Biomarker Discovery for Cervical Cancer

Natalia I. Govorukhina

Cover design: Janna Bystrykh

The work described in this thesis was performed in the Department of Analytical Biochemistry Centre of Pharmacy at the University of Groningen, The Netherlands in collaboration with Department of Gynecologic Oncology UMCG, Groningen.

Printed by: Proefschriftmaken.nl

RIJKSUNIVERSITEIT GRONINGEN

Biomarker Discovery for Cervical Cancer

Proefschrift

ter verkrijging van het doctoraat in de
Wiskunde en Natuurwetenschappen
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
vrijdag 21 december 2007
om 14:45 uur

door

Natalia I. Govorukhina

geboren op 8 januari 1956
te Kopejsk, USSR

Promotores: Prof. dr. R. Bischoff
Prof. dr. A. G. J. van der Zee

Beoordelingscommissie: Prof. dr. Ch. Huber
Prof. dr. S. Verpoorte
Prof. dr. R. Vonk

ISBN: 9789036732222

Contents

Chapter I.

Introduction.

Chapter I.I
Introduction to Cervical Cancer

1. Incidence of Cervical cancer

According to the International Agency for Research on Cancer, gynecological cancers accounted for 19% of the 5.1 million estimated new cancer cases, the 2.9 million cancer deaths and the 13 million 5-year prevalent cancer cases among women in the world in 2002 [1]. Global estimations for cervical cancer arrive at 468,000 new cases and 233,000 deaths in 2000 year [2]. The frequency of occurrence of cervical cancer differs significantly between countries, from 0.4% in Israel to 5.3% in Colombia [3] of all cancers. More than 80% of all cases occur in developing countries [1], which emphasizes the need for considerable prophylactic efforts, especially in third-world countries. However, the risk of death in developed countries is also high. In the USA, cervical cancer accounts for 1.3% of all deaths due to cancer among women [4]. In the Netherlands 583 new cervical cancer cases were diagnosed in 2003, compared to 644 in 2002 (http://www.ikcnet.nl/). The incidence of cervical cancer in the Netherlands has been decreasing at a rate of approximately 2% annually since 1990. This is mostly due to the nation-wide program for early diagnosis of cancer.

2. Diagnosis

Cervical cancer originates from dysplastic lesions of various degrees. Diagnosis is largely based on biopsy of the epithelium of the cervix. For screening purposes Papanicolau and Trout (1941) introduced the smear test for cytomorphological analysis of epithelial cells [5], scraped from the cervix, which is still in use till now (known in Netherland as KOPAC-B). In many developed countries, population-based screening by Pap smear tests are routine procedures, organized and financed by the respective governments. Many third-world countries have no such preventive and prophylactic measures. The histological cervical intraepithelial neoplasia (CIN) scoring system was introduced by Richart in the 1960's [6]. The third kind of test is based on the Bethesda classification [7], which attempts to distinguish cases which are more- or less- likely to progress into serious (aggressive) types of epithelial lesions. Table 1 compares the different histological and cytological scoring systems and nomenclatures [8].

Table 1. Histological and cytological abnormalities (nomenclature and scoring systems)

Histology		**Cytology**	
Dysplasia	**CIN**	**Bethesda**	**Papanicolaou**
Normal	Normal	Within normal limits	Pap 1
Benign atypia	Inflammatory atypia	Benign cellular changes	Pap 1
Atypical cells	Squamous atypia	ASCUS	Pap 2
Mild Dysplasia	CIN I	Low-grade SIL	Pap 3A1
Moderate Dysplasia	CIN II	High-grade SIL	Pap 3A2
Severe Dysphasia	CIN III	High-grade SIL	Pap 3B
Carcinoma in situ	CIN III	High-grade SIL	Pap 4
(Microinvasive) cancer	(Microinvasive) cancer	(Microinvasive) cancer	Pap 5

SIL:squamous intra-epithelial lesion.

To classify the different stages of cervical cancer, the International Federation of Gynecology and Obstetrics (FIGO) has agreed on a five stage division [9] (Table2).

Table 2. FIGO classification (staging)

FIGO stage	Description
0	Carcinoma in situ (preinvasive carcinoma)
I	Cervical carcinoma confined to uterus
IA	Invasive carcinoma diagnosed only by microscopy. All microscopically visible lesions even with superficial invasion are stage IB
IA1	Stromal invasion no greater than 3.0 mm in dept and 7.0 mm or less in horizontal spread
IA2	Stromal invasion more than 3.0 mm and not more than 5.0 mm with a horizontal spread of 7.0 mm or less
IB	Clinically visible lesion confined to the cervix or microscopic lesion greater than IA2
IB1	Clinically visible lesion 4.0 cm or less in greatest dimension
IB2	Clinically visible lesion more than 4.0 cm in dimension
II	Tumour invades beyond the uterus but not to the pelvic wall or to lower third of the vagina
IIA	Without parametrial invasion
IIB	With parametrial invasion
III	Tumour extends to pelvic wall and/or involves lower third of vagina and/or causes hydronephrosis or non-functioning kidney
IIIA	Tumour involves lower third of vagina; no extension to pelvic wall
IIIB	Tumour extends to pelvic wall and/or causes hydronephrosis or non-functioning kidney
IVA	Tumour invades mucosa of bladder or rectum and/or extends beyond the true pelvis
IVB	Distant metastasis

Table 3.

Life-table analysis of 291 patients with all stages of cervical carcinoma (follow-up, 38,9 months)

Age	% 5-year survival
≤45	56
>45	64
FIGO stage	
I	72
II	42
III or IV	24

Early or advanced stage of patients are treated differently [10, 11, 12]. Survival rates are close to 100% for stage IA and much worse for stage III and IV [13], Table 3 [14].

3. Human Papilloma Virus (HPV)

HPV is a sexually transmitted virus which has the potential to cause cervical cancer. However, in the large majority of women the infection is cleared, so in fact cervical cancer is a rare complication of HPV infection. HPVs are non-enveloped double stranded DNA viruses. Their DNA is 8 kb in size and in circular form. The DNA encodes a long control region without protein coding sequences, early proteins (E1-E8) and late proteins (L1-L2). DNA is packaged within a capsid shell made of the major and minor capsid proteins L1 and L2, respectively. Purified L1 protein has the property of self-assembly into an empty shell [15]. The virus can multiply episomally, (which is known for HPV 16 strain) or in an integrated form with the preference differing from one strain to another [16]. Consequently, viral load of different strains does not necessarily represent severity of infection, because the integrated form with less viral load can cause more persistent disease than episomal one with higher viral load. Similar to other integrating viruses, the HPV seems to be capable of insertional mutagenesis and deregulation of proto-oncogenes or tumor suppressors. There is one report described consistent upregulation of the MYC proto-oncogene (v-myc myelocytomatosis viral oncogene homolog (avian)) [16].

At present more than 100 different HPV types have been identified. They differ significantly in carcinogenic potential [17]. Therefore, detection of a papilloma virus infection does not necessarily imply a high risk of cervical cancer. In most cases, HPV infections are transient, with 70% of newly infected individuals clearing the virus within 1 year and 90% of them showing no trace of HPV within 2 years [18, 19, 20]. Persistent infection causes the greatest risk and is probably enhanced with high-risk types of HPV (16 and 18). Recent studies in the United Kingdom showed that 34% of infected women carried these high-risk variants of HPV. Risk of having a persistent infection differed also between the analyzed social groups. It was higher in non-white women, in unmarried women or women cohabiting, in hormonal contraceptive users and in current smokers [21]. In Durango, Mexico, only 4.8% of all women were found to be HPV-positive. However, 75% of them carried the high-risk HPV strains, 16 and 18 [22].

The E5, E6 and E7 proteins are mostly responsible for the development of the disease. At present, two HPV proteins, E6 and E7, were found to interact and block p53 and pRB in infected cells [23]. Since these are two major regulators of cell cycle progression and apoptosis, infected cells appear to gain a proliferative advantage over non-infected cells. The full list of targets for E6 and E7 is much broader and includes proteins involved in DNA replication and control of cell division [24].

4. Markers in use/trial for HPV and or cervical cancer

It is important to realize that different biomarkers may serve different goals (e.g. markers for early detection such as HPV DNA detection in cervical scrapings, markers for premalignant lesions likely to progress such as Ki-67 in CIN lesions or markers for monitoring response to therapy such as serum SCC). Dysplastic cells show increased cell cycling. Therefore, markers of cell cycle progression might be a logical choice as biomarkers for cervical cancer. Ki-67 is one of the known antigens related to cell cycle progression [25]. Though its function is not known, it is measured with antibodies. Numerous studies confirmed the diagnostic value of this marker in identifying HPV infection and the extent of the cervical neoplasia [26]. Other markers are based on the idea that infected cells are more actively cycling, therefore they are enriched in S-,G2 and M-phase proteins such as PCNA, Cdc6, Mcm5 etc [27]. Others proposed staining for the cell senescence marker p16^{INK4a} [28], or for chromosomal aberrations such as aneuploidy [29]. Aberrant methylation of tumor suppressor genes is another concept-driven approach towards the discovery of potential cancer-related markers [30].

A separate group of markers are those directly related to the presence of HPV. Examples are the E6 and E7 proteins and squamous cell carcinoma antigen (SCC). SCC is a serine protease inhibitor [31] and it is not viral protein, though it is induced upon HPV infection. Approximately 60% of the patients with cervical cancer show elevated levels of serum SCC when all FIGO stages were included [32]. SCC levels increase with the stage of the disease and 90% of women test positive when at an advanced stage [33, 34]. Monitoring SCC levels during a patient’s treatment, for example, with radiotherapy, provides useful information for further management of the disease [35].

Another potential group of biomarkers are disease-induced proto-oncogenes. An example is the HCCR protein originally identified as human cervical cancer oncogene [36]. Its level in blood was shown to be elevated in hepatocellular carcinoma and breast cancer, however, its role in cervical cancer (despite the original name) remains unclear. Some biomarkers are related to the immune response and/or suppression thereof. Transforming growth factor β1 (TGF- β1) is a well-known cytokine with immunosuppressive activity. Its downstream target PAI1 appeared to be a good prognostic marker for cervical cancer, whereas TGFβ1 itself was a poor indicator [37].

Normal cervical epithelium is not keratinized. However, infected cells undergo complex changes of expressions of keratins, which affect cell functions and interactions. Keratinization is therefore an essential component of the disease. The spectrum of expressed keratins in normal and infected cervical epithelium was studied over the last 20 years and early reports were summarized in 1985 [38]. Recently Shadeo et al. [39] use a genome-wide SAGE (Serial analysis of gene expression) tagging technique. They found one isoform of keratin (KRT6A) as one of a few proteins forming a specific expression signature of neoplastic cervical epithelium at the mRNA level. In general, however, expression of keratins is extremely complex, so it is not very reproducible from one study to another. Recently, keratins were suggested as valuable markers not only for HPV- but also for Epstein Barr Virus-related tumors [40].

5. Vaccination against HPV

HPV infections may escape the immune response because they are often restricted to epithelial cells. Only about 60% of infected women carried detectable antibodies against HPV [41]. The first trial of a monovalent vaccine targeting the L1 protein 6 years ago showed good protection against new infections and the development of lesions [42].

Vaccination trials with a bivalent vaccine against L1 protein from HPV 16 and 18 was reported in 2004 [43]. Last year the U.S. Food and Drug Administration issued a license for a quadrivalent vaccine against L1 protein from types 6, 11, 16 and 18 [15]. This triggered initiation of national and worldwide programs of vaccination. In the Netherlands, the minister of health recently suggested a full vaccination program for young women, which should reduce the risk of infection to a minimum. This program has not yet been implemented to date. A recent survey shows that most but not all pediatricians support the idea of a large-scale vaccination campaign [44]. It must be noted that vaccination can reduce infection with the above-mentioned types of HPV to about 20-70% in one vaccination campaign (calculated based on a 90% efficiency of the vaccine, as reported by Markowitz et al. [15]). Such a program is thus not absolutely effective and must be seen as a long-term measure. Other types HPV must likely also be included in vaccine production to increase coverage. Therefore routine preventive screening and early diagnosis remain important.

In conclusion: a variety of biomarkers, serving different purposes have been described in cervical cancer. The ideal biomarker with a high sensitivity and specificity, also present in premalignant and early stage cervical neoplasia which can be determine in easy available body fluids such as serum, is not yet available.

References.

[1]. Sankaranarayanan R, Ferlay J. Worldwide burden of gynaecological cancer: the size of the problem. Best Pract Res Clin Obstet Gynaecol. 2006;20(2):207-25.

[2]. Parkin DM, Bray FI, Devesa SS. Cancer burden in the year 2000. The global picture. Eur.J.Cancer 2001;37(suppl.8):64-66.

[3]. Parkin DM, Pisani P, Ferlay J. Estimates of the worldwide incidence of 25 major cancers in 1990. Int.J. Cancer 1990;80:827-841.

[4]. Jemal A, Murray T, Ward E., et al., Cancer statistics CA. Cancer J Clin. 2005;55:10-30.

[5]. Papanicolau GN, Traut HF. The diagnostic value of vaginal smears in carcinoma of the uterus. Am. J. Obstet. Gynecol. 1941;42:193-206.

[6]. Richart RM. A theory of cervical carcinogenesis. Obstet. Gynecol. Surv. 1969;24:874-879.

[7]. Solomon D, Davey D, Kuman R., et al., The 2001 Bethesda System: terminology for reporting results of cervical cytology. JAMA 2002;287:2114-2119.

[8]. Nuijhuis ER, Reesink-Peters N, Wisman GB., et al., An overview of innovative techniques to improve cervical cancer screening. Cell Oncol. 2006;28(5-6):233-246.

[9]. Benedet JL, Bender H, Jones H, Ngan HYS, Pecorelli S. Staging classifications and clinical practice guidelines of gynaecologic cancers. International Journal of Gynaecology and obstetrics 2000;70:207-312.

[10]. Landonio F, Maneo A, Colombo A., et al., Randomised study of radical surgery versus radiotherapy for stage Ib- IIa cervical cancer. Lancet 1997;350:535-540.

[11]. Lu KH, Burke TV. Early cervical cancer. Curr. Treat. Options. Oncol. 2000;1:147-155.

[12]. Green JA, Kirwan JM, Tierney JF., et al., Susvival and recurrence after concomitant chemotherapy and radiotherapyfor cancer of uterin cervix:a systematic review and meta-analysis. Lancet 2001;358:781-786.

[13]. Waggoner S.E. Cervical Cancer. Lancet 2003;361:2217-2225.

[14]. Burger RA, Monk BJ, Kurosaki T, Anton-Culver H, Vasilev SA, Berman ML, Wilczynski SP. Human papillomavirus type 18: association with poor prognosis in early stage cervical cancer. J Natl Cancer Inst. 1996 Oct 2;88(19):1361-1368.

[15]. Markowitz LE, Dunne EF, Saraiya M, Lawson HW, Chesson H, Unger ER. Centers for Disease Control and Prevention (CDC); Advisory Committee on Immunization Practices (ACIP). Quadrivalent Human Papillomavirus Vaccine: Recommendations of the Advisory Committee on Immunization Practices (ACIP). MMWR Recomm Rep. 2007 Mar 23;56(RR-2):1-24.

[16]. Woodman CB, Collins SI, Young LS. The natural history of cervical HPV infection: unresolved issues. Nat Rev Cancer 2007;7(1):11-22.

[17]. Naucler P, Ryd W, Tornberg S, Strand A, Wadell G, Hansson BG, Rylander E, Dillner J. HPV type-specific risks of high-grade CIN during 4 years of follow-up: A population-based prospective study. Br J Cancer 2007 Jul 2;97(1):129-132.

[18]. Franco EL, Villa LL, Sobrinho JP. Epidemiology of acquisition and clearance of cervical human papillomavirus infection in woman from high-risk area for cervical cancer. J. Infect. Dis. 1999;180:1415-1423.

[19]. Molano M., Van den BA, Plummer M., et al., Determinants of clearance of human papillomavirus infections in Colombian women with normal cytology: a population-based, 5-year follow-up study. Am. J. Epidemiol. 2003;158:486-494.

[20]. Moscicki AB, Shiboski S, Broering J., et al., The natural history of human papillomavirus infection as measured by repeated DNA testing in adolescent and young woman. J. Pediatr. 1998;132:277-284.

[21]. Cotton SC, Sharp L, Seth R, Masson LF, Little J, Cruickshank ME, Neal K, Waugh N. Lifestyle and socio-demographic factors associated with high-risk HPV infection in UK women.Br J Cancer 2007 Jul 2;97(1):133-139.

[22]. Sanchez-Anguiano LF, Alvarado-Esquivel C, Reyes-Romero MA, Carrera-Rodriguez M. Human papillomavirus infections in women seeking cervical Papanicolaou cytology of Durango, Mexico: prevalence and genotypes. BMC Infect Dis. 2006:6-27.

[23]. Hanahan D, Weinberg RA. The hallmarks of cancer. Cell 2000;100:57-70.

[24]. Yim E-K, Park J-S. Biomarkers in Cervical Cancer. Biomarker Insights 2006;2:215-225.

[25]. Gerdes J, Lemke H, Baisch H., et al., Cell cycle analysis of a cell proliferation associated human nuclear antigen defined by the monoclonal antibody Ki-67. J.Immunol. 1984;133:710-1715.

[26]. al Saleh W, Delvenne P, Greimers R., et al., Assessment of Ki-67 antigen immunostaining in squamous intraepithelial lesions of the uterine cervix. Correlation with the histologic grade and human papillomavirus type. Am. J. Clin. Pathol. 1995;10:154-160.

[27]. Williams GH, Romanowski L., Morris et al. Improved cervical cancer smear assessment using antibodies against proteins that regulate DNA replication. Proc. Natl. Acad. Sci. USA 1998;95:4932-14937.

[28]. Klaes R., Benner A., Friedrich T., et al., p16INK4a immunohistochemistry improves interobserver agreement in the diagnosis of cervical intraepithelial neoplasia. Am J surg Pathol. 2002;26:1389-1399.

[29]. Monsonego J, Valensi P, Zerat L., et al., Simultaneous effects of aneuploidy and oncogenic human papillomavirus on histological grade of cervical intraepithelial neoplasia. Br J Obstet Gynaecol. 1997;104:723-727.

[30]. Herranz M, Esteller M. DNA methylation and histone modifications in patients with cancer: potential prognostic and therapeutic targets. Methods Mol Biol. 2007;361:25-62.

[31]. Suminami Y, Kishi F, Sekiguchi K., et al., Squamous cell carcinoma antigen is a new member of the serine protease inhibitors. Biochem. Biophys Res. Commun. 1991;181:51-58.

[32]. Farghaly SA. Tumor markers in gynecologic cancer. Gynecol. Obstet. Invest. 1992;(34):65-72.

[33]. Gaarnestroom KN, Bonfrer JMG, Kenter GG., et al., Clinical value of pre-treatment serum Cyfra 21-1, tissue polypeptide antigen, and squamous cell carcinoma antigen levels in patients with cervical cancer. Cancer 1995;76:807-813.

[34]. Duk JM, Groenier KH, De Bruijn HWA, et al., Pretreatment serum squamous cell carcinoma antigen: a newly identified prognostic factor in early-stage cervical carcinoma. J Clin Oncol. 1996;14:111-118.

[35]. Hong JH, Tsai CS, Chang JT., et al., The prognostic significance of pre-and post-treatment SCC levels in patients with squamous cell carcinoma of the cervix treated by radiotherapy. Int. J. Radiat. Oncol. Biol. Phys. 1998;41:823-830.

[36]. Chung YJ, Kim JW. Novel oncogene HCCR: its diagnostic and therapeutic implications for cancer. Histol Histopathol. 2005 Jul;20(3):999-1003.

[37]. Hazelbag S, Kenter GG, Gorter A, Fleuren GJ. Prognostic relevance of TGF-beta1 and PAI-1 in cervical cancer.Int J Cancer 2004 Dec 20;112(6):1020-1028.

[38]. Puts JJ, Moesker O, Kenemans P, Vooijs GP, Ramaekers FC. Expression of cytokeratins in early neoplastic epithelial lesions of the uterine cervix. Int J Gynecol Pathol. 1985;4(4):300-313.

[39]. Shadeo A, Chari R, Vatcher G, Campbell J, Lonergan KM, Matisic J, van Niekerk D, Ehlen T, Miller D, Follen M, Lam WL, Macaulay C. Comprehensive serial analysis of gene expression of the cervical transcriptome.BMC Genomics. 2007 Jun 1;8(1):142

[40]. Chauhan SC, Kumar D, Bell MC, Koch MD, Verma M. Molecular markers of miscellaneous primary and metastatic tumors of the uterine cervix. Eur J Gynaecol Oncol. 2007;28(1):5-14.

[41]. Carter JJ, Koutsky LA, Hughes JP., et al., Comparison of human papillomavirus types 16, 18 and 6 capsid antibody responses following incident infection. J. Infect.Dis. 2000;181:1911-1919.

[42]. Harro CD, Pang YY, Roden RB. Safety and immunogenicity trial in adult volunteers of a human papilloma virus 16 L1 virus like particle vaccine. J.Natl Cancer Inst. 2001;93:284-292.

[43]. Harper DM, Franco EL, Wheeler C., et al., Efficacy of a bivalent L1 virus-like particle vaccine in prevention of infection with human papillomavirus types 16 and 18 in young women: A rondomized controlled trial. Lancet. 2004;364:1757-1765.

[44]. Daley MF, Liddon N, Crane LA, Beaty BL, Barrow J, Babbel C, Markowitz LE, Dunne EF, Stokley S, Dickinson LM, Berman S, Kempe A. A national survey of pediatrician knowledge

and attitudes regarding human papillomavirus vaccination. Pediatrics 2006 Dec;118(6):2280-2289.

Chapter I.II
Sample Preparation of Body Fluids for Proteomics Analysis

Natalia Govorukhina and Rainer Bischoff

This is an updated version of a book chapter entitled: Sample preparation of body fluids for proteomics analysis; in: Proteomics of human body fluids: Principles, Methods, and Applications, Ch. 2, 31-71 (2007); (Visith Thongboonkerd, Ed.), Humana Press (Totowa, New Jersey, USA).

1. Introduction

1.1. Proteomics of Human Body Fluids

The analysis of human body fluids constitutes one of the most important approaches to the diagnosis of disease and in following therapeutic interventions. Human body fluids carry information about the status of the organism that may help in the recognition of physiological misbalances when overt pathological symptoms are not yet present. Analyzing the constituents of body fluids presents a number of challenges, the most difficult being the discrimination between variability in composition caused by an ongoing disease process and natural variability. The composition of body fluids varies due to endogenous, possibly pathological, processes and many environmental influences such as diet and life style and the way the organism deals with them (e.g., metabolism and detoxification). This variability is most obvious when one is analyzing samples from different persons (cross-sectional studies) but is also present, albeit to a lesser extent, when one is analyzing samples from the same person over a given time period (longitudinal studies). Variability cannot be avoided but may be reduced by careful selection of the study population. At any rate, the discovery of disease-related changes in the composition of body fluids requires the study of a significant number of samples from patients and controls and a careful statistical interpretation of the results.

From an analytical chemistry point of view, body fluids constitute highly complex biological samples containing cells, proteins, peptides, and many metabolites. Thus, preparation of body fluids is unavoidable prior to determining the concentration or amount of a given set of constituents. Sample pretreatment and all further downstream steps will affect the ultimate analytical result and must therefore be carefully controlled and validated. It is not easy to give a general overview of sample pretreatments for body fluids, since each of them requires an adapted protocol, which in turn needs to be tailored to certain groups of analytes. In this chapter we focus on sample pretreatments for the analysis of proteins and peptides in serum. Although serum is just one example of a body fluid, albeit an important one, we will use it to highlight general principles of sample pretreatments that have a bearing on other kinds of body fluids.

The first step after taking a blood sample from a patient is to treat it in a way that makes it suitable for storage and subsequent analysis. A common initial step is to separate blood cells from soluble components, for example, by low-speed centrifugation. During sampling and centrifugation, it is pivotal to avoid disruption or activation of cells, notably hemolysis of red blood cells (which is shown by an orange to red color of the supernatant) and activation of platelets, which may lead to degranulation. The remaining supernatant, the blood plasma, may be stored as such in the case anticoagulants were added during collection to prevent blood clotting. Alternatively, blood clotting may be allowed or induced by leaving plasma at room temperature for a few hours.

Deciding whether to store plasma or serum for subsequent analyses is important. Although plasma is easier to prepare, it requires the presence of efficient anticoagulants for long-term stability. The components of the coagulation, fibrinolytic, and complement systems are all sensitive to contact with unnatural surfaces, such as plastic containers, glass, or injection needles, and there is a risk of activating these systems during processing steps (e.g., during chromatography or solid-phase extraction). The preparation of serum requires coagulation of the plasma, which is a complex biochemical process that may be difficult to control. In most hospital or laboratory settings, coagulation is effectuated at room temperature for 1 to 4 h. During this time the endogenous coagulation system is activated, leading to a cascade of proteolytic events that results in the formation of a fibrin-containing blood clot, which is usually removed by centrifugation.

It is obvious that activating a proteolytic system can have serious consequences for subsequent proteomic analyses, and some authors have noted that the coagulation time affects the resulting serum [1,3] However, proteolytic events associated with coagulation are highly controlled due to the sequence specificity of the major proteases (thrombin, factor Xa) and their well-defined location in the coagulation pathway (factor VIIIa, factor XIa) (Figure 1).

It is thus not clear whether the coagulation time affects the final composition of the proteome significantly, but there are indications that the lower molecular weight part, the so-called peptidome, is altered (Schulz-Knappe, personal communication).

In our initial studies, which applied tryptic digestion prior to LC-MS analysis (the shotgun approach), we have not observed major changes in the resulting profiles (Figure 2).

It is, however, important to validate this sample processing step carefully within the context of the overall analytical scheme (e.g., the complete protein vs the shotgun and peptidomics approach), because coagulation time is not well controlled in most laboratory or hospital settings and experience shows that it is hard to impose strict rules on hospital personnel with respect to this parameter. Finally, for retrospective studies on already acquired and stored serum samples, it is not possible to influence this step; thus the decision here is whether to include these samples in the analysis or not. Chapter IV of this thesis will deal in detail with the effect of clotting time on the serum protein profile.

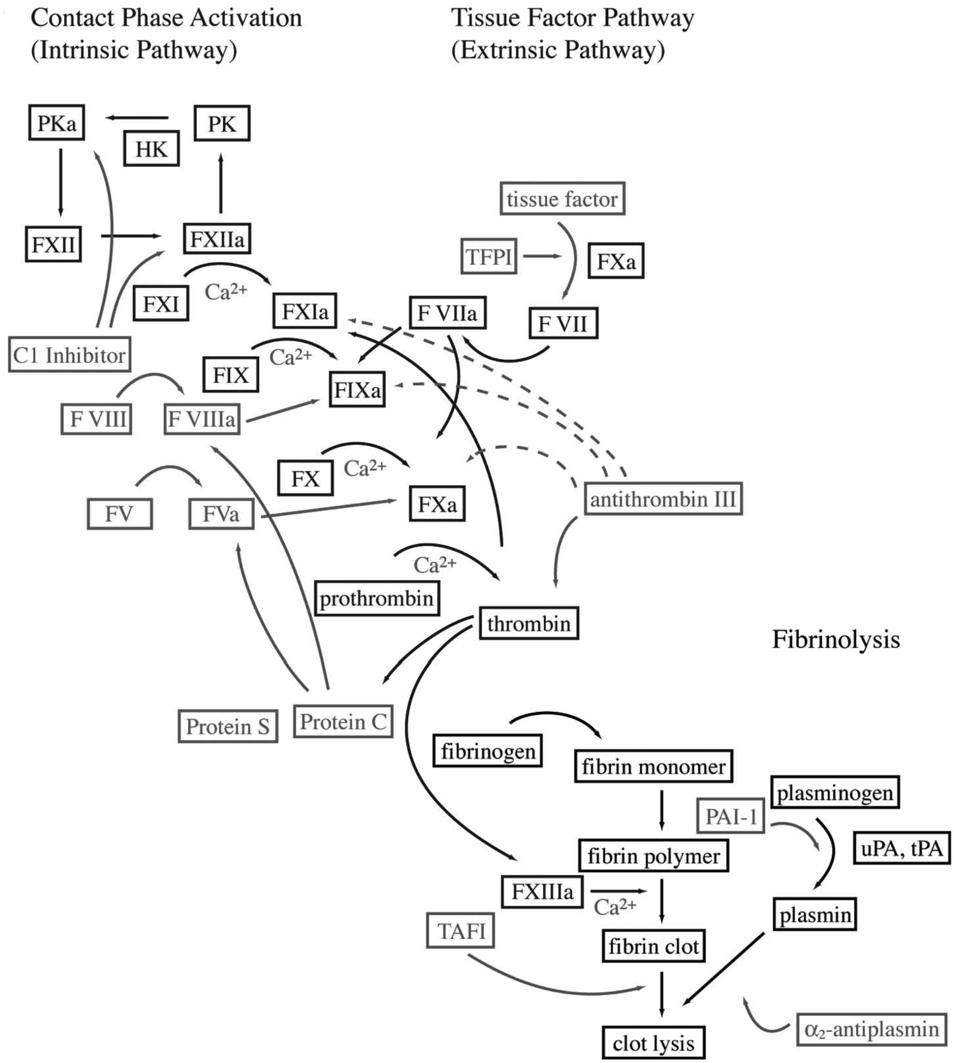

Figure 1. Overview of the intrinsic and extrinsic coagulation pathways. Both pathways are activated during preparation of serum from blood plasma. Proteolytic activity associated with coagulation may affect the profile of low-molecular-weight proteins and peptides used for peptidomics analysis. F, factor; PK, protein kinase; PKa, protein kinase A; TAFI, tissue angiogenesis factor inhibitor; TFPI, tissue factor pathway inhibitor; TPA, tissue plasminogen activator; UPA, urokinase plasminogen activator. Reproduced with permission from Tapper H, Herwald H. Modulation of hemostatic mechanisms in bacterial infectious diseases. *Blood* 2000;96:2329–2337.

