogen fragments. Logistic regression analyses in the training set yielded a classification model with a performance comparable to those reported in previously performed independent validation studies. As these moderate performances most likely originate from the highly heterogeneous nature of breast cancer, selection of breast cancer subgroups for comparison with healthy controls is expected to improve results of future diagnostic SELDI-TOF MS studies.

Acknowledgements

This study was supported by a grant of the Dutch Cancer Society (project NKI 2005-3412).

References

1. Jemal A, Siegel R, Ward E, *et al*: Cancer statistics, 2008. CA Cancer J Clin 58: 71-96, 2008.
2. Ries L, Melbert D, Krapcho M, Stinchcomb D, Howlader N, Horner M, *et al*: SEER Cancer Statistics Review, 1975-2005, National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2005/ - based on November 2007 SEER data submission, posted on the SEER website, 2008.
3. Stieber P, Molina R, Chan DW, *et al*: Clinical evaluation of the Elecsys CA 15-3 test in breast cancer patients. Clin Lab 49: 15-24, 2003.
4. Banks RE, Dunn MJ, Hochstrasser DF, *et al*: Proteomics: new perspectives, new biomedical opportunities. Lancet 356: 1749-1756, 2000.
5. Becker S, Cazares LH, Watson P, *et al*: Surfaced-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) differentiation of serum protein profiles of BRCA-1 and sporadic breast cancer. Ann Surg Oncol 11: 907-914, 2004.
6. Belluco C, Petricoin EF, Mammano E, *et al*: Serum proteomic analysis identifies a highly sensitive and specific discriminatory pattern in Stage 1 breast cancer. Ann Surg Oncol 14: 2470-2476, 2007.
7. Hu Y, Zhang S, Yu J, *et al*: SELDI-TOF-MS: the proteomics and bioinformatics approaches in the diagnosis of breast cancer. Breast 14: 250-255, 2005.
8. Laronga C, Becker S, Watson P, *et al*: SELDI-TOF serum profiling for prognostic and diagnostic classification of breast cancers. Dis Markers 19: 229-238, 2003.
9. Li J, Zhang Z, Rosenzweig J, *et al*: Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. Clin Chem 48: 1296-1304, 2002.
10. Li J, Orlandi R, White CN, *et al*: Independent validation of candidate breast cancer serum biomarkers identified by mass spectrometry. Clin Chem 51: 2229-2235, 2005.
11. Mathelin C, Cromer A, Wendling C, *et al*: Serum biomarkers for detection of breast cancers: a prospective study. Breast Cancer Res Treat 96: 83-90, 2006.
12. Vlahou A, Laronga C, Wilson L, *et al*: A novel approach toward development of a rapid blood test for breast cancer. Clin Breast Cancer 4: 203-209, 2003.
13. Goncalves A, Esterni B, Bertucci F, *et al*: Postoperative serum proteomic profiles may predict metastatic relapse in high-risk primary breast cancer patients receiving adjuvant chemotherapy. Oncogene 25: 981-989, 2006.
14. Pusztai L, Gregory BW, Baggerly KA, *et al*: Pharmacoproteomic analysis of prechemotherapy and postchemotherapy plasma samples from patients receiving neoadjuvant or adjuvant chemotherapy for breast carcinoma. Cancer 100: 1814-1822, 2004.
15. Heike Y, Hosokawa M, Osumi S, *et al*: Identification of serum proteins related to adverse effects induced by docetaxel infusion from protein expression profiles of serum using SELDI ProteinChip system. Anticancer Res 25: 1197-1203, 2005.
16. Ransohoff DF: Lessons from controversy: ovarian cancer screening and serum proteomics. J Natl Cancer Inst 97: 315-319, 2005.
17. Villanueva J, Philip J, Chaparro CA, *et al*: Correcting common errors in identifying cancer-specific serum peptide signatures. J Proteome Res 4: 1060-1072, 2005.
18. Gast MC, Bonfrer JM, van Dulken EJ, *et al*: SELDI-TOF MS serum protein profiles in breast cancer: assessment of robustness and validity. Cancer Biomark 2: 235-248, 2006.
19. Hu J, Coombes KR, Morris JS, *et al*: The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales. Brief Funct Genomic Proteomic 3: 322-331, 2005.
20. Karsan A, Eigl BJ, Flibotte S, *et al*: Analytical and preanalytical biases in serum proteomic pattern analysis for breast cancer diagnosis. Clin Chem 51: 1525-1528, 2005.
21. van Winden AWJ, Gast MCW, Beijnen JH, *et al*: Validation of previously identified serum biomarkers for breast cancer with SELDI-TOF MS: a case control study. BMC Med Genomics 2: 4, 2009.
22. Sorlie T, Perou CM, Tibshirani R, *et al*: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci USA 98: 10869-10874, 2001.
23. Perou CM, Sorlie T, Eisen MB, *et al*: Molecular portraits of human breast tumours. Nature 406: 747-752, 2000.
24. Bertucci F, Birnbaum D and Goncalves A: Proteomics of breast cancer: principles and potential clinical applications. Mol Cell Proteomics 5: 1772-1786, 2006.

