Disease Markers 19 (2003,2004) 33–39
IOS Press

Development of proteomic patterns for detecting lung cancer[1]

Xueyuan Xiao, Danhui Liu, Ying Tang, Fuzheng Guo, Liang Xia, Jin Liu and Dacheng He *
Key laboratory for Cell Proliferation and Regulation Biology Ministry of Education, Beijing Normal University, Beijing 100875, China

Abstract. Lung cancer is at present the number one cause of cancer death and no biomarker is available to detect early lung cancer in serum samples so far. The objective of this study is to find specific biomarkers for detection of lung cancer using Surface Enhanced Laser Desorption/Ionization (SELDI) technology. In this study, serum samples from 30 lung cancer patients and 51 age-and sex-matched healthy were analyzed by SELDI based ProteinChip reader, PBSII-C. The spectra were generated on WCX2 chips and protein peaks clustering and classification analyses were performed utilizing Biomarker Wizard and Biomarker Patterns software packages, respectively. Three protein peaks were automatically chosen for the system training and the development of a decision classification tree. The constructed model was then used to test an independent set of masked serum samples from 15 lung cancer patients and 31 healthy individuals. The analysis yielded a sensitivity of 93.3%, and a specificity of 96.7%. These results suggest that the serum is a capable resource for detection of specific lung cancer biomarkers. SELDI technique combined with an artificial intelligence classification algorithm can both facilitate the discovery of better biomarkers for lung cancer and provide a useful tool for molecular diagnosis in future.

1. Introduction

Lung cancer is one of the most prevalent cancers and is also the leading cause of cancer death in the world. The 88% mortality rate for non-small cell lung carcinomas (NSCLC) has remained remarkably unchanged since 1985 despite improvement of detection and treatment of lung cancer [1]. The key problem is NSCLC generally resistant to chemotherapy or radiotherapy. So the operation has been mainly selected to treat NSCLC patients in clinic yet. Although putative tumor suppressor genes involved in NSCLC have been characterized [1–4], such as DAL-1, Lc19, p21, p63, and some proteins were observed over expressed in lung cancer, such as, metallothionein , hCG, SP1 and CEA [5–7], these candidate genes and proteins are all lack of specificity and no biomarkers have been able to detect early lung cancer based on a serum sample so far. It is clear that discovery of lung cancer-associated soluble biomarkers in serum for early detection becomes an urgent task in clinic.

In theory, study of the alternations of dynamic Proteomics in cells will let us know the tiny changes for cancers at the early stage, and specific biomarkers for particular cancers should be identified by comparing the protein profiles between patient samples and normal control. To achieve this goal, the study of serum proteome provides a turning point for biomarker discovery. Whereas two-dimensional electrophoresis (2-DE) and mass spectrometry (MS) had been the indispensable tool for proteins study, some new technologies for detection of biomarker were developed as well more recently [8,9]. Among them is ProteinChip technology coupled with SELDI-TOF-MS (surfaced enhanced laser desorption/ionization time of flight mass spectrometry) technology [10]. This innovative technique is

[1]This work was supported by the Science Technology Key Project of Ministry of Education (Grant No:272006) and the Major State Basic Research Project (Grant No.G1999053901) and the National High Technology Research and Development Program (Grant No. 2002AA232031).

*Corresponding author: Dacheng He, Key laboratory for Cell Proliferation and Regulation Biology Ministry of Education, Beijing Normal University, Universities' of Confederated Institute for Proteomics, Beijing 100875, China. Tel.: +86 10 62208439; Fax: +86 10 62209729; E-mail: dhe@bnu.edu.cn.

ISSN 0278-0240/03,04/$17.00

much fast, has a high-throughput capability, needs very little amounts of sample, and can analyze the complex biological mixtures directly. The efficacy of the SELDI technology for discovery of prostate cancer and ovarian cancer protein biomarker in serum [11,12], as well as breast cancer protein biomarker in nipple aspirate fluids and transitional cell carcinoma of the bladder in urine have recently been demonstrated by Paweleta and Vlahou [13,14]. In this paper, a set of serum samples was analyzed by SELDI technology and a discriminatory proteomic pattern for lung cancer was detected by artificial intelligence data analysis algorithm.