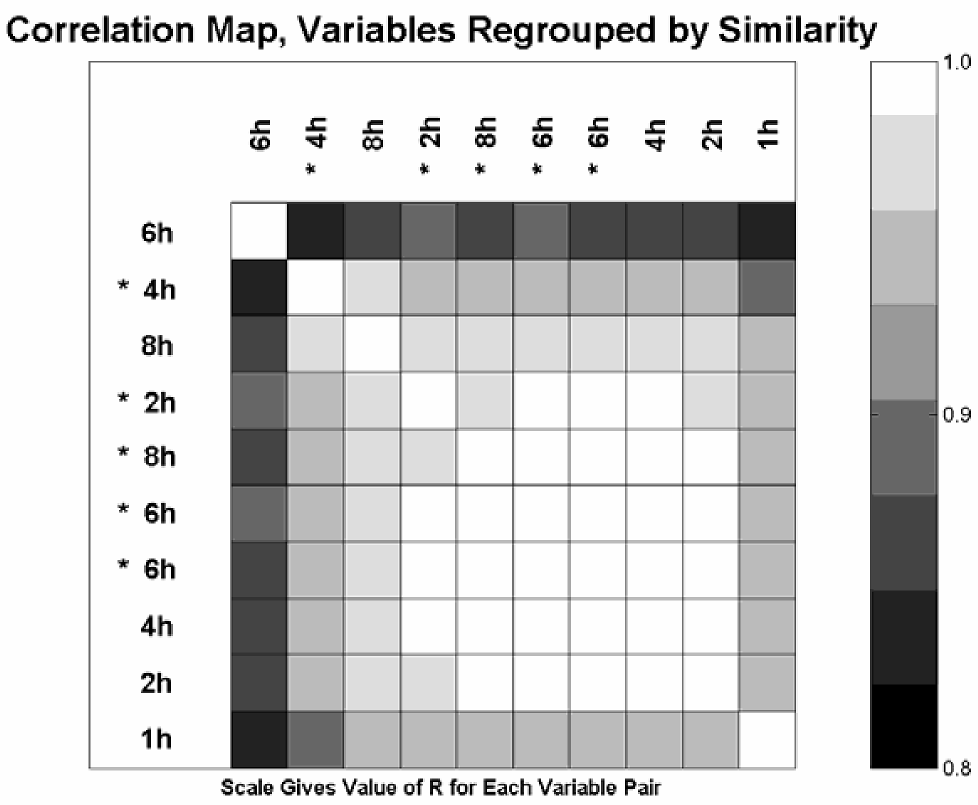

Figure 2. Correlation map of LC-MS data sets obtained from the analysis of serum (male) after depletion and tryptic digestion. Coagulation at room temperature was allowed to proceed for 1, 2, 4, 6, or 8 h. As indicated in this plot, high correlation of all data sets was found, with correlation coefficients above 0.8 in each case.

The way samples are initially treated determines in part what kind of analytes can be detected and quantified. Although this is true for any kind of analyte, it is particularly critical for the analysis of proteins and peptides, which are susceptible to degradation, precipitation, chemical modification (e.g. oxidation), adsorption to the walls of containers, and so on. Establishing a well-controlled and reproducible sampling procedure is therefore critical for any study involving human body fluids [4-7]. The proteomes in body fluids differ significantly from intracellular or tissue-derived proteomes, which are the subject of most proteomics studies. Systemic body fluids, like blood, sample the whole organism and give an average picture of the physiological state of that organism at a given point in time. Notably, blood contains a few high-abundance proteins that are to a large extent produced and secreted by the liver. In contrast, urine is a much more dilute body fluid that samples the metabolic end products from blood. Its composition is greatly influenced by the status of the kidneys. Although every body fluid presents particular challenges with respect to sample pretreatment, it is fair to say that blood is one of the most difficult body fluids to analyze.

In the following, we will highlight a number of options for sample pretreatment prior to proteomics analysis. Our focus will be directed at serum, but the principles are applicable to other body fluids. We will try to emphasize that there are strategic choices to be made early on in the analytical procedure that will determine the final result.

1.2. Methodological Overview

There is no single approach to proteomics in body fluids. It is likely that the comprehensive analysis of the proteome of any given body fluid is still beyond our reach despite great methodological advances in recent years. A major challenge is the concentration range of proteins in most body fluids, which spans more than 11 orders of magnitude [8] (Figure 3).

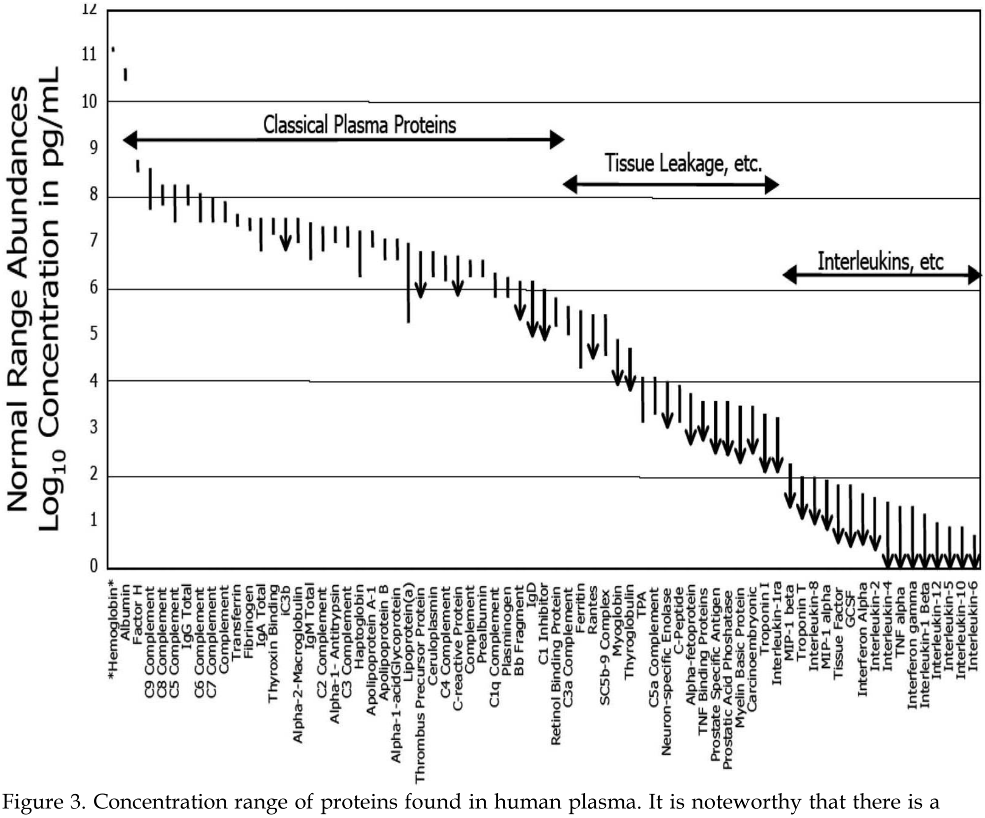

Figure 3. Concentration range of proteins found in human plasma. It is noteworthy that there is a difference of more than 11 orders of magnitude between the most concentrated and the very low-abundance proteins. Reproduced with permission from Anderson NL, Anderson NG. The Human Plasma Proteome: history, character, and diagnostic prospects. *Mol Cell Proteomics* 2002;1:845–867.

Furthermore, it is difficult to predict the number of proteins in body fluids owing to processing events (like the generation of smaller proteins and peptides from larger precursors, posttranslational modifications, and the fact that proteins can enter body fluids by well-defined pathways like secretion) as well as cell and tissue turnover as a result of necrosis or apoptosis. From a methodological point of view, the proteome of a body fluid may be roughly divided into high- and low-molecular-weight compartments (Figure 4).

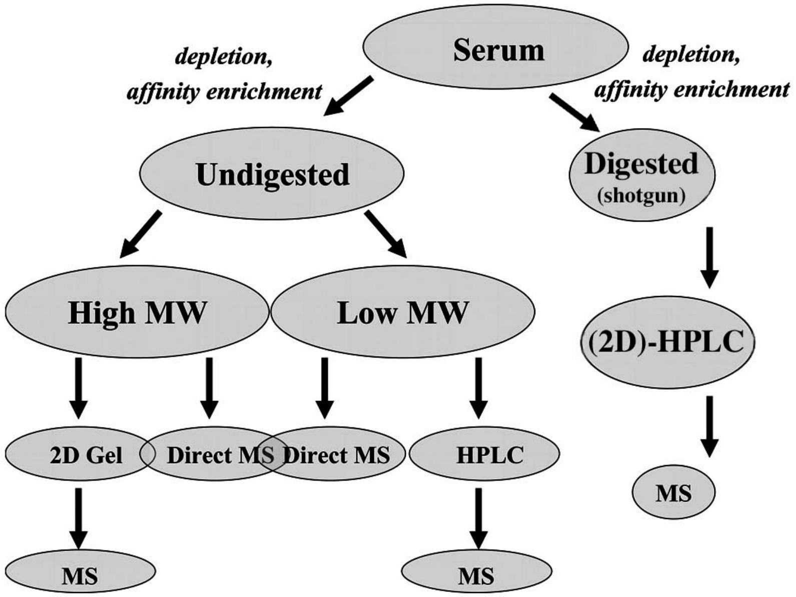

Figure 4. Schematic overview of different approaches of sample preparation. The main initial difference is whether undigested or trypsin-digested proteins are being analyzed. A second level of differentiation comes into play when one is deciding whether to analyze the high- or the low-molecular-weight (MW) fraction of the proteome. High-molecular-weight proteins are best analyzed by 2D electrophoresis, whereas peptides are more amenable to high-performance liquid chromatography (HPLC) coupled to mass spectrometry.

This discrimination is rather arbitrary and is mostly defined by the size-dependent separation method used for prefractionation.

Body fluids have been prefractionated by ultrafiltration with membranes of various cutoff values [9-11]. Although ultrafiltration appears to be an easy separation methodology with an apparently clear-cut separation mechanism, its application to complex biological samples shows that discrimination between proteins above and below the nominative cutoff of the membrane is never complete [11]. Effects such as adsorption of proteins to the membrane, the generation of a polarization layer close to the membrane surface, and deformation of the pores in relation to the *g*-force can all affect filtration.

Ultrafiltration is also performed on a large scale in patients with renal insufficiency, and membrane material has been the source for many studies of bioactive proteins and peptides below approximately 20 kDa [12-14]. Although the kidney itself is performing much more complex tasks than just ultrafiltration, urine may be considered an ultrafiltrate of blood and thus has a considerably lower concentration of high molecular- weight proteins.

An elegant combination of ultrafiltration and chromatography is based on restricted access materials (RAMs), which have an adsorbing internal pore surface and a non-adsorbing external surface [15,16]. The pore diameter in most RAMs is about 6 nm, which corresponds roughly to a cutoff value of 20 kDa. RAM chromatography has been integrated into analytical systems with the goal of analyzing the low-molecular-weight part of the proteome of blood diafiltrates [17,18] or artificial cerebrospinal fluid [19]. Although a clear enrichment of the low-molecular-weight fraction was observed, a considerable amount of albumin was still present even after RAM chromatography.

The decision whether to work with undigested proteins or to digest proteins with trypsin prior to further analyses is of principal importance in an analytical scheme (Figure 4). Performing separations of very complex mixtures of proteins is difficult, owing to the wide range of physicochemical properties and the possibility that proteins will denature, aggregate, or even precipitate under separation conditions. The most universally applicable separation method for proteins is 2D polyacrylamide gel electrophoresis (2D-PAGE), whereby all proteins are denatured from the beginning and kept in a denatured state throughout separation. This reduces the risk of aggregation and precipitation with subsequent loss of proteins as well as that of proteolysis. There is no comparable universal chromatographic method, and it is thus necessary to develop an appropriate fractionation scheme for groups of proteins or individual proteins. However, as 2D-PAGE has limitations with respect to low-molecular-weight proteins of 10 to 20 kDa and hydrophobic or basic proteins, alternatives are being developed.

One approach is based on the so-called shotgun method, whereby the complete protein mixture is digested with trypsin (other proteases are conceivable as well for this purpose but are not widely used) and the generated peptides are separated by 2D or 3D high-performance liquid chromatography (HPLC) [20-23]. Shotgun proteomics has the advantage of overcoming many of the difficulties related to very hydrophobic or otherwise intractable proteins at the expense of rendering the separation problem quite daunting. Assuming that serum contains about 10^5 different protein forms, each of which generates 50 tryptic peptides, one has to deal with a mixture of about 5 x 10^6 peptides to be separated. Fortunately, not all peptides need to be separated into single peaks, and not all peptides of each protein need to be identified by mass spectrometry to trace them back to the protein of origin. A drawback of the shotgun method is that not all regions of a protein are covered by the analysis, which may mean that some possibly relevant posttranslational modifications or processed forms are missed. Nevertheless, the excellent separation capacity of HPLC for peptides compared with complete proteins and the much easier identification of peptides by tandem mass spectrometry have accelerated the use of shotgun proteomics in the biomarker discovery area. The daunting separation problem posed by this approach has also driven recent new developments in HPLC stationary phase chemistry and technology that increase separation efficiency and reduce analysis time [24-28].

The presence of a few high-abundance proteins in body fluids such as albumin has driven developments to deplete these proteins specifically and thus to increase the loading capacity by a factor of 5 to 10 [29-35] or even 30 to 50, depending on the number of proteins that are depleted (e.g. ProteoPrep® 20 Plasma Immunodepletion kit, www.Sigma-Aldrich.com removes the 20 most abundant serum proteins). In addition to increasing the protein or peptide load, depletion also augments the capacity to detect peptides derived from lower abundance proteins [36]. It has been shown, however, that depletion of high-abundance proteins does not proceed without the loss of some low-molecular-weight proteins and peptides [37]. It is thus important to decide whether to deplete or rather try to design a fractionation strategy that deals with high-abundance proteins such as albumin or immunoglobulins in another way [38].

Arguably, very low-abundance proteins in the ng/mL to pg/mL range cannot be detected in complex protein mixtures such as serum even after depletion. Many regulatory proteins such as cytokines or some of the known tumor-specific markers reside in this concentration range and are presently measured by immunological methods. To reach into this lower concentration range, it is often necessary to enrich a given set of proteins by affinity chromatography using highly selective antibodies or group-specific ligands like lectins. The use of protein-specific antibodies limits the scope of the analysis to those proteins that are recognized. Group-specific affinity ligands such as lectins or antibodies directed at a common structural element such as phosphotyrosine represent a compromise between the comprehensive proteomics approach, which often fails to detect low-abundance proteins, and the highly specific methods. For example, lectins have been applied to the discovery of tumor-specific glycoprotein markers, since tumor cells often produce proteins with aberrant glycosylation patterns [39-41]. Lectins have also been used to enrich glycoproteins from complex protein mixtures or glycopeptides from tryptic digests of such mixtures [42-45]. In general, targeted approaches require a hypothesis concerning the role that different kinds of proteins may play in a given disease in order to chose appropriate affinity ligands for enrichment.

2. Sample Preparation

2.1. Preparation of Plasma and Serum

Between two fundamentally different compartments of the blood, namely, blood cells and the actual fluid, most clinical analyses are done on derivatives of the fluid, like plasma or serum. Discussion continues on whether serum or plasma should be used, but this may also depend on the general practice of the hospital that provides samples for analysis, notably, whether samples are analyzed from existing collections. Preparation of plasma requires addition of anticoagulants, such as EDTA, citrate, and/or heparin, whereas serum contains no extra additives. Serum lacks components of the coagulation system, such as thrombin and fibrinogen, since they are part of or become entrapped in the blood clot and are thus removed from the serum. In addition, other proteins or peptides that have some binding affinity to the clot may be partially depleted. Being a proteolytic process, coagulation generates peptide fragments from larger proteins that may especially affect the composition of the peptidome. Interestingly, comparison of plasma made with EDTA, citrate, or heparin also shows variation in protein composition [46].

Sampling blood for plasma or serum preparation is routine in most hospital laboratories and a reasonably standardized procedure is in place using commercial reagents and materials. However, most laboratory technicians and nurses are not aware of the specific requirements of proteomics and thus need to be informed. Very restrictive standard operating procedures (SOPs), notably with respect to the coagulation time and conditions, are often difficult to follow in routine hospital operations. Biomarkers discovered thus far therefore need to be robust enough to be useful in a routine clinical laboratory, and very unstable proteins or peptides are probably not of interest in the long run.

Recently, Schulte et al. [2] reported that a considerable number of peptides were found in serum but not in human plasma. The authors suggest that these peptides appeared as a result of a clotting-related proteolytic activity. This might be indicative of artifacts generated as a result of the clotting reaction, which is disturbing with respect to peptidomics. The authors therefore propose to use human plasma for this purpose.

An example of the preparation of plasma for biomarker discovery by Peptidomics® (Schulz-Knappe, personal communication) involves taking a blood sample from a superficial vein of the cubital region. The blood sampling procedure should not take longer than 1 min, and EDTA is used as the anticoagulant. Prior to collection, the first sample (approximately 2.5 mL) is discarded. To remove platelets, the sample is centrifuged at 2,000*g* for 10 min. The final plasma sample (approximately 1.5 mL) should be frozen within 30 min after being taken and stored at -80°C. Serum is made by letting a fresh blood sample coagulate (with or without thrombin as activator) and either filtering it through a gel or collecting the liquid fraction after centrifugation. Although their hypothesis has not been proven, Sorace and Zhan [47] suggest that variations in coagulation might be a significant factor in obscuring clinical proteomics data sets. The source of variation can be both technical and natural. Schulte et al. [2] found that a naturally occurring Val-34 to Leu mutation in the activation peptide of factor XIII (FXIIIA) not only affected the process of blood clotting but also correlated with a lower incidence of myocardial infarction and ischemic stroke and an increased risk of hemorrhagic stroke. According to our results (see chapter IV of this thesis), different clotting times ranging from 1 to 8 h in the preparation of serum samples resulted in highly correlated liquid chromatography-mass spectrometry (LC-MS) data sets when analyzing serum proteins after depletion and trypsin digestion. Correlation coefficients above 0.8 were found for all samples after selecting the 37 top information-rich m/z traces using the CODA component detection algorithm [48] (Figure 2).

2.2. Removal of High-Abundance Proteins

As the presence of abundant proteins in most biofluids used for diagnostic purposes decreases the capacity of analytical methods to detect low-abundance proteins or peptides, a range of approaches has been developed to reduce the total amount of protein. Blood serum is a complex mixture of thousands of proteins and peptides. However, few of the serum proteins are present in extremely high amounts compared with the rest of the serum components. (human serum albumin [HSA] constitutes 57–71% and γ-globulins 8–26% of the total of all human serum proteins). The 10 most abundant proteins account for 97% of all the protein content in plasma [8]. In a recent publication, it is stated that the search for specific markers occurs in a fraction of less than 1% of all plasma proteins [49].

Removal of these high-abundance proteins increases the loading capacity of the analytical system by a factor of 5 to 10 and thus improves the detection of low-abundance proteins. Several affinity columns are presently on the market based on dye ligands or antibodies for albumin removal and protein A or G for the removal of immunoglobulins [50]. Technically simple approaches that allow processing of multiple samples in parallel based on HSA- and IgG-binding spin columns or filters have been developed [51]. For HSA binding, two types of stationary phases are generally used: (1) those based on dye ligands such as Cibacron-Blue and derivatives thereof [52], and (2) those based on specific antibodies against human serum albumin [51] raised in mammals (IgG) but also in chickens (IgY), as recently described [31]. HSA was also successfully removed by affinity capture on immobilized phage-derived peptides [53]. Recently, a synthetic peptide derived from protein G was used for HSA affinity chromatography and depletion of HSA from human plasma [54]. The column could easily be regenerated with alkaline treatment owing to the stability of the peptide, and its specificity and capacity were quite high. However, the column is presently not commercially available. Removal of IgG is exclusively done by well-established methods based on immobilized protein

A, protein G, or protein L, owing to their high affinity and selectivity [55-60]. Comparative studies of HSA- and IgG-binding columns based on Poros® polystyrene-divinylbenzene beads (Applied Biosystems) [32] as well as on Mimetic Blue (ProMetic BioSciences) and HiTrap Blue (Amersham Biosciences) for HSA removal have been performed [33].

We tested several approaches specifically to reduce the level of high-abundance proteins in serum based on either specific antibodies, dye ligands (for albumin), or protein A or G (for γ-globulins) [30]. Analysis by sodium dodecyl sulfate (SDS)-PAGE (Figure 5) and LC-MS after tryptic digestion of the remaining proteins (Figure 6), showed that reduction with albumin-directed antibodies was most effective, albeit not complete [36].

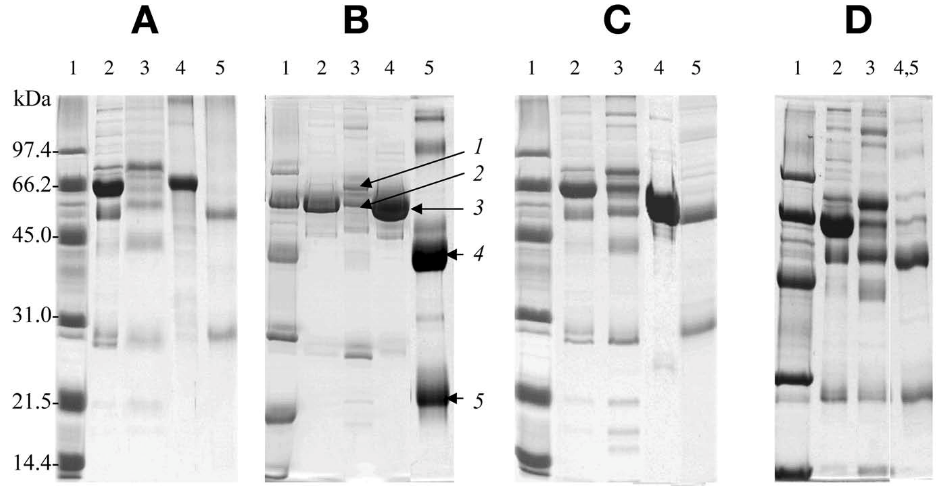

Figure 5. Depletion of albumin and γ-globulins from human serum. In each lane 8 to 10 µg of protein were loaded, and gels were stained with Coomassie Blue G-250. (A) POROS Anti-HSA and POROS Protein G columns. (B) HiTrap Blue and HiTrap Protein G columns. (C) Merck Albumin Removal column and HiTrap Protein G column. (D) Aurum Serum Protein column. Lanes: 1, standards; 2, crude serum; 3, depleted serum; 4, bound protein eluted from albumin-depleting columns; and 5, bound γ-globulins eluted from columns. Labeled protein bands: 1, serotransferrin; 2, α_1-antitrypsin; 3, albumin; 4, 5, γ-globulins, heavy and light chains, respectively. Reproduced with permission from Govorukhina NI, Keizer-Gunnink A, van der Zee AGJ, de Jong S, de Bruijn HWA, Bischoff R. Sample preparation of human serum for the analysis of tumor markers: comparison of different approaches for albumin and [gamma]-globulin depletion. *J Chromatogr A* 2003;1009:171–178.

In our initial studies, which applied tryptic digestion prior to LC-MS analysis (the shotgun approach), we have not observed major changes in the resulting profiles (Figure 2).

A more recently introduced multiple affinity removal column, which depletes certain high-abundance proteins (albumin, IgG, IgA, transferrin, haptoglobin, and α$_1$-antitrypsin [50]), proved to be most effective in our hands and provided more reproducible results during LC-MS analysis regarding retention times and peak areas than previously evaluated methods [61,62]. Similar results were recently published for 2D gel electrophoresis (Figure 7) [50]. Recently immunoaffinity materials have been introduced that remove the 20 most abundant proteins (ProteoPrep® 20 Plasma Immunodepletion kit/ spin columns (www.Sigma-Aldrich.com)) or the 14 most abundant proteins (Hu-14 Spin Cartridges (www.home.agilent.com)).

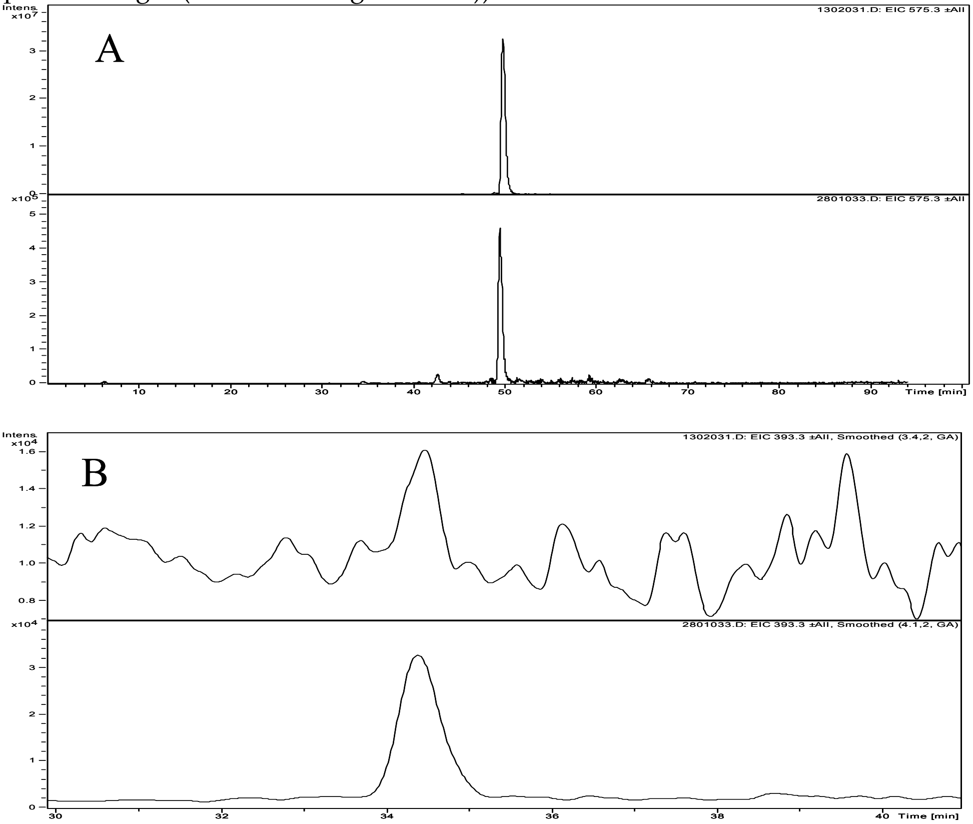

Figure 6. Efficiency (A) and selectivity (B) of albumin removal from human serum using an anti-albumin immunoaffinity column. (A) Extracted ion chromatogram of *m/z* =575.3 (doubly charged molecular ion of peptide LVNEVTEFAK; positions 41–50 in human serum albumin) of tryptic digests of human serum (upper trace; peak height 3.2 x 10^7) or of human serum after depletion with an anti-albumin immunoaffinity column (lower trace; peak height 4.6 x 10^5). (B) Extracted ion chromatogram of *m/z* x 393.3 (doubly charged molecular ion of peptide IVDLVK; positions 193–198 in human α$_1$-antitrypsin) of tryptic digests of human serum (upper trace; peak height 16,052) or of human serum after depletion with an antialbumin immunoaffinity column (lower trace; peak height 32,607). Note the much cleaner detection of this peptide fragment after depletion and the increased overall peak height. Reproduced with permission from Bischoff R, Luider TM. Methodological advances in the discovery of protein and peptide disease markers. *J Chromatogr B* 2004;803:27–40.

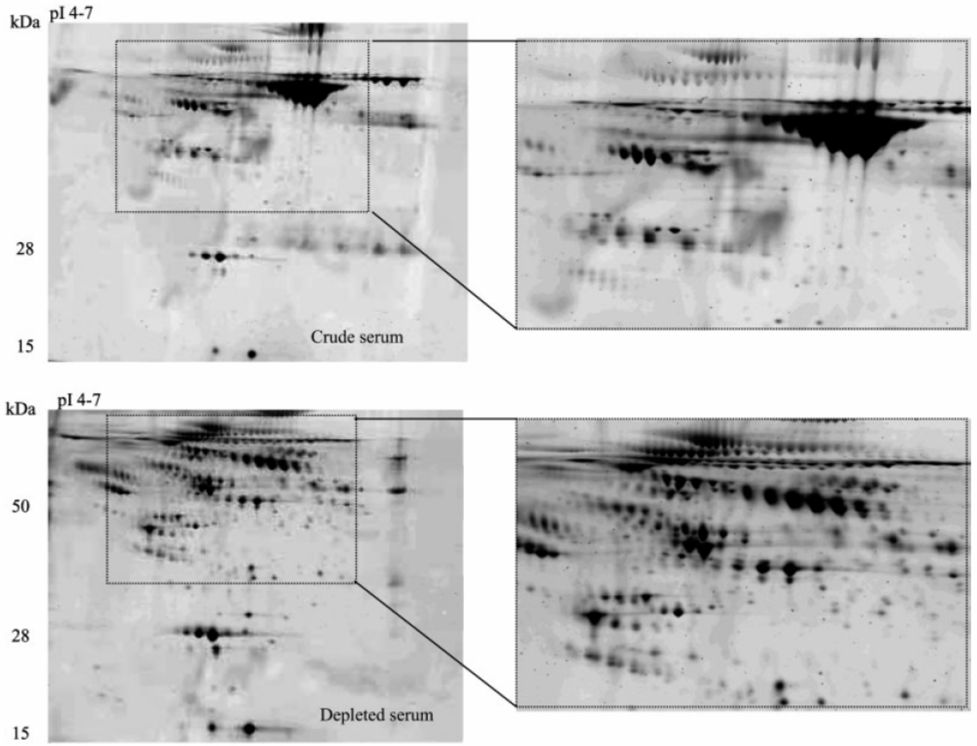

Figure 7. 2D electrophoresis of crude (70 µg protein) and depleted (6-protein multiple affinity removal column, 100 µg protein) serum samples. On the right side is a zoom view of the area containing albumin. Reproduced with permission from Bjorhall K, Miliotis T, Davidsson P. Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. *Proteomics* 2005;5:307–317.

Removal of high-abundance proteins by ultrafiltration through cellulose filters with a cutoff of 30 kDa proved to be less successful [11]. Many known "landmark" proteins of low molecular weight (<30 kDa) were missing upon 2D gel electrophoresis. Ultrafiltration has, however, the advantage of allowing one to concentrate the low-molecular-weight fraction of the proteome and was found to be useful for analysis of low-molecular-weight proteins (LMWs) by LC-MS after prefractionation by strong cation-exchange HPLC [10]. To prevent binding of LMWs to serum carriers, particularly albumin, 20% acetonitrile was used. In the resulting fraction, the authors could identify 314 unique proteins including cytokines, growth factors, and transcription factors, which are proteins of low abundance that are very difficult to detect by other methods without serum depletion. This method of sample preparation at a cutoff of 10 kDa was recently used to analyze the LMW fraction of pooled serum from patients with ovarian cancer by nanoLC-electrospray ionization-fourier transform ion cyclotron resonance (ESI-FT-ICR)-MS and analyzed statistically [9].

A quite different set of methods uses electrophoretic approaches to fractionate complex samples and to separate high-abundance proteins from those of low abundance. The basic principle is based on preparative isoelectric focusing and/or free-flow electrophoresis in solution, whereby the crude sample is prefractionated in a specially designed chamber according to the different electrophoretic mobilities or isoelectric points of the proteins. The HSA-rich fraction was discarded, and other fractions were pooled or analyzed separately. The method was originally reported more than 10 years ago [63] and is still in limited use [64-66]. Some commercially available systems can be used for electrophoretic prefractionation (e.g., the Zoom IEF fractionator® [Invitrogen] or the system produced by Weber).

There is one particular problem associated with the removal of serum albumin and globulins. These proteins appear to fulfill the function of carriers for less abundant proteins [37,67]. This is especially critical for LMWs, since they can escape kidney clearance only when bound to high-molecular-weight carrier proteins. Many of these LMWs are found to be associated with the development of cancer and could therefore be extremely important biomarkers (see ref. 37). Binding of LMWs to high-abundance, high-molecular-weight proteins may be used advantageously based on a two-step procedure, whereby abundant carrier proteins are first specifically bound to the corresponding affinity column followed by elution of the bound LMWs using a gradient [37,67].

2.3. Targeted Enrichment of Individual Proteins or Protein Families

Since many disease-specific biomarkers are likely of low to very low abundance in body fluids, it is a major challenge to detect them using profiling methods. Reaching the required concentration sensitivity often requires complex, well-designed protocols of enrichment and fractionation that are rather time-consuming. A concept of a multi-dimensional fractionation system (MDFS) was recently proposed and discussed [68]. Based on a given hypothesis about the disease mechanism, it is therefore often advisable to use targeted, affinity-based methods for enrichment prior to analysis. A combination of proteomics technology with targeted enrichment that does not require a very "sharp" hypothesis is based on group-specific ligands like lectins [69-70] in case of glycoproteins, or activity-based probes (ABPs) in the case of proteases or other enzymes [71-77].

2.3.1. Lectins

Glycosylation of proteins is a posttranslational modification that is easily affected by cellular growth conditions. Modifications of glycosylation patterns are therefore often observed in fast growing cancerous cells compared with their quiescent counterparts [39,78]. Analysis of the carbohydrate portion of proteins is a rather complex task, since the glycosyl moiety is usually a branched chain polymer with an enormous variety in length, composition, and complexity. Studies of glycoproteins can be divided into two types: 1) identification of the proteins and their glycosylation sites and 2) the more demanding analysis of the structure of the attached glycosyl residues themselves.

Glycosyl residues can be linked to the protein core via asparagine (*N*-linked glycans) or bound via serine or threonine (*O*-linked glycans). For *N*-linked glycans, *N*-acetylglucosamine (GlcNAc) is the first monosaccharide in the chain, whereas *N*-acetylgalactosamine (GalNAc) is most often found for *O*-linked glycans. In addition to being potentially interesting as biomarkers, failure of proper glycosylation can cause severe abnormalities [79].

Although the exact structure of glycosylated proteins varies considerably, probably all known glycoproteins can be enriched by lectin affinity chromatography [80]. There are several commercially available lectin affinity columns, which differ in specificity and are used widely in early stages of the isolation of glycoproteins [81]. The specificity of many lectins is known (Table 1), allowing the rational design of complementary enrichment schemes.

Table 1: Selected lectins with their specificities.

Lectin	**Specificity**
Concanavalin A (ConA)	glucosyl and mannosyl residues of N–linked oligosaccharides
Wheat germ agglutinin (WGA)	chitobiose core (di-N-acetylglucosamine) and N-acetylneuraminic acid
Peanut agglutinin (PNA)	T antigen (Galb1-3GalNAc) found in O-glycans of mucin-type proteins
Aleuria aurantia (AAL)	L-fucose–containing oligosaccharides
Galectines	N-acetyllactoseamine (LacNAc)-containing glycans found in both N- and O-glycans

Recently five lectins, concanavalin A (Con A), wheat germ agglutinin (WGA), jacalin, lentil lectin (LCA), and peanut lectin (PNA) were tested for capturing glycoproteins from human serum [41]. At first, the authors depleted the high-abundance proteins with a multiple affinity depletion column followed by enrichment on single- or multiple-lectin affinity columns. The enriched proteins were eluted with buffers containing specially selected sugars. The resulting fractions were analyzed by LC-MS after digestion with trypsin. Figure 8 gives an example of how the analysis of proteins in serum can be focused to a subset containing a fucose residue by prior enrichment on a column containing the fucose-specific Lotus tetragonolobus agglutinin (LTA) [43].

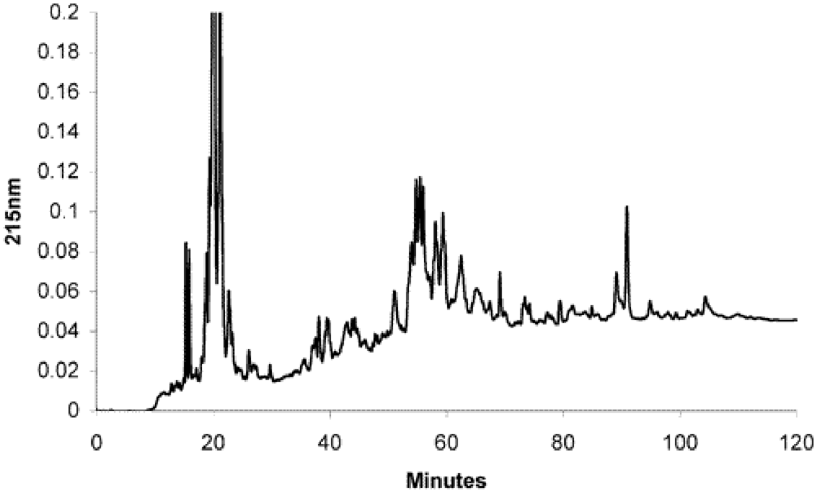

Figure 8. Enrichment of fucose-containing peptides in serum proteins after tryptic digestion. An LTA lectin affinity column was used to select the peptides followed by deglycosylation and reversed-phase HPLC. Reproduced with permission from Xiong L, Andrews D, Regnier F. 2003. Comparative proteomics of glycoproteins based on lectin selection and isotope coding. *J Proteome Res* 2003;2:618–625.

2.3.2. Activity-Based Profiling of Proteases

Standard proteomics techniques give information about the relative abundance of proteins and possibly posttranslational modifications. In most cases, however, these techniques do not provide information about biological activity. In recent years another branch of proteome analysis has developed to tackle this problem with the development of affinity-labeling techniques, generally called activity-based protein profiling (ABPP) [82-85]. This line of research focuses on profiling the activity of families of enzymes like the various types of proteases. A derivative of this work is to use affinity ligands, like protease inhibitors, to enrich classes of proteins based on their activity [76]. Arguably, it is the activity of enzymes that is involved in disease development and that may therefore serve as biomarkers rather than the abundance, since most enzymes are present as inactive proforms that are activated upon appropriate (or inappropriate) stimuli.

ABPs have been described for cysteine proteases [84,86-88], serine hydrolases, including serine proteases [73,89], and also metalloproteases [71,90-92]. In most cases the labels contain biotin, which allows one not only to visualize but also to isolate the labeled proteins. Even *in vivo* labeling, for example, in tissue biopsies or cells in culture is feasible [84]. By employing this strategy, sample depletion for abundant proteins can be bypassed as long as the inhibitors or other affinity ligands are sufficiently specific and nonspecific binding to the support materials can be minimized. Figure 9 gives an example of how strongly some proteins may bind to materials used for the immobilization of affinity ligands and how nonspecific binding may be overcome by chemical derivatization of the surface.

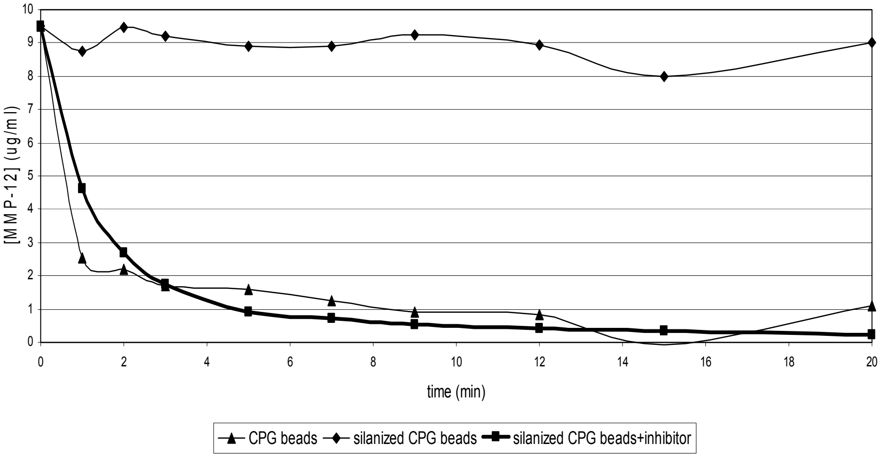

Figure 9. Binding of the catalytic domain of matrix metalloprotease-12 (MMP-12) to unmodified controlled porosity glass beads (▲), silanized with a diol layer (◆) or silanized and then coupled to a reversible MMP inhibitor (■) (Dr. Robert Freije, unpublished data).

Nonspecific binding to affinity ligands or the surfaces on which they have been immobilized makes stringent controls necessary. For example, preparing non-functionalized "control" materials and competing with the affinity interaction by adding an excess of ligand to the binding buffer are common ways of assessing specificity. Some authors have also denatured the proteins by heat treatment prior to binding as a control. As an example, a complex mixture of extracellular serine hydrolases was successfully identified by MS after the hydrolases were enriched by affinity chromatography. In addition, resolution of protein bands on SDS-PAGE was improved upon deglycosylation of the enriched enzymes [93,94].

Successful affinity-based profiling requires suitable affinity ligands. To address a wider range of proteins based on their activity, combinatorial chemistry approaches have been developed [95]. A new promising concept of *in vivo* click chemistry protein labeling utilizing the copper(I)-catalyzed azide-alkyne cycloaddition reaction was recently introduced [96]. However, this method is mostly limited to catalytically active proteins, such as enzymes. A great number of protein-protein interactions, however, remain largely unexplored.

2.4. Protein Chip Technology (SELDI-TOF-MS)

Since natural body fluids are too complex to analyze directly, investigators are in constant search of new techniques of subfractionation prior to MS. Most often prefractionation is done by LC or electrophoresis, but simple adsorption/washing/desorption methods are finding more widespread use, as they are rather fast and can be automated more easily. In classical matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS), prefractionated protein samples are digested with trypsin prior to analysis (peptide mass fingerprinting). In this version all peptides are indiscriminately deposited on the MALDI target plate and entrapped in the light-absorbing matrix.

The central idea of surface-enhanced laser desorption/ionization (SELDI) is to use adsorptive surfaces, mainly based on well-known chromatographic principles, to bind a subfraction of proteins from a sample and to analyze the bound proteins or peptides by MALDI-MS. By varying the adsorptive surface, different groups of proteins can be bound and analyzed. This technology has more recently been further developed and commercialized under the trade name ProteinChip® (Ciphergen Biosystems, Palo Alto, CA) and has found widespread application, notably in the medical and clinical research community [97]. The original mass spectrometer was a simple linear MALDI-time of flight (TOF) system, but interfaces have now been developed that allow coupling of the ProteinChip technology to tandem mass spectrometers of the quadrupole- TOF hybrid type.

The key components of this technology are Protein Chip Arrays and the Protein Chip Reader. The array comprises a set of different surfaces, such as a hydrophobic, hydrophilic, or metal chelate to which the biological samples are added. The unbound proteins are washed away, and the bound fraction is subjected to MS analysis. Optionally, it is possible to digest the bound proteins with trypsin to facilitate their identification. However, since most chips bind many diverse proteins, interpretation of the results after trypsin digestion is not always obvious because it is not straightforward to link the observed peptides back to the proteins that gave rise to an increased or decreased peak in the original spectrum. As a recent example of this approach, cystatin C was suggested as a biomarker in the diagnosis of Creutzfeld-Jakob disease [98]. Direct fragmentation of the high-molecular-weight ions to obtain sequence information for identification would be most advantageous. This has recently been facilitated due to ion activation mechanisms such as Electron Capture/Transfer Dissociation or Infrared Multiphoton Dissociation [99-104].

As with any mass spectrometric method dealing with highly complex mixtures, there is a competition between different molecules to ionize (also known as *ion suppression*). It is thus unlikely that the mass spectrum obtained from a ProteinChip will give a true representation of the proteins or peptides adsorbed on the chip. Most applications of SELDI-TOF-MS to body fluids therefore generate rather simple mass spectra, which can be easily analyzed. A number of applications of SELDI-TOF to biological samples, notably plasma, serum, or urine, have shown that samples taken from patients differ significantly from those from healthy controls or from patients with other kinds of disease, opening the possibility of discriminating patient groups and performing early diagnosis of, for example, ovarian or breast cancer [105-106]. However, recent efforts to reproduce these results have met with limited success, and the jury is still out on whether this fairly straightforward approach to sample preparation will lead to clinically relevant results [107-109].

Probably one of the most impressive studies using SELDI-TOF in recent years was the detection of an antiviral factor secreted by CD8 T-cells upon infection with HIV-1 in immunologically stable patients, which was identified as a member of the α-defensin family by subsequent isolation and protein sequencing [110]. This factor had been known since 1986 but had eluded identification for 15 years [111].

2.5. Automated Sample Preparation Using Magnetic Beads

The automation of sample preparation in light of increasing sample throughput and reproducibility is an important aspect of clinical proteomics. In analogy to the previously described ProteinChips, it is possible to prepare samples by adsorption/washing/desorption on magnetic beads (or other kinds of beads). Magnetic beads are an effective tool for fast concentration of diluted samples and for the crude separation of proteins and peptides prior to MS analysis. Magnetic beads are widely used in automated immunoassays, cell purification, and more recently, the detection of bacterial pathogens. Mostly, the assay is targeted at individual proteins, like prostate-specific antigen (PSA), which is captured with a biotinylated anti-PSA antibody (anti-F-PSA-M30-IgG) and subsequently bound to streptavidin-coated magnetic beads [112]. This approach, however, is targeted to a specific protein and is not applicable to proteomic studies in a broad context. In another example, magnetic nanoparticles modified with vancomycin were used to trap Gram-positive bacteria [113]. The method was able to detect *Staphylococcus aureus* in a 3-mL urine sample at a concentration of 7 x 10^4 CFU/mL (colony-forming units) by MALDI-TOF-MS. A combination of affinity trapping with MS has also been successful in detecting bacterial and viral infections based on immobilized lectins [114,115]. Application of magnetic beads to clinical proteomics has emerged only recently mainly based on adapted liquid handling systems [116]. Serum was precipitated with ethanol to remove larger proteins, and the remaining polypeptides in the supernatant were bound to reversed-phase super-paramagnetic silica beads. The washed and eluted peptides were profiled by MALDI-TOF/TOF-MS with the possibility of performing partial sequencing and identification by MS/MS. Four hundred polypeptides were detected in 50 µL serum (range 0.8–15 kDa), and discrimination between samples from brain tumor patients and healthy controls was 96.4% based on a learning algorithm. The number of examples of protein enrichment with magnetic beads has been steadily increasing over the last years [117-120].

2.6. Analysis of Other Body Fluids

Sample preparation is equally important for proteomics and peptidomics in other body fluids, and many of the methods and considerations developed for serum are suitable. Urine is probably the second most relevant body fluid after blood for general proteome studies, owing to its availability. Urine is in fact filtered blood plasma, so it might be representative of the protein spectrum of blood, but with lower protein concentrations. In normal conditions the kidney restricts passage of plasma proteins above approx 40 kDa during filtration in the glomeruli. Proximal renal tubules reabsorb filtered proteins and degrade them. Total amounts of secreted protein per voiding vary from 1 to 10 mg, whereas in pathology, the protein concentrations can dramatically increase [121,122]. Urine collects the metabolic end products of the organism destined for excretion, and its composition is therefore more variable than that of serum or plasma. In particular, the composition and concentration of proteins and peptides in urine are strongly affected by nutrition, the day/night cycle, and the health status of the kidneys. It is thus important to try to control and document these parameters as accurately as possible.

Although protein concentration in urine is much lower than in serum (approximately 1000-fold) and filtration takes place in the kidneys, albumin is still the major protein. Proteome maps of human urinary proteins were recently constructed after LC-MS analysis of trypsin-digested unfractionated urine [123,124], by 2D electrophoresis after acetone precipitation [125], and after depletion of high-abundance proteins (albumin and IgG) by ultrafiltration and 2D electrophoresis [126]. An equivalent of urine, human hemofiltrate, was also analyzed by restricted access chromatography to select the peptidome followed by 2D HPLC and MS [17,18]. Urine has furthermore been analyzed by capillary electrophoresis coupled to electrospray ionization (CE-MS) [127,128]. Combined with new analysis software, this analytical method is presently under further investigation. Normalization of the data obtained to an internal standard that takes biological variation into account is critical for urine [128]. This has been common practice in clinical chemistry for a long time, and creatinine is widely used for this purpose. However, whether creatinine is also a suitable normalization standard for proteomics and peptidomics studies in urine is questionable. Normalization based on the total protein content or the area under the curve of the HPLC-UV trace may be preferable. It remains to be seen whether the urinary spectrum of proteins and peptides can be successfully used to detect human diseases short of those related to the kidney or general inflammation.

Urine samples should be collected under sterile conditions, cooled down, and treated with protease inhibitors. The next steps of sample pretreatment vary from one publication to the next. For example, proteins can be concentrated by precipitation with trifluoroacetic acid followed by centrifugation [129]. The resulting sample can be further applied to 1D or 2D electrophoresis, or subjected to solid-phase extraction and trypsin digestion followed by LC-MS analysis. Pieper et al. [126] compared urinary proteomes of healthy and renal cell carcinoma patients. Initially, cooled samples with added protease inhibitors were cleared by centrifugation and concentrated by membrane filtration. Samples were further desalted and fractionated by gel filtration on Superdex G-75. The resulting sample of more than 30 kDa proteins was passed through a depleting column specific for albumin, IgG, and α-1-acid glycoprotein. The final comparative analysis was done by 2D electrophoresis and mass spectrometry. Urine samples were recently used for comparative studies of normal and lung cancer patients [130]. The collected urine samples were first desalted by gel filtration (PD-10 columns) followed by lyophilization. The pellet was resuspended in phosphate buffer, extracted with methanol/ chloroform, and precipitated with trichloroacetic acid/acetone to remove organic acids and lipids. Finally, the sample was fractionated with HPLC and 1D or 2D electrophoresis followed by MALDI-MS and MS-MS. The image analysis of the gels demonstrated a quite impressive number of protein spots, but albumin and IgG were still quite abundant. Whereas easily obtainable body fluids such as blood, urine, saliva, or tears are samples of first choice for human proteomic studies, more specialized samples are frequently used to evaluate the condition of a given organ system. It is implied that a sample taken closer to the diseased organ will show changes in protein composition that are more closely related to the disease than those occurring in blood or urine.

Recently, the proteomics of bronchoalveolar lavage fluid (BALF) was reviewed [131,132]. Bronchoalveolar lavage samples the epithelial lining of the lung and is frequently analyzed in cases of severe respiratory diseases (e.g. chronic obstructive pulmonary disease, severe asthma, pulmonary fibrosis). BALF consists of a soluble part often used for biomarker analysis and cells derived from lung tissue or blood (alveolar macrophages, lymphocytes, neutrophils, and eosinophils). It is noteworthy that most proteins found in BALF correspond to abundant plasma proteins, indicating "plasma leakage" into the alveolar space owing to the lavage procedure. A map of the BALF proteome showed up to 1400 different proteins on a 2D gel [131], with some proteins at higher concentrations than in serum or plasma. These proteins are likely directly derived from the lung. Removing albumin as the most abundant protein in BALF by RAM chromatography allows one to process larger volumes and thus to detect lower abundance components (Figure 10).

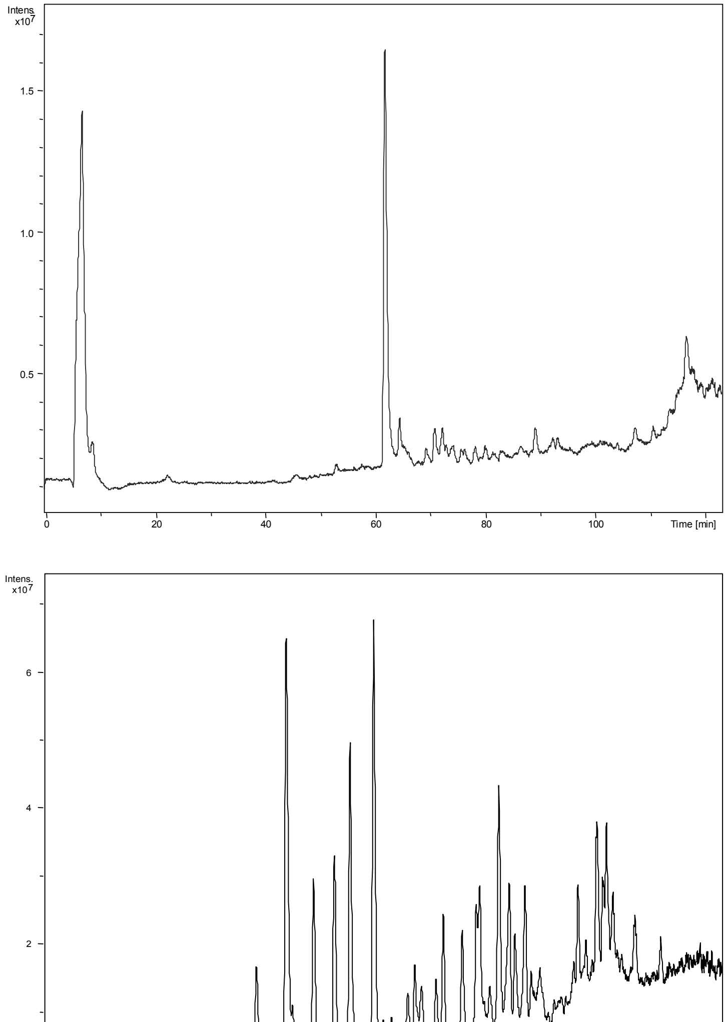

Figure 10. Sample preparation of bronchoalveolar lavage fluid (BALF) by restricted access material (RAM) chromatography. The upper panel shows the reversed-phase HPLC analysis of 10 µL BALF (major peak is albumin), and the lower panel shows the analysis of 1 mL BALF after sample preparation (Dr. Begona Barroso, unpublished data).

3. The Linkage to Separation Methods and Mass Spectrometry

The analysis of complex proteomes requires that dedicated and effective sample preparation be followed by high-resolution separation to reduce complexity to a level that can be handled by MS in terms of protein or peptide ionization, identification, and quantification. A wide range of separation methods has been applied to proteins and peptides, and it is beyond the scope of this chapter to review them all. The main purpose of the ensuing sections is to highlight how sample preparation of body fluids affects the downstream separation procedures and MS. To this end, two of the major separation methods will be highlighted, notably 2D electrophoresis for whole proteins and HPLC for the low-molecular-weight fraction of proteins and peptides or for protein digests.

3.1. 2D Gel Electrophoresis of Proteins in Body Fluids

The presence of a few high-abundance proteins in most body fluids poses a problem for 2D electrophoresis. 2D gels have a limited loading capacity of some hundred micrograms of total protein, which makes the detection of medium- to low-abundance proteins difficult if not impossible unless the sample is prefractionated. Considering that albumin in serum represents 40 mg/mL of the 80 mg/mL total protein concentration and that established tumor markers circulate at concentrations of a few ng/mL or even less, it is clear that applying, for example, 500 µg of total protein to a gel (corresponding to approximately 6 µL of serum) will yield only about 10 pg (approximately 0.12 fmol for a protein of 50 kDa) of a given tumor marker in the original sample. Even assuming a recovery of 100%, this is clearly below the detection level of any protein staining technique and definitely an amount that cannot be identified by in-gel digestion and MS. Without any enrichment or prefractionation, 2D electrophoresis will not be able to reveal proteins at concentrations much below the µg/mL range, an area that is largely occupied by well-known plasma or serum proteins that are likely not relevant as disease-specific biomarkers (Figure 3) [8,133].

One way to enhance the capacity of 2D electrophoresis to detect proteins at lower concentrations is to remove high-abundance proteins selectively. An affinity column developed to deplete the six most abundant proteins from serum resulted in a reduction in total protein by about a factor of 10 [50]. The effect of this removal step on 2D electrophoresis can be appreciated in Fig. 7, which shows that after depletion a range of proteins become visible. However, increasing the loading capacity by a factor of 10 is not sufficient to reach into the ng/mL concentration range for complex body fluids like serum.

Another strategy to cover more of the low-abundance proteins is to enrich some of them specifically. This is naturally at the expense of losing the overview over the proteome and thus potentially missing relevant markers. Enrichment depends critically on the selection of appropriate ligands in combination with stationary phases of low nonspecific protein binding. Both requirements are not easily met but it is often the elimination of nonspecific binding that poses the greatest problems. As an example of the effect of nonspecific binding, Figure 9 shows the binding kinetics of a metalloprotease to controlled porosity glass beads (with or without silanization to render the surface more hydrophilic) and to the silanized beads containing an immobilized metalloprotease inhibitor.

Prefractionation of the sample is another option to reduce complexity prior to 2D electrophoresis, at the expense, however, of having to run multiple gels for a single sample. This is often not a viable option owing to the work-intensive nature of 2D gels. An alternative is to select narrow pH ranges to visualize only that part of the proteome that does not coincide with the high-abundance proteins. Unfortunately, most serum proteins have similar isoelectric points between pH 5 and 6 (*see* Figure 7), making fractionation difficult. More recently, prefractionation by preparative in-solution isoelectric focusing has emerged as a first step in body fluid analysis prior to 2D electrophoresis and also chromatography [134]. This approach has the advantage that proteins are fractionated based on a clear-cut physicochemical parameter, their isoelectric point, but multiple fractions still need to be analyzed, meaning that an efficient, preferably automated method should be used downstream [135]. The possibilities of prefractionating [136] proteins in depleted serum/plasma on a newly developed “Macroporous Reversed-Phase C18 High-Recovery Protein Fractionation HPLC column (mRP)” prior digestion and analysis (Agilent, www.agilent.com/chem, Agilent technologies publication 5989-250EN) will be described in chapter VI.

3.2. LC-MS of Proteins and Peptides in Body Fluids

Based on the above discussion, it is obvious that 2D electrophoresis is not the method of choice for analyzing large series of clinical samples in quest of new disease-specific markers. Consequently, other methods have been sought to reduce the workload of 2D gels, methods that make use of automated equipment. In addition to the already described direct combination of sample preparation on protein chips or magnetic beads with MS, there is increasing interest in the combination of online sample preparation with LC (LC-MS). In the following two examples, a focus on the low-molecular weight part of the body fluid proteome (also referred to as the *peptidome*) and the shotgun proteomics approach requiring proteolytic digestion will be highlighted. Indeed, there are many possibilities of integrating sample preparation with the ensuing separation, but these two approaches may serve as examples.

3.2.1. Peptidomics

Dividing the proteome of body fluids into a high- and a low-molecular weight fraction (the so-called peptidome [14,137-140]) is an approach to detect lower abundance small proteins and peptides. Although it is restricted to a certain molecular weight range, the peptidome contains extensive information about processes in the organism that may be relevant for diagnosis and follow-up of therapy. This is a deliberate choice of sample pretreatment, eliminating most of the high-abundance serum or plasma proteins. An additional advantage of focusing on the molecular weight region below 15 to 20 kDa is that these molecules are more easily separated and recovered by reversed-phase HPLC (RP-HPLC), which is the preferred method for coupling to MS.

There are a number of techniques that allow elimination of the fraction of the proteome above approximately 20 kDa, such as ultrafiltration, precipitation with acids or organic solvents or the combination of ultrafiltration with adsorption chromatography (e.g., RAM chromatography) [15,16]. Full integration of all analytical steps in an automated system is often desirable for biofluid analysis in a clinical or biomedical environment to increase throughput, reduce the need for skilled personnel, and increase reproducibility. Furthermore, documentation is often facilitated by using an integrated, fully automated analytical system. Combining the "unit operations" of sample pretreatment, separation, and detection in the case of peptidomics was achieved in a system described by Wagner et al. [17] and further developed by Machtejevas et al. [18]. Figure 11 shows the instrumental setup combining selection of the peptidome from human hemofiltrate by RAM chromatography followed by prefractionation on a strong cation exchanger and finally separation by RP-HPLC.

Although this setup was not coupled online to a mass spectrometer, analysis of selected fractions after RP-HPLC by MALDI-TOF-MS showed that complexity of the original hemofiltrate had been reduced to such a level that most of the fractions contained one major peptide or small protein (Figures 12 and 13).

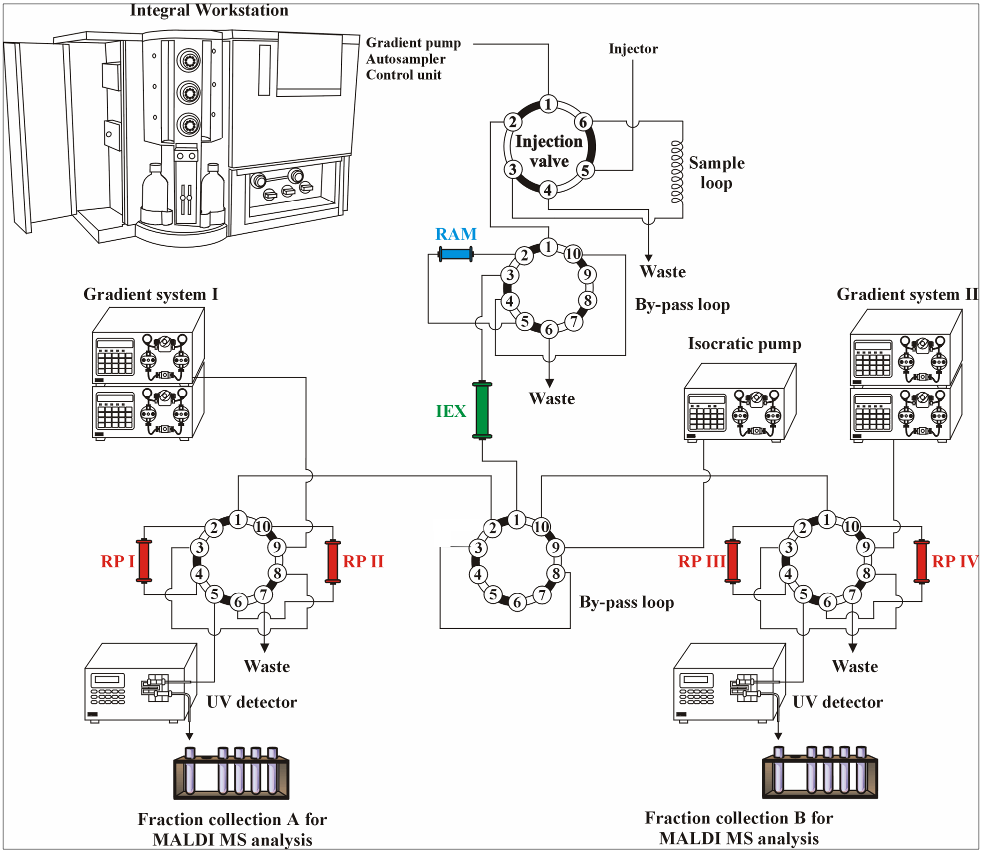

Figure 11. Schematic diagram of an online comprehensive 2D HPLC system including an integrated sample preparation step by restricted access chromatography. Strong cation-exchange HPLC is used in the first dimension (IEX) followed by rapid reversed-phase (RP) HPLC on four nonporous particle-packed columns working in parallel. Reproduced with permission from Machtejevas E, John H,Wagner K, et al. Automated multi-dimensional liquid chromatography: sample preparation and identification of peptides from human blood filtrate. *J Chromatogr B* 2004;803:121–130.

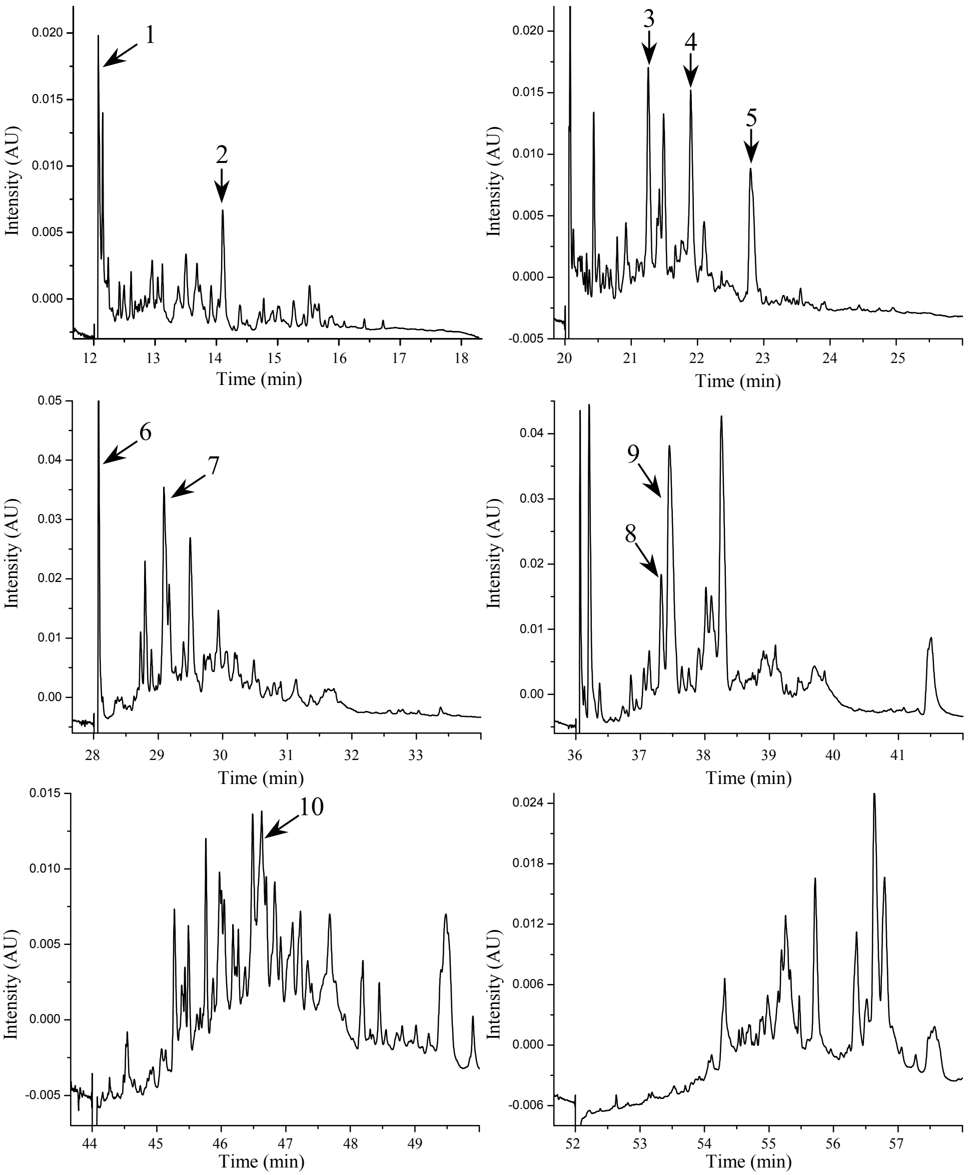

Figure 12. Selected reversed-phase chromatograms from human hemofiltrate processed through the integrated, multidimensional chromatography system shown in Fig. 11. Numbered and marked peak fractions 1 to 10 were selected for MS analysis. Reproduced with permission from Machtejevas E, John H, Wagner K, et al. Automated multi-dimensional liquid chromatography: sample preparation and identification of peptides from human blood filtrate. *J Chromatogr B* 2004;803:121–130.

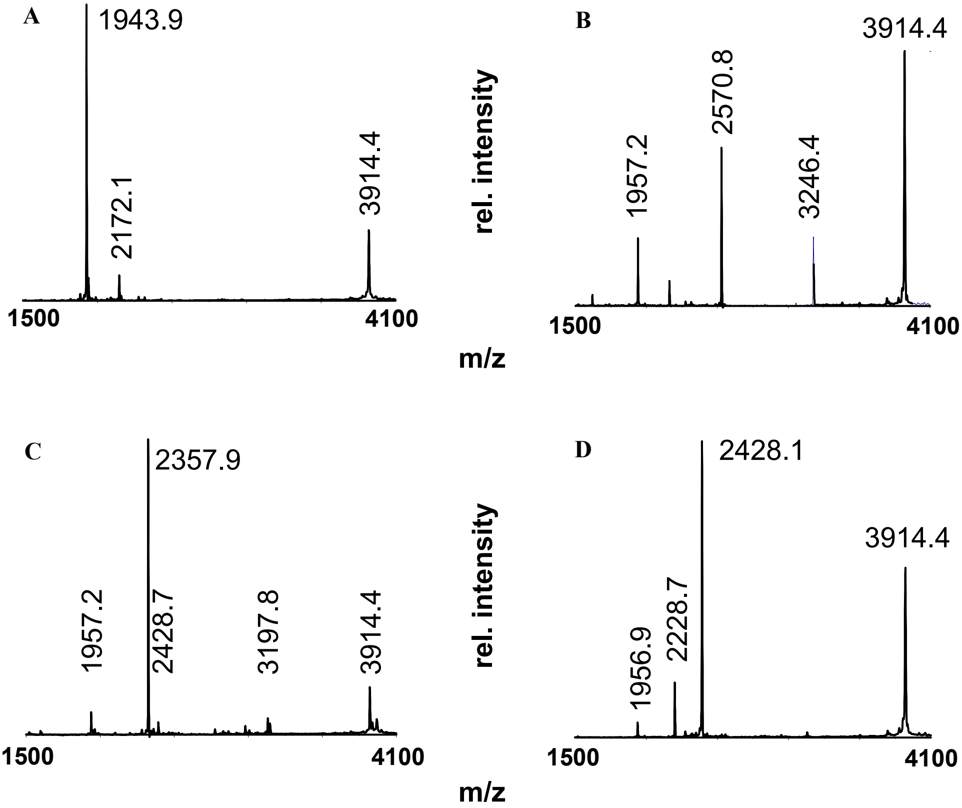

Figure13. MALDI-TOF mass spectra of selected peaks from the reversed-phase HPLC fractions shown in Fig. 12. Spectra (A), (B), (C), and (D) correspond to peak fractions 4, 7, 8, and 9, respectively. Measurement was performed in the linear mode with positive ionization using a matrix of α-cyano-4-hydroxycinnamic acid mixed with fucose. The peak at *m*/*z* =3914.4 is an internal standard. Reproduced with permission from Machtejevas E, John H, Wagner K, et al. Automated multi-dimensional liquid chromatography: sample preparation and identification of peptides from human blood filtrate. *J Chromatogr B* 2004;803:121–130.

3.2.2. Shotgun Proteomics

Applying HPLC separation to the high-molecular-weight region of the proteome requires prior proteolytic digestion, since most complex proteins are not stable under the denaturing conditions of RP-HPLC and are thus not quantitatively recovered. In addition, it is not possible to identify proteins based on their molecular weight only, and fragmentation of large proteins is impossible in most commonly employed mass spectrometers.

This has led to the shotgun proteomics approach, whereby the entire proteome is first digested with trypsin followed by one- or two-dimensional chromatographic separations of the peptides [20-23]. Most of the observations about high-abundance proteins masking those of lower abundance that were made with respect to 2D electrophoresis also apply to the shotgun method. This is partially because of the limited loading capacity of the chromatographic columns, especially when nanoLC-MS (loadability in the ng–µg range) is the final analytical step. Removing high-abundance proteins by affinity chromatography prior to digestion is one way of increasing the loading of medium- and low-abundance proteins. The effect of depletion of albumin on the detection of a tryptic peptide derived from α_1-antitrypsin, whose concentration in serum is about 20-fold lower, shows the clear improvement even for a protein that is still considered to belong to the high-abundance class (Fig. 6). Depleting high-abundance proteins and notably albumin can also have drawbacks, however, as albumin is known to bind and carry numerous compounds including small proteins and peptides that may be co-depleted. Zhou et al. [37] showed, for example, that albumin binds some 210 proteins and peptides, which could be identified in the albumin-containing fraction after depletion. In some of our own studies, we observed that a small protein added as an internal standard (horse heart Cytochrome C) at pmol/µL concentrations was co-depleted to leave about 25% [61] of the original amount when the six most abundant proteins were removed by affinity chromatography. It is important to note that although depletion removes some other proteins and peptides, this seems to be rather reproducible, at least when judged from the results with horse heart Cytochrome C (the relative standard deviation of peak areas without normalization is 10–30%) and our recent data with a large series of serum samples from prostate cancer patients [28].

The enormously complex mixture generated by shotgun proteomics usually requires multiple separation dimensions prior to MS. This has inspired researchers to develop novel ways of performing “coupled column” HPLC multidimensional protein identification technology (MudPIT) [21] or the integrated setup outlined in Fig. 11. Because analysis of individual fractions from earlier dimensions in such a multidimensional HPLC approach requires 12 to 24 h per sample, more efficient and rapid separation methods are under development. One of them is based on reducing the particle diameter of the chromatographic materials to 1-2 µm, which results in rather high backpressures that can only be delivered by special HPLC equipment [141-144]. Other approaches are based on monolithic materials that support very high linear flow rates at pressures amenable to common HPLC equipment [24,145-149] or the use of pellicular stationary phases also at elevated temperatures [150-153]. These developments show that in addition to the major advances in MS, which have made proteomics as we know it today possible, there is also considerable activity in the fields of sample preparation and separation methodology. It is only through integration of these unit operations into an analytical strategy that the challenges of body fluid analysis can be tackled and possibly the entire dynamic range covered. Much remains to be done.

4. Conclusions

Proteomics of body fluids is a rapidly expanding field driven by the search for better biomarkers for disease diagnosis, follow-up on therapy and evaluation of the response of patients to newly developed pharmaceuticals. The analysis of body fluids has a long tradition in clinical chemistry and serves to support decision making by clinicians in many respects. Because of recent methodological developments in separation science, MS, and bioinformatics, there has been a surge of efforts to apply them to biomarker discovery, often focusing on biomarker patterns rather than individual molecules. Sample preparation, the indispensable and very critical first step in an analytical method, has attracted less attention, and its relevance is often underestimated. As outlined in this chapter, the approach to sample preparation is an important decision of strategic relevance for the ensuing analyses. It is therefore pivotal to weigh the pros and cons of each approach in light of the final goal. We hope that the overview given in this chapter will guide the reader in this complex methodological field.

Acknowledgments

We thank Dr. Theo Reijmers (Leiden University, Division of Analytical Biosciences) for calculations related to Fig. 2 and Dr. Begona Barroso (Astellas BV, Leiderdorp) for contributing the data for Fig. 10. We are also very grateful to Dr. Peter Schulz-Knappe (BioVision, Hannover, Germany) for communicating unpublished information for use in this chapter.

References

[1]. Pusch W, Flocco MT, Leung SM, Thiele H, Kostrzewa M. Mass spectrometrybased clinical proteomics. Pharmacogenomics 2003;4:463–476.

[2]. Schulte I, Tammen H, Schulz-Knappe P, Selle H. Peptides in body fluids and tissues as markers of disease. Exp Rev Mol Diagn. 2005;5:145–157.

[3]. Tammen H, Schulte I, Hess R, Menzel C, Kellmann M, Mohring T, Schulz-Knappe P. Peptidomic Analysis of Human Blood Specimens: Comparison Between Plasma Specimens and Serum by Differential Peptide Display. Proteomics 2005;5:3414-3422.

[4]. Drake RR, Cazares LH, Corica A, Malik G, Schwegler EE, Libby AE, Wright GL, Semmes OJ, Adam BL. Qualit control, preparation and protein stability issues for blood serum and plasma used in biomarker discovery and proteomic profiling assays. Bioprocessing Journal 2004;43-49.

[5]. Hsieh SY, Chen RK, Pan YH, Lee HL. Systematical Evaluation of the Effects of Sample Collection Procedures on Low-Molecular-Weight Serum/Plasma Proteome Profiling. Proteomics 2006;6(10):3189-3198.

[6]. Timms JF, Arslan-Low E, Gentry-Maharaj A, Luo Z, T'jampens D, Podust VN, Ford J, Fung ET, Gammerman A, Jacobs I, Menon U. Preanalytic Influence of Sample Handling on SELDI-TOF Serum Protein Profiles. Clin. Chem. 2007;53(4):645-656.

[7]. Villanueva J, Philip J, Chaparro CA, Li Y, Toledo-Crow R, DeNoyer L, Fleisher M, Robbins RJ, Tempst P. Correcting Common Errors in Identifying Cancer-Specific Serum Peptide Signatures. J Proteome. Res. 2005;4:1060-1072.

[8]. Anderson NL, Anderson NG. The Human Plasma Proteome: history, character, and diagnostic prospects. Mol Cell Proteomics 2002;1:845–867.

[9]. Johnson KL, Mason CJ, Muddiman DC, Eckel JE. Analysis of the low molecular weight fraction of serum by LC-dual ESI-FT-ICR mass spectrometry: precision of retention time, mass, and ion abundance. Anal Chem. 2004;76:5097–5103.

[10]. Tirumalai RS, Chan KC, Prieto DA, Issaq HJ, Conrads TP, Veenstra TD. Characterization of the low molecular weight human serum proteome. Mol Cell Proteomics 2003;2:1096–1103.

[11]. Georgiou HM, Rice GE, Baker MS. Proteomic analysis of human plasma: failureof centrifugal ultrafiltration to remove albumin and other high molecular weightproteins. Proteomics 2001;1:1503–1506.

[12]. Schulz Knappe P, Schrader M, Standker L., et al., Peptide bank generated by large-scale preparation of circulating human peptides. J Chromatogr A. 1997;776:125–132.

[13]. Schulz Knappe P, Raida M, Meyer M, Quellhorst EA, Forssmann WG. Systematic isolation of circulating human peptides: the concept of peptide trapping. Eur J Med Res. 1996;1:223–236.

[14]. Raida M, Schulz-Knappe P, Heine G, Forssmann WG. Liquid chromatography and electrospray mass spectrometric mapping of peptides from human plasma filtrate. J Am Soc Mass Spectrom. 1999;10:45–54.

[15]. Racaityte K, Lutz ESM, Unger KK, Lubda D, Boos KS. Analysis of neuropeptide Y and its metabolites by high-performance liquid chromatographyelectrospray ionization mass spectrometry and integrated sample clean-up with a novel restricted-access sulphonic acid cation exchanger. J Chromatogr A. 2000; 890:135–144.

[16]. Boos KS, Grimm CH. High-performance liquid chromatography integrated solidphasextraction in bioanalysis using restricted access precolumn packings. TrAC Trends Anal Chem. 1999;18:175–180.

[17]. Wagner K, Miliotis T, Marko-Varga G, Bischoff R, Unger KK. An automated online multidimensional HPLC system for protein and peptide mapping with integrated sample preparation. Anal Chem. 2002;74:809–820.

[18]. Machtejevas E, John H, Wagner K., et al., Automated multi-dimensional liquid chromatography: sample preparation and identification of peptides from human blood filtrate. J Chromatogr B. 2004;803:121–130.

[19]. Rieux L, Bischoff R, Verpoorte E, Niederlander H A G. Restricted-Access Material-Based High-Molecular-Weight Protein Depletion Coupled on-Line With Nano-Liquid Chromatography-Mass Spectrometry for Proteomics Applications. Journal of Chromatography A 2007;1149:169-177.

[20]. Wolters DA, Washburn MP, Yates JR, III. An automated multidimensional protein identification technology for shotgun proteomics. Anal Chem. 2001;73:5683–5690.

[21]. Washburn MP, Wolters D,Yates JR. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol. 2001;19:242–247.

[22]. McDonald WH, Yates JR, III. Shotgun proteomics and biomarker discovery. Dis Markers 2002;18:99–105.

[23]. Shen Z, Want EJ, Chen W, Keating W, Nussbaumer W, Moore R, Gentle TM, Siuzdak G. Sepsis Plasma Protein Profiling With Immunodepletion, Three-Dimensional Liquid Chromatography Tandem Mass Spectrometry, and Spectrum Counting. J. Proteome. Res. 2006;5:3154-3160.

[24]. Barroso B, Lubda D, Bischoff R. Applications of monolithic silica capillary columns in proteomics. J Proteome Res. 2003;2:633–642.

[25]. Strittmatter EF, Ferguson PL, Tang K, Smith RD. Proteome analyses using accurate mass and elution time peptide tags with capillary LC time-of-flight mass spectrometry. J Am Soc Mass Spectrom 2003;14:980–991.

[26]. Shen Y, Tolic N, Masselon C., et al., Ultrasensitive proteomics using highefficiency on-line micro-SPE-nanoLC-nanoESI MS and MS/MS. Anal Chem. 2004;76:144–154.

[27]. Adkins JN, Varnum SM, Auberry KJ, et al. Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. Mol Cell Proteomics 2002;1:947–955.

[28]. Horvatovich PL, Govorukhina NI, Reijmers TH, van der Zee AGJ, Suits F, Bischoff R. Chip-LC-MS for label-free profiling of human serum. Electrophoresis, accepted.

[29]. Li C, Lee KH. Affinity depletion of albumin from human cerebrospinal fluid using Cibacron-blue-3G-A-derivatized photopatterned copolymer in a microfluidic device. Anal Biochem 2004;333:381–388.

[30]. Govorukhina NI, Keizer-Gunnink, van der Zee AGJ, de Jong S, de Bruijn HWA, Bischoff R. Sample preparation of human serum for the analysis of tumor markers: comparison of different approaches for albumin and [gamma]-globulin depletion. J Chromatogr A. 2003;1009:171–178.

[31]. Hinerfeld D, Innamorati D, Pirro J, Tam SW. Serum/plasma depletion with chicken immunoglobulin Y antibodies for proteomic analysis from multiple mammalian species. J Biomol Tech. 2004;15:184–190.

[32]. Greenough C, Jenkins RE, Kitteringham NR, Pirmohamed M, Park BK, Pennington SR. A method for the rapid depletion of albumin and immunoglobulin from human plasma. Proteomics 2004;4:3107–3111.

[33]. Fountoulakis M, Juranville JF, Jiang L., et al., Depletion of the high-abundance plasma proteins. Amino Acids 2004;27:249–259.

[34]. Chromy BA, Gonzales AD, Perkins J., et al., Proteomic analysis of human serum by two-dimensional differential gel electrophoresis after depletion of high-abundant proteins. J Proteome Res. 2004;3:1120–1127.

[35]. Bjorhall K, Miliotis T, Davidsson P. Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. Proteomics 2005;5:307–317.

[36]. Bischoff R, Luider TM. Methodological advances in the discovery of protein and peptide disease markers. J Chromatogr B. 2004;803:27–40.

[37]. Zhou M, Lucas DA, Chan KC, et al. An investigation into the human serum “interactome”. Electrophoresis 2004;25:1289–1298.

[38]. Solassol J, Marin P, Demettre E., et al., Proteomic detection of prostate-specific antigen using a serum fractionation procedure: potential implication for new lowabundance cancer biomarkers detection. Anal Biochem. 2005;338:26–31.

[39]. Troyer DA, Mubiru J, Leach RJ, Naylor SL. Promise and challenge: markers of prostate cancer detection, diagnosis and prognosis. Dis Markers 2004;20:117–128.

[40]. Baldus SE, Engelmann K, Hanisch FG. MUC1 and the MUCs: a family of human mucins with impact in cancer biology. Crit Rev Clin Lab Sci. 2004;41:189–231.

[41]. Yang Z, Hancock WS. Approach to the comprehensive analysis of glycoproteins isolated from human serum using a multi-lectin affinity column. J Chromatogr A. 2004;1053:79–88.

[42]. Xiong L, Regnier FE. Use of a lectin affinity selector in the search for unusual glycosylation in proteomics. J Chromatogr B. 2002;782:405–418.

[43]. Xiong L, Andrews D, Regnier F. Comparative proteomics of glycoproteins based on lectin selection and isotope coding. J Proteome Res. 2003;2:618–625.

[44]. Schulenberg B, Beechem JM, Patton WF. Mapping glycosylation changes related to cancer using the Multiplexed Proteomics technology: a protein differential display approach. J Chromatogr B. Analyt Technol Biomed Life Sci. 2003;793:127–139.

[45]. Geng M, Zhang X, Bina M, Regnier F. Proteomics of glycoproteins based on affinity selection of glycopeptides from tryptic digests. J Chromatogr B. Biomed Sci Appl. 2001;752:293–306.

[46]. Drake R, Cazares L, Corica A., et al., Quality control, preparation and protein stability issues for blood serum and plasma used in biomarker discovery and proteomic profiling assays. Bioprocessing J. 2004 July/August:43–49.

[47]. Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics 2003;4:24-36.

[48]. Windig W, Phalp JM, Payne AW. A Noise and background reduction method for component detection in liquid chromatography/mass spectrometry. Anal Chem. 1996;68:3602–3606.

[49]. Zolg JW, Langen H. How industry is approaching the search for new diagnostic markers and biomarkers. Mol Cell Proteomics 2004;3:345–354.

[50]. Bjorhall K, Miliotis T, Davidsson P. Comparison of different depletion strategies for improved resolution in proteomic analysis of human serum samples. Proteomics 2005;5:307–317.

[51]. Wang YY, Chan DW,Wang YY, Cheng P. A simple affinity spin tube filter method for removing high-abundant common proteins or enriching low-abundant biomarkers for serum proteomic analysis. Proteomics 2003;3:243–248.

[52]. Gianazza E, Arnaud P. Chromatography of plasma proteins on immobilized Cibacron Blue F3-GA. Mechanism of the molecular interaction. Biochem J. 1982;203:637–641.

[53]. Sato AK, Sexton DJ, Morganelli LA., et al., Development of mammalian serum albumin affinity purification media by peptide phage display. Biotechnol Prog. 2002;18:182–192.

[54]. Baussant T, Bougueleret L, Johnson A., et al., Effective depletion of albumin using a new peptide-based affinity medium. Proteomics 2005;5:973–977.

[55]. Bjorck L, Kronvall G. Purification and some properties of streptococcal protein G, a novel IgG-binding reagent. J Immunol. 1984;133:969–974.

[56]. Akerstrom B, Bjorck L. A physicochemical study of protein G, a molecule with unique immunoglobulin G-binding properties. J Biol Chem. 1986;261:10240–10247.

[57]. Akerstrom B, Brodin T, Reis K, Bjorck L. Protein G: a powerful tool for binding and detection of monoclonal and polyclonal antibodies. J Immunol. 1985;135:2589–2592.

[58]. Guss B, Eliasson M, Olsson A., et al., Structure of the IgG-binding regions of streptococcal protein G. EMBO J. 1986;5:1567–1575.

[59]. Fahnestock SR. Cloned streptococcal protein G genes. Trends Biotechnol. 1987;5:79–83.

[60]. Roque AC, Taipa MA, Lowe CR. An artificial protein L for the purification of immunoglobulins and Fab fragments by affinity chromatography. J Chromatogr A. 2005;1064:157–167.

[61]. Govorukhina NI, Reijmers TH, Nyangoma SO, van der Zee AGJ, Jansen RC, Bischoff R. Analysis of Human Serum by Liquid Chromatography-Mass Spectrometry: Improved Sample Preparation and Data Analysis. Journal of Chromatography A 2006;1120:142-150.

[62]. Dekker LJ, Bosman J, Burgers PC, van Rijswijk A, Freije R, Luider T, Bischoff R. Depletion of high-abundance proteins from serum by immunoaffinity chromatography: A MALDI-FT-MS study Journal of Chromatography B 2007;847:65-69.

[63]. Horvath ZS, Corthals GL, Wrigley CW, Margolis J. Multifunctional apparatus for electrokinetic processing of proteins. Electrophoresis 1994;15:968–971.

[64]. Pang L, Fryksdale BG, Chow N, Wong DL, Gaertner AL, Miller BS. Impact of prefractionation using Gradiflow on two-dimensional gel electrophoresis and protein identification by matrix assisted laser desorption/ionization-time of flightmass spectrometry. Electrophoresis 2003;24:3484–3492.

[65]. Rothemund DL, Locke VL, Liew A, Thomas TM, Rylatt DB, Wasinger V. Depletion of the highly abundant protein albumin from human plasma using the Gradiflow. Proteomics 2003;3:279–287.

[66]. Weber G, Bocek P. Recent developments in preparative free flow isoelectric focusing. Electrophoresis 1998;19:1649–1653.

[67]. Mehta AI, Mehta AI, Ross S., et al., Biomarker amplification by serum carrier protein binding. Dis Markers 2004;19:1–10.

[68]. Lee HJ, Lee EY, Kwon MS, Paik YK. Biomarker discovery from the plasma proteome using multidimensional fractionation proteomics.Curr Opin Chem Biol. 2006;10(1):42-49.

[69]. Steel LF, Mattu TS, Mehta A., et al., A proteomic approach for the discovery of early detection markers of hepatocellular carcinoma. Dis Markers 2001;17:179–189.

[70]. Block TM, Comunale MA, Lowman M., et al., Use of targeted glycoproteomics toidentify serum glycoproteins that correlate with liver cancer in woodchucks andhumans. Proc Natl Acad Sci USA. 2005;102:779–784.

[71]. Saghatelian A, Jessani N, Joseph A, Humphrey M, Cravatt BF. Activity-based probes for the proteomic profiling of metalloproteases. Proc Natl Acad Sci U SA. 2004;101(27):10,000–10,005.

[72]. Ovaa H, Kessler BM, Rolen U, Galardy PJ, Ploegh HL, Masucci MG. Activitybased ubiquitin-specific protease (USP) profiling of virus-infected and malignant human cells. Proc Natl Acad Sci U SA. 2004;101(8):2253–2258.

[73]. Liu YS, Patricelli MP, Cravatt BF. Activity-based protein profiling: the serine hydrolases. Proc Natl Acad Sci U S A 1999;96:14694–14699.

[74]. Kumar S, Zhou B, Liang F, Wang WQ, Huang Z, Zhang ZY. Activity-based probes for protein tyrosine phosphatases. Proc Natl Acad Sci USA. 2004;101:7943–7948.

[75]. Jessani N, Cravatt BF. The development and application of methods for activitybased protein profiling. Curr Opin Chem Biol. 2004;8:54–59.

[76]. Freije JR, Bischoff R. Activity-based enrichment of matrix metalloproteinases using reversible inhibitors as affinity ligands. J Chromatogr A. 2003;1009:155–169.

[77]. Falgueyret JP, Black WC, Cromlish W., et al., An activity-based probe for the determination of cysteine cathepsin protease activities in whole cells. Anal Biochem. 2004;335:218–227.

[78]. Xiong L, Regnier FE. Use of a lectin affinity selector in the search for unusual glycosylation in proteomics. J Chromatogr B. 2002;782:405–418.

[79]. Grunewald S, Matthijs G, Jaeken J. Congenital disorders of glycosylation: a review. Pediatr Res. 2002;52:618–624.

[80]. Rudiger H, Gabius HJ. Plant lectins: Occurrence, biochemistry, functions and applications. Glycoconjugate J. 2001;18:589–613.

[81]. Wiener MC,Van Hoek AN, Wiener MC. A lectin screening method for membrane glycoproteins: application to the human CHIP28 water channel (AQP-1). Anal Biochem. 1996;241:267–268.

[82]. Nazif T, Bogyo M. Global analysis of proteasomal substrate specificity using positional-scanning libraries of covalent inhibitors. Proc Natl Acad Sci U SA. 2001;98:2967–2972.

[83]. Jeffery DA, Bogyo M. Chemical proteomics and its application to drug discovery. Curr Opin Biotechnol. 2003;14:87–95.

[84]. Greenbaum D, Baruch A, Hayrapetian L., et al., Chemical approaches for functionally probing the proteome. Mol Cell Proteomics 2002;1:60–68.

[85]. Adam GC, Cravatt BF, Sorensen EJ. Profiling the specific reactivity of the proteome with non-directed activity-based probes. Chem Biol. 2001;8:81–95.

[86]. Bogyo M, Verhelst S, Bellingard-Dubouchaud V, Toba S, Greenbaum D. Selective targeting of lysosomal cysteine proteases with radiolabeled electrophilic substrate analogs. Chem Biol. 2000;7:27–38.

[87]. Borodovsky A, Ovaa H, Kolli N., et al., Chemistry-based functional proteomics reveals novel members of the deubiquitinating enzyme family. Chem Biol. 2002;9:1149–1159.

[88]. Faleiro L, Kobayashi R, Fearnhead H, Lazebnik Y, Faleiro L. Multiple species of CPP32 and Mch2 are the major active caspases present in apoptotic cells. EMBO J. 1997;16:2271–2281.

[89]. Kidd D, Liu Y, Cravatt BF. Profiling serine hydrolase activities in complex proteomes. Biochemistry 2001;40:4005–4015.

[90]. Freije JR, Klein T, Ooms JA, Franke JP, Bischoff R. Activity-based matrix metallo-protease enrichment using automated, inhibitor affinity extractions J.Proteome.Res. 2006;5:1186-1194.

[91]. Freije JR, Bischoff R. The use of affinity sorbents in targeted proteomics Drug Discovery Today: Technologies. 2006;3:5-11.

[92]. Coon JJ, Ueberheide B, Syka JEP, Dryhurst DD, Ausio J, Shabanowitz J, Hunt DF. Protein Identification Using Sequential Ion/Ion Reactions and Tandem Mass Spectrometry. PNAS. 2005;102:9463-9468.

[93]. Jessani N, Cravatt BF. The development and application of methods for activitybased protein profiling. Curr Opin Chem Biol. 2004;8:54–59.

[94]. Jessani N, Liu Y, Humphrey M, Cravatt BF. Enzyme activity profiles of the secreted and membrane proteome that depict cancer cell invasiveness. Proc Natl Acad Sci USA. 2002;99:10,335–10,340.

[95]. Speers AE, Adam GC, Cravatt BF. Activity-based protein profiling in vivo using a copper(i)-catalyzed azide-alkyne [3 + 2] cycloaddition. J Am Chem Soc. 2003;125:4686–4687.

[96]. Speers AE, Cravatt BF. Profiling enzyme activities in vivo using click chemistry methods. Chem Biol. 2004;11:535–546.

[97]. Petricoin EF, Zoon KC, Kohn EC, Barrett JC, Liotta LA. Clinical proteomics: translating benchside promise into bedside reality. Nat Rev Drug Discov. 2002;1:683–695.

[98]. Sanchez JC, Guillaume E, Lescuyer P, et al. Cystatin C as a potential cerebrospinal fluid marker for the diagnosis of Creutzfeldt-Jakob disease. Proteomics 2004;4:2229–2233.

[99]. Patrie SM, Charlebois JP, Whipple D., et al., Construction of a hybrid quadrupole/ fourier transform ion cyclotron resonance mass spectrometer for versatile MS/MS above 10 kDa. J Am Soc Mass Spectrom. 2004;15:1099–1108.

[100]. Horn DM, Zubarev RA, McLafferty FW. Automated de novo sequencing of proteins by tandem high- resolution mass spectrometry. Proc Natl Acad Sci USA. 2000;97:10313–10317.

[101]. Ge Y, ElNaggar M, Sze SK., et al., Top down characterization of secreted proteins from Mycobacterium tuberculosis by electron capture dissociation mass spectrometry. J Am Soc Mass Spectrom. 2003;14:253–261.

[102]. Horn DM, Ge Y, McLafferty F. Activated Ion Electron Capture Dissociation for Mass Spectral Sequencing of Larger (42 KDa) Proteins. Analytical Chemistry 2000;72:4778-4784.

[103]. Patrie SM, Charlebois JP, Whipple D, Kelleher NL, Hendrickson CL, Quinn JP, Marshall AG, Mukhopadhyay B. Construction of a Hybrid Quadrupole/Fourier Transform Ion Cyclotron Resonance Mass Spectrometer for Versatile MS/MS Above 10 KDa. Journal of the American Society for Mass Spectrometry 2004;15:1099-1108.

[104]. Sleno L, Volmer DA. Ion Activation Methods for Tandem Mass Spectrometry. J Mass Spectrom. 2004;39:1091-1112.

[105]. Vlahou A, Gregory B, Wright J., et al., A novel approach toward development of a rapid blood test for breast cancer. Clin Breast Cancer 2003;4:203–209.

[106]. Petricoin EF, Ardekani AM, Hitt BA., et al., Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–577.

[107]. Diamandis EP. Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations. Mol Cell Proteomics 2004;3:367–378.

[108]. Diamandis EP. OvaCheck: doubts voiced soon after publication. Nature 2004;430:611.

[109]. Ekblad L, Baldetorp B, Ferno M, Olsson H, Bratt C. In-Source Decay Causes Artifacts in SELDI- TOF MS Spectra. J. Proteome Res. 2007;6(4):1609-1614.

[110]. Zhang L, Yu W, He T., et al., Contribution of human alpha-defensin 1, 2, and 3 to the anti-HIV-1 activity of CD8 antiviral factor. Science 2002;298:995–1000.

[111]. Walker CM, Moody DJ, Stites DP, Levy JA CD8+ lymphocytes can control HIV infection in vitro by suppressing virus replication. Science 1986;234:1563–1566.

[112]. Peter J, Unverzagt C, Krogh TN, Vorm O, Hoesel W. Identification of precursor forms of free prostate-specific antigen in serum of prostate cancer patients by immunosorption and mass spectrometry. Cancer Res. 2001;61:957–962.

[113]. Lin YS, Weng MF, Chen YC, Tsai PJ. Affinity capture using vancomycin-bound magnetic nanoparticles for the MALDI-MS analysis of bacteria. Anal Chem. 2005;77:1753–1760.

[114]. Bundy JL, Fenselau C. Lectin and carbohydrate affinity capture surfaces for mass spectrometric analysis of microorganisms. Anal Chem 2001;73:751–757.

[115]. Bundy J, Fenselau C. Lectin-based affinity capture for MALDI-MS analysis of bacteria. Anal Chem 1999;71:1460–1463.

[116]. Villanueva J, Philip J, Entenberg D., et al., Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. Anal Chem. 2004;76:1560–1570.

[117]. Hong CY, Chen YC. Selective enrichment of ochratoxin A using human serum albumin bound magnetic beads as the concentrating probes for capillary electrophoresis/electrospray ionization-mass spectrometric analysis. J Chromatogr A. 2007 Aug 3;1159(1-2):250-255.

{118}. Chen, W.Y.; Chen Y.C. Acceleration of microwave-assisted enzymatic digestion reactions by magnetite beads. Anal Chem. 2007;79(6):2394-2401.

[119]. Dubois M, Becher F, Herbet A. Ezan E. Immuno-mass spectrometry assay of EPI-HNE4, a recombinant protein inhibitor of human elastase. Rapid Commun Mass Spectrom. 2007; 21(3):352-358.

[120]. West-Norager M, Kelstrup CD, Schou C, Hogdall EV, Hogdall CK, Heegaard NH. Unravelling in vitro variables of major importance for the outcome of mass spectrometry-based serum proteomics.J Chromatogr B Analyt Technol Biomed Life Sci. 2007,847(1):30-37.

[121]. Celis JE, Rasmussen HH, Vorum H., et al., Bladder squamous cell carcinomas express psoriasin and externalize it to the urine. J Urol. 1996;155:2105–2112.

[122]. Marshall T, Williams KM, Marshall T. Clinical analysis of human urinary proteins using high resolution electrophoresis methods. Electrophoresis 1998;19:1752–1770.

[123]. Spahr CS, Davis MT, McGinley MD., et al., Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. I. Profiling an unfractionated tryptic digest. Proteomics 2001;1:93–107.

[124]. Davis MT, Spahr CS, McGinley MD., et al., Towards defining the urinary proteome using liquid chromatography-tandem mass spectrometry. II. Limitations of complex mixture analyses. Proteomics 2001;1:108–117.

[125]. Thongboonkerd V, McLeish KR, Arthur JM, Klein JB. Proteomic analysis of normal human urinary proteins isolated by acetone precipitation or ultracentrifugation. Kidney Int. 2002;62:1461–1469.

[126]. Pieper R, Gatlin CL, McGrath AM., et al., Characterization of the human urinary proteome: a method for high-resolution display of urinary proteins on twodimensional electrophoresis gels with a yield of nearly 1400 distinct protein spots. Proteomics 2004;4:1159–1174.

[127]. Wittke S, Fliser D, Haubitz M., et al., Determination of peptides and proteins in human urine with capillary electrophoresis-mass spectrometry, a suitable tool for the establishment of new diagnostic markers. J Chromatogr A. 2003;1013:173–181.

[128]. Kemperman RF, Horvatovich PL, Hoekman B, Reijmers TH, Muskiet FA, Bischoff R. Comparative Urine Analysis by Liquid Chromatography-Mass Spectrometry and

Multivariate Statistics: Method Development, Evaluation, and Application to Proteinuria. J. Proteome Res. 2007;6:194-206.

[129]. Pang JX, Ginanni N, Dongre AR, Hefta SA, Opitek GJ. Biomarker discovery in urine by proteomics. J Proteome Res. 2002;1:161–169.

[130]. Tantipaiboonwong P, Sinchaikul S, Sriyam S, Phutrakul S, Chen ST. Different techniques for urinary protein analysis of normal and lung cancer patients. Proteomics 2005;5:1140–1149.

[131]. Wattiez R, Falmagne P. Proteomics of bronchoalveolar lavage fluid. J Chromatogr B 2005;815:169–178.

[132]. Noel-Georis I, Bernard A, Falmagne P, Wattiez R. Database of bronchoalveolar lavage fluid proteins. J Chromatogr B. 2002;771:221–236.

[133]. Horvatovich P, Govorukhina N, Bischoff R. Biomarker Discovery by Proteomics: Challenges Not Only for the Analytical Chemist. Analyst 2006;131:1193-1196.

[134]. Righetti PG, Castagna A, Antonioli P, Boschetti E. Prefractionation techniques in proteome analysis: the mining tools of the third millennium. Electrophoresis 2005;26:297–319.

[135]. Rieux L, Bischoff R, Verpoorte E, Niederlander HAG. Restricted-Access Material-Based High-Molecular-Weight Protein Depletion Coupled on-Line With Nano-Liquid Chromatography-Mass Spectrometry for Proteomics Applications. Journal of Chromatography A. 2007;1149:169-177.

[136]. Martosella J, Zolotarjova N, Liu H, Nicol G, Boyes BE. Reversed-Phase High-Performance Liquid Chromatographic Prefractionation of Immunodepleted Human Serum Proteins to Enhance Mass Spectrometry Identification of Lower-Abundant Proteins. J Proteome. Res. 2005;4:1522-1537.

[137]. Schulz-Knappe P, Zucht HD, Heine G, Jurgens M, Hess R, Schrader M. Peptidomics: the comprehensive analysis of peptides in complex biological mixtures. Comb Chem High Throughput Screen 2001;4:207–217.

[138]. Schulz-Knappe P, Schrader M. Peptidomics in biomarker and drug discovery. Curr Drug Discov. 2003;21–24.

[139]. Schrader M, Schulz-Knappe P. Peptidomics technologies for human body fluids. Trends Biotechnol. 2001;19:S55-60.

[140]. Heine G, Zucht HD, Jürgens M., et al., High-resolution peptide mapping of cerebrospinal fluid: a novel concept for diagnosis and research in central nervous system diseases. J Chromatogr B. 2002;782:353–361.

[141]. MacNair JE, Patel KD, Jorgenson JW. Ultrahigh-pressure reversed-phase capillary liquid chromatography: isocratic and gradient elution using columns packed with 1.0-micron particles. Anal Chem. 1999;71:700–708.

[142]. MacNair JE, Lewis KC, Jorgenson JW. Ultrahigh-pressure reversed-phase liquid chromatography in packed capillary columns. Anal Chem. 1997;69:983–989.

[143]. MacNair JE, Opiteck GJ, Jorgenson JW, Moseley MA, III. Rapid separation and characterization of protein and peptide mixtures using 1.5 microns diameter nonporous silica in packed capillary liquid chromatography/mass spectrometry. Rapid Commun Mass Spectrom. 1997;11:1279–1285.

[144]. Mellors JS, Jorgenson JW. Use of 1.5-micron porous ethyl-bridged hybrid particles as a stationary-phase support for reversed-phase ultrahigh-pressure liquid chromatography. Anal Chem. 2004;76:5441–5450.

[145]. Xiong L, Zhang R, Regnier FE. Potential of silica monolithic columns in peptide separations. J Chromatogr A. 2004;1030:187–194.

[146]. Xie S, Allington RW, Svec F, Frechet JMJ. Rapid reversed-phase separation of proteins and peptides using optimized 'moulded' monolithic poly(styrene-codivinylbenzene) columns. J Chromatogr A. 1999;865:169–174.

[147]. Walcher W, Toll H, Ingendoh A, Huber CG. Operational variables in highperformance liquid chromatography-electrospray ionization mass spectrometry of peptides and proteins using poly(styrene-divinylbenzene) monoliths. J Chromatogr A. 2004;1053:107–117.

[148]. Kimura H, Tanigawa T, Morisaka H, et al. Simple 2D-HPLC using a monolithic silica column for peptide separation. J Separation Sci. 2004;27:897–904.

[149]. Hennessy TP, Boysen RI, Huber MI, Unger KK, Hearn MTW. Peptide mapping by reversed-phase high-performance liquid chromatography employing silica rod monoliths. J Chromatogr A. 2003;1009:15–28.

[150]. Eschelbach JW, Jorgenson JW. Improved Protein Recovery in Reversed-Phase Liquid Chromatography by the Use of Ultrahigh Pressures. Anal. Chem. 2006;78:1697-1706.

[151]. Stoll DR, Cohen JD, Carr PW. Fast, Comprehensive Online Two-Dimensional High Performance Liquid Chromatography Through the Use of High Temperature Ultra-Fast Gradient Elution Reversed-Phase Liquid Chromatography. Journal of Chromatography A 2006;1122:123-137.

[152]. Wang X, Barber WE, Carr PW. A Practical Approach to Maximizing Peak Capacity by Using Long Columns Packed With Pellicular Stationary Phases for Proteomic Research. Journal of Chromatography A 2006;1107:139-151.

[153]. Xiang Y, Liu Y, Lee ML. Ultrahigh Pressure Liquid Chromatography Using Elevated Temperature. Journal of Chromatography A 2006;1104:198-202.

Chapter I.III

Label-Free Proteomics of Serum

N.I. Govorukhina, P. Horvatovich and R. Bischoff

Book chapter in: Functional Proteomics, the Humana Press (Totowa, New Jersey, USA), in press (2007).

1. Introduction

The comparative analysis of serum samples from patients and healthy controls requires highly standardized operating procedures that produce reproducible data [1]. The generated data need to be processed in a manner to bring the significant, disease-related differences in protein or peptide profiles forward and to reduce non-related noise [2]. Processed data have to be analyzed in a statistically rigorous fashion and subjected to both statistical and biological validation.

In this chapter we present a protocol to perform proteomics of serum samples obtained from cancer patients but the protocol is generic enough to be also applicable to sera from patients with other diseases. This is obviously just one way of proceeding and there are quite a number of other approaches, some of which can be found in this series. In order to enhance the concentration sensitivity of this method, we remove high-abundance proteins by immunoaffinity chromatography. We have recently shown that this can be done efficiently and with high repeatability [3,4]. The subsequent trypsin digestion step and reversed-phase HPLC-MS analysis are performed in a reproducible manner and controlled with standard samples at regular intervals. Concentration sensitivity of this method is app. 0.5 µM for the added Cytochrome C. In order to enhance concentration sensitivity further, it is optional to include an additional protein separation step. We describe the use of a recently developed reversed-phase material that can be run at 80°C [5].

Although we touch upon data processing and statistical analysis, we cannot go into the methodological details due to limited space. We refer the reader to the cited references as well as to a dedicated book in this series focusing on bioinformatics.

2. Materials

2.1. Depletion of the 6 most abundant proteins on a Multiple Affinity Removal column

1. Store serum samples at -80ºC in aliquots.
2. Buffer A (# 5185-5987, Agilent, Palo Alto, California, USA).

3. Buffer B (#5185-5988, Agilent, Palo Alto, California, USA).
4. 0.22 µm spin filters (Part # 5185-5990, Agilent).
5. Multiple Affinity Removal column (Agilent, 4.6 x 50 mm, Part # 5185-5984 Palo Alto, California, USA).

2.2. SDS-PAGE

1. All chemicals for Polyacrylamide gel were from BIO-RAD (Bio-Rad, www.biorad.com)
2. PageRuler ™ Prestained Protein Ladder (Fermentas, #SM0671)
3. Coomassie Brilliant Blue R concentrate (Sigma, www.sigmaaldrich.com).

2.3. HPLC-MS

1. Atlantis™ dC 18 (1.0 x 150 mm, 3 µm) column for cap-LC-MS, (Waters, Milford, Massachusetts, USA, www.waters.com).
2. Atlantis™ dC 18 in-line trap column for cap-LC-MS (Waters, Milford, Massachusetts, USA, www.waters.com).
3. Chip for chip-LC-MS with a 40 nl trap column (75 µm × 11 mm) and a 75 µm × 43 mm analytical column both containing C-18SB-ZX 5 µm chromatographic material (Cat. # G4240-62001, Agilent, Palo Alto, California, USA). The chip is equipped with a nanoelectrospray tip of 2 mm length with conical shape: 100 µm OD × 8 µm ID.
4. Micro BCA™ protein assay reagent kit (www.piercenet.com).
5. Sequencing-grade modified trypsin (Promega, Cat# V5111, U.S.A.).
6. Acetonitrile HPLC-S (ACN) gradient-grade was from (Biosolve, Valkenswaard, The Netherlands).
7. Formic acid, FA, 98-100% pro analysis (Cat# 1.00264.1000., Merck, Darmstadt, Germany).
8. Ultra-pure water (conductivity: 18.2 MΩ), Maxima System (Elga Labwater, Ede, The Netherlands).

2.4. Pre-fractionation of depleted serum on an mRP-C18 Macroporous Reversed-Phase column

1. Macroporous Reversed-Phase mRP-C18 column (Agilent, 4.6 x 50mm, Part # 5188-5231 Palo Alto, California, USA).
2. Trifluoroacetic Acid, TFA, sequencing grade (# 28902, Pierce).
3. Ultra-pure water (conductivity: 18.2 MΩ), Maxima System (Elga Labwater, Ede, The Netherlands).
4. Urea (#084K0063, Sigma, www.sigmaaldrich.com)
5. Glacial acetic acid (Cat#1.00063.1000, Merck, Darmstadt, Germany)
6. Solvent A for mRP column (97% water/0.1% TFA)
7. Solvent B for mRP column (97% AcN/0.1% TFA)

3. Methods

3.1. Preparation of Samples

1. Mix 20 µL of crude serum with 80 µL of buffer A (Agilent). Filter through 0.22 µm spin filters at 13000g and 4˚C for 10 min to remove particulates.
2. Inject 80 µL (80% of the total amount of diluted crude serum) on a Multiple Affinity Removal column for depletion according to the manufacturer's instructions (with detection at 280 nm using the following timetable: 0-9 min, 100% buffer A (0.25 mL/min); 9.0-9.1 min, linear gradient 0-100 B % (1 mL/min), 9.1-12.5 min, 100% buffer B (1 mL/min); 12.5-12.6 min, linear gradient 100-0% buffer B (1 mL/min); 12.6-20 min, 100% buffer A (1 mL/min). Removal of abundant proteins, as described above, was performed on a LaChrom HPLC System (Merck Hitachi, www.merck.com) or on an AKTA FPLC system (GE Healthcare).
3. Collect the flow-through fraction (depleted serum collected between 2-6 min) of a total volume of appr. 1 mL.
4. Determine protein concentrations with the Micro BCA™ protein assay reagent kit (www.piercenet.com) and calculate for an average protein molecular weight of 50 kDa. Use BSA as the calibration standard.
5. Digest 100 µL (~10% of the total amount, which corresponds to ~8 µg or 160 pmol of total protein) of depleted serum with trypsin (1:20 wt/wt enzyme to substrate) at 37°C overnight with shaking at 400 rpm.

3.2. SDS-PAGE

1. SDS PAGE was performed in a Mini-Protein III cell (Bio-Rad, www.biorad.com) using 12% gels with 0.1% SDS according to the manufacturer's instructions.
2. Boil samples with sample buffer containing 0.02 M DTT for 1 min, cool down and apply directly to the gel.
3. Perform staining with Coomassie Brilliant Blue R concentrate (Sigma, www.sigmaaldrich.com) diluted and used as instructed by the manufacturer.

3.3. HPLC-MS

3.3.1. Capillary- and chip-LC-MS

1. All LC-MS analyses were performed on an Agilent 1100 capillary (cap) HPLC system coupled on-line to an SL ion-trap mass spectrometer (www.home.agilent.com). In the case of cap-LC-MS, the instrument was equipped with an Atlantis™ dC 18 (1.0 x 150 mm, 3 µm) column that was protected by an Atlantis™ dC 18 in-line trap column. 40 µL of the pretreated (depleted and digested) serum corresponding to ~8 µg or 160 pmol of total protein digest (calculation based on a 50kDa protein) were injected. An autosampler (cat. # G1367A) equipped with a 100 µL injection loop was used for cap-LC-MS. For chip-LC-MS the same mass spectrometer was used but

equipped with a microfluidics (chip-cube) interface (cat. # G4240A) including a chip microfluidic device. The injected sample amount was 0.25 µg (3.4 - 5.1 µL; 5 pmol) of depleted, trypsin-digested serum, 10-times diluted with 0.1% aq. FA. Injections were performed with an autosampler (Agilent, cat. n° G1389A) equipped with an injection loop of 8 µL (this includes also the dead volume up to the trapping column). In both case, the autosampler was temperature-controlled using a cooler (cat. # G1330A) maintaining the samples at 4°C.

The HPLC system for cap-LC-MS had the following additional modules: capillary pump (cat. #, G1376A), solvent degasser (cat. #, G1379A), UV detector (cat. # G1314A) and column holder (cat. #, G1316A). The sample was injected and washed in the back-flush mode for 30 min (0.1% aq. FA and 3% acetonitrile at a flow-rate of 50 µL/min). Peptides were eluted in a linear gradient from 0 to 70% (0.5%/min) acetonitrile with 0.1% formic acid at a flow-rate of 20 µL/min. After each injection the in-line trap and the analytical column were equilibrated with eluent A for 20 min prior to the next injection.

The chip-LC-MS system contained the following additional modules: nanopump (cat. n° G2226A), capillary loading pump and solvent degasser. The sample was injected and washed in the back-flush mode for 4 min (0.1% aq. FA, 2 µL/min) and then the on-chip trapping column was switched in-line with the analytical column on the microfluidics device. For these separations, the same eluents A and B as for the cap-LC-MS system were used at a flow-rate of 0.3 µL/min. After elution for 6 minutes with eluent A, a linear gradient from 0 to 50% eluent B at 0.5 %/min followed by a step gradient from 50 to 70% at 1 %/min of eluent B was run. 70% eluent B was maintained for 10 min. After each injection, the in-line trap and the analytical column were equilibrated with eluent A for 20 min at 2 and 0.3 µL/min, respectively.

2. In the MS acquisition parameters only the ionization voltage and the use of nebulizer gas were different between the two systems (1800-2000 V of ionization voltage and no use of nebulizer gas for chip-LC-MS; 16.0 psi N_2 nebulizer gas and 3500 V of ionization voltage for cap-LC-MS). The following general settings were used for mass spectrometry during LC-MS: drying gas: 6.0 L/min N_2, skimmer: 40.0 V, cap. exit: 158.5 V, Oct. 1: 12.0V, Oct. 2: 2.48 V, Oct. RF: 150 Vpp (Voltage, Peak Power Point), Lens 1: -5.0 V, Lens 2: -60.0 V, Trap drive: 53.3, T^o: 325^o, Scan resolution: enhanced (5500 m/z per second scan speed). Target mass: 600. Scan range: 100-1500 m/z. Spectra were saved in centroid mode. LC-MS chromatographic data were analyzed with Bruker Data Analysis software, version 2.1 (Build 37).

3.3. Data Processing

The original Bruker Daltonics LC-MS data files were converted into ASCII-format with the Bruker data analysis software. For further data analysis Matlab (version 7.2.0.232 (R2006a), Mathworks, Natick, Massachusetts, USA) and the PLS toolbox (version 3.5.2, Eigenvector Research Inc., Wenatchee, Washington, USA) were used. Centroid data were smoothed and reduced using a normalized two-dimensional Gaussian filter with rounding of the nominal m/z ratios to 1 m/z (the original data had a resolution of 0.1 m/z). After meshing the data files of all chromatograms, they were time-aligned to a reference data file using Correlation Optimized Warping (COW) based on Total Ion Currents (TICs) constructed from signals in the range 100-1500 m/z.

A modified M-N rule was applied for peak detection by first calculating a median local baseline using a sliding window technique separately for each m/z trace. A median window size of 1200 data points, corresponding to 20.84 min for chip-LC-MS and 20.17 min for cap-LC-MS, was used with a moving rate of 10 points and a minimum median value of 200 counts. According to the M-N rule, a threshold of M-times the local baseline was used and a peak was assigned if, within one m/z trace, the signal exceeded this threshold for at least N consecutive points. For each detected peak the m/z value, the mean retention times of the three highest measured intensities (within the same peak reduced by the local baseline) were stored in a peak list created for every chromatogram.

In order to combine the peak lists from different samples with each other, one-dimensional peak matching was achieved by using the sliding window technique, in which the same m/z traces were evaluated for peaks that are proximate in time (step size 0.1 min; search window 1.0 min; maximal accepted standard deviation for all retention times within a group of matched peaks was 0.75 min). Missing peak locations were filled with extracted local signals reduced with the local baseline at a given m/z retention time location. The generated peak matrix, created from the peak lists of the individual samples, consisted of a peak(row)-sample(column)-intensity(value) matrix. This peak matrix was used for multivariate statistical analysis.

A Nearest Shrunken Centroid (NSC) supervised classification algorithm in conjunction with leave-one-out cross-validation (LOOCV) was applied to select the most discriminating compounds. The selected compounds were then subjected to autoscaled Principal Component Analysis (PCA) and visualized using biplots of the first two principal components. All data processing and statistical analyses were done on a personal computer equipped with a dual core +3800 MHz AMD 64 X2 processor equipped with 4 GB of RAM. Figure 1 shows an example of data obtained by chip- and capillary-LC-MS.

3.4. Pre-fractionation of depleted serum by reversed-phase HPLC on an mRP column at 80°C

1. Add to ~300µg (about 300µL) of depleted serum 0.48 g urea and 13µL of glacial acetic acid, according to the manufacturer's instructions (www.agilent.com/chem/bioreagents).

2. Add solvent A to a final volume of 1mL and inject the total volume with a 1mL loop onto the column.

3. Fractionate at 80 °C (pH <5.0) at a flow rate of 0.75mL/min with UV detection at 280nm.
4. Run gradient from 3 to 30% B in 6 min, to 55% solvent B in 40 min and up to 100% in 53 min.
5. Collect fractions of 0.75mL (see Figure 2a).
6. Compare fractions after pre-fractionation by SDS-PAGE (in our case pair-wise before and after medical treatment for each patient) (see Figure 2b).

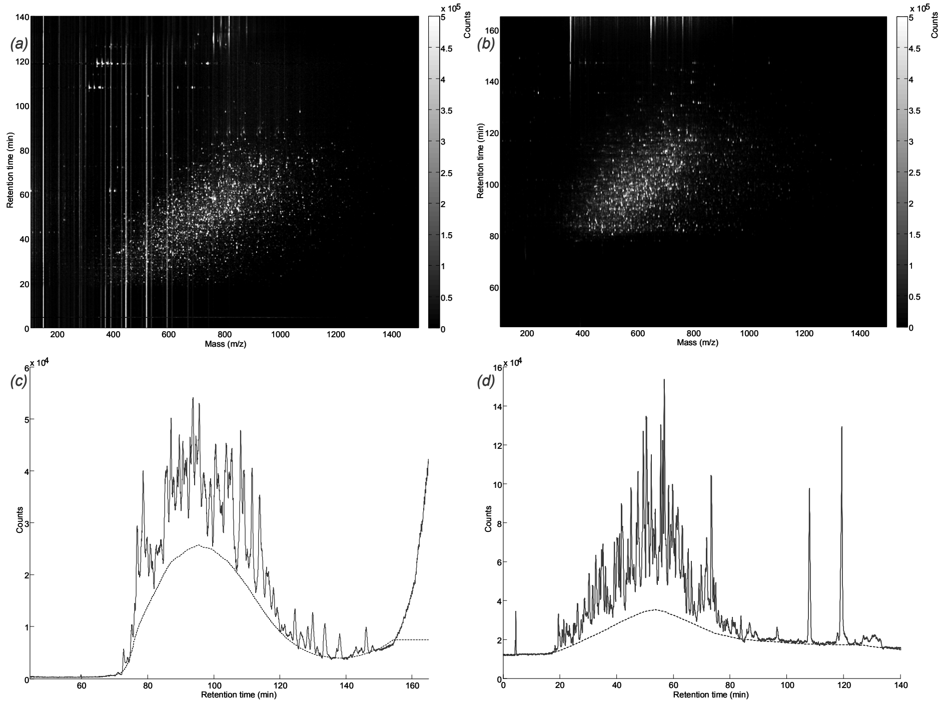

Figure 1. Raw LC-MS data of depleted and trypsin-digested serum analyzed by chip- *(a)* or capillary LC-MS *(b)* represented as "heat map". The horizontal axis represents the m/z values in amu and the vertical axis shows the retention time in min. Peak intensity is coded as indicated (white: high; black: low). Panels *(c)* and *(d)* show the same data in the conventional representation as Total Ion Chromatograms (TICs). The dashed lines depict the calculated baseline. Data were collected in centroid mode and meshed using a data reduction of 1:10.

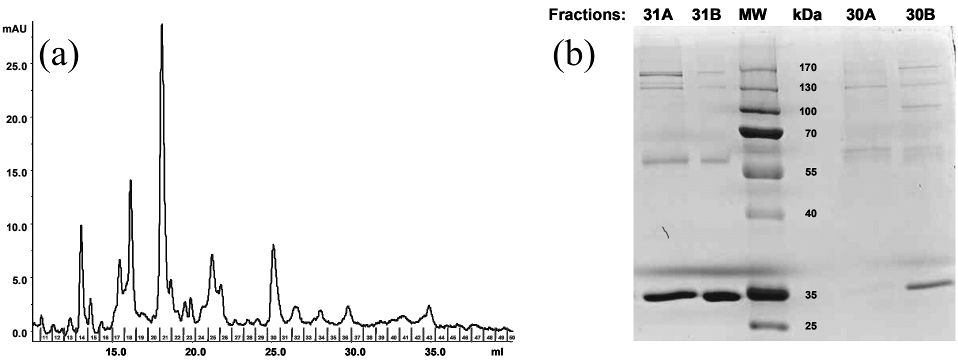

Figure 2. (a) 300 µg of depleted serum were pre-fractionated on an mRP column at 80°C (example of a sample from a cervical cancer patient before treatment). (b) 12% SDS-PAGE of serum from a cervical cancer patient before "A" and after "B" medical treatment. 30A and 31A: fractions 30 and 31 from the mRP column of patient serum before medical treatment; 30B and 31B: fractions 30 and 31 of the serum from the same patient after treatment. Note the clear difference at about 35 kDa in fraction 30.

4. Notes

Prefractionating proteins in depleted serum/plasma on a newly developed "Macroporous Reversed-Phase C18 High-Recovery Protein Fractionation HPLC column (mRP)" prior to digestion is reproducible and enables high resolution to be achieved. High protein recoveries allow the complexity of the sample to be significantly reduced. Robustness and high recovery mRP fractionation makes higher-quality protein identification by coupled LC-MS methods [5]. Based on our experiences and the work of Martosella *et al.* [5], elevated mRP column temperature (80 degrees C) is a very critical operational parameter, while poor control of the temperature could result in poor reproducibility and bad chromatographic resolution.

The performance of the described methodology was evaluated by comparing the ability of cap- and chip-LC-MS to find discriminating features. For this purpose, 5 serum samples, spiked with 21 pmol of horse heart Cytochrome C in 2 µL serum, were analyzed next to 5 non-spiked serum samples. Due to losses during immunoaffinity depletion of high-abundance proteins, the actual amount of Cytochrome C that was analyzed was 4.2 pmol [3], corresponding to about 3% (molar) of the total protein content. The raw data obtained were subjected to data processing as described [6], followed by supervised classification and selection of discriminating features using the Nearest Shrunken Centroid (NSC) algorithm [7]. The shrinkage parameter was optimized using a "leave one out" cross validation (LOOCV), strategy with the aim of reaching the lowest cross validation error. Although we applied a rather low threshold (M = 2, N = 5) for peak picking, which introduced more noise in the peak list, a large domain of shrinkage showed no cross-validation error (0.90-29.51 for chip- and 0.61-16.80 for cap-LC-MS, Figures 3a and b, respectively), indicating a robust classification model. Evaluating the 16 most discriminating features selected at shrinkages of 10 and 8.5 for chip- and cap-LC-MS, respectively, resulted in 6 different peptides. Six peptides selected from the chip-LC-MS and 5 of the 6 peptides selected from the cap-LC-MS data corresponded to *in-silico* predicted tryptic peptides of horse heart Cytochrome C. Figure 3 shows that correct discrimination between spiked versus non-spiked serum samples was easily possible based on the selected peaks (Figures 3c and d). PCA analysis of the selected features (Figures 3e and f) revealed that almost all variability in the data can be explained by Principal Component 1 (99% for chip- and 98% for cap-LC-MS). Visualization of the Extracted Ion Chromatograms (EICs) of some of the selected peaks (Figure 4) confirmed that highly discriminating peaks had been correctly found within the complex mixture of digested serum proteins. Figure 4 shows also the generally good time alignment using COW.

The results show that the integration of nanoLC into microfluidic devices enables quantitative profiling studies to find protein expression differences using ~30 times less sample with higher separation efficiency compared to capillary LC.

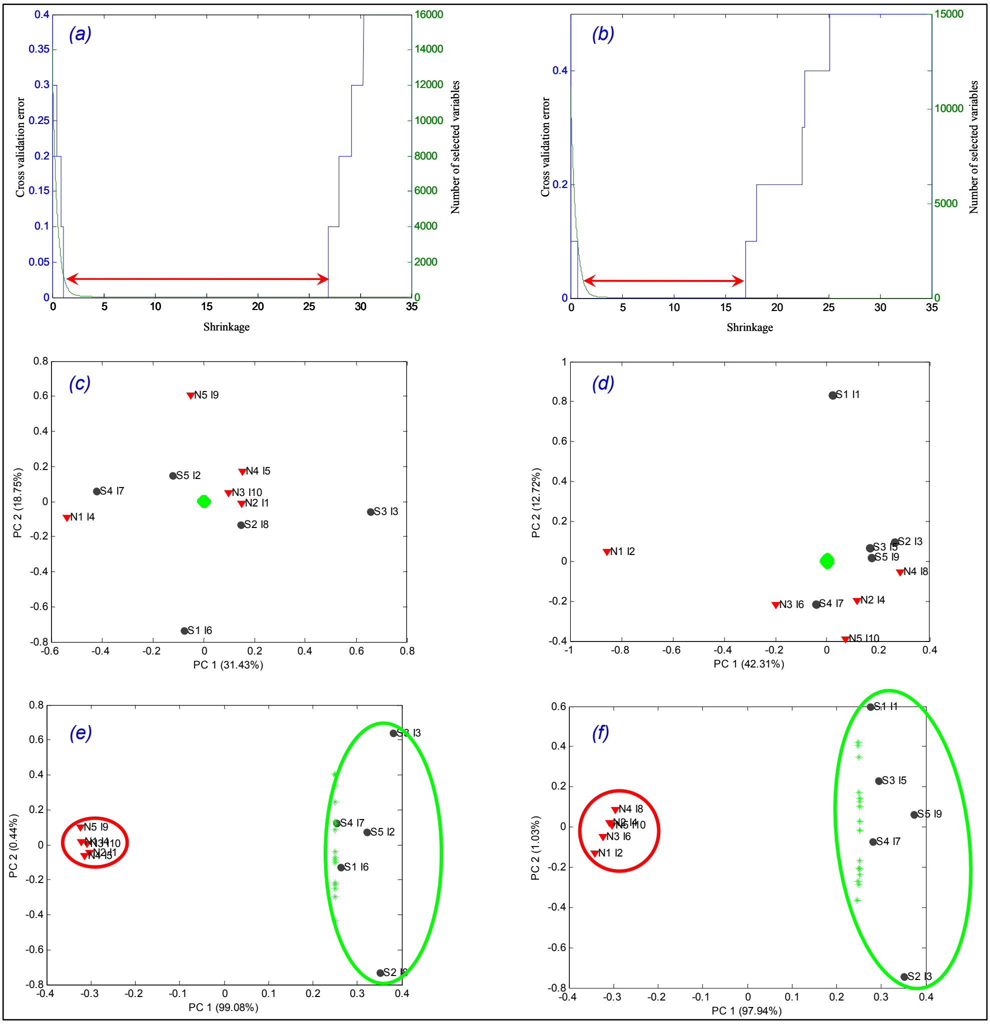

Figure 3. Representation of the "leave-one out" cross-validation (LOOCV) error and the number of selected variables as a function of the shrinkage for chip-LC-MS *(a)* and cap-LC-MS *(b)*. The selected variables, where the shrinkage domain has no cross-validation error, are indicated with arrows. For these domains, the selected variables enabled a perfect separation of the two classes (see panels *(e)* and *(f)*). PCA plots using all peaks obtained with M = 2, N = 5 for chip-LC-MS *(c)* and cap-LC-MS *(d)*, (14091 for chip-LC-MS and 11256 for cap-LC-MS) did, however, not allow discrimination between the classes. In the figures PC 1 and PC 2 refer to the Principal Component axis 1 and 2.

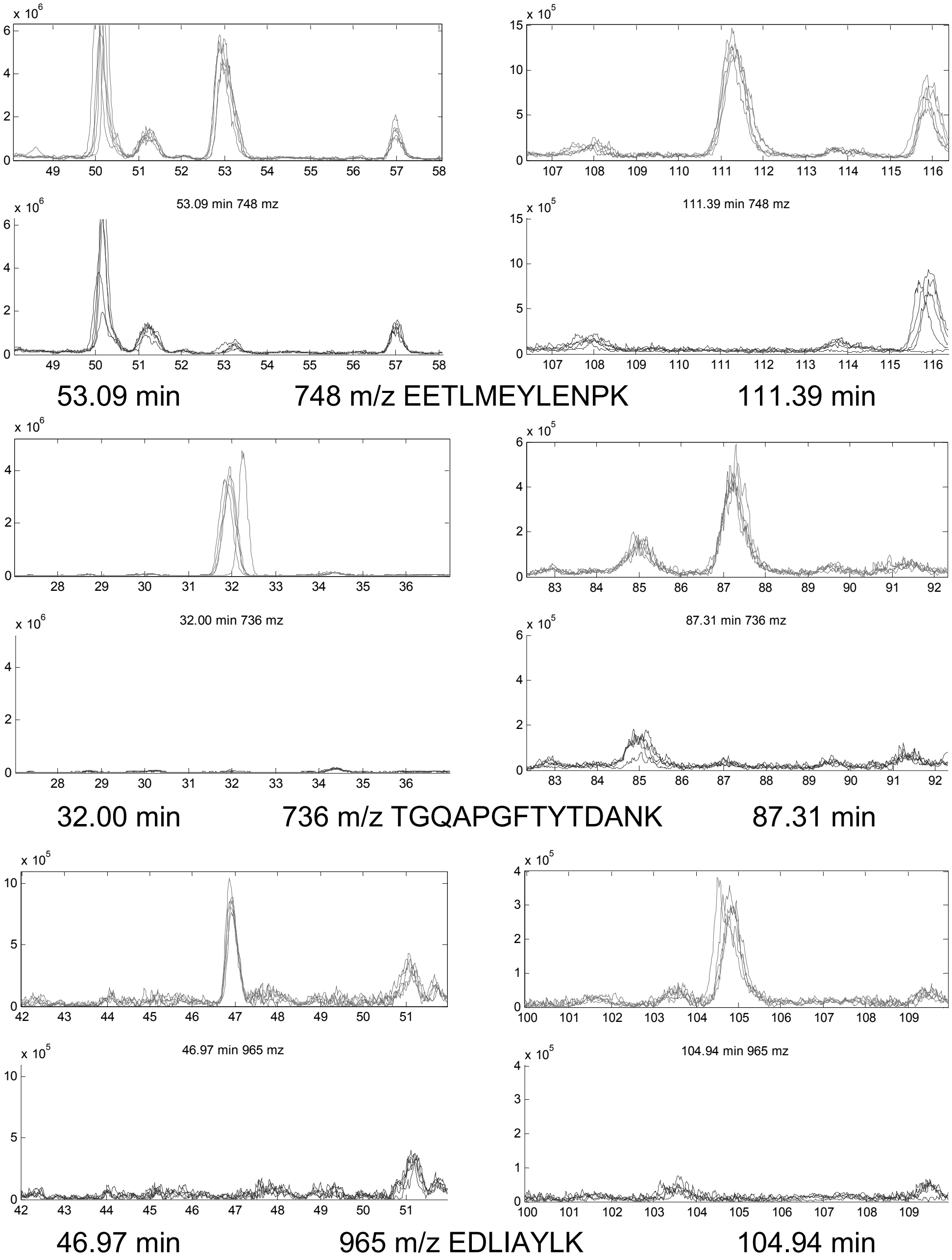

Figure. 4. Examples of Extracted Ion Chromatograms (EICs) of NSC-selected peaks corresponding to tryptic fragments of horse heart Cytochrom C from datasets obtained with chip-LC-MS (left) and cap-LC-MS (right). The green upper traces were obtained from spiked, the blue lower traces were obtained from non-spiked samples.

5. References

[1]. Villanueva J, Philip J, Chaparro CA, Li Y, Toledo-Crow R, DeNoyer L, Fleisher M, Robbins RJ, Tempst P. Correcting Common Errors in Identifying Cancer-Specific Serum Peptide Signatures. J Proteome. Res. 2005;4:1060-1072.

[2]. Kemperman RF, Horvatovich PL, Hoekman B, Reijmers TH, Muskiet FA, Bischoff R. Comparative Urine Analysis by Liquid Chromatography-Mass Spectrometry and Multivariate Statistics: Method Development, Evaluation, and Application to Proteinuria. J. Proteome Res. 2007;6:194-206.

[3]. Govorukhina NI, Reijmers TH, Nyangoma SO, van der Zee AGJ, Jansen RC, Bischoff R. Analysis of human serum by LC-MS: improved sample preparation and data analysis. J. Chromatogr. A. 2006;110:42-150.

[4]. Dekker LJ, Bosman J, Burgers PC, van Rijswijk A, Freije R, Luider T, Bischoff R. Depletion of High-Abundance Proteins From Serum by Immunoaffinity Chromatography: A MALDI-FT-MS Study. J. Chromatogr. B. 2007;847:65-69.

[5]. Martosella J, Zolotarjova N, Liu H, Nicol G, Boyes BE. Reversed-Phase High-Performance Liquid Chromatographic Prefractionation of Immunodepleted Human Serum Proteins to Enhance Mass Spectrometry Identification of Lower-Abundant Proteins. J Proteome. Res. 2005;4:1522-1537.

[6]. Horvatovich P, Govorukhina NI, Reijmers TH, van der Zee AGJ, Bischof R. Evaluation of HPLC-chip/MS platform for label-free profiling for biomarker discovery. Electrophoresis 2007; submitted.

[7]. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression.. Proc. Natl. Acad. Sci. USA. 2002;99:6567-6572.

Chapter I.IV
Outline of the thesis

Blood (plasma or serum) and urine can be easily sampled from patients and or healthy volunteers and are therefore often the first choice when trying to discover novel biomarkers or biomarker patterns to diagnose cancer and other diseases. There are, however, drawbacks in using a blood plasma or serum such as the masking of low-abundance by high-abundance proteins and the possible effect of sampling and sample handling procedures (e.g. different times and conditions for blood clotting). A review of the current techniques dealing with blood proteomics in blood is given in **Chapter I.I**. Practical details of the protocol used for analyzing blood serum is given in **Chapter I.III**.

The LC-MS method to analyze serum is presented in **Chapter II**. Two sample preparation methods were tested in terms of their efficiency to deplete high-abundance serum proteins and how they affect the repeatability of the LC–MS analysis. The first method comprised depletion of human serum albumin (HSA) on a column that contained dye ligand and immunoglobulin G (IgG) on an immobilized protein support followed by tryptic digestion, fractionation by cation-exchange chromatography, trapping on a C18 column and reversed-phase LC–MS. The second method included depletion of the six most abundant serum proteins based on multiple immunoaffinity chromatography followed by tryptic digestion, trapping on a C18 column and reversed-phase LC–MS.

Chapter III describes an application of a miniaturized, microfluidics-based LC-MS system (chip-LC-MS). It is shown that chip-LC-MS has at least twice the resolution of the standard capillary LC-MS method described in **Chapter II**. Two protocols were compared side-by-side. As a control, some samples were spiked with horse heart Cytochrome C. Further statistical analysis allowed discrimination between control and spiked samples.

Since blood is a multifunctional, complex physiological fluid, its state and composition will change under the influence of external factors. In **Chapter IV** the influence of clotting time on protein composition of serum samples was studied. It appeared that variation in clotting time between 1 and 8 hours had only a minor effect on the overall serum composition. However, the concentration of fibrinopeptides varied significantly, as expected, since they are directly involved in the clotting process.

Chapter V describes a more comprehensive approach to evaluate the influence of various pre-analytical parameters on the serum proteome. We applied a factorial design to assess the importance of seven factors, including the level of hemolysis, the clotting time, and storage conditions.

Chapter VI describes the analysis of serum samples from cervical cancer patients before and after treatment using the methodology mentioned above, followed by data processing and statistical data analysis. While we did not discover major changes in the serum proteome using this method, some changes in proteins composition were found in samples taken before and after medical treatment, the significance of which are being further investigated. It is thus demonstrated that the described methods are applicable to highly complex body fluids such as serum and that further studies into the relevance of the discovered changes of the serum proteome are warranted.

Chapter II.

Analysis of human serum by liquid chromatography–mass spectrometry: Improved sample preparation and data analysis

N.I. Govorukhina, T.H. Reijmers, S.O. Nyangoma, A.G.J. van der Zee, R.C. Jansen and R. Bischoff

Published in Journal of Chromatography A, v.1120, 142-150 (2006).

Abstract

Discovery of biomarkers is a fast developing field in proteomics research. Liquid chromatography coupled on line to mass spectrometry (LC–MS) has become a powerful method for the sensitive detection, quantification and identification of proteins and peptides in biological fluids like serum. However, the presence of highly abundant proteins often masks those of lower abundance and thus generally prevents their detection and identification in proteomics studies. To perform future comparative analyses of samples from a serum bank of cervical cancer patients in a longitudinal and cross-sectional manner, methodology based on the depletion of high-abundance proteins followed by tryptic digestion and LC–MS has been developed. Two sample preparation methods were tested in terms of their efficiency to deplete high-abundance serum proteins and how they affect the repeatability of the LC–MS data sets. The first method comprised depletion of human serum albumin (HSA) on a column that contained dye ligand and immunoglobulin G (IgG) on an immobilized protein support followed by tryptic digestion, fractionation by cation-exchange chromatography, trapping on a C18 column and reversed-phase LC–MS. The second method included depletion of the six most abundant serum proteins based on multiple immunoaffinity chromatography followed by tryptic digestion, trapping on a C18 column and reversed-phase LC–MS. Repeatability of the overall procedures was evaluated in terms of retention time and peak area for a selected number of endogenous peptides showing that the second method, besides being less time consuming, gave more repeatable results (retention time: <0.1% RSD; peak area: <30% RSD). Application of an LC–MS component detection algorithm followed by principal component analysis (PCA) enabled discrimination of serum samples that were spiked with horse heart Cytochrome C from non-spiked serum and the detection of a concentration trend, which correlated to the amount of spiked horse heart Cytochrome C to a level of 5 pmol Cytochrome C in 2 μL original serum.

1. Introduction

Various methods have been applied in recent years for the discovery of biomarkers or biomarker patterns of major human diseases, especially for various types of cancer [1-13]. Surface-enhanced laser desorption ionization mass spectrometry (SELDI-MS), which combines on-chip sample preparation with mass spectrometric analysis, has become a popular method [14], although more recent results question the viability of this approach [15,16]. Liquid chromatography coupled on line to mass spectrometry (LC–MS) is one of the most widely used analytical methods with applications going beyond proteomics and biomarker discovery. It has the advantage of combining powerful separation by one- or multi-dimensional chromatography with the exquisite selectivity and sensitivity of modern mass spectrometers. The complexity of a typical LC–MS data set reaches 10^8 data points per sample at a resolution of 0.1 amu in the m/z domain and a chromatographic run time of approximately 2 h (7000 data points). It is thus pivotal to apply data pre-processing algorithms to reduce data complexity and multivariate statistics to reduce dimensionality to be able to compare data sets obtained from longitudinal or cross-sectional patient studies comprising significantly less samples than the number of original variables in the data.

Sample preparation is an often underestimated problem in comparative biomarker analysis. Notably in serum there are a few highly abundant proteins that will prevent detection of many minor proteins present in the sample. Since it is unlikely that high-abundance proteins like albumin or transferrin will be biomarkers for specific diseases, it is necessary for biomarker discovery methodology to detect and quantify proteins present at lower concentrations. One way to reduce serum complexity is chromatographic removal of the most abundant proteins. In human serum, the most abundant proteins are albumin and γ-globulins. Earlier [17], we tested different depletion strategies to reduce the level of abundant proteins based on either specific antibodies, dye ligands (for albumin) [18] or Protein A and G (for γ-globulins) [19] and [20]. Other approaches based, for example, on ultrafiltration showed lower selectivity for these target proteins but allowed on the other hand to concentrate the sample [21]. Co-depletion of proteins and peptides is a concern when employing such depletion strategies [22].

LC–MS is particularly adapted to the separation and detection of peptides. This has triggered development of the so-called "shotgun" proteomics approach [23]. Contrary to proteomics based on two-dimensional polyacrylamide gel electrophoresis (2D-PAGE), where the proteins are separated prior to tryptic digestion, trypsin digestion precedes the separation step. The shotgun approach results therefore in extremely complex mixtures of peptides presenting a challenge to any separation methodology. Very high efficiency separation systems have recently been applied to this problem, allowing identification of a wide range of proteins in plasma or serum [13]. A new approach developed recently combines the use of immunoaffinity depletion with reversed-phase separation of whole proteins at elevated temperatures to reduce sample complexity prior to identification of proteins in human serum [24]. This approach allowed, for example, to identify hepatocyte growth factor, which is present at a level of 20 ng/mL.

In this study, we combined efficient depletion of high-abundance proteins with LC–MS-based shotgun proteomics followed by data pre-processing to select information-rich chromatographic traces based on the CODA algorithm [25]. This was followed by multivariate statistical analysis of the selected traces by principal component analysis (PCA) to evaluate the performance of the method to discriminate samples. Initially we tested two approaches for the preparation of human serum for LC–MS analysis. Removal of abundant proteins was performed by dye ligand or antibody-based multiple-affinity chromatography, which eliminates the six most abundant serum proteins. In the case of dye ligand chromatography, a two-dimensional chromatographic system was employed consisting of strong cation exchange followed by reversed-phase HPLC. Proteins remaining after the multiple-affinity removal step were directly digested and analyzed by reversed-phase LC–MS. The performance of each approach was assessed in terms of repeatability of retention times and peak areas for a selected number of endogenous peptides showing that the repeatability of the second method was superior. The discriminatory power of this method was assessed by adding decreasing amounts of horse heart Cytochrome C to the original serum proving that a concentration trend was correctly represented in the first principal component after non-supervised data pre-processing and PCA down to a lower level of 50 pmol Cytochrome C in 20 µL serum.

2. Experimental

Serum samples were obtained from the Department of Gynecological Oncology (University Medical Center Groningen, The Netherlands) and stored at −80 °C in aliquots. All intermediate fractions were stored at −20 °C. To develop and optimize the analytical procedure, serum from a single cervical cancer patient with a squamous cell carcinoma antigen-1 (SCCA-1) [26] level of 2.2 ng/mL (determined by ELISA) was used.

2.1. Sample preparation of human serum

2.1.1. Method 1

Two hundred and forty microliters of diluted crude serum (60 µL of crude serum mixed with 180 µL of 20 mM NaH_2PO_4 pH 7.0) were depleted on a 1 mL Bio-Rad Aurum Serum Protein (www.biorad.com) column according to the manufacturer's instructions. Protein concentrations were determined with the Micro BCA™ Protein assay reagent kit (www.piercenet.com) and calculated for an average protein molecular weight of 50 kDa. BSA was used as the calibration standard.

An equivalent of 100 µg total protein of depleted serum were digested with trypsin (1:20, w/w enzyme to substrate) at 37 °C overnight (sequencing grade modified trypsin from Promega, Cat# V5111, USA). One hundred micrograms of digested, depleted serum were fractionated by strong cation exchange HPLC (Poly SEA 2.0 mm × 150 mm, 5 µm, 300 Å column, Michrom BioResources, Auburn, CA, USA) operated at 0.3 mL/min on a Beckman Gold HPLC system (www.beckman.com). The salt gradient ranged from 0 to 1 M KCl with a slope of 10 mM/min. The mobile phase comprised two buffers: A: 5 mM KH_2PO_4/H_3PO_4 pH 3, 25% acetonitrile and B: 5 mM KH_2PO_4/H_3PO_4 pH 3, 25% acetonitrile, 1 M KCl. The following fractions were collected: #1 (0–50 mM) KCl, 17–21 min, #2 (50–100 mM) KCl, 22–26 min, #3 (100–150 mM) KCl, 27–31 min and #4 (150–200 mM) KCl, 32–36 min. Samples were concentrated by vacuum centrifugation (Speed Vac, Univapo 150H, UniEquip, Martinsried, Germany) to approximately 1 mL to remove acetonitrile and fractions #3 and #4 were pooled together because of their low concentration of peptides. The concentrated fractions were passed through a Macro Trap 200 hydrophilic C18 silica cartridge (ODS-AQ; 3 mm × 8 mm; Michrom, USA) for desalting and further concentration by loading at 0.3 mL/min followed by a column wash with 2 mL of 5% acetonitrile, 0.1% formic acid in H_2O. Peptides were eluted with 0.5 mL 70% acetonitrile, 0.1% formic acid at a flow rate of 1 mL/min. Acetonitrile was evaporated by vacuum centrifugation (Speed Vac) and the final volume was adjusted to 150 µL. Pure acetonitrile and pure formic acid were added to reach final concentrations of 5% acetonitrile and 0.1% formic acid, respectively. All LC–MS analyses were performed on an Agilent 1100 capillary HPLC system coupled on-line to an SL ion-trap (www.home.agilent.com) equipped with an Atlantis™ dC 18 (1.0 mm × 150 mm, 3 µm) column (www.waters.com). Forty microliters of the pretreated fractions corresponding to ~8 µg or 160 pmol of total protein digest (calculation based on a 50 kDa protein) were injected. Peptides were eluted in a linear gradient from 0 to 70% (0.5%/min) acetonitrile with 0.1% formic acid at a flow rate of 20 µL/min.

2.1.2. Method 2

Eighty microliters (80% of the total amount of diluted crude serum (20 µL of crude serum mixed with 80 µL of buffer A (Agilent)) were depleted on a Multiple Affinity Removal column (Agilent, 4.6 mm × 50 mm, Part # 5185–5984) according to the manufacturer's instructions.The flow-through fraction (depleted serum collected between 2 and 6 min) of a total volume of approximately 1 mL was collected. Protein concentrations were determined as in Method 1. One hundred microliters (~10% of the total amount, which corresponded to ~7 µg or 140 pmol of total protein considering a molecular weight of 50 kDa) of depleted serum were digested with trypsin (1:20, w/w enzyme to substrate) under the same conditions as described in Method 1. All LC–MS analyses were performed on the identical LC–MS system (Agilent 1100 capillary HPLC; SL ion-trap mass spectrometer) except that an in-line trap column was used (Atlantis™ dC 18, 3 µm, 2.1 mm × 20 mm Guard column (www.waters.com)). One hundred microliters depleted and digested serum were injected and washed in the back-flush mode for 40 min (0.1% aq. formic acid and 3% acetonitrile at a flow rate of 50 µL/min) and eluted on-line to the analytical column (Atlantis™ dC 18, 1.0 mm × 150 mm, 3 µm column (www.waters.com)). Gradient conditions were identical to Method 1.

2.2. Polyacrylamide gel electrophoresis (SDS–PAGE)

SDS–PAGE was performed in a Mini-Protein III cell (Bio-Rad, www.biorad.com) using 12% gels with 0.1% SDS according to the manufacturer's instructions. Staining was performed with Coomassie Brilliant Blue R concentrate (Sigma, www.sigmaaldrich.com) diluted and used as prescribed by the manufacturer.

2.3. Mass spectrometry

The following conditions were used for mass spectrometry during LC–MS. Nebulizer gas: 16.0 psi N_2, drying gas: 6.0 L/min N_2, skimmer: 40.0 V, cap. exit: 158.5 V, Oct. 1: 12.0 V, Oct. 2: 2.48 V, Oct. RF: 150 Vpp (Voltage, Peak Power Point), Lens 1 : −5.0 V, Lens 2: −60.0 V, Trap drive: 53.3, T°: 325°, Scan resolution: Enhanced, 5500 m/z per second scan speed. Target mass: 600. Scan range: 100–1500 m/z. Spectra were saved in centroid mode. LC–MS chromatographic data were analyzed with Bruker Data Analysis software, version 2.1 (Build 37).

2.4. Repeatability study

A serum sample from a cervical cancer patient was treated six times with Method 1 and five times with Method 2. In Method 2 crude sample of the same patient was also spiked with horse heart Cytochrome C (Sigma, www.sigmaaldrich.com) before depletion (210 pmol Cytochrome C per 20 µL of original serum) and the procedure repeated four times.

2.5. Standard addition of Cytochrome C

Horse heart Cytochrome C was added to 20 µL of the original serum over a range of 25 pmol–1.26 nmol, of which 10% were subjected to the final LC–MS analysis to evaluate the discriminatory capacity of Method 2. Cytochrome C was alternatively digested with trypsin and added in the same amounts to depleted and trypsinized serum (2 µL equivalent) prior to LC–MS to evaluate whether the depletion procedure affected the recovery of Cytochrome C.

2.6. Data analysis

2.6.1. Data pre-processing

The original Bruker Daltonics LC–MS data files were converted into ASCII-format with the Bruker Data Analysis software. The original m/z ratios (0.1 amu resolution) were combined into 1 amu bins by rounding the m/z ratios off to the closest integer values. This reduced the amount of data by almost a factor 10 but, more importantly, avoided misclassification of m/z traces due to the fact that centroid sampling of the original mass spectra introduced a slight error that could lead to misalignment of m/z traces. In order to classify the individual m/z traces with respect to their "information content", the component detection algorithm (CODA) developed by Windig et al. was applied [25]. This algorithm compares the raw chromatograms with their smoothed (using a moving average) and standardized versions. The difference between the raw and smoothed chromatogram is small for high-quality chromatograms (and the "CODA quality score" with a value ranging between 0 and 1, is high), while the opposite is true for chromatograms containing mainly background noise and/or spikes. By setting a threshold the user can define down to which level of quality m/z traces will be considered for the subsequent statistical analysis. In our study mass traces with a quality value higher than 0.98 were retained for further analysis meaning that only about 45 very-high-quality chromatograms were considered. The total ion current (TIC) was calculated from all mass traces or from the CODA-selected high-quality ones. The latter represented rather well the main characteristics of the original TIC, while background noise was essentially eliminated. Information of the peaks present in this rather conservative class of mass traces turned out to be sufficient for the detection of interpretable patterns in the data, for example, separation of spiked from non-spiked samples.

2.6.2. Multivariate statistical analysis

To perform multivariate statistical data analysis of multiple LC–MS samples, information present in the union set of all selected high-quality mass traces was used. For each high-quality mass trace in the union set of mass traces, the peak with the highest intensity was obtained and entered in a peak list. This peak list was further analyzed using principal component analysis.

PCA is a widely used statistical technique that enables the search for and visualization of patterns present in highly multivariate datasets [27]. In this study mainly biplots were used to analyze the available LC–MS peak lists. All statistical data analysis calculations were performed in the MATLAB programming environment (version 3.5.1, release 13).

2.7. Protein identification

1D nanoLC–ESI–MS–MS analysis was performed on an integrated nanoLC system (Agilent) comprising a binary gradient pump with a cooled autosampler, an auxilary pump for loading and washing the trap column, a column switching module configured for trap plus analytical capillary column, and a Q-Star XL API mass spectrometer (Applied Biosystems, MDS Sciex, Framingham, USA) fitted with nano-LC sprayer and operated under Analyst QS 1.1 control. Injected samples were first trapped and desalted isocratically on an LC-Packings PepMap C18 μ-Precolumn Cartridge (5 μm, 300 μm I.D. × 1 mm; Dionex, Sunnyvale, CA, USA) for 5 min with 0.1% formic acid delivered by the auxillary pump at 10 μL/min after which the peptides were eluted from the trap column and separated on an analytical C18 capillary column (5 cm × 75 μm, Atlantis) connected in-line to the mass spectrometer, at 250 nL/min using a 90 min gradient of 5–50% acetonitrile in 0.1% formic acid.

The QStar XL mass spectrometer was operated in information-dependent acquisition (IDA) mode. In MS mode, ions were screened from *m/z* 300 to 1500, and MS–MS spectra were acquired from *m/z* 50 to 2000 (pulsing mode on). In standard acquisition mode, each acquisition cycle was comprised of a 1s MS and a 1s MS–MS scan. In IDA mode, the four most intense peaks were selected and MS–MS spectra acquired when their intensities exceeded 30 counts. In product ion mode, MS–MS spectra were acquired for selected precursor ions (*m/z* 619.2, 694.4, 753.5, 682.5 and 909.6) without threshold restriction. Acquired MS–MS spectra were searched against the SwissProt/Trembl database with a mass tolerance of 1.1 Da for the precursor and 0.15 Da for the obtained fragment ions. A hit was considered significant when the score exceeded 2.0, which corresponds to a confidence interval of more than 99%.

3. Results

3.1. Preparation of human serum for LC–MS analysis

Depletion of high-abundance proteins is one way to enhance the capability of proteomic methods to detect subtle changes in the protein profile of human serum. Previously we reported on the efficacy of several depletion columns to remove albumin and γ-globulins [17]. Method 1 (Figure 1), used here as a first approach, is partially based on a previously published protocol, which was extended and optimized in the current work. Method 1 relies on dye-ligand affinity chromatography to remove albumin and Protein A to remove IgG.

The subsequent steps comprise trypsin digestion and strong cation exchange chromatography (SCX) to fractionate the sample prior to LC–MS analysis. This method was repeated six-times in order to evaluate the overall repeatability. To this end 10 endogenous peptides from three cation-exchange fractions that covered the entire retention time range of eluting peptides in serum were selected and their respective extracted ion chromatograms integrated. While repeatability in terms of retention times was satisfactory (RSD < 0.8%), peak areas differed over a wide range (RSD between 12 and 160%) (see Table 1).

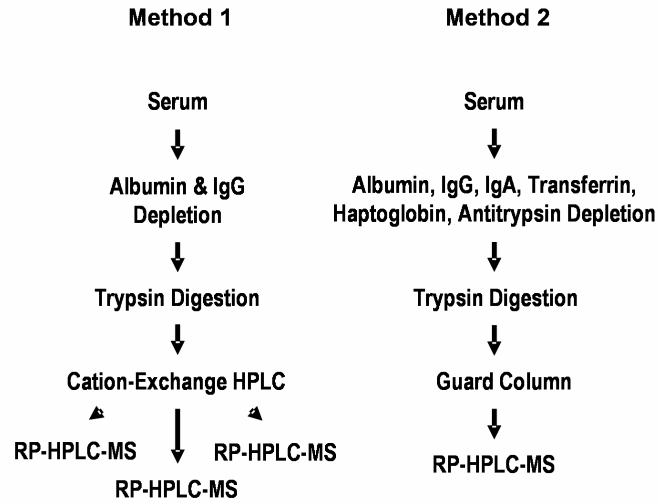

Figure. 1. Schematic description of two methods for sample preparation and analysis of human serum. The main differences between the methods are that Method 1 uses dye ligand affinity chromatography and Protein A for depletion of albumin and IgG (Aurum column, BioRad) while Method 2 employs a multiple affinity removal column (Agilent) based on a mixture of antibodies and Protein A. Method 1 includes a strong cation-exchange prefractionation step, while this step was omitted in Method 2.

We argued that the low repeatability with respect to peak area of some peptides was due to the cation-exchange prefractionation step, since fraction collection from a highly complex chromatogram of partially resolved peaks can easily lead to arbitrary cutting of component peaks and thus to major quantitative differences. This effect can be depicted in Figure 2A showing six repetitions of the final LC–MS analysis using Method 1 clearly indicating two groups of three chromatograms. In order to overcome this limitation, we improved the efficiency of the depletion step by employing a multiple-affinity removal column (removes albumin, IgG, IgA, transferrin, haptoglobin, and α1-antitrypsin) and eliminating the cation-exchange chromatography step (Figure 1, Method 2). Visual inspection of the chromatograms showed already that this approach was superior in terms of repeatability (Figure 2B).

Neither retention times nor peak areas were normalized. In bold (Table 1): Peaks detected in both methods (*m/z* 619.2 and 694.4: doubly charged ions of peptides [DLATVYVDVDVLK and VSFLSALEEYTK, respectively] from apolipoprotein A (accession number (P02647, SwissProt/Trembl); *m/z* 682.5: triply charged ion of [LLLQQVSLPELPGEYSMK] and *m/z* 753.5: doubly charged ion of [AAQVTIQSSGTFSSK] from alpha-2-macroglobulin precursor (P01023); *m/z* 909.6: doubly charged ion of [SNLDEDIIAEENIVSR] from human complement C3 precursor (P01024).

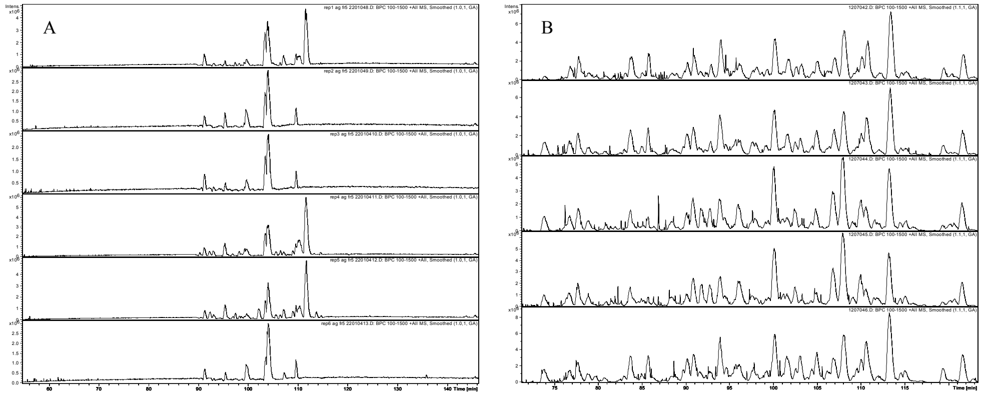

Figure. 2. LC–MS analyses of trypsin-digested serum samples prepared by Method 1 (A) or Method 2 (B). Six repetitions of fraction 2 of the strong cation-exchange HPLC pre-fractionation step (Method 1) are shown in comparison with five repetitions of Method 2.

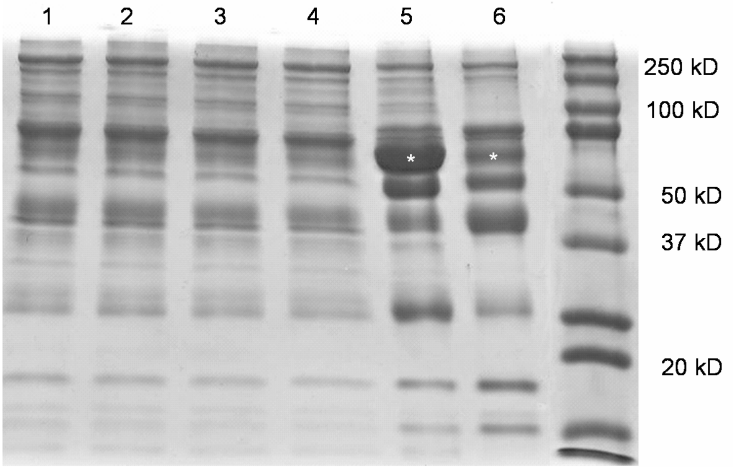

Figure. 3. SDS–PAGE analysis of human serum prior to (lane 5) and after depletion using the multiple affinity removal column (lanes 1–4) or dye ligand/Protein A affinity chromatography (lane 6). Albumin is labelled with an asterisk.

Table 1.Repeatability of Method 1 (six repetitions) and Method 2 (five repetitions) (see Figure 2) in terms of retention time and peak area for a selected number of endogenous peptides

m/z	**Average RT (min)**	**%RT dev.**	**Average area**	**% Area dev.**
Method 1. CEX fraction 1				
772.1	100.72	0.25	4.9×10^6	36.8
472.2	106.20	0.23	3.0×10^7	85.5
552.9	108.95	0.79	9.7×10^5	104.8
713.8	111.62	0.24	5.4×10^7	50.0
682.5	**115.25**	**0.26**	$\mathbf{7.3 \times 10^6}$	**157.5**
Method 1. CEX fraction 2				
619.2	**99.75**	**0.05**	$\mathbf{2.1 \times 10^7}$	**24.0**
694.4	**103.95**	**0.08**	$\mathbf{4.1 \times 10^7}$	**11.9**
753.5	**111.62**	**0.10**	$\mathbf{9.9 \times 10^7}$	**113.2**
909.6	**114.37**	**0.12**	$\mathbf{2.5 \times 10^6}$	**94.1**
Method 1. CEX fraction 3				
525.1	99.34	0.50	2.2×10^7	73.8
Method 2.				
756.7	109.40	0.06	4.5×10^7	13.1
753.5	**110.78**	**0.04**	$\mathbf{3.3 \times 10^7}$	**13.7**
909.6	**113.54**	**0.08**	$\mathbf{3.8 \times 10^6}$	**27.5**
682.5	**114.20**	**0**	$\mathbf{8.3 \times 10^7}$	**14.1**
619.2	**99.90**	**0.05**	$\mathbf{4.4 \times 10^7}$	**20.1**
694.4	**104.20**	**0.04**	$\mathbf{2.4 \times 10^7}$	**9.1**

An overview of the efficiency of depletion by both methods was obtained by SDS–PAGE showing that Method 2 (multiple depletion) was more efficient in depleting high-abundant serum proteins (Figure. 3).

This is in agreement with recently published data using 2D gel electrophoresis [28]. While Method 1 removed approximately 70% of total serum protein, Method 2 removed 90–95% according to determination of the total protein content after depletion. This allowed an increased loading capacity of the remaining digested proteins of 10–20-fold on the reversed-phase column.

3.2. Repeatability

As indicated in Figure 2, Method 2 resulted in a better repeatability in terms of retention times (Method 1: <0.8% RSD; Method 2: <0.1% RSD). This may be due to the use of an in-line trap column for sample clean-up and focusing. It is noteworthy that retention time differences of five repetitive LC–MS runs ranged from 0 to 6 s in the case of Method 2 without any alignment of the chromatograms. Importantly, Method 2 proved also to be significantly more repeatable with respect to the observed peak areas for identical, selected endogenous peptide peaks (Method 1: 12–160% RSD; Method 2: 10–30% RSD). This was particularly for those peptides that showed large standard deviations with Method 1 (Table 1). This may be attributed to the significantly reduced number of sample preparation steps and in particular elimination of the fraction collection step by strong cation-exchange HPLC.

To estimate the repeatability of the developed methods on a more global scale, CODA was applied to all measured replicates. So instead of limiting ourselves to 6–10 univariate repeatability measures, all CODA-selected mass traces were used to compare the methods. Figure 4 displays the calculated CODA quality scores for the different mass traces of the replicates prepared according to sample preparation Method 1 (top) or Method 2 (bottom).

Since the image plots of the first, second and third SCX fraction gave similar results, only the quality values of the second fraction are visualized. While for Method 2 the quality values of the different replicates are higher and very similar over the whole m/z range, Method 1 shows far less repeatable quality values. Replicates 2, 3 and 6 (Method 1) differ significantly from replicates 1, 4 and 5 (see also Figure 2). This result is confirmed after application of PCA to the peak lists generated from the high quality mass traces (CODA score > 0.98) selected by CODA. In the scores plot (not shown here) the replicates separate into two groups: a group containing replicates 2, 3 and 6 and a group with replicates 1, 4 and 5. Such analytical variability may interfere with detecting patterns of samples when analyzing patient sera or indicate false clusters.

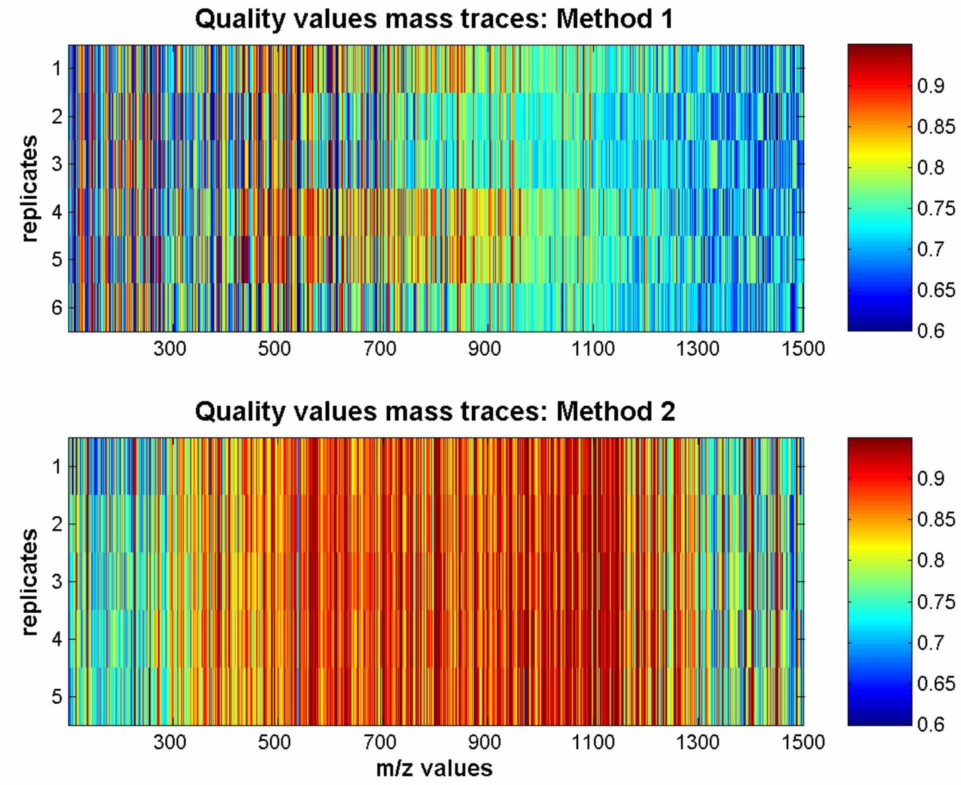

Figure. 4. Image plots showing the colour-coded CODA quality scores of the different mass traces (*m/z* 100–1500). High-quality values are red (in web version) and low-,quality values are blue (in web version). The image plot at the top displays quality values for the 6 replicates (the second fraction is displayed) using Method 1 (see Figure 2A). The image plot at the bottom shows the quality values of 5 replicates measured using sample preparation Method 2 (see Figure 2B).

3.3. Data analysis

In view of trying to discover differences in the abundance of peptides amongst samples from cancer patients at various stages of disease, it is pivotal that baseline variations are kept to a minimum due to methodological variability. Thus, Method 2 was chosen for further work. In an effort to evaluate the discriminatory capacity of the analytical and data analysis methodology, 210 pmol of horse heart Cytochrome C were added to 20 μL of the original serum sample and analyzed with Method 2 as described above. LC–MS data was acquired for in total four replicate samples spiked with Cytochrome C and five non-spiked replicates. Peak lists for all 9 LC–MS data files were obtained after application of the binning algorithm and usage of CODA. Figure 5 summarizes the results of PCA of the union peak list. The biplot shows besides the scores of the nine samples (4 spiked samples = red diamonds, five normal samples = blue squares) also the loadings of the high quality mass traces (45 mass traces = green triangles) that were used for the statistical analysis. In Figure 5A the spiked samples are clearly separated from the normal (non-spiked) samples. Especially *m*/*z* traces 483, 748 and 965 contribute significantly to the discrimination of the spiked from the normal samples (Figure 5B). All of these *m*/*z* values are related to tryptic peptides of Cytochrome C (the *m*/*z* values of 483 and 965 correspond to doubly and singly charged ions of the same peptide).

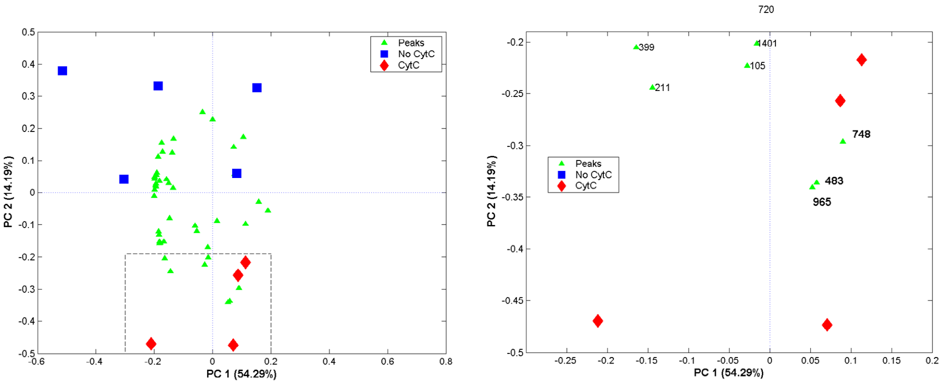

Figure 5. (A) Principal component analysis (PCA) biplot of trypsin digested, depleted human serum samples with 210 pmol spiked horse heart Cytochrome C in 20 µL serum (red diamonds) and without Cytochrome C (blue squares) after selection of high quality mass traces with CODA using a threshold of 0.98. Note that the recovery of Cytochrome C after depletion is only 20–25% (see Table 2). The used high quality mass traces are indicated by the green triangles. (B) Enlarged part of the boxed area in (A) showing that peaks at *m*/*z* 483, 965 and 748 contribute strongly to the discrimination between spiked and non-spiked. All of these peaks are derived from Cytochrome C.

Since depletion of high-abundance proteins, notably albumin, has been reported to lead to co-depletion of other proteins, peptides and metabolites that are bound to albumin [21,22], we investigated the recovery of the added Cytochrome C from 20 µL serum of a cervical cancer patient (same patient serum as used before) after depletion. For comparison, an equal amount of a tryptic digest of Cytochrome C was added to the depleted and digested sample just prior to LC–MS analysis. Both analyses were repeated nine-times and the average peak areas for five selected *m*/*z* traces from peptides related to horse heart Cytochrome C were compared. A comparison of peak areas of Cytochrome C added to the original serum or added just prior to LC–MS analysis, revealed that 19–27% of spiked Cytochrome C were recovered following multiple affinity removal of high-abundance proteins (Table 2). Recoveries on the same order were also found when 10–63 pmol of Cytochrome C were spiked into 20 µL serum. The repeatability of the peak areas of spiked Cytochrome C was between 12 and 26% RSD (not shown) and thus within the same range as relative standard deviations for other endogenous peptide peaks (see Table 1). Repeatability of the initial depletion step itself was assessed to be better than 5% RSD in terms of peak area of the flow-through fraction of the multiple affinity removal column.

Table 2. Recovery of horse heart Cytochrome C (Method 2) added to human serum prior to depletion (210 pmol in 20 µL serum) of high-abundance proteins using the multiple affinity removal column (Agilent; depletes albumin, IgG, IgA, transferrin, haptoglobin, and α_1-antitrypsin) based on extracted ion chromatograms (results based on 9 independent experiments with and without addition of Cytochrome C)

m/z	**Peak area**	**Recovery %**	**RSD % (*n* = 9)**
302.7	1233449	19.1	6.1
604.4	12538882	27.3	8.6
482.8	27618370	26.0	10.8
748.4	39544944	26.4	7.0
1495.7	504101	20.3	7.5
		Average recovery 23.8	Average RSD 8.0

To investigate whether loss of Cytochrome C was due to direct binding to the affinity column or mediated through binding to high-abundance proteins, similar amounts of Cytochrome C were dissolved in Buffer A and applied to the multiple affinity removal column. Practically all Cytochrome C was recovered showing clearly that binding to high-abundance proteins, most likely albumin, was responsible for the loss. Therefore, the actual amounts of Cytochrome C applied to LC–MS analysis in the comparative spiking studies shown in Figure 5 were about five times lower than the amount originally added to the serum sample. These results confirmed the observation that high-abundance proteins may act as “molecular sponges”, which bind and transport low molecular weight proteins or peptides, an effect that needs to be corrected for when using spiked internal standards.

In an effort to evaluate whether Method 2 in combination with CODA and PCA could detect a trend in concentration of a protein in serum, different amounts of horse heart Cytochrome C (25 pmol–1.26 nmol) were added to 20 µL of the original serum sample before multiple depletion and 10% of the tryptic digest was subjected to LC–MS as described above. In Figure 6 multivariate statistical data analysis (PCA) results are displayed of LC–MS datasets treated with CODA to generate peak lists based on high-quality mass traces (CODA score > 0.98). These results show that principal component 1 (PC1), which describes about 89% of the variability in the data, follows the concentration trend generated by the added Cytochrome C down to a level of 5 pmol in the equivalent of 2 µL serum (recalculated for an injection of 10% of the original sample). Variability reflected in PC2 (approximately 7%) is not correlated with the concentration of Cytochrome C. The serum sample containing 2.5 pmol follows the trend described by PC1 but is separated from the other samples mainly in PC2 most likely due to variability introduced by the analytical procedure. PCA analysis showed again that traces at *m/z* 483 and 748, corresponding to peptides derived from Cytochrome C, contributed most significantly to the observed trend (see Figure 6). Method 2 combined with *m/z* trace selection by CODA and PCA thus correctly identified 2 *m/z* values that correlated with the observed trend in PC1 and the introduced concentrations of Cytochrome C added to the serum.

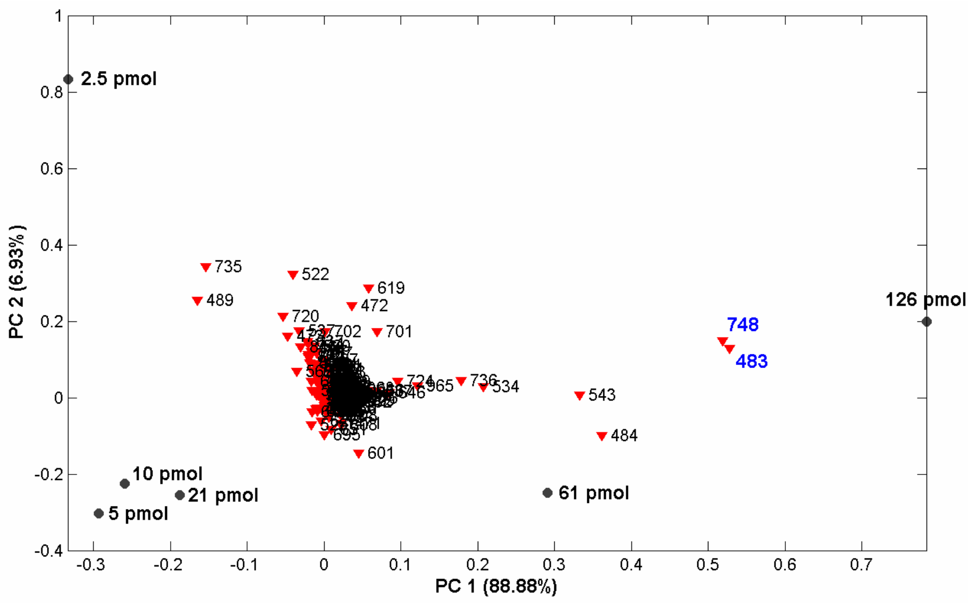

Figure 6. Principal component analysis (PCA) biplot of human serum samples spiked with increasing amounts of horse heart Cytochrome C (25, 50, 100, 210, 610 and 1260 pmol in 20 µL original serum, recalculated for an injected amount corresponding to 10% of the original 2µL). Peaks (*m/z* values) used for PCA (selected by CODA) are shown as red triangles. Peaks with *m/z* 483 and 749 contribute strongly to the observed trend in PC1. Both peaks are derived from Cytochrome C.

In order to evaluate whether there is a good correlation between the detected m/z traces and the amount of added Cytochrome C, a linear correlation analysis was performed showing that the peak areas of the extracted ion chromatograms at 483 and 748 amu correlated well with the added amounts over the studied concentration range (R^2 = 0.97–0.99).

4. Discussion

A procedure for the depletion of high-abundance proteins from human serum for subsequent analysis by shotgun proteomics using LC–MS has been described. The overall procedure showed a repeatability of 10–30% for peak area and better than 0.1% for retention times without the use of internal standards or alignment of the chromatograms. Data were pre-processed using CODA at a quality score threshold of 0.98 that selects about 45 of the most "information rich" m/z traces. This allowed the reconstruction of a TIC which is highly similar to the original raw data, though, with drastically reduced background noise. PCA of pre-processed LC–MS datasets obtained after spiking different amounts of horse heart Cytochrome C into the original serum (range 25 pmol–1.26 nmol in 20 µL serum) revealed the trend in Cytochrome C concentrations in Principal Component 1, which described 89% of the variability in the data. Determination of recoveries for spiked Cytochrome C after depletion using a multiple affinity removal column depleting albumin, IgG, IgA, transferrin, haptoglobin, and α_1-antitrypsin showed that only 20–25% of the added protein were recovered and that this loss was due to the presence of high-abundance proteins and not to direct binding to the affinity column. The effect of protein co-depletion needs therefore to be taken into account when adding internal standards to serum and likely also to other complex biological samples. Our data indicate that internal standardization using added marker peptides is a viable alternative to stable-isotope labeling for quantitative, comparative proteomics but that care needs to be taken to account for limited recoveries. Operating according to a strictly standardized procedure is mandatory. Other groups have also shown that reliable quantitative results in shotgun proteomics of complex peptide mixtures can be obtained based on reproducible peak areas [29,30]. However, repeatability may still be improved using the so-called "Global Internal Standard" strategy based on stable-isotope labeled samples [31], an approach that is under investigation.

While our approach shows the feasibility of combining an LC–MS-based method including depletion of high-abundance proteins and trypsin digestion with data pre-processing and multivariate statistics to discover trends in concentrations of proteins in complex mixtures like serum, the concentration sensitivity is not sufficient to reach into the range of known tumor markers (ng/mL range). Improvements will therefore have to be made in sample preparation with the goal of being able to treat a larger volume of serum. Our analysis eventually used only 2 µL of the original serum sample, while it is possible to obtain 1mL without difficulty. Due to the high remaining protein content even after depletion (approximately 4–5 mg/mL), it is pivotal to start with a "preparative" separation method and to analyze the prefractionated sample. We are presently investigating prefractionation strategies at the protein level prior to tryptic digestion to reach a better concentration sensitivity. The described method is currently being applied to comparative cross-sectional and longitudinal studies with samples from healthy subjects and cervical cancer patients at different stages of disease to evaluate its suitability to classify different groups of patients.

References

[1]. Hanash S, Brichory F, Beer D A proteomic approach to the identification of lung cancer markers. Dis. Markers 2001;17(4):295-300.

[2]. Paweletz CP, Trock B, Tsangaris M, Magnant C, Liotta LA., et al., Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer. Dis. Markers 2001;17(4):301-307.

[3]. Wisman GB, Knol AJ, Helder MN, Krans M, de Fries EG, Hollema H, de Jong S, van derZee AGJ. Telomerase in relation to clinicopathologic prognostic factors and survival in cervical cancer. Int. J. Cancer 2001;91:658-664.

[4]. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572-577.

[5]. Rai AJ, Zhang Z, Rosenzweig J, Shih IeM, Pham T, Fung ET, Sokoll LJ, Chan D.W. Proteomic approaches to tumor marker discovery. Arch. Pathol. Lab. Med. 2002;(126):1518-1526.

[6]. Ardekani AM, Liotta LA, Petricoin EF 3rd. Clinical potential of proteomics in the diagnosis of ovarian cancer. Expert. Rev. Mol. Diagn. 2002;2:312-320.

[7]. Grossklaus DJ, Smith JA, Shappell SB, Coffey CS, Chang SS, Cookson MS. The free/total prostate-specific antigen ratio (%fPSA) is the best predictor of tumor involvement in the radical prostatectomy specimen among men with an elevated PSA. Urol. Oncol. 2002;(7):195-198.

[8]. Whitehouse C, Solomon E. Current status of the molecular characterization of the ovarian cancer antigen CA125 and implications for its use in clinical screening. Gynecol. Oncol. 2003;88:152-157.

[9]. DeSouza L, Diechl G, Rodrigues MJ, Guo J, Romaschin AD, Colgan TJ, Siu KWM. Search for cancer markers from endometrial tissues using differentially labeled tags iTRAQ and cICAT with multidimensional liquid chromatography and tandem mass spectrometry. J. Proteome Res. 2005;4:377-386.

[10]. Koomen JM, Li D, Xiao L-ch, Liu TC, Coombes KR, Abbruzzese J, Kobayashi R. Direct tandem mass spectrometry reveals limitations in protein profiling experiments for plasma biomarker discovery. J Proteome Res. 2005;4:972-981.

[11]. Fortier M-H, Bonneil E, Goodley P, Thibault P. Integrated microfluidic device for mass spectrometry-based proteomics and its application to biomarker discovery programs. Anal. Chem. 2005;77:1631-1640.

[12]. Tammen H, Schulte I, Hess R, Menzel C, Kellmann M, Mohring T, Schulz-Knappe P. Peptidomic analysis of human blood specimens: comparison between plasma specimens and serum by differential peptide display. Proteomics 2005;5:3414-3422.

[13]. Jacobs JM, Adkins JN, Qian WJ, Liu T, Shen Y, Camp DG, Smith RD. Utilizing human blood plasma for proteomic biomarker discovery. J. Proteome. Res. 2005;4:1073-1085.

[14]. Petricoin EF, Liotta LA. SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. Curr. Opin. Biotechnol. 2004;15:24-30.

[15]. Diamandis EP. OvaCheck: Doubts Voiced Soon After Publication. Nature 2004;430:611.

[16]. Diamandis EP. Proteomic Patterns to Identify Ovarian Cancer: 3 Years on. Expert. Rev. Mol Diagn. 2004;4:575-577.

[17]. Govorukhina NI, Keizer A, van der Zee AGJ, de Jong S, de Bruijn HW, Bischoff R. Sample preparation of human serum for the analysis of tumor markers. Comparison of different approaches for albumin and gamma-globulin depletion. J. Chromatogr. A. 2003;1009:171-178.

[18]. Gianazza E, Arnaud P. Chromatography of plasma proteins on immobilized Cibacron Blue F3-GA. Mechanism of the molecular interaction. Biochem. J. 2001;1982:129-136.

[19]. Akerstrom B, Brodin T, Reis K, Bjorck LJ. Protein G: a powerful tool for binding and detection of monoclonal and polyclonal antibodies. Immunology 1985;135:2589-2592.

[20]. Guss B, Eliasson M, Olsson A, Uhlen M, Frej A, Jornvall H, Flock U, Lindberg M. Structure and evolution of the repetitive gene encoding streptococcal protein G. EMBO J. 1986;5:1567-1575.

[21]. Tirumalai RS, Cha KC, Prieto DA, Issaq HJ, Conrads TP, Veenstra TD. Characterization of the low molecular weight human serum proteome. Mol. Cell. Proteomics 2003;2:1096-1103.

[22]. Zhou M, Lucas DA, Chan KC, Issaq HJ, Petricoin 3rd EF, Liotta L.A., Veenstra TD, Conrads TP. An investigation into the human serum "interactome". Electrophoresis 2004;25:1289-1298.

[23]. Washburn MP, Wolters D, Yates JR 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat. Biotechnol. 2001;19:242-247.

[24]. Martosella J, Zolotarjova N, Liu H, Nicol G, Boyes BI. Reversed-phase high-performance liquid chromatographic prefractionation of immunodepleted human serum proteins to enhance mass spectrometry identification of lower-abundant proteins. J. Proteome Res. 2005;4:1522-1537.

[25]. Windig W, Meuzelaar HL. Nonsupervised numerical component extraction from pyrolysis mass spectra of complex mixtures. Anal Chem. 1984 Nov;56(13):2297-2303.

[26]. Esajas MD, Duk JM, de Bruin HW, Aalders JG, Willemse PH, Sluiter W, Pras B, ten Hoor K, Hollema H, van der ZeeAJG. J. Clinical value of routine serum squamous cell carcinoma antigen in follow-up of patients with early-stage cervical cancer. Clin. Oncol. 2001;19:3960-3966.

[27]. Jackson JE. A User's Guide to Principal Components. Wiley, New York (1991).

[28]. K. Bjorhall, T. Miliotis and P. Davidsson, Proteomics 2005;5:307-317.

[29]. Curry S. Beyond expansion: structural studies on the transport roles of human serum albumin. Vox Sang. 2002;83:315-319.

[30]. Wiener MC, Sachs JR, Deyanova EG, Yates NA. Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures. Anal. Chem. 2004;76:6085-6096.

[31]. Chakraborty A, Regnier FE. Global internal standard technology for comparative proteomics. J. Chromatogr. A. 2002;949:173-184.

Chapter III.

Chip-LC-MS for label-free profiling of human serum

P. Horvatovich, N.I. Govorukhina, T.H. Reijmers, A.G.J. van der Zee, F. Suits, R. Bischoff

Submitted for publication for a special issue of Electrophoresis on Biomarker Discovery (accepted)

Abstract

The discovery of biomarkers in easily accessible body fluids such as serum is one of the most challenging topics in proteomics, requiring highly efficient separation and detection methodologies. Here we present the application of a microfluidics-based LC-MS system (chip-LC-MS) to the label-free profiling of immunodepleted, trypsin-digested serum in comparison to conventional capillary LC-MS (cap-LC-MS). Both systems proved to have a repeatability of ~20% relative standard deviation (RSD) for peak area, all sample preparation steps included, while repeatability of the LC-MS part by itself was less than 10% RSD for the chip-LC-MS system. Importantly, the chip-LC-MS system had a two fold higher resolution in the LC dimension and resulted in a lower average charge state of the tryptic peptide ions generated in the electrospray ionization (ESI) interface when compared to cap-LC-MS.

In order to characterize both systems for their capability to find discriminating peptides in trypsin-digested serum samples, five out of ten individually prepared, identical sera were spiked with horse heart Cytochrome C. A comprehensive data processing methodology was applied, including two-dimensional smoothing, resolution reduction, peak picking, time alignment and matching of the individual peak lists to create an aligned peak matrix amenable for statistical analysis. Statistical analysis by supervised classification and variable selection showed that both LC-MS systems could discriminate the

two sample groups. However, the chip-LC-MS system allowed the assignment of 55% of the overall signal to selected peaks, whereas this was only 32% for the cap-LC-MS system.

1. Introduction

Biomarker discovery through proteomics requires an intricate collaboration between medical, analytical and data sciences [1]. From an analytical point of view, it is a complex procedure and success is dependent on accurate quantification and the precise assignment of properties, such as retention time and m/z ratio, to individual molecules in highly diverse samples, while maintaining a comprehensive view of their compositions. Most biomarker discovery studies use easily accessible body fluids like urine, plasma or serum as starting materials. These samples are, however, extremely complex and cover a wide concentration range, which for serum can reach 10-12 orders of magnitude [2].

Fast analytical screening methods such as surface enhanced laser desorption ionization time-of-flight mass spectrometry (SELDI-TOF-MS), which combines on-chip sample preparation with mass spectrometric analysis, are frequently used in biomarker research [3], although some recent results question the reliability of this approach [4,5]. Present analytical research is trying to meet the challenge of more precise quantification and the coverage of a wider concentration range. The accuracy of relative quantification may be improved by stable-isotope labeling techniques, which have the advantage that the labeled proteins are mixed and subjected to the same sample preparation procedure and final LC-MS analysis, thus compensating for the variance occurring during these analytical steps. For *in vitro* cell culture experiments, amino acids labeled with different stable-isotopes (SILAC) are introduced through metabolic labeling [6,7], a procedure that is not applicable to most clinical samples such as body fluids. Chemical labeling provides an alternative option to introduce stable-isotopes into proteins. ICAT™ and iTRAQ™ are examples of two widely used reagents that react with thiol- or primary amine functionalities [8-10].

Label-free proteomics is frequently applied in biomarker research, since large numbers of samples need to be analyzed, increasing the cost of stable-isotope labeling. For label-free profiling, it is preferable to first use the MS-mode over the more commonly employed different information-dependent MS/MS techniques. Profiling in MS-mode results in more accurate quantification of minor peaks, because present-day mass analyzers are not fast enough to measure all the components of a complex mixture during information-dependent MS/MS routines. In cases where the sample is first digested with trypsin followed by reverse-phase HPLC coupled on-line to mass spectrometry (LC-MS) (the so-called "shotgun" approach [11-13]), nano-LC systems with columns of 50-100 µm inner diameter (ID) are standard for protein identification [14]. However, they have rarely been employed in large-scale biomarker discovery studies, due to their lack of robustness and poor repeatability over long time periods. Because of these shortcomings, nano-LC-MS has found its main application in situations where only a few runs need to be compared, such as for protein identification from minute sample amounts. As a result, most biomarker discovery studies use columns of 0.3 to 1 mm ID (flow rates of 1-50 μL/min) for the label-free profiling of complex biological fluids [11,16].

The reliability of nano-LC-MS systems may be improved by integrating the trap and separating columns in a microfludics device (chip-LC-MS) [17], thus avoiding connections of varying dead volumes and the risk of leakage. Such devices also allow the integration of the electrical contact point into the chip, thereby avoiding the need for metal-coated spray needles which have limited long-term stability. In the present study, we show that chip-LC-MS compares favorably with a previously used capillary LC-MS (cap-LC-MS) system (1 mm ID column) [16] in terms of repeatability for both retention time and peak area. Human serum samples, depleted of the six most abundant proteins using a commercially available affinity column followed by trypsin digestion, were used to evaluate the chip-LC-MS system. The discrimination power of the system was compared with the cap-LC-MS system by adding horse heart Cytochrome C to the serum. The data generated by chip-LC-MS were treated in a comprehensive manner through automatic data processing and statistical analysis, revealing that chip-LC-MS is well suited for comparative, label-free profiling of highly complex samples. Moreover discrimination between samples of variable compositions was possible.

2. Experimental

2.1. Chemicals

Acetonitrile HPLC-S (ACN) gradient grade was from Biosolve (Valkenswaard, The Netherlands), formic acid (FA) 98-100% pro analysis was a product of Merck (Darmstadt, Germany) and ultra-pure water (conductivity: 18.2 MΩ) was obtained from a Maxima System (Elga Labwater, Ede, The Netherlands). Equine heart Cytochrome C was from Fluka (cat n° 30397, primary accession number in Swiss Prot database P00004, Buchs, Switzerland). The Cytochrome C standard protein solution with a concentration of 42 nmol/mL was stored at -20 ºC and prepared in 0.1% aqueous formic acid (FA). Buffers A and B used for immunoaffinity depletion were products of Agilent (cat nº 5185-5987 and 5185-5988, Palo Alto, California, USA).

2.2. Sample Preparation

Serum samples were obtained from the Department of Gynecological Oncology (University Medical Center, Groningen, The Netherlands) and stored at −80°C in aliquots. All intermediate fractions were stored at −20°C. To develop and optimize the analytical procedure, serum from one cervical cancer patient was used.

20 µL of serum were mixed with 80 µL of Buffer A and 80 µL of the diluted serum were depleted on a Multiple Affinity Removal column (Agilent, 4.6 mm×50 mm, cat. n° 5185–5984; 1 mL/min flow rate) on an Akta FPLC system with cooled autosampler (4°C) and fraction collector according to the manufacturer's instructions [20]. The flow-through fraction of a total volume of ~1 mL was collected. Protein concentrations were determined with the Micro BCA™ protein assay reagent kit (Pierce, cat. n° 23235, Illinois, USA) and calculated for an average protein molecular weight of 50 kDa. BSA supplied with the reagent kit was used as calibration standard. 330 µL(~33% of the total amount) of depleted serum were digested with trypsin (1:20, w/w enzyme to substrate) at 37°C overnight (sequencing grade modified trypsin from Promega, cat. n° V5111, Madison, Wisconsin, USA). When Cytochrome C was spiked into serum, its recovery was shown to be 23.8±8.0% after the depletion [16].

2.3. HPLC-MS

2.3.1. Cap-LC-MS

Two HPLC-MS systems (Agilent) were used for analysis of the depleted, trypsin-digested serum. One was equipped with a conventional capillary column, the second with a chip-cube interface. Using the conventional capillary system, 7 µg (15-20 µL) (equivalent to 140 pmol of total protein assuming a molecular weight of 50 kDa) of protein were analyzed with an Agilent 1100 series HPLC system containing an Atlantis™ dC 18 in-line trap column (Waters, Milford, Massachusetts, USA, 3 µm, 2.1 mm × 20 mm guard column, cat. n° 186001381) embedded in a Universal Sentry Guard Holder assembly (Waters, cat. n° 186000262) and an analytical capillary column of the same material (Waters, 1.0 mm × 150 mm, 3 µm, cat. n° 186001283) coupled on-line to an MSD-Trap-SL ion-trap mass spectrometer (cat. n° G2445A). The autosampler (cat. n° G1367A) was equipped with a 100 µL injection loop and a thermostated cooler (cat. n° G1330A) maintaining the samples at 4°C. The HPLC system had the following additional modules: capillary pump (cat. n°, G1376A), solvent degasser (cat. n°, G1379A), UV detector (cat. n° G1314A) and column holder (cat. n°, G1316A). The sample was injected and washed in the back-flush mode for 30 min (0.1% aq. FA and 3% acetonitrile at a flow-rate of 50 µL/min). For elution of the retained peptides to the analytical column, eluent A (0.1% aq. FA) and B (acetonitrile with 0.1% FA) were used at a flow rate of 20 µL/min as follows: eluent A (15 min); linear gradient from 0 to 50% eluent B at 0.5%/min and then from 50 to 70% at 1%/min. The column was finally washed with 70% eluent B for 10 minutes. After each injection, the in-line trap and the analytical column were equilibrated with eluent A for 20 min prior to the next injection.

2.3.2. Chip-LC-MS

The nano-LC-MS system was equipped with a microfluidics (chip-cube) interface (cat. n° G4240A) including a chip (cat. n° G4240-62001) with a 40 nL trap column (75 µm × 11 mm) and a 75 µm × 43 mm analytical column both containing C-18SB-ZX 5 µm chromatographic material. The interface contains a nanoelectrospray tip (2 mm length with conical shape: 100 µm OD × 8 µm ID) and was coupled on-line to an MSD-Trap-SL ion-trap mass spectrometer. The injected sample was 0.25 µg (3.39-5.05 µL) (5 pmol) of depleted, trypsin-digested serum diluted ten fold with 0.1% aq. FA. Injections were performed with an autosampler (Agilent, cat. n° G1389A) equipped with an injection loop of 8 µl (this includes the dead volume up to the trapping column) and a thermostated cooler maintaining the samples in the autosampler at 4°C during the analysis. The interface was connected to an Agilent 1100 series HPLC system containing the following additional modules: nanopump (cat. n° G2226A), capillary pump and solvent degasser. The sample was injected and washed in the back-flush mode for 4 min (0.1% aq. FA, 2 µL/min) and then the trapping column was switched in-line with the analytical column. For these separations the same eluents A and B as for the cap-LC-MS system were used at a flow rate of 0.3 µL/min. After elution for 6 minutes with eluent A, a linear gradient from 0 to 50% eluent B at 0.5 %/min followed by a gradient from 50 to 70% at 1 %/min of eluent B was run. 70% eluent B was maintained for 10 min. After each injection the in-line trap and the analytical column were equilibrated with eluent A for 20 min at 2 and 0.3 µL/min, respectively.

2.3.3. Mass spectrometry settings

The following settings were common for both systems for ESI and mass analysis on the LC/MSD-Trap-SL mass spectrometer: Drying gas: 6.0 L/min N_2, Skimmer: 40 V, Cap. exit: 136.0 V, Oct. 1: 12.0 V, Oct. 2: 1.74 V, Oct. RF: 150.0 V, Lens 1: -5.0 V, Lens 2: -60.0 V, Trap drive: 53.3, T°: 325 °C, Scan resolution: enhanced, 5500 m/z per second scan speed, ICC target: 30 000, max. accumulation time: 15 000 µs. Scan range: 100-1500 m/z. Each scan was filtered by a Gaussian acquisition filter (width 0.1 m/z). Rolling average: average of 2 spectra. Spectra were saved in profile mode. For cap-LC-MS, a nitrogen nebulizer gas pressure of 2.0 psi and an ionization voltage of 3500V were used. For the chip-LC-MS, the ionization voltage ranged between 1800 and 2000V.

2.4. Data analysis and multivariate statistics

2.4.1. Data analysis for method evaluation

LC-MS data were first analyzed with the data analysis software of the LC/MSD Trap, version 3.3 (build 146) (Bruker Daltonics, Bremen, Germany).

2.4.2. Data (pre-)processing

For (pre-)processing and multivariate statistical analysis of the original Bruker Daltonics HPLC-MS data, the files were converted into ASCII-format with the Bruker data analysis software and saved in centroid mode. For further data analysis Matlab (version 7.2.0.232 (R2006a), Mathworks, Natick, Massachusetts, USA) and the PLS toolbox (version 3.5.2, Eigenvector Research Inc., Wenatchee, Washington, USA) were used.

Centroid data were smoothed and reduced using a normalized two-dimensional Gaussian filter, with rounding of the nominal m/z ratios to 1 m/z (the original data had a resolution of 0.1 m/z). In the retention time dimension, no data reduction was performed. This meshing procedure reduced the number of available data points by roughly a factor 10 and corrected for shifting m/z values as a result of different loadings of the ion-trap during elution of abundant peptides, a phenomenon that is common for ion-trap mass spectrometers [21,22]. After meshing the data files of all the chromatograms, they were time-aligned (warped) to a reference data file using Correlation Optimized Warping (COW) [23] based on TICs constructed from signals in the range 100-1500 m/z.

A modified M-N rule was applied for peak detection by first calculating a median local baseline using the sliding window technique separately for each m/z trace. A median window size of 1200 data points, corresponding to 20.84 min for chip-LC-MS and 20.17 min for cap-LC-MS, was used with a moving rate of 10 points and a minimum median value of 200 counts. According to the M-N rule, a threshold of M times the local baseline was used, and a peak was assigned if the signal exceeded this threshold for at least N consecutive points within one m/z trace [24]. For each detected peak, the m/z value and the mean retention times of the three highest measured intensities within the same peak reduced by the local baseline were stored in a peak list created for every chromatogram.

We used a similar approach as Radulovic *et al.* [24] to obtain optimal settings for M and N. Different values for M (1.5-4) and N (4-8) were applied to two blank LC-MS runs and two LC-MS runs of depleted, trypsin-digested serum samples. At the settings used, the ratio between the number of peaks (between 5 and 100 min for the chip-LC-MS and 60-155 min for the cap-LC-MS runs) in the samples relative to the blank chromatograms was highest, and a minimal number of peaks were extracted from the noise in the blank chromatogram (M = 2 and N = 5 in our case).

In order to combine the peak lists from different samples with each other, one-dimensional peak matching was achieved by using the sliding window technique, in which the same m/z traces were evaluated for peaks that are proximate in time (step size 0.1 min; search window 1.0 min; maximal accepted standard deviation for all retention times within a group of matched peaks was 0.75 min). Missing peak allocation was performed by extracting the background-subtracted local signal of the given m/z trace at the given retention time. The generated peak matrix, created from the peak lists of the individual samples, consisted of a peak(*row*)-sample(*column*)-intensity(*value*) matrix. This peak matrix was used for multivariate statistical analysis. All data preprocessing work was done on a personal computer equipped with a dual core +3800 MHz AMD 64 X2 processor equipped with 4 GB of RAM.

2.4.3. Classification and multivariate statistical methods

To select the most discriminating peaks between spiked and non-spiked serum samples, the Nearest Shrunken Centroid (NSC) classification algorithm was applied [25,26]. NSC regularizes data whereby class-specific centroids are "shrunk" toward the overall (non-class-specific) centroid, which has the effect of eliminating the influence of the most weakly discriminating peaks, thereby reducing the risk to overfit the data [27]. This algorithm selects those peaks that are most relevant for the discrimination of the predefined classes. NSC is used in conjunction with leave-one-out cross-validation (LOOCV) to find the shrinkage value that gives the minimal LOOCV error [28]. In LOOCV, one observation per class is iteratively omitted from the data set that is used to construct the classification model, which is then used to classify the omitted observation. The selected peaks at the highest shrinkage value giving the lowest LOOCV error were employed for construction of the final classification model. The selected peaks were then subjected to autoscaled Principal Component Analysis (PCA) and visualized using biplots of the first two principal components [29].

3. Results and discussion

3.1. Qualitative comparison of chip-LC-MS and cap-LC-MS

3.1.1. Background ions

Figure 1 shows a heat map of LC-MS data obtained with chip-LC-MS and cap-LC-MS at different stages of data processing. This figure conveys some of the similarities and differences of the data generated by the two methods when analyzing the same depleted, trypsin-digested serum sample.

A remarkable feature of the chip-LC-MS data is that clear, straight lines at constant m/z values can be observed (Figure 1a). The majority of these peaks belongs to 3 types of ion series corresponding to a family of ubiquitous air contaminants, the polydimethylcyclosiloxanes. The ion series with rounded m/z values of 371.5, 445.5, 519.5, 593.5, 667.5, 741.5 and 815.5 after data reduction corresponding originally to 371.8, 445.4, 519.2, 593.1, 667.0, 741.0 and 814.9 can be assigned to an oligomer with 4-10 monomers, respectively [30]. The ions with rounded m/z values of 355.5 and 429.5 (originally 355.6 and 429.4) correspond to polydimethylcyclosiloxanes that lost one methyl group, due to fragmentation occurring during ion transfer between the capillary and the skimmer lens [31]. The ion series with rounded m/z values of 461.5, 536.5, 610.5 and 683.5 (originally 462.3, 536.2, 610.1 and 684.1) are ammonia adducts of the corresponding polydimethylcyclosiloxanes. Beside these ions, the phthalate anhydride cation gave also raise to an intense background peak with a rounded m/z value of 150.5 (149.7), the ion at 279.5 (279.8) corresponds to the plasticizer di-*n*-butyl-phthalate, and the ion at 371.5 m/z (371.8) to another plasticizer (di-(2-ethylhexyl)-adipate) [32]. It has been shown that the intensity of some of these ions can be significantly reduced using an ESI source that is pressurized with nitrogen supplemented with clean artificial air as used with other nano-LC-MS systems [30]. However, it is also possible to remove the signals corresponding to these ions during data pre-processing by calculating the local medium background with the sliding window technique (see Figures 1c and 1e). Figure 1c shows the signal obtained from the raw data after background subtraction while Figure 1e shows the subtracted background clearly indicating that the vertical lines were recognized as background signals. For samples analysed with the cap-LC-MS system no such contaminants were observed.

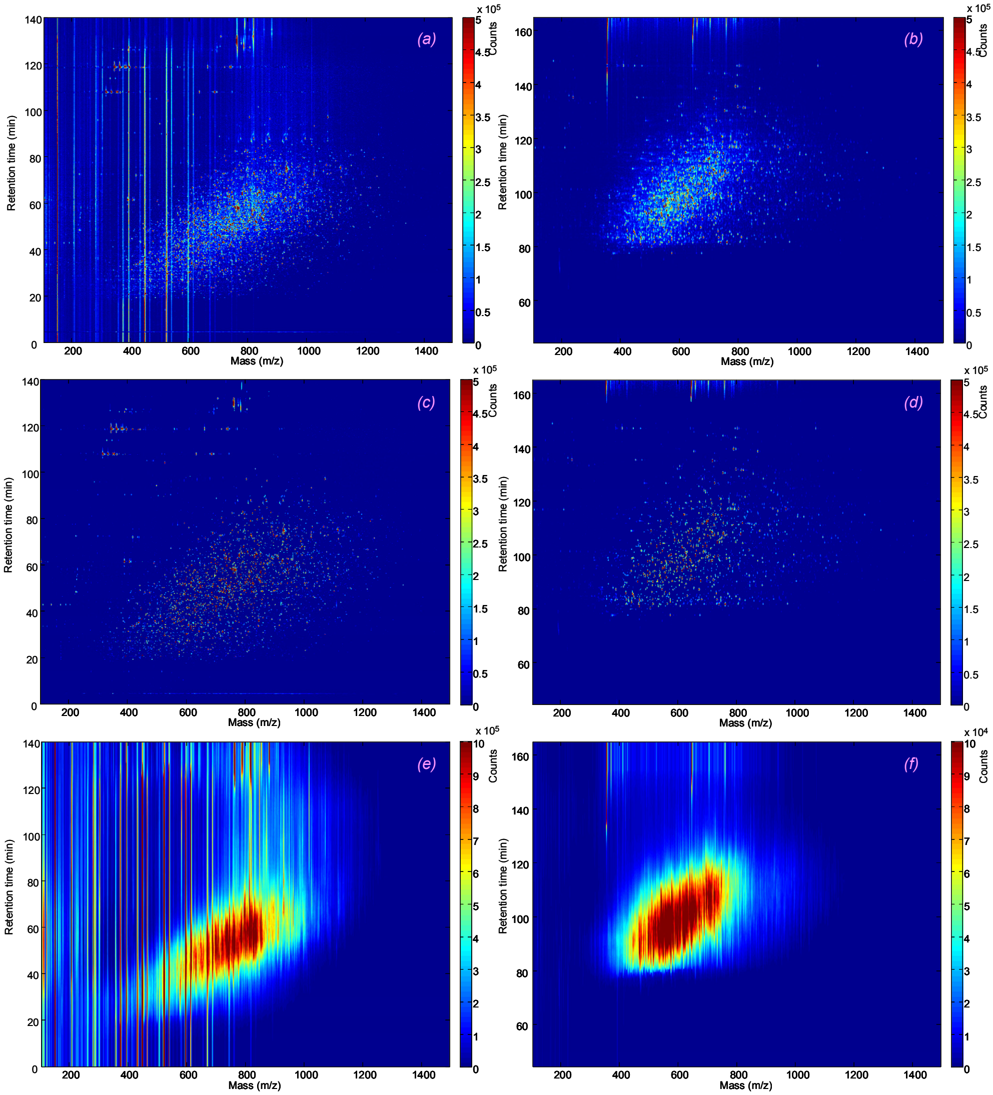

Figure 1. Visualization of LC-MS data at different stages of data pre-processing of depleted and trypsin-digested serum analyses obtained by chip-LC-MS and cap-LC-MS. Smoothed data obtained with chip-LC-MS *(a)* and cap-LC-MS *(b)* after application of a Gaussian filter with data reduction of 1:10 of the centroided raw data. LC-MS data after applying an M-N rule (for M = 2, N = 5) for peak detection for chip- *(c)* and cap-LC-MS *(d)*. Representation of the local median baseline obtained with chip-LC-MS *(e)* and cap-LC-MS *(f)* by applying a sliding window technique (window size 1200 points, minimum value 200 points, moving step size of 10 points).

Background ions from the polydimethylcyclosiloxanes might serve as indicators of ion suppression effects due to the eluting peptides. This would indicate that such effects might also affect co-eluting peptides, thus influencing quantification. However, no such effects were observed in the case of chip-LC-MS at retention times where significant amounts of peptides elute. Several papers reported ion suppression of hydrophobic, non-charged, weakly surface-active compounds at comparatively high flow rates (2-100 µL/min), while ion suppression was absent at lower nL/min flow rates (5-50 nL/min) [33-35]. It has been shown that surface-active compounds or compounds charged before the electrospray process are enriched at the droplet surface, while compounds having the opposite properties are located in the bulk of the droplets [33-35]. Oligosaccharides and glycoproteins are therefore poorly ionized and less well detected by cap-LC-MS with ESI ionisation than in the nano-ESI regime [33,36]. The flow rate of 300 nL/min applied in chip-LC-MS is in the intermediate range between regular ion spray and nanospray and still not optimal to avoid ion suppression. Running the chromatographic separation at a flow rate of 10 nL/min, at which ion suppression ceases, thus requires further miniaturization [34].

3.1.2. Charge state

The average charge state of peptides in the chromatogram of the chip-LC-MS run is lower than for cap–LC-MS. This can be observed in Figures 1a-d, where most of the peptide ions are in the range between 400-800 m/z for cap-LC-MS and between 500-1000 for chip-LC-MS. Furthermore, for chip-LC-MS the peaks show a wider spread in the m/z dimension, which makes better use of the available mass range of the ion-trap mass analyzer. The lower average charge state cannot be explained simply by the lower voltage used in case of the chip interface (~1800 V) as compared to the regular ion spray interface (3500V), which uses N_2 as nebulizer gas. As the distance between the end of the sprayer needle and the counter electrode is shorter for the chip interface (3.5 mm) than for the regular ion spray interface (15 mm), this results in calculated electric field strengths of 0.514 V/m for the chip and 0.233 V/m for the regular ion spray. Most likely the higher average charge state observed for cap-LC-MS is partially due to the triboelectric effect generated by the nebulizer gas and the liquid droplets. This phenomenon has been used to produce ions at high velocity (100 m/sec) with pneumatic assistance, without the use of an external electric field [37]. However, since the flow rates in the applied ion spray interface are lower, this phenomenon may just lead to the observed shift in charge state. Another effect that may contribute to the observed shift is the higher occurrence of electrical discharge at the higher electric field strength in the chip-LC-MS interface, which results in the production of electrons. The combination of these electrons with the positively charged droplets would also lead to the observed reduction of the average charge state [38]. Further investigations are needed to clarify the underlying physical mechanisms.

To give an overall assessment of the degree of charge-state reduction, the baseline was calculated using a sliding window technique in the m/z dimension. The observed ellipsoid-shaped area with an elevated baseline is caused by peaks that are non-resolved by the LC-MS system and that cannot be discriminated from the background by data pre-processing (Figures 1e-f). This background is completely absent in blank chromatograms, indicating that it is related to the overall resolution of the LC-MS system (data not shown). The centers of these regions reflect the average charge state of the ionized peptides, which is ~630 m/z for cap-LC-MS and ~815 m/z for chip-LC-MS.

3.1.3. Chromatographic Resolution

Since both methods make use of the same ion-trap mass analyzer, there was no difference in resolution in the m/z dimension. However, we observed an improved resolution in the retention time dimension for the chip-LC-MS module, despite the shorter column length. The mean full-width-at-half-maximum (FWHM) for peaks was 0.32±0.05 min for chip-LC-MS and 0.62±0.07 min for cap-LC-MS based on 10 randomly chosen peaks. The better chromatographic performance of the chip-LC-MS system is difficult to explain. A trapping column for sample injection in-line coupled to the analytical column was used in both systems but with a different type of C18 reverse-phase chromatographic material (Atlantis™ dC 18 3 µm for the cap-LC-MS and C-18SB-ZX 5 µm for the chip-LC-MS; for details see Material and Methods). The considerable difference in column length, diameter, and applied flow rate also influenced the observed resolution. The gradient parameter ($Q = \frac{(t_G F)}{L d_c^2}$, where t_G is the gradient time in min, F the flow rate in mL/min, L the column length in cm and d_c the column internal diameter in cm) introduced by Snyder [15] has an effect on resolution. These values were 13.33 and 124.03 for the cap-LC-MS and chip-LC-MS respectively. This large difference may explain the observed increase in resolution for the chip-LC-MS system.

The overall peak capacity of an LC-MS system is determined by the resolution of the mass analyzer and the LC system. Due to improved chromatographic resolution, the overall peak capacity for the chip-LC-MS system doubled compared to cap-LC-MS. This capacity is also better used because of the wider spread of signals in the m/z dimension, which enabled better repartition and separation of the compounds. This means that taking a 100 min gradient time and an m/z range of 1400 for both system into account, the calculated uniform peak capacity of the cap-LC-MS system is 11980 and 23211 for the chip-LC-MS system (calculated by using the measured FWHM values in the retention time dimension and an average 2.5–times m/z at the FWHM in the mass dimension allowing for 5% peak overlap while producing the tightest arrangement of the peaks). Concerning the effective resolution space (space where most of the peaks can be found) obtained for the studied samples, the useful range is 95 min for both system and 400 versus 500 m/z for the cap- and chip-LC-MS, respectively, resulting in 3251 and 7875 of peaks.

Figures 2a and 2b represent the TICs of the raw data and of the calculated baselines for chip- and cap-LC-MS.

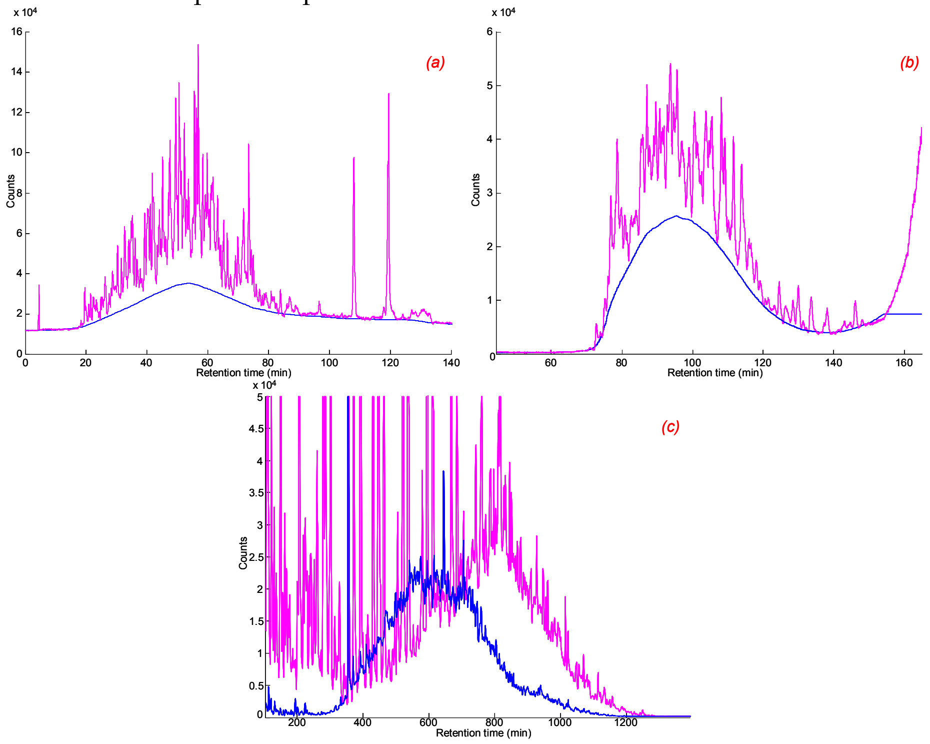

Figure 2. TIC chromatogram of raw data and of the baseline obtained with chip-LC-MS *(a)* and cap-LC-MS *(b)*. TIC of the local median baseline obtained with the sliding window technique from the chromatogram of chip-LC-MS (blue) and cap-LC-MS (purple) *(c)*.

The part of the overall TIC signal that can be explained based on peaks selected by the M-N rule is 55% for chip-LC-MS and 32% for cap-LC-MS, indicating that a larger portion of the acquired data can be assigned to actual peaks in the case of the chip-LC interface using the same data pre-processing method. Visualization of the local baseline gives an indication of the distribution of the non-resolved peaks, which confirms that there is a lower amount of non-resolved peaks in the case of the chip-LC-MS system (Figures 1e-f). Using rather conservative parameters for peak picking (M = 3 and N = 8), the number of extracted features was 4216 (between 60 – 155 min and 300 – 1500 m/z) for cap-LC-MS and 11746 (between 5 – 100 min and 300 – 1500 m/z) for chip-LC-MS. Because of two-dimensional smoothing, data reduction, which combines isotopic peaks into a single peak distributed over several m/z traces, and with the presence of multiple charge states, 2-8 features in the peak list corresponded on average to one peptide. This gives the chip-LC-MS system the capacity to detect between 1468-5873 peptides while the cap-LC-MS system can detect between 527-2108 peptides in the studied sample.

3.2. Repeatability and Quantification

3.2.1. Repeatability

Comparative biomarker studies rely on the repeatability of the analytical method. To test the repeatability of the overall method including sample preparation and LC-MS, 10 serum samples of the same patient were independently prepared by immunoaffinity depletion and trypsin digestion. All samples were analyzed by LC-MS on the same day. Six peptides from 3 high-abundance proteins (apolipoprotein A-I [accession number: P02647, SwissProt/Trembl], α-2-macroglobulin precursor [P01023], human complement C3 precursor [P01024]) with m/z values of 756.7, 753.5, 909.6, 682.5, 619.2 and 694.4 m/z [16] were selected to test the analytical repeatability for both cap-LC-MS and chip-LC-MS. Manual integration of peak areas was first performed using extracted ion chromatograms (EICs) for m/z values taken at the maximum intensity of each selected peptide peak ±0.5 m/z, in order to account for shifts occurring in the ion-trap mass analyzer during peak elution. Retention time variation was small in both cases, although the cap-LC-MS system showed a slightly higher standard deviation (SD) (mean value of 0.29 min ± 0.15) than chip-LC-MS (mean value of 0.04 min ± 0.01). For peak area, practically the same relative standard deviation (RSD) was obtained for the cap-LC-MS and chip-LC-MS systems (21.87±7.60% and 21.73±7.51%, respectively) showing that the microfluidics-based chip-LC-MS system is equally well-suited for comparative label-free peptide profiling studies as the cap-LC-MS system, while requiring about 30 times less sample. The repeatability of ~20% in peak area is in agreement with an earlier study using a comparable cap-LC-MS system [16]. Repeatability of chip-LC-MS by itself showed that most of the analytical variability was due to the upstream sample preparation steps, since 5 repeated injections of the same sample on the same day resulted in a repeatability of 5.19±2.48% RSD in peak area.

Despite the good retention time repeatability of cap-LC-MS and chip-LC-MS, it is necessary to align the chromatograms to avoid incorrect peak matching between runs, especially in peak-rich regions. Since retention time shifts are minor and the TICs are highly similar, Correlation Optimized Warping (COW) [23] was used with success to reduce the RSD of the 6 peptide peaks to 0.07±0.03 min for cap-LC-MS and 0.02±0.01 min chip-LC-MS[*].

The peak area repeatability of both cap-LC-MS and chip-LC-MS was improved from 21.87±7.60% and 21.73±7.51%, respectively, to 19.79±7.72% and

[*] The ion 757 was omitted from the RSD calculations, since peak mismatching occurred.

14.52±3.91%[*] RSD by automated data pre-processing. Considering the chip-LC-MS part only, repeatability improved from 5.19±2.48% to 3.50±0.96% for the 5 repeated injections of the same sample. The increased repeatability is probably due to the fact that data reduction and Gaussian filtering smoothed out some of the variability of the mass detector.

Assessing repeatability based on only 6 peaks does not give a comprehensive view of the variability of the LC-MS profiling system. Automatic data processing using a higher threshold (M = 3, N = 8) for peak picking to exclude the majority of noