RESEARCH ARTICLE

Acute phase serum amyloid A in ovarian cancer as an important component of proteome diagnostic profiling

Sergei A. Moshkovskii[1], Maria A. Vlasova[1], Mikhail A. Pyatnitskiy[1], Olga V. Tikhonova[1], Metanat R. Safarova[2], Oleg V. Makarov[2] and Alexander I. Archakov[1]

[1] Institute of Biomedical Chemistry, Moscow, Russia
[2] Russian State Medical University, Moscow, Russia

Received: December 15, 2005
Revised: August 1, 2006
Accepted: September 13, 2006

In the context of serum amyloid A (SAA) identification as ovarian cancer marker derived by SELDI-MS, its serum levels were measured by immunoassay in different stages of ovarian cancer, in benign gynecological tumors, and in healthy controls. In addition, SELDI-TOF-MS spectra were obtained by protocol optimized for the SAA peak intensity. SELDI data on small proteins (5.5–17.5 kDa) and SAA immunoassay data were combined with cancer antigen (CA)125 data in order to study the classification accuracy between cancer and noncancer by support vector machine (SVM), logistic regression, and top scoring pair classifiers. Although an addition of SAA immunoassay data to CA125 data did not significantly improve cancer/noncancer discrimination, SVM applied to combined biomarker data (CA125 and SAA immunoassay variables plus 48 SELDI peak variables) yielded the best classification rate (accuracy 95.2% *vs.* 86.2% for CA125 alone). Notably, most of discriminatory peaks selected by the classifiers have significant correlation with the major known peaks of SAA (11.7 kDa) and transthyretin (13.9 kDa). Acute phase serum amyloid A (A-SAA) was proved to be an important member of cancer discriminatory protein profile. Among the eight known ovarian cancer SELDI profile components, A-SAA is the most relevant to molecular pathogenesis of cancer and it has the highest degree of up-regulation in disease.

Keywords:
Biomarker / Ovarian cancer / SELDI / Serum amyloid A / Serum proteome

1 Introduction

Proteomics has recently announced a new approach to molecular diagnosis of disease, in particular, of those cancers that lack early detection in consequence of their evident clinical manifestation only at advanced stages. Of various proteome methodologies, plasma/serum profiling on SELDI-MS chips became the most popular tool for disease biomarker discovery and new pattern-based diagnosis [1]. Beginning from initial experience [2], these approaches based on statistical modeling SELDI-MS profiles have shown excellent results in terms of diagnosis specificity and sensitivity.

At the same time, low reproducibility of the marker masses obtained by different scientific groups that investigated the same disease condition, as well as the lack of marker identity information causes some skepticism among experts [3]. Actually, despite the possible major component depletion, the SELDI protein chip approach just skims the cream off the serum proteome and skips an overwhelming majority of "traditional" cancer markers derived from molecular biology and oncoimmunology background. Among them, for ovarian cancer, there are fetal glycoproteins such as

Correspondence: Dr. Sergei A. Moshkovskii, Center of Proteomics, Institute of Biomedical Chemistry, 10 Pogodinskaya Str., Moscow 119121, Russia
E-mail: sergei.moshkovskii@ibmc.msk.ru
Fax: +7-495-245-0857

Abbreviations: AIC, Akaike information criterion; **A-SAA**, acute phase serum amyloid A; **CA,** cancer antigen; **LR**, logistic regression; **RFE,** recursive feature elimination; **SAA,** serum amyloid A; **SVM**, support vector machine; **TSP,** top scoring pair; **TTR,** transthyretin

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

cancer antigen (CA)125, cytokines, *e.g.*, M-CSF [4], specific proteases, such as prostasin and kallikreins [5], with their serum levels being below the sensitivity of up-to-date plasma MS profiling.

Anyway, now it is obvious that discriminatory SELDI peak identification is highly desirable. The first work which identified three biomarkers by means of SELDI technology was published in 2002 [6]. These authors identified transferrin (79 kDa) and an immunoglobulin heavy chain (54 kDa) as down-regulated proteins in ovarian cancer and a haptoglobin fragment (9.2 kDa) as an up-regulated product. Another work which deciphers the discriminatory peaks in large-scale SELDI profiling ovarian cancer sera [7] reports three protein products, truncated transthyretin (TTR) and apolipoprotein A being down-regulated in cancer, while a fragment of inter-α-trypsin inhibitor heavy chain H4 being up-regulated. An identification of further discriminatory proteins for ovarian cancer has recently been reported [8]. In this paper, five diagnostic peaks found before [9] were identified as TTR (13.9 kDa), truncated TTR (12.9 kDa; as in [7]), beta-hemoglobin (15.9 kDa), apolipoprotein A1 (28 kDa; as in [7]), and transferrin (79 kDa; as in [6]). All these biomarkers are decreased in patients with ovarian cancer except beta-hemoglobin which is, in contrast, increased. Notably, all the biomarkers in [8] behave similarly in patients with low malignant (*i.e.*, benign) and malignant ovarian tumors.

Although almost all biomarkers mentioned above fall in the class of acute-phase and host-response proteins and changes in their levels seem to result from the inflammation during cancer invasion, the independent validation of their marker properties yielded significant specificity and sensitivity improvement compared with CA125 marker alone.

With modification of profiling conditions, we could identify another MS profile component characteristic of ovarian cancer, acute phase serum amyloid A (A-SAA) [10]. Retrospective evaluation of the MS serum/plasma profiling studies of ovarian cancer without protein identification [2, 6, 9, 11] reveals the A-SAA mass as a marker only in one case [11]. As the above-mentioned biomarkers described in [7], the A-SAA also is a common acute phase protein, and it has been a target for a lot of clinical and basic research due to its unique properties.

Serum amyloid A (SAA) is a very ancient and highly conservative family of defense proteins, the family being found in vertebrates including fish [12] and even in echinoderm [13]. Functionally, SAA are divided into inducible acute phase SAAs (A-SAA) and constitutive SAA (C-SAA). Human SAA protein family encounters acute phase SAA1 and SAA2, rarely expressed SAA3 and constitutive SAA4 [14]. SAA1 and SAA2 share a very high identity and only differ in 7–8 amino acid residues of 104 mature protein residues.

As early as several decades ago it was shown that both human A-SAAs are extremely induced during acute phase of inflammation, due to infection or trauma, such that their total plasma level may increase 1000-fold from the initial level of approximately 10 μg/mL. Structure, regulation and functions of mammalian SAA have been reviewed in detail [15]. Briefly, A-SAAs elicit a remarkably wide range of protective functions during inflammatory processes. It has lipid-related functions and replaces apolipoprotein A from HDL facilitating the cholesterol uptake by inflamed and damaged tissue. At the same time, A-SAA has some anti-inflammatory effects. It is important that the SAA can induce extracellular matrix degrading enzymes [16], including some matrix metalloproteinases, which play a role in tissue repair processes, but, on the other hand, are involved in pathogeny of inflammatory diseases such as rheumatoid arthritis [17] and in cancer progression [18].

Increased serum SAA levels are described for different cancers. For example, some recent works estimate the serum SAA levels after its detection in MS profiles of lung cancer [19] and renal cancer [20]. However, no data on combining SAA measurement with approved cancer markers is provided, and also its role in cancer development and progression is unclear. In our previous work [10], a proteomic identification of SAA in ovarian cancer plasma was provided without determining exact SAA blood levels in this cancer and its combination with other markers.

In this context, we continued the study of serum SAA levels in ovarian cancer. First of all, we measured SAA levels in an expanded set of different stages of ovarian cancer and benign tumor and normal sera. These data were combined with approved marker CA125 measurement and SELDI protein profiling to evaluate contribution of SAA level and its MS signal to sensitivity and specificity of different statistical classifiers which could distinguish between ovarian cancer sera and noncancer sera. Discriminatory peaks determined by the classifiers were studied in context of those SELDI peaks which were already identified. Finally, known proteins which composed SELDI-based ovarian cancer discriminatory profiles were discussed in context of their serum levels and biological function in cancer.

2 Materials and methods

2.1 Subjects

All patient-related specimens were collected under institutional ethics recommendations. Disease blood specimens were collected preoperatively from 34 women with epithelial ovarian cancer. Further, samples were taken from 14 women with benign ovarian tumors and 17 women with uterine myoma. As healthy controls, specimens from 26 normal women were used which were withdrawn during regular gynecological examination. Tumor histology was determined in operative biopsies. More detailed subject information is provided in Table 1. Subjects involved in this study were distinct from those of our previous work [10] with the exception of some controls whose blood were taken repeatedly.

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 1. Subject set descriptive statistics: case data, serum CA125 and SAA levels

Subject group	Histology	Number of subjects	Age, mean (min–max)	CA125 (U) mean ± SD/*median* (min–max)	SAA (mg/L) mean ± SD/*median* (min–max)
Ovarian cystadenocarcinoma, early stage	Serous	5	47 (26–70)	230 ± 251	748 ± 688
	Mucinous	2		*139*	*708*
				(33–826)	(10–1760)
Ovarian cystadenocarcinoma, late stage	Serous	22		780 ± 1021	1080 ± 1489
	Mucinous	5		*466*	*247*
				(2–4177)	(5–4465)
	Total, cancer	34		667 ± 957	981 ± 1376
				257	*333*
				(2–4177)	(5–4465)
Ovarian cystadenoma	Serous	9	44 (19–73)	33 ± 30	565 ± 1393
	Mucinous	2		*21*	*31*
	Other	3		(2–106)	(2–4767)
Uterine myoma		17	44 (22–56)	56 ± 35	13 ± 8
				45	*11*
				(14–140)	(4–38)
Healthy controls		26	41 (21–59)	15 ± 7	44 ± 53
				14	*26*
				(5–41)	(3–184)
	Total, noncancer	57		32 ± 30	163 ± 717
				20	*17*
				(2–140)	(2–4767)

2.2 SELDI-TOF processing of serum samples

All solvents for MS were of HPLC grade (Merck, Germany). In order to optimize sera processing conditions for A-SAA signal acquisition, NP20, SAX, WCX, H4 protein chips (Ciphergen Biosystems, USA) were used. Optimized protocol includes usage of normal phase NP20 protein chips for A-SAA-directed serum profiling. Each serum was diluted 1:10 with deionized water. One microliter of each diluted serum was applied on chip, air-dried, and washed twice with deionized water according to the manufacturer's manual. Spots were then air-dried once more and coated by two 0.5 μL additions of matrix solution that was a saturated solution of CHCA (Ciphergen) in 0.5% v/v TFA, 50% v/v ACN diluted two-fold by the same solvent (50% saturated CHCA). Chips were analyzed using the SELDI-TOF Protein Biology System II (PBS II) mass-spectrometer (Ciphergen). Spectra were collected automatically in 5500–70 000 Da range with laser intensity 230, detector sensitivity 9, and 90 laser shots *per* spot. Spectra were calibrated using external calibration with Peptide Standard kit (Ciphergen), equine cytochrome c (12 361 Da), whale sperm myoglobin (17 200 Da), and BSA (66 431 Da). All samples were processed at least in duplicate to obtain reproducible result.

2.3 Processing SELDI data

The mass interval used for processing was 5500 to 17 500 Da. Peak clustering was performed using Biomarker Wizard™ software (Ciphergen Biosystems) with the following settings: S/N (first pass) 10, S/N (second pass) 5, minimum peak threshold 0%, and mass error 0.2%. Peak mass and intensity were exported to the MS Excel table, and peak intensities from each duplicate spectra were averaged.

2.4 Immunoassays

For measurement of CA125 concentration, CA125 EIA enzyme immunometric assay (CanAg, Canada) was used. To determine SAA concentration, Human SAA immunoassay Kit (Biosource, USA) was used. All procedures were carried out according to the manufacturer's protocols. All samples were processed at least twice, preferably, in different serum dilution (for high SAA levels) until converging results were reached.

2.5 Statistical analysis

Statistical analysis comprised processing spectral data which consisted of intensity values of 48 peak clusters identified by Biomarker Wizard software (Section 2.3; see also Supporting

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

materials; http://www.ibmc.msk.ru/departments_en/LDP/mosh_support.xls) alone or in combination with two biomarker ELISA data, by two commonly used classification methods such as support vector machine (SVM) [21] and logistic regression (LR) [22] to differentiate between cancer and noncancer patients.

2.5.1 SVM and recursive feature elimination (RFE)

SVM is a supervised machine learning algorithm, which has been widely applied to various biological problems. Its main idea is to search for an optimal hyperplane that separates a given set of binary labeled data with maximum margin between two classes. In our study, prior to running SVM we performed feature selection in order to improve overall classifier performance. As a feature selection technique RFE algorithm [23] was used. RFE is a sequential backward feature elimination method based on SVM. It was proposed for feature selection in microarray analysis and was shown to perform well [23]. At first, algorithm starts with all the variables. At each iteration, SVM is trained with the existing variables and then a variable, which received minimum weight, is removed from the data. This procedure continues until all features are ranked according to the removal order.

In our study, for SVM classifier we used linear kernel and a grid search for tuning the only cost parameter C. We analyzed the performance of SVM classifier with features selected by RFE. When applying RFE, we chose the feature subset, which showed the best performance (estimated by ten-fold cross validation) on the whole dataset.

2.5.2 LR and Akaike information criterion (AIC)

As another approach to predict ovarian cancer from multiple markers, we used LR which estimates the probability of having ovarian cancer. LR is a statistical regression model, which is used when the dependent variable is binary and the independents are of any type. As in the case of the SVM we used a feature selection method in order to increase classifier's performance. In case of feature selection for LR, we employed stepwise model selection based on AIC [24]. AIC criterion tries to examine the complexity of the model together with goodness of its fit to the sample data, and to produce a measure which is a trade-off between the two. Applying forward stepwise AIC-based method, we used resultant features when the algorithm had converged. In the final model each marker was included as a linear term.

2.5.3 Estimation of classifier performance

In order to obtain a distribution of classification accuracy, the ten-fold crossvalidation was run 100 times with both SVM and LR. Use of crossvalidation is a standard technique to estimate classifier generalization ability, *i.e.*, to correctly classify previously unseen examples, when testing dataset is not available or if the whole dataset is relatively small. In k-fold crossvalidation, the original sample is partitioned into K subsamples. Of the K subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $K-1$ subsamples are used as training data. The crossvalidation process is then repeated K times (the folds), with each of the K subsamples used exactly once as the validation data. The K results from the folds are averaged to produce a single estimation. We used ten-fold crossvalidation as recommended in [25]. Also for each classifier, its sensitivity and specificity was calculated.

2.5.4 Top scoring pairs (TSP)

This is a relatively new, promising approach (TSP) for classification and feature selection [26, 27]. It is based only on relative levels of peak intensities and generates simple and easily interpretable decision rules. Classifier searches for pairs of peaks, which have consistently different intensity ranks on two classes (for details please refer to the original paper). For the classifier training and leave-one-out crossvalidation, publicly available TSP software was downloaded from web-site of method inventors (http://www.bme.jhu.edu/~actan/KTSP/) and used with default settings.

2.5.5 Correlation analysis

For analysis of correlation between intensities of different peak clusters, we used Spearman rank-order correlation coefficient. This coefficient does not make the assumption that variables are normally distributed, and it also may be a better indicator if the relationship between two variables is not linear. Correlation was considered statistically significant at $p<0.01$. Correlations that had $p>0.01$ were considered to be statistically insignificant and were forced to have Spearman correlation coefficient equal to zero.

2.5.6 Software

We implemented the whole scheme of applying SVM with RFE and LR with AIC to our data using open-source R language (www.r-project.org). Also all other statistical computations were done with this software.

3 Results

3.1 SAA plasma levels

Serum levels of SAA measured by immunoassay in all involved sera samples were shown to be extremely elevated in many cancer samples, which is consistent with our previous data [10]. Descriptive statistics for SAA concentration as well as for standard marker CA125 levels among the patient groups is set forth in Table 1 and the distribution of these levels in cancer and noncancer subjects is shown in Fig. 1. Individual SAA levels for every serum studied are

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

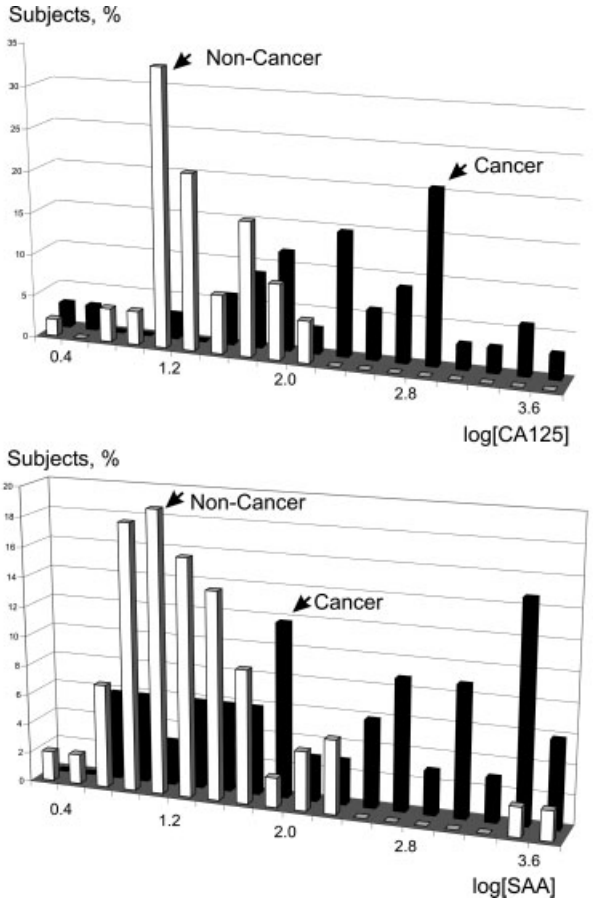

Figure 1. Histograms for the CA125 and SAA level distribution in 34 cancer patients and 57 noncancer subjects of the study set. A common logarithm transformation is applied to CA125 and SAA concentrations in arbitrary units and g/L, respectively.

shown in Fig. 2. It is typical that SAA level increases in disease abruptly rather than gradually and it was described [15] that this protein is often up-regulated by two to three orders of magnitude. Notably, in our data, SAA levels in healthy population are higher than it was reported [20] and amount to 180 mg/L. Probably, this may result from transient inflammatory processes. Cancer-associated increase in SAA level does not seem to be stage-specific, because out of seven early stage cancer sera four are shown to have extremely high SAA levels (Fig. 2). On analyzing case histories of cancer patients with and without high SAA levels, we did not find any specific explanation of such divergence. The reasons for this may be complicated and include differences in tumor vascularity, patient cytokine status, *etc*.

As shown in Fig. 2, two of 14 benign ovarian tumor sera have very high SAA content, both of them being serous cystadenomas. It can be speculated that these high SAA levels may be prognostic in context of near malignization, but this hypothesis needs further confirmation. We also included in the study some patients with uterine myoma. It is known that, in this condition, some slight increase of CA125 over reference level (38–40 U) is observed which sometimes allows suspecting silent malignancy as a false result (see Table 1). In contrast, these sera have very low SAA levels.

In terms of biomarker properties when SAA level is considered alone to distinguish cancer *versus* noncancer, its sensitivity is only 50% (17 detected cancers out of a total 34) at specificity of 96.5% (55/57 noncancers; cut-off level about 0.3 g/L).

3.2 Direct SELDI-MS detection of SAA in serum

In our initial work, SAA served as a discriminatory peak for cancer when plasma thermostable fractions were applied to strong anion exchanger (SAX) surface of SELDI-chips [10]. A thermostable fraction is not quite useful due to its time-consuming preparation and low compatibility with MS as well as some artifacts resulting from heating. Thus, we tried to select a method for SAA signal MS collection using non-processed serum SELDI measurement. The approach should be fast and also effective for SAA detection. After SELDI spectra collection of serum samples containing predetermined SAA levels using different protein chip surfaces (data not shown), best results in terms of SAA peak intensities were obtained using a surprisingly simple protocol with normal phase protein chip NP20 (see Section 2). Using this protocol, focused on 11.52–11.68 kDa peak collection [10], all sera involved in the study were processed.

First of all, it was important to determine the sensitivity of serum SAA detection by SELDI-MS. After processing spectra using Ciphergen Protein chip software to detect the peaks with default parameters, the peak of 11.68 kDa was detected in all sera beginning from the determined SAA level of 333 mg/L. In other words, all 19 sera with SAA level higher than 0.3 g/L have a well-defined peak in the region of interest (Fig. 3). There were no samples in the interval between 333 and 184 mg/L, however, the sensitivity of this SELDI method should be stated as at least 0.3 g/L SAA (shown by dotted line in Fig. 2). The MS SAA determination was not quantitative due to variation in absolute peak intensities inherent in serum spectra and also because of quick signal saturation after 1 g/L level (slight or lack of difference between 1 and 3 g/L). Thus, SELDI-MS can qualitatively detect SAA levels in serum which are about 0.3 g/L or higher. This threshold content corresponds to 2.5×10^{-5} M, but the real molar sensitivity is at least twice lower due to the existence of at least two major forms of every SAA protein [28]. It is necessary to note that this threshold is close to SAA reference levels reported for different inflammation-related conditions [29, 30], thereby direct MS determination should be considered as a possible method for fast screening of elevated SAA level in blood.

Despite to the pour mass accuracy of SELDI-TOF-MS, we could calibrate these spectra quite precisely using external standard of close mass, *e.g.*, 17.2 kDa whale myoglobin and

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

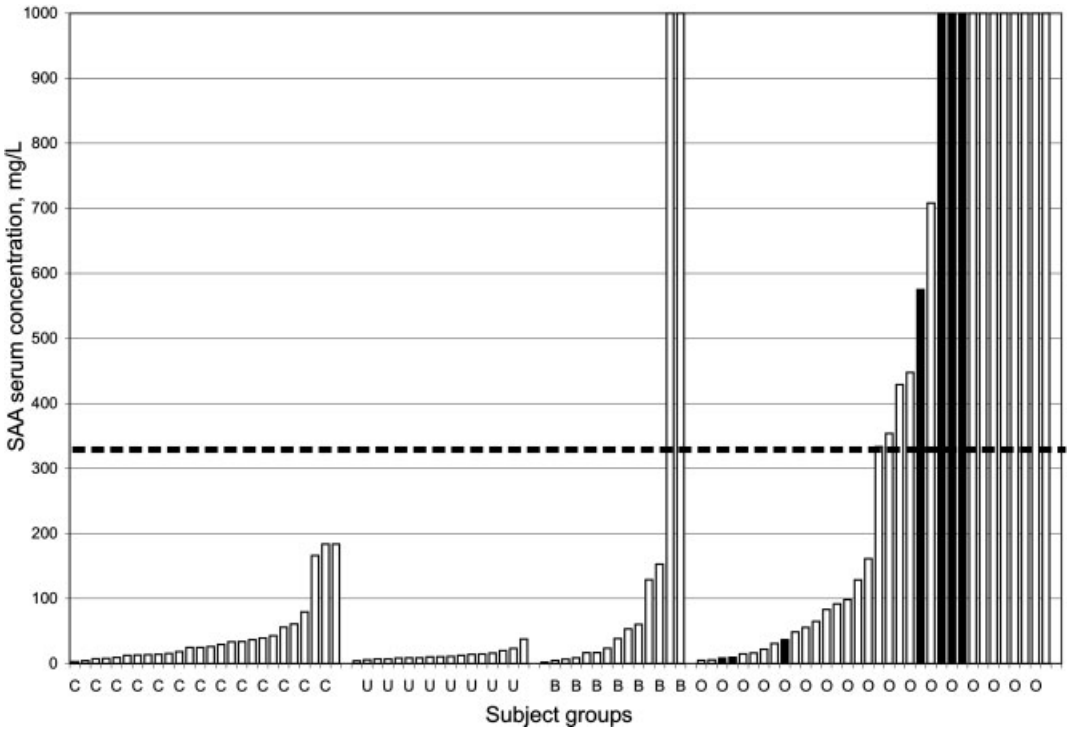

Figure 2. A-SAA levels in sera of subjects involved in the study. Letters at horizontal axis indicate subject group: O, ovarian cancer; B, benign ovarian tumor; U, uterine myoma; C, healthy control. Columns of the graph are sorted by magnitude within every subject group. Exact values higher than 1 g/L are not shown. A dotted line indicates the approximate cut-off value for optimal cancer detection in sera set of the study, which is similar to the concentration threshold for A-SAA detection in these samples by SELDI-MS. Columns which correspond to early stage cancer are filled.

its two-charged peak, such that mass tolerance for SAAs did not exceed 3–5 Da. In this context, all but two of 17 such spectra contained peak of 11 683 Da accompanied with equally or less intensive peak of 11 526 Da, $Arg^1(-)$-form (Fig. 4A). They should be predicted as 1-1/1-5 (former 1α/1β*) alleles of SAA1 gene. In four cases, besides mentioned peak pair we detected additional peaks of 11 629/11 472 Da, *i.e.*, 2–1 (former 2α) allele of SAA2 gene. One of the spectra contained the following masses: 11 699/11 543 Da and 11 647/11 492 Da (Fig. 4B). Due to about 16 Da difference in comparison with more frequent forms, these peaks may be suggested to be methionine-oxidized forms of SAA1-1/1-5 and SAA2-1, although this hypothesis requires further validation. Finally, one of the samples comprised peaks of 11 655/11 497 Da (Fig. 4C) which, obviously, were forms of SAA 1-3 (former 1γ).

In general, these data suggest that SAA1-1/1-5 are major alleles of SAA1 in studied population (16 out of 17). Prevalence of SAA1 peaks in the spectra indicate that serum level of SAA1 in ovarian cancer is higher than that of SAA2. These differences observed between alleles did not create difficulties for MS profile classification because they did not exceed the 0.3% range accepted for peak combination [31].

3.3 Stringent processing of normal phase SELDI spectra

Early works on so-called "low-resolution" diagnostic serum profiling using a SELDI linear mode detector usually used whole spectra for discrimination. As a result, some very low *m/z* values such as 168 and 321 Da were accepted for discrimination between cancer and noncancer [2, 32]. It is obvious from general theory of MALDI process that the low mass region of such spectra is saturated by artifact peaks which are hardly useful for biological consideration [33]. Nonsupervised use of classifying algorithms without physical and biochemical considerations should lead to an appearance of low intensive or low mass discriminatory *m/z* markers. In contrast, identified and validated cancer discriminatory peaks derived from direct MALDI/SELDI serum profiles are highly intensive peaks, all but one of them being higher than 9 kDa (Table 2). In this respect, we apply more stringent conditions for SELDI peak determination (see Section 2.3) to all 91 normal phase spectra obtained for further development of models classifying them as cancer and noncancer. As a result, 48 *m/z* values were determined for data mining in 5500–17 500 Da interval; each of them gives a clear, highly intensive (S/N>5) peak in at least one sample.

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

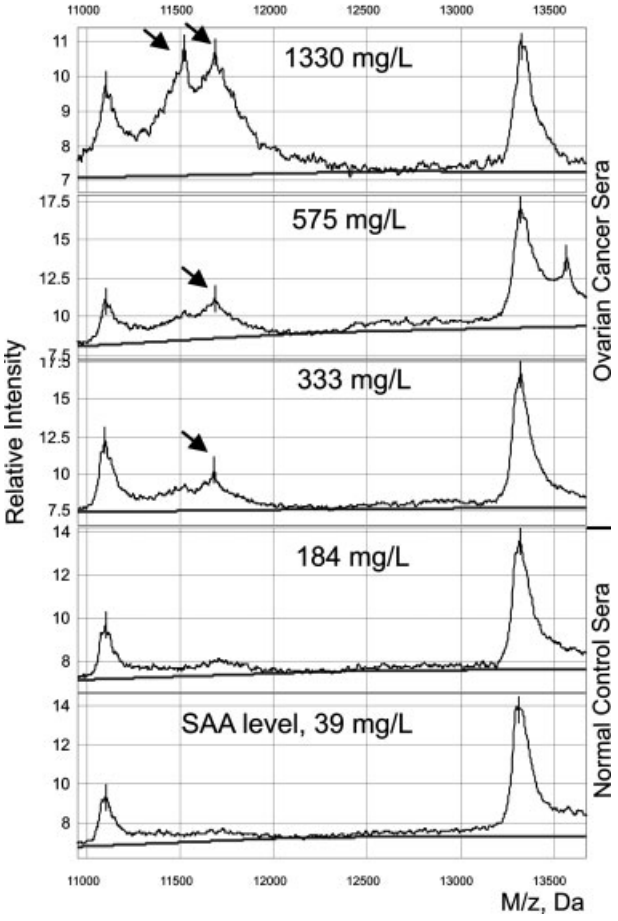

Figure 3. Sensitivity of serum SAA detection by SELDI-MS. Mass-spectra of 10 × diluted serum collected using NP-20 chip are shown as well as the corresponding SAA levels determined by ELISA. Arrows show the SAA peaks.

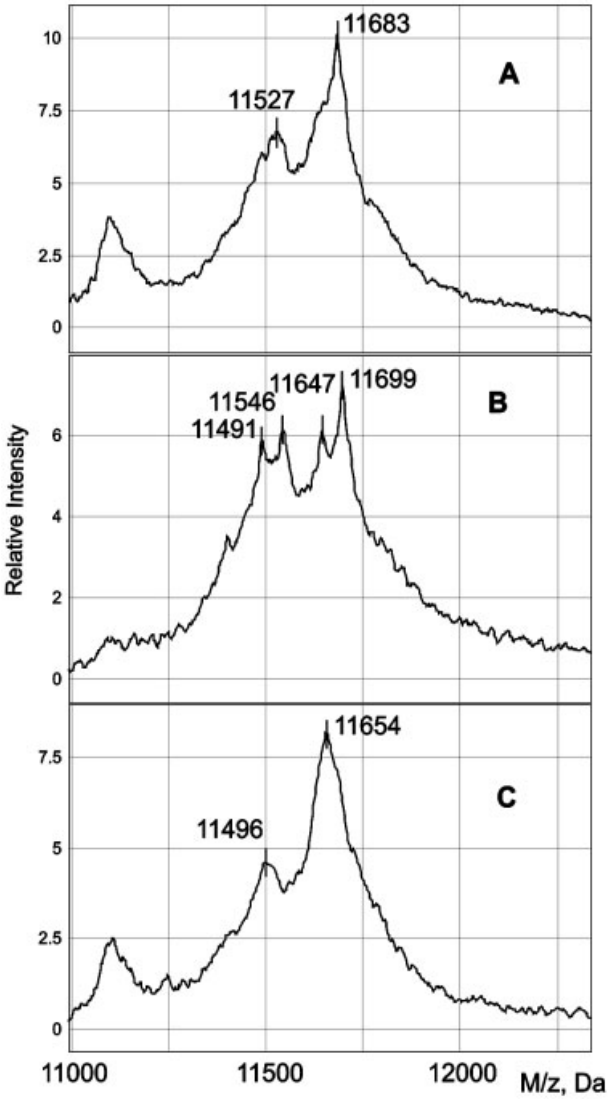

Figure 4. SAA peaks from SELDI-TOF spectra of selected serum samples. A: peak pair predicted as SAA1-1/1-5, full-length and Arg[1]-truncated form (minus approximately 156 Da); B: peak pairs predicted as SAA1-1/1-2 and SAA2-1, full-length and R1-truncated form, all methionine-oxidized (plus approximately 16 Da); C: peak pair predicted as SAA1-3, full-length and Arg[1]-truncated form.

Thus, we try to ensure that every mass to be considered as discriminatory relates to certain compound and is not a physical artifact or matrix derivative. Peak lists obtained in such manner and divided to cancer and noncancer classes (for reference see Supporting Materials; http://www.ibmc.msk.ru/departments_en/LDP/mosh_support.xls) were subjected to different statistical classifiers, as described below.

3.4 TSP classifier for SELDI spectral data

TSP classifier was recently suggested for mRNA microarray data integration and marker gene identification [27]. This algorithm classifies samples using differences between signal intensities rather that these intensities themselves, thus, yielding TSP of marker variables or several such pairs (*k*-TSP). Being guided by the proposal that spectral data are somewhat similar to microarray data, we applied *k*-TSP to our set of 91 sample mass-spectra. The algorithm returned only one TSP which is intensity of 11 681 Da peak minus intensity of 13 769 Da peak. If this difference is positive the algorithm's decision is "cancer" and if it is negative the decision is "noncancer." Although the percentage of correct cancer/noncancer discrimination by this TSP (*i.e.*, method accuracy) is as low as 79.9% even in leave-one-out cross-validation, the highly relevant biological meaning of features selected by the algorithm is remarkable. Indeed, the peak of about 11 681 Da is a major form of A-SAA [10] which was shown to be up-regulated in cancer, while the second peak of about 13 769 Da is most likely a native form of TTR which was shown to be down-regulated in cancer [34].

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 2. Overview of identified members of diagnostic SELDI-based serum profiles in ovarian cancer

IPI[a)]	Name	Discriminatory mass(es) (kDa)	Details	Level change in cancer	Reference level, normal blood[b)]	Reference to identification	Reference to similar discriminatory mass
IPI00294193	Inter-α-trypsin inhibitor heavy chain H4	3.27	Fragment	Up	N/A for the fragment	[7]	
IPI00641737	Haptoglobin	9.2	N/A (probably, α1-chain)	Up	0.88 g/L (1.7×10^{-5} M for one α1-chain *per* molecule)	[6]	
IPI00552578	SAA(1)	11.52, 11.68	Major and Arg^1-truncated forms	Up	0.01 g/L[c)] (8.6 ± 10^{-7} M)	[10]	[11]
IPI00022432	TTR	12.8 (12.9), 13.9	Major form and fragment	Down	0.26 g/L (1.9×10^{-5} M)	[7, 8]	This paper
IPI00654755	Hemoglobin beta	15.9	Major form	Up	N/A	[8]	
IPI00021841	Apolipoprotein A1	28.0	Major form	Down	1.4 g/L (5.0×10^{-5} M)	[7, 8]	
N/A	Immuno-globulin heavy chain	52	N/A	Down	9.7 g/L (1.7×10^{-4} M)	[6]	
IPI00022463	Transferrin	79	Major form	Down	2.3 g/L (2.9×10^{-5} M)	[7, 8]	

a) International Protein Index [38].
b) Protein levels are taken from recent interlaboratory study [39] except SAA.
c) [15].

Thus, the (*k*-)TSP classifier seems perspective and merits attention in terms of marker selection in proteome profiling. The pair selected from our data (SAA-TTR) generated a new binary variable, *i.e.*, positive or negative difference between intensities of said peaks. This variable was then combined with ELISA data for further sample classification.

3.5 Correlation analysis of SELDI spectral data

Important and advantageous feature of MS-based analyses is a possibility to observe multiple modified forms of the proteins [35]. In this connection, we calculated correlation between intensities of disease-relevant protein peaks attributed to A-SAA and TTR and intensities of other peaks of *m/z* region 5500–17 500 Da involved. Despite the fact, that TTR mass selected by TSP algorithm was 13 769 Da, we took the neighbor more intensive peak of 13 871 Da associated to *S*-cysteinylated TTR [34] which was a major plasma form in nonredundant conditions. Surprisingly, it was shown that as many as 19 of 47 peaks (40%) in that range significantly correlated with A-SAA peak and 13 of 47 peaks correlated with TTR peak (Fig. 5). Notably, the regions of correlation to these two proteins overlapped only in four points, thus, yielding virtually independent peak groups indicated as "SAA group" and "TTR group." These regions of correlation most likely comprise various modified or truncated forms of A-SAA and TTR, some of them being well-known and observed in MALDI-MS profiles [7, 10, 34, 35].

3.6 SVM and LR for cancer/noncancer classification of combined proteome data

Different types of data generated in our study were subjected to SVM and LR classifiers to estimate which of the data types are optimal for cancer/noncancer discrimination. Thus, all samples were classified based on (i) CA125 level as a standard marker; (ii) ELISA data, *i.e.*, CA125 level plus SAA level; (iii) MS data, *i.e.*, intensity values for 48 SELDI-MS *m/z* peaks as described above; (iv) combined ELISA and MS data; and (v) ELISA data combined with binary TSP-based variable which was a simplified and formalized representation of MS data. Characteristics of SVM and LR classifiers for these data types which may be compared using crossvalidation based confidence interval are set forth in Table 3. The first important observation from the table is that ELISA measurement of SAA level in addition to CA125 level gains no significant advantage though several cancer cases have a high SAA level with normal CA125. For MS data classification, SVM had better performance than LR and the accuracy of SVM in the case of MS data was significantly better than for CA125 alone (89.5% *vs.* 86.1%, respectively). In contrast, LR classifier worked significantly better than SVM when binary TSP variable was used in addition to ELISA data, and TSP essentially improved the classification accuracy (90.7% *vs.* 85–86% for CA125 alone). Thus, the TSP feature requires further investigation because it may be more reproducible between different MS machines and populations studied than the whole spectral data due to substantial data

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

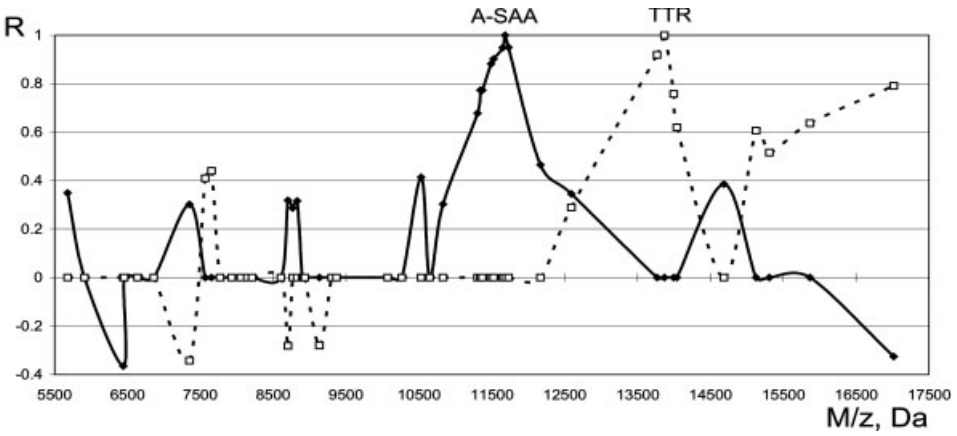

Figure 5. Analysis of correlation between intensities of identified SELDI-MS profile peaks (A-SAA and TTR major forms) and intensities of other peaks of SELDI-MS profile. *R* is a Spearman rank correlation coefficient. Correlation between A-SAA peak (11 681 Da) intensity and other discriminatory peak intensities is shown by solid line and filled diamonds, and correlation between TTR peak (13 870 Da) intensity and other discriminatory peak intensities is shown by dotted line and empty squares. Correlations that had $p>0.01$ were considered to be statistically insignificant and were forced to have Spearman correlation coefficient equal to zero.

Table 3. Performance of SVM and LR classifiers in discriminating between cancer and noncancer sera estimated using 100 runs of ten-fold crossvalidation

Method	Accuracy (%)	Specificity (%)	Sensitivity (%)
SVM (CA125)[a)]	86.17 ± 0.70[b)]	98.82 ± 0.28	64.68 ± 1.67
SVM (ELISA)[c)]	86.42 ± 0.70	96.48 ± 0.49	70.27 ± 1.75
SVM RFE (MS)[d)]	89.47 ± 0.66	93.34 ± 0.70	83.57 ± 1.59
SVM (TSP(MS)+ELISA)[e)]	86.74 ± 0.70	92.84 ± 0.80	77.46 ± 1.61
SVM RFE (ELISA+MS)[f)]	95.23 ± 0.43	98.07 ± 0.35	90.84 ± 1.08
LR (CA125)	85.05 ± 0.70	95.62 ± 0.54	67.53 ± 1.72
LR (ELISA)	86.58 ± 0.69	94.30 ± 0.58	74.19 ± 1.66
LR (MS)	86.03 ± 0.74	87.52 ± 1.28	83.71 ± 1.47
LR AIC (TSP(MS)+ELISA)	90.74 ± 0.59	96.87 ± 0.47	81.24 ± 1.41
LR AIC (ELISA+MS)	91.93 ± 0.58	92.73 ± 0.71	90.69 ± 1.12

a) CA125 level was the only variable used.
b) Confidence interval at $p<0.05$.
c) Two ELISA variables included CA125 and SAA levels.
d) Variables included 48 peak intensity values from SELDI data.
e) Two ELISA variables were combined with a single binary feature derived by TSP from spectral data.
f) The variable set included 48 SELDI variables and 2 ELISA variables.

simplification. Finally, the best classification accuracy (95.2%) between cancer and noncancer samples was obtained using SVM and ELISA + MS combined data.

3.7 Discriminatory spectral variables used by different classifiers

From the point of view of correlation analysis performed with intensity of A-SAA and TTR major peaks, it was interesting whether peak groups correlated with those peaks relate to discriminatory peaks selected by the classifiers from MS data. This relationship is illustrated in Table 4, which shows that a majority of these discriminatory peaks (16 out of 21) correlate with A-SAA or TTR peak and some of them represent multiple forms of these marker products. One apparent *m/z* peak which is independent from A-SAA and TTR is a peak with *m/z* of about 10.2 kDa. This *m/z* value is selected for discrimination between cancer and noncancer spectra by four different classification methods (Table 4). A high consistency between discriminatory peaks obtained by different methods indicates that the classifications have biological meaning and are not casual.

4 Discussion

MS-based blood protein profiling for cancer diagnostic purposes was initiated from software-based classification of spectra which was presented as a "black box". Lack of certain marker identity in the profile had originally generated an opinion that discriminatory peaks are novel, cancer cell-derived peptides [1]. At the same time, there was a lot of skepticism against the protein profile approach which resulted from a priori low sensitivity of the method and unknown nature of discriminators [3]. Early bioinformatics works related to SELDI spectra classification actually included low mass peaks in the discriminatory pattern [32], these peaks being possibly suggested as matrix-derived artifacts.

Subsequently, it was realized that MS diagnostic plasma profile is not a black box, but, more likely, an array of plasma proteins subjected to MS which are measured simultaneously in fast and rather inexact manner. Recent studies of SELDI serum and plasma profiling identified several proteins that formed cancer discriminatory masses [6–8, 10]. These SELDI-derived markers for ovarian cancer are summarized in Table 2. The first observation from this Table is

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

Table 4. Discriminatory peaks selected by SVM, LR and TSP classifiers from spectral data (MS) and combined spectral and ELISA data

Classifier				SVM RFE (MS)	SVM RFE (MS + ELISA)	LR AIC (MS)	LR AIC (MS + ELISA)	TSP (MS)
#	*m/z* (Da)	R_{11681}[a)]	R_{13870}[b)]					
1	5675	0.35	–	+	+	+		
2	6441	−0.36	–	+				
3	6454	–	–	+	+	+		
4	7566	–	0.41	+			+	
5	7651	–	0.44			+		
6	7769	–	–		+			
7	8208	–	–		+			
8	8766	0.29	–			+		
9	8829	0.32	–			+		
10	10 265	–	–	+	+	+	+	
11	11 304	0.68	–	+				
12	11 370	0.77	–			+	+	
13	11 649	0.95	–	+				
14	11 681	1.00	–	+		+		+
15	11 728	0.95	–	+	+			
16	12 168	0.47	–	+	+	+	+	
17	13 769	–	0.92		+	+		+
18	13 870	–	1.00	+	+	+	+	
19	14 685	0.39	–	+	+			
20	15 310	–	0.52	+	+			
21	17 017	−0.33	0.79	+	+			
	CA125			N/A	+	N/A	+	N/A

a) R_{11681} is a Spearman correlation coefficient between intensity of the 11 681 (±0.2%) Da peak (A-SAA) and other peak intensities of the profile over 91 samples involved. Only significant *R* values are shown ($p<0.01$). Discriminatory peaks of "A-SAA group" are filled with a grid.

b) R_{13870} is a Spearman correlation coefficient between intensity of the 11 870 (±0.2%) Da peak (TTR) and other peak intensities of the profile over 91 samples involved. Only significant *R* values are shown ($p<0.01$). Discriminatory peaks of "TTR group" are filled with a gray color.

that although many low *m/z* (peptide) discriminatory masses were reported, none of them except inter-α-trypsin inhibitor-derived peptide as identified. Moreover, in two ovarian cancer works exploring serum peptidome, no matching discriminatory peptide masses were shown [2, 32]. This intangibility of discriminatory peptides supports the proposal that some of them are of artifact nature.

In our study, which uses a small protein range for MS-based classification, most of the discriminatory peaks are shown to belong to A-SAA and TTR correlation groups probably representing multiple forms of these disease-related serum components. Discriminatory profiles contain only one obvious unknown protein with *m/z* of about 10.2 kDa which is down-regulated in cancer. ELISA measurement of SAA and TTR without distinguishing their forms may not improve cancer diagnostics in comparison with CA125 marker measurement, while its combination with MS-based estimate of these protein forms yields higher diagnostic accuracy at least in this study.

It should be noted that, among protein SELDI markers of Table 2, A-SAA is quite distinct. First, its normal plasma level is 10- to 100-fold lower than levels of other marker candidates. Second, pathological increase in its level is more expressed and it may be said to change in qualitative manner (100-fold and more) whereas other identified marker levels have a significant, but relatively weak change [7, 8]. Interestingly, apolipoprotein A1 down-regulation is directly connected with A-SAA up-regulation as the latter replaces ApoA1 from blood high density lipoproteins in inflammation [36]. Furthermore, of the SELDI-derived markers, the SAA attracts special interest because, according to recent studies making link between inflammation and cancer [37], it may be not just a host-response protein, but also a product directly involved in cancer progression. In particular, it has been shown that SAA was highly expressed in malignant colon epithelium [14] leading to a suggestion that blood SAA levels in cancer may be derived not only from liver, but also from tumor tissue.

5 Concluding remarks

The black box concept of cancer *versus* normal serum/plasma classification based on direct MS, *e.g.*, SELDI-TOF, protein profiling is extensively expanded by identification of these

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim

profile components. Cancer and noncancer SELDI profiles of small proteins obtained in this study were classified based on peaks related to A-SAA and TTR forms. The best classification scheme achieved herein included the use of SVM algorithm on combined data of SELDI spectra and CA125 level. Among the eight known ovarian cancer SELDI profile components, A-SAA is the most relevant to cancer molecular pathogenesis and, at the same time, it has the highest degree of up-regulation in disease conditions.

This work was supported by the "Proteomics for Medicine" program of Russian Academy of Medical Science and by the Federal Agency of Science and Innovation (FASI), Russia.

6 References

[1] Petricoin, E. F., Liotta, L. A., *Curr. Opin. Biotechnol.* 2004, *15*, 24–30.

[2] Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J. *et al.*, *Lancet* 2002, *359*, 572–577.

[3] Diamandis, E. P., *Clin. Chem.* 2003, *49*, 1272–1275.

[4] Jacobs, I. J., Menon, U., *Mol. Cell Proteomics* 2004, *3*, 355–366.

[5] Yousef, G. M., Diamandis, E. P., *Biol. Chem.* 2002, *383*, 1045–1057.

[6] Rai, A. J., Zhang, Z., Rosenzweig, J., Shih, I. *et al.*, *Arch. Pathol. Lab Med.* 2002, *126*, 1518–1526.

[7] Zhang, Z., Bast, R. C., Jr., Yu, Y., Li, J. *et al.*, *Cancer Res.* 2004, *64*, 5882–5890.

[8] Kozak, K. R., Su, F., Whitelegge, J. P., Faull, K. *et al.*, *Proteomics* 2005, *5*, 4589–4596.

[9] Kozak, K. R., Amneus, M. W., Pusey, S. M., Su, F. *et al.*, *Proc. Natl. Acad. Sci. USA* 2003, *100*, 12343–12348.

[10] Moshkovskii, S. A., Serebryakova, M. V., Kuteykin-Teplyakov, K. B., Tikhonova, O. V. *et al.*, *Proteomics* 2005, *5*, 3790–3797.

[11] Vlahou, A., Schorge, J. O., Gregory, B. W., Coleman, R. L., *J. Biomed. Biotechnol.* 2003, *2003*, 308–314.

[12] Jensen, L. E., Hiney, M. P., Shields, D. C., Uhlar, C. M. *et al.*, *J. Immunol.* 1997, *158*, 384–392.

[13] Santiago, P., Roig-Lopez, J. L., Santiago, C., Garcia-Arraras, J. E., *J. Exp. Zool.* 2000, *288*, 335–344.

[14] Gutfeld, O., Prus, D., Ackerman, Z., Dishon, S. *et al.*, *J. Histochem. Cytochem.* 2006, *54*, 63–73.

[15] Uhlar, C. M., Whitehead, A. S., *Eur. J. Biochem.* 1999, *265*, 501–523.

[16] O'Hara, R., Murphy, E. P., Whitehead, A. S., FitzGerald, O., Bresnihan, B., *Arthritis Res.* 2000, *2*, 142–144.

[17] O'Hara, R., Murphy, E. P., Whitehead, A. S., FitzGerald, O., Bresnihan, B., *Arthritis Rheum.* 2004, *50*, 1788–1799.

[18] Miyoshi, A., Kitajima, Y., Kido, S., Shimonishi, T. *et al.*, *Br. J. Cancer* 2005, *92*, 252–258.

[19] Howard, B. A., Wang, M. Z., Campa, M. J., Corro, C. *et al.*, *Proteomics* 2003, *3*, 1720–1724.

[20] Tolson, J., Bogumil, R., Brunst, E., Beck, H. *et al.*, *Lab. Invest.* 2004, *84*, 845–856.

[21] Vapnik, V. N., *Statistical Learning Theory*, Springer-Verlag, New York 1998.

[22] Skates, S. J., Horick, N., Yu, Y., Xu, F. J. *et al.*, *J. Clin. Oncol.* 2004, *22*, 4059–4066.

[23] Guyon, I., Weston, J., Barnhill, S., Vapnik, V., *Machine Learning* 2002, *46*, 389–422.

[24] Venables, W. N., Ripley, B. D., *Modern Applied Statistics with S*, 4th Edn., Springer-Verlag, New York 2002.

[25] Wu, B., Abbott, T., Fishman, D., McMurray, W. *et al.*, *Bioinformatics* 2003, *19*, 1636–1643.

[26] Tan, A. C., Naiman, D. Q., Xu, L., Winslow, R. L., Geman, D., *Bioinformatics* 2005, *21*, 3896–3904.

[27] Xu, L., Tan, A. C., Naiman, D. Q., Geman, D., Winslow, R. L., *Bioinformatics* 2005, *21*, 3905–3911.

[28] Kiernan, U. A., Tubbs, K. A., Nedelkov, D., Niederkofler, E. E., Nelson, R. W., *FEBS Lett.* 2003, *537*, 166–170.

[29] Casl, M. T., Coen, D., Simic, D., *Eur. J. Clin. Chem. Clin. Biochem.* 1996, *34*, 31–35.

[30] Cicarelli, L. M., Perroni, A. G., Zugaib, M., de Albuquerque, P. B. *et al.*, *Mediators Inflamm.* 2005, *2*, 96–100. DOI:

[31] Ressom, H. W., Varghese, R. S., Abdel-Hamid, M., Eissa, S. A. *et al.*, *Bioinformatics* 2005, *21*, 4039–4045.

[32] Zhu, W., Wang, X., Ma, Y., Rao, M. *et al.*, *Proc. Natl. Acad. Sci. USA* 2003, *100*, 14666–14671.

[33] Sorace, J. M., Zhan, M., *BMC Bioinformatics* 2003, *4*, 24.

[34] Gericke, B., Raila, J., Sehouli, J., Haebel, S. *et al.*, *BMC Cancer* 2005, *5*, 133.

[35] Fung, E. T., Yip, T. T., Lomas, L., Wang, Z. *et al.*, *Int. J. Cancer* 2005, *115*, 783–789.

[36] Rohrer, L., Hersberger, M., von Eckardstein, A., *Curr. Opin. Lipidol.* 2004, *15*, 269–278.

[37] Pikarsky, E., Porat, R. M., Stein, I., Abramovitch, R. *et al.*, *Nature* 2004, *431*, 461–466.

[38] Kersey, P. J., Duarte, J., Williams, A., Karavidopoulou, Y. *et al.*, *Proteomics* 2004, *4*, 1985–1988.

[39] Haab, B. B., Geierstanger, B. H., Michailidis, G., Vitzthum, F. *et al.*, *Proteomics* 2005, *5*, 3278–3291.

© 2007 WILEY-VCH Verlag GmbH & Co. KGaA, Weinheim