2. Materials and methods

2.1. Serum samples

Blood serum samples from patients diagnosed with lung cancer were procured from the Department of Respiratory, Capital University Affiliated Beijing Tiantan Hospital. After obtaining informed consent from the patient, the sample was collected into a 10cc serum separator vacutainer tube and laid up at 4 °C for 1 hour, and then centrifuged 20 min at 4000 rpm. The serum was distributed into 100 μl aliquots and stored frozen at -80°C. Qualified age-and gender-matched healthy sera were obtained from the State Sport and Physical Culture Administrator, and the processing, collection and storage protocols for these samples were exactly the same as above mentioned of the patient's.

2.2. Patients and healthy control

Specimens from three groups of patients were used in this study: (a) 23 patients were adenocarcinomas; (b) 21 patients were squamous cell carcinomas; (c) 11 patients were small-cell lung cancer. The healthy group coming from general survey of health were used as normal control and consisted of 47 males and 34 females ranging in age 40–70 years.

2.3. SELDI protein profiling

Two type chips, H4 and WCX2 were initially evaluated to determine which affinity chemistry provides the best serum profiles in terms of number and resolution of proteins. The WCX2 chip was observed to give the best results. WCX2 chips were pretreated with 5 μl of 10 mM HCl on each array and stayed at room temperature for 10 min. The chips were rinsed 3 times with 10 ml M-Q-water in a conical tube and then put into a bioprocessor (Ciphergen Biosystems, Inc.), which is a device that allows application of larger volumes of serum to each chip array. The bioprocessor was washed and shaken on a platform shaker at a speed of 250 rpm for 5 min with 200 μl of Binding buffer (100 mM NaAc, pH 4) in each well. This was repeated once more, and each time the binding buffer was discarded by inverting the bioprocessor on a paper towel. Each serum sample (lung cancer patient and healthy control) for SELDI analysis was separately prepared by vortexing 8 μl of serum with 12 μl of 8 M urea with 1% CHAPS (3-[(3-cholamidopropyl)-1-propanesulfonic acid) in PBS in a 1.5 ml microfuge tube at 4 °C for 10 min. Pippet 20 μl of the serum/urea samples to a, washing the sample tube with 20 μl 1M urea with 0.125% CHAPS. Then pippet the solution to another 1.5 ml fresh microfuge tube, and then pooled two serum samples. Add 40 ul samples to a fresh microfuge that contained 30 μl Cibacron Blue 3GA immobilized on 4% beaded agarose, vertex at 4 °C for 15 minutes, centrifuge the tube at 1000 g for 30 seconds, pippet supernatant into a fresh tube. Add 20 μl of the 1 M urea with 0.125% CHAPS to that microfuge tube contained Cibacron Blue beaded agarose, vertex at 4 °C for 15 minutes, centrifuge the tube at 1000 g for 30 seconds, pippet supernatant into a fresh tube. Pool the 2 supernatant. The volume was approximately 55 μl. Pippet the diluted serum samples into a 1.5 ml microfuge tube and used binding buffer (100 mM NaAc, pH 4) to make 1:3 dilution, and placed on ice until applied to a protein chip array. 100 μl of the diluted sample were applied to each well, and the bioprocessor was sealed and shaken on a platform shaker at a speed of 250 rpm for 35 min. The samples were discarded, and each well was washed with the washing buffer (100 mM NaAc, pH 4) three times. The chips were removed from the bioprocessor, washed with 1 mM Hepes quickly, air dried. 0.5 μl of a saturated solution of the EAM sinapinic acid in 50% (v/v) acetonitrile, 0.5% trifluoroacetic acid was applied onto each chip array twice, letting the array surface air dry between each sinapinic acid application. Chips were stored in the dark at room temperature until SELDI analysis. Chips was placed in the Protein Biological System IIC mass spectrometer reader (Ciphergegn Biosystems, Inc.) and time-of-flight spectra were generated by averaging 128 laser shots collected in the positive mode at laser intensity 195, detector sensitivity 9 and the optimization range was from 3000 to 50000 Da, high mass was 200000Da. Mass accuracy was calibrated externally using the All-in-one peptide molecular mass standard (Ciphergen Biosystems, Inc.).

2.4. Peak detection

Peak detection was performed using Ciphergen SELDI software version 3.0.2. The optimizing mass range of 3000–50,000Da was selected for analysis because this range contained the majority of the resolved protein/peptides. Peak detections involved baseline subtraction, mass accuracy calibration and automatic peak detection. The settings used for this study were as follows: for peak detection, the signal/noise was 3, minimum peak threshold was 30%; for cluster completion, the cluster mass was 2%, the signal/noise for the second pass was 2%.

2.5. Decision tree classification

Construction of the decision tree classification algorithm was performed by Ciphergen Biomarker Pattern software version 4.0. using a training data set consisting of 90 samples (40 lung cancer patients serum and 50 normal samples). The setting used for this study were as follows: the levels of target variable details was 2, minimum value was 0, and the tree type was classification, root nodes was 2 cases, Classification tree was selected Gini, using one rule at a time in the form of a peak intensity. The splitting is defined by the intensity the analyzed peaks. For example, whether mass A has an intensity less than or equal to the "X" splits the data set into two nodes, a left node for "yes" and a right node for "no". This splitting process will stop if terminal nodes for further splitting have no gain. The validity of this decision tree was then challenged with a blinded test data set consisting of 15 lung cancer patients and 30 normal controls which were independent of the training samples.

2.6. Statistical analysis

Comparing with normal control group, if the biomarkers were none or low expressed in patient group, the sensitivity was defined as the ratio of the lung cancer patients that did not have or low expressed the biomarkers to the total number of pathologically confirmed lung cancer patients included in the study. Specificity was defined as the ratio of the individuals that high expressed protein peaks and did not have lung cancer, to the total number of individuals without lung cancer.

3. Results

3.1. Reproducibility

The reproducibility of mass location and mass intensity between chips was determined using the pooled normal serum quality control sample. 22 peaks were selected manually throughout the range 3000Da–30000Da. For better comparability, the 22 peaks were chosen to have good resolution and better ratio of signal/noise. The average coefficient of variance (CV) based on five normal pooling human sera for intensities of 22 peaks were lower than 20%. There was little variation with day-to-day sampling and instrumentation or chip variations (Fig. 1).

3.2. Detection of lung cancer

Using SELDI software program approximately125 peaks/spectrum was detected in 1kDa-30kDa mass range. No single peak was identified that alone could completely separate two groups. These 125 peaks identified in the training set were then used to construct the decision tree classification algorithm. The classification algorithm selected three masses between 1400Da and 8500Da (8122, 1452 and 1610Da) to generate 4 terminal nodes (Fig. 2). Other four masses 2953, 5335, 6105 and 5902Da split the data set in the same way as 8122Da but with variable importance score (Fig. 3). Four masses among them between 3–10 kDa were showed in Fig. 4. Once the algorithm identifies the most discriminatory peaks, the classification rule is quite simple. If an unknown sample has a peak at mass 8122Da, the peak intensity is lower than 0.514, then the sample is placed to the left branch node, otherwise to the right branch node. The cases in each branch node were then confirmed or reclassified at the second layer following the same process while 1452Da and 1610Da is chosen to split the subset. Finally, all the 90 cases including lung cancer patients and healthy individuals could be classified in the four terminal nodes, but the percent of correct was not completely.

A summary of the classification results from the 4 terminal nodes was presented in Table 1A. The sensitivity and specificity of this classification system in blind test set were presented in Table 1B.

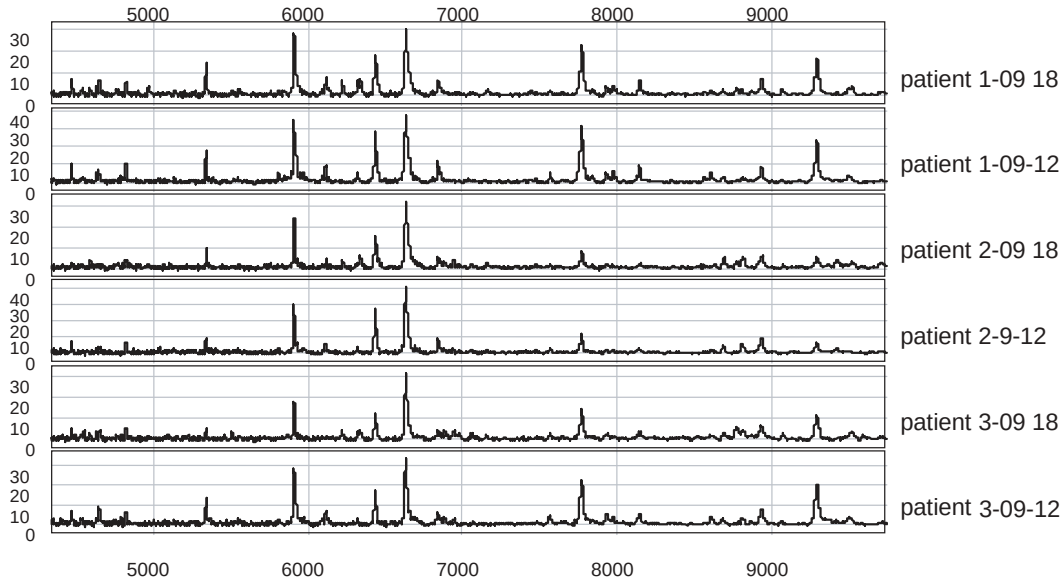

Fig. 1. A part of SELDI spectra of three patients (patient 1, patient 2 and patient 3) on WCX2 proteinchip and generated on different days (September 12th,18th,2002). "x" axis was molecular weight of peaks, and "y" axis was intensity of peaks.

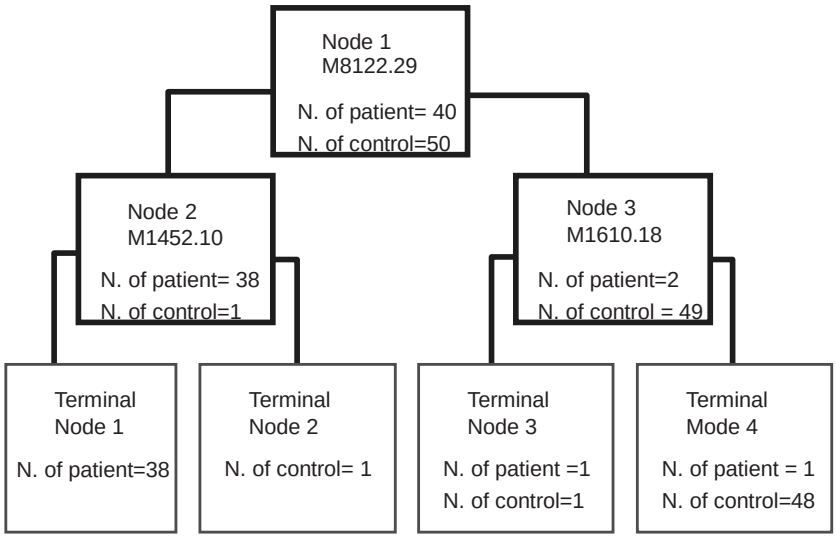

Fig. 2. Classification of lung cancer vs normal samples by the decision tree alogorithm. The left branch node after the first layer is the cases of peak intensity under or equal to 0.514, the right one is higher than 0.514. The cutoff point for 1452 mass was "equal or lower than 4.389" and M1610 was "equal or lower than 0.115". N represents the number of samples. M represents the molecular weight.

4. Discussion

The current common approaches for diagnosis of lung cancer in clinic are largely based on X-ray and computed tomography, but these kinds of methods have been insufficient in the detection of very small lesion. In many cases, the definite diagnosis for lung cancer was mainstay of pathological diagnosis on biopsy, which is however invasive and not suitable for all kinds of lung cancer. Efforts to shape the future in prevention, diagnosis, and treatment monitoring of lung cancer may largely depend on the discovery of specific biomarkers capable of distinguishing and characterizing lung cancer, subtypes, and different stages.

Complex serum proteomic patterns could reflect the underlying pathological state of an organ such as lung.

Molecular Weight	Score	
8122.29	**100.00**	
5902.29	**83.73**	
2953.44	**79.96**	
6105.99	**76.34**	
5335.73	**74.99**	

Fig. 3. Variable Importance of other masses in the first node.

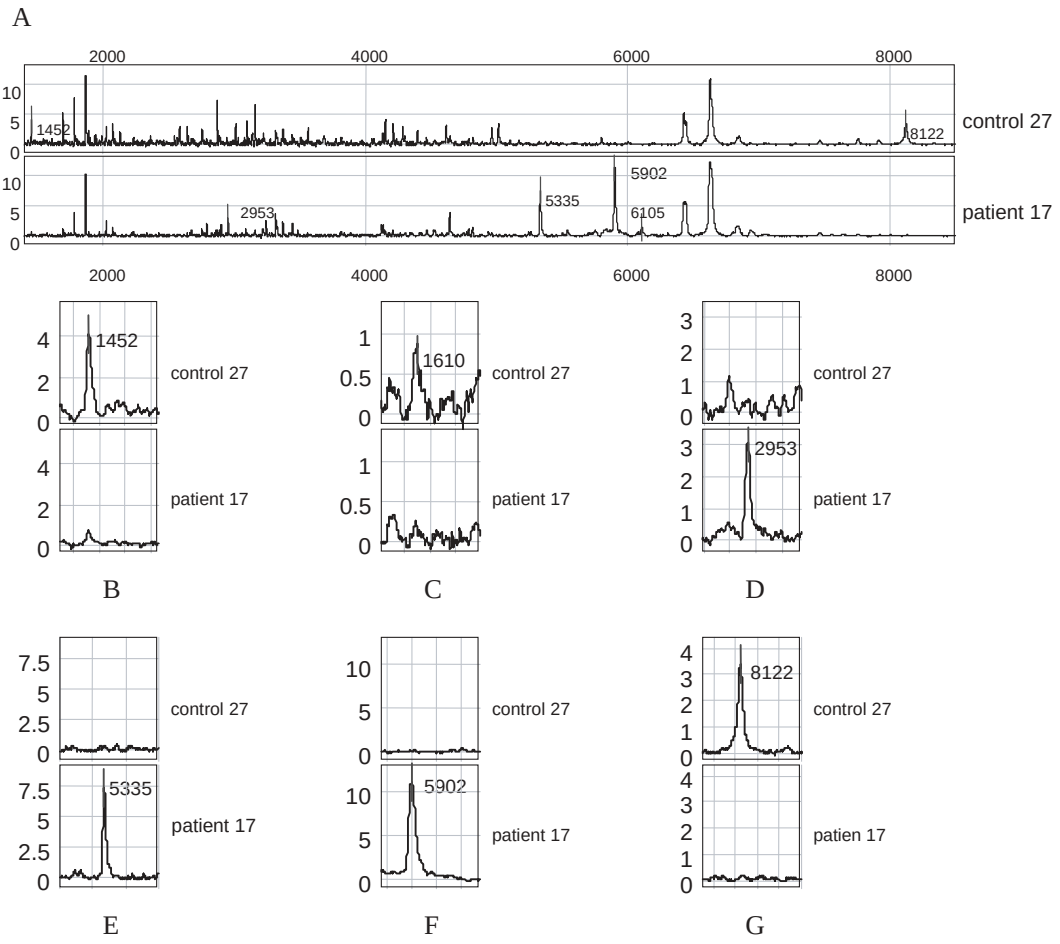

Fig. 4. "A" was partial spectrum of one patient and one control sample shown on WCX-2 chip. X-axis was molecular weight of peak, y-axis was intensity of peak. "B-G" were locally enlargement of spectrum "A".

This hypothesis is supported by many reports recently. Petricoin, et al. [12] reported that hydrophobic proteinchips were used to screen serum samples from ovarian cancer patients and a discriminator pattern was generated consisting of five protein masses of 534, 989, 2111, 2251, and 2456Da. The PPV for this biomarker pattern achieved 94% in differentiating ovarian cancer from benign ovarian disease and healthy unaffected women. A similar classification pattern consisting of nine protein masses for prostate cancer was discovered by Wright, et al. [11] using IMAC-Cu proteinchips, and reached a PPV of 96%. In our previous study, we

Table 1
Decision tree classification of cell lung cancer test set and blind test

		A. Training set		
Group	Total cases	Percent correct	Correct cases	Error cases
Normal	50	98.0	49	1
Cancer	40	97.5	39	1
		B. Blind test set		
Group	Total cases	Percent correct	Sensitivity	Specificity
Normal	31	96.7 (30/31)	–	96.7%
Cancer	15	93.3 (14/15)	93.3%	–

tried several methods to treat sera and obtained similar sets of biomarkers for lung cancer on different types of protein chips when using SELDI technology. We found that sera treated with Cibacron Blue 3GA immobilized on 4% beaded agarose could give much better results than with size column or anion exchange column, etc. Our results in this report were derived from 136 samples, including 55 lung cancer and 81 normal samples, using SELDI technique with WCX-2 proteinchips and Cibacron Blue 3GA immobilized on 4% beaded agarose. The classification tree was established to discriminate the lung cancer cases from the healthy when using three masses 8122Da, 1452Da and 1610Da for biomarker pattern. When this model was tested with a blind set, independent from the training set, it yielded a sensitivity of 93.3% and the specificity was 96.7%. Although this technology is rather young, the achievements mentioned above, as well as the data we presented in this paper, have provided higher sensitivity and specificity than any single biomarker currently used in clinic. The establishment and further complement of these patterns are speedily forming a solid foundation for the prospective "molecular diagnosis" technology that has a potential to play a major role in clinic.

It is noticed that most of biomarkers for lung cancer we got are limited by the technique optimum range to low-molecular-weight proteins or peptides. They were not being identified by this technology, moreover, their origination could be from the host organ, the cancer, or even metabolic products. They could still be valuable diagnostic tools as long as they are "specific" for the corresponding disease. We should not expect that all diseases show their specific markers in serum, but an increased number of diseases will do particularly if we can improve the ability to detect non-abundant proteins and generate more powerful software to analyze the obtained data. The specificity of our protein profiling classification algorithm is also being challenged with pulmonary diseases, such as pneumonia, to assure the specificity of this biomarker pattern for lung cancer. The current study is compromised by relatively small sample sizes. Further studies with extended scale and with the attempt to distinguish of four major subtypes of lung cancer, namely squamous cell lung carcinomas, adenocarcinomas, large cell carcinomas and small cell lung carcinomas are under investigation in our laboratory. This study is expected to further confirm and/or improve the sensitivity and specificity, and facilitate the diagnosis for different subtype of lung cancer as well.

Acknowledgments

We thank Dr.Ligong Du and Xiuping Wei for healthy and lung cancer serum samples collection, and also thank Dr. Stanley Zhang for technical assist.

References

[1] Y.K. Tran, O. Bögler, K.M. Gorse, I. Wieland, M.R. Green and I.F. Newsham, A novel member of the NF2/ERM/4.1 superfamily with growth suppressing properties in lung cancer, *Cancer Research* **59** (1999), 35–43.

[2] K.K. Wang, N. Liu, N. Radulovich, D.A. Wigle, M.R. Johnston, F.A. Shepherd, M.D. Minden and M.S. Tsao, Novel candidate tumor marker genes for lung adenocarcinoma, *Oncogene* **21** (2002), 7598–7604.

[3] T. Shoji, F. Tanaka, T. Takata, K. Yanagihara, Y. Otake, N. Hanaoka, R. Miyahara, T. Nakagawa, Y. Kawano, S. Ishikawa, H. Katakura and H. Wada, Clinical significance of p21 expression in non-small-cell lung cancer, *J Clin Oncol* **20** (2002), 3865–3871.

[4] G. Pelosi, F. Pasini, S. C. Olsen, U. Pastorino, P. Maisonneuve, A. Sonzogni, F. Maffini, G. Pruneri, F. Fraggetta, A. Cavallon, E. Roz, A. Iannucci, E. Bresaola and G. Viale, p63 immunoreactivity in lung cancer: yet another player in the development of squamous cell carcinomas? *J Pathol* **198** (2002), 100–109.

[5] G. Fontanini, S. Vignati, D. Bigini, G.R. Merlo, A. Ribecchini, C.A. Angeletti, F. Basolo, R. Pingitore and G. Bevilacqua, Human non-small cell lung cancer: p53 protein accumulation is an early event and persists during metastatic progression, *J Pathol* **174** (1994), 23–31.

[6] J. Slodkowska, M. Szturmowicz, D. Giedronowicz, P. Rudainski, A. Sakowicz and W. Kupis, Trophoblastic markers and CEA in non-small cell lung cancer: the comparison studies of tumour cells expression and serum concentration, *Rocz Akad Med Bialymst* **42**(suppl 1) (1997), 190–198.

[7] S. Theocharis, C. Karkantaris, T. Philipides, E. Agapitos, A. Gika, A. Marglil, C. Kittas and A. Koutselinis, Expression of metallothionein in lung carcinoma: correlation with histological type and grade, *Histopathology* **40** (2002), 143–151.
[8] M.B. Jones, H. Krutzsch, H. Shu, Y. Zhao, L.A. Liotta, E. C. Kohn and E.F. 3rd. Petricoin, Proteomic analysis and identification of new biomarkers and therapeutic targets for invasive ovarian cancer, *Proteomics* **2** (2002), 76–84.
[9] T.C.W. Poon and P.J. Johnson, Proteome analysis and its impact on the discovery of serological tumor markers, *Clin Chim Acta* **313** (2001), 231–239.
[10] T.W. Hutchens and T.T. Yip, New desorption strategies for the mass spectrometric analysis of macromolecules, *Rapid Commun Mass Spectrom* **7** (1993), 576–580.
[11] B.L. Adam, Y. Qu, J.W. Davis, M.D. Ward, M.A. Clements, L.H. Cazares, O.J. Semmes, P.F. Schellhammer, Y. Yasui, Z. Feng and G.L. Wright, Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men, *Cancer Research* **62** (2002), 3609–3614.
[12] E.F. Petricoin III, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn and L.A. Liotta, Use of proteomic patterns in serum to identify ovarian cancer, *The Lancet* **359** (2002), 572–577.
[13] C.P. Paweleta, B. Trock, M. Pennanen, T. Tsangaris, C. Magnant, L.A. Liotta and E.F. Petricoin III, Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer, *Dis Marker* **17** (2001), 301–307.
[14] A. Vlahou, P.F. Schellhammer, S. Mendrinos, K. Patel, F. I. Kondylis, L. Gong, S. Nasim and G.L. Wright, Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine, *American Journal of Pathology* **158** (2001), 1491–502.