25. Hugli TE: Human anaphylatoxin (C3a) from the third component of complement. Primary structure. J Biol Chem 250: 8293-8301, 1975.
26. Bohana-Kashtan O, Ziporen L, Donin N, *et al*: Cell signals transduced by complement. Mol Immunol 41: 583-597, 2004.
27. Sahu A, Sunyer JO, Moore WT, *et al*: Structure, functions, and evolution of the third complement component and viral molecular mimicry. Immunol Res 17: 109-121, 1998.
28. de Bruijn MH and Fey GH: Human complement component C3: cDNA coding sequence and derived primary structure. Proc Natl Acad Sci USA 82: 708-712, 1985.
29. Nettesheim DG, Edalji RP, Mollison KW, *et al*: Secondary structure of complement component C3a anaphylatoxin in solution as determined by NMR spectroscopy: differences between crystal and solution conformations. Proc Natl Acad Sci USA 85: 5036-5040, 1988.
30. Gabay C and Kushner I: Acute-phase proteins and other systemic responses to inflammation. N Engl J Med 340: 448-454, 1999.
31. Carli M, Bucolo C, Pannunzio MT, *et al*: Fluctuation of serum complement levels in children with neuroblastoma. Cancer 43: 2399-2404, 1979.
32. Gminski J, Mykala-Ciesla J, Machalski M, *et al*: Immunoglobulins and complement components levels in patients with lung cancer. Rom J Intern Med 30: 39-44, 1992.
33. Maness PF and Orengo A: Serum complement levels in patients with digestive tract carcinomas and other neoplastic diseases. Oncology 34: 87-89, 1977.
34. Lee IN, Chen CH, Sheu JC, *et al*: Identification of complement C3a as a candidate biomarker in human chronic hepatitis C and HCV-related hepatocellular carcinoma using a proteomics approach. Proteomics 6: 2865-2873, 2006.
35. Habermann JK, Roblick UJ, Luke BT, *et al*: Increased serum levels of complement C3a anaphylatoxin indicate the presence of colorectal tumors. Gastroenterology 131: 1020-1029, 2006.
36. Ward DG, Suggett N, Cheng Y, *et al*: Identification of serum biomarkers for colon cancer by proteomic analysis. Br J Cancer 94: 1898-1905, 2006.
37. Miguet L, Bogumil R, Decloquement P, *et al*: Discovery and identification of potential biomarkers in a prospective study of chronic lymphoid malignancies using SELDI-TOF-MS. J Proteome Res 5: 2258-2269, 2006.
38. Shin S, Cazares L, Schneider H, *et al*: Serum biomarkers to differentiate benign and malignant mammographic lesions. J Am Coll Surg 204: 1065-1071, 2007.
39. Han KQ, Huang G, Gao CF, *et al*: Identification of lung cancer patients by serum protein profiling using surface-enhanced laser desorption/ionization time-of-flight mass spectrometry. Am J Clin Oncol 31: 133-139, 2008.
40. Hamad OA, Ekdahl K, Lambris JD and Nilsson B: Complement activation triggered by thrombin receptor-activated platelets. Mol Immunol 44: 180, 2007.
41. Banks RE, Stanley AJ, Cairns DA, *et al*: Influences of blood sample processing on low-molecular-weight proteome identified by surface-enhanced laser desorption/ionization mass spectrometry. Clin Chem 51: 1637-1649, 2005.
42. Song J, Patel M, Rosenzweig CN, *et al*: Quantification of fragments of human serum inter-alpha-trypsin inhibitor heavy chain 4 by a surface-enhanced laser desorption/ionization-based immunoassay. Clin Chem 52: 1045-1053, 2006.
43. Fung ET, Yip TT, Lomas L, *et al*: Classification of cancer types by measuring variants of host response proteins using SELDI serum assays. Int J Cancer 115: 783-789, 2005.
44. Villanueva J, Shaffer DR, Philip J, *et al*: Differential exoprotease activities confer tumor-specific serum peptidome patterns. J Clin Invest 116: 271-284, 2006.
45. Timms JF, Arslan-Low E, Gentry-Maharaj A, *et al*: Preanalytic influence of sample handling on SELDI-TOF serum protein profiles. Clin Chem 53: 645-656, 2007.
46. Zhou L, Cheng L, Tao L, *et al*: Detection of hypopharyngeal squamous cell carcinoma using serum proteomics. Acta Otolaryngol 126: 853-860, 2006.
47. Engwegen JY, Helgason HH, Cats A, *et al*: Identification of serum proteins discriminating colorectal cancer patients and healthy controls using surface-enhanced laser desorption ionisation-time of flight mass spectrometry. World J Gastroenterol 12: 1536-1544, 2006.
48. Chang SJ, Hou MF, Tsai SM, *et al*: The association between lipid profiles and breast cancer among Taiwanese women. Clin Chem Lab Med 45: 1219-1223, 2007.
49. Van Lenten BJ, Reddy ST, Navab M, *et al*: Understanding changes in high density lipoproteins during the acute phase response. Arterioscler Thromb Vasc Biol 26: 1687-1688, 2006.
50. Zhang Z, Bast RC Jr, Yu Y, *et al*: Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. Cancer Res 64: 5882-5890, 2004.
51. Steel LF, Shumpert D, Trotter M, *et al*: A strategy for the comparative analysis of serum proteomes for the discovery of biomarkers for hepatocellular carcinoma. Proteomics 3: 601-609, 2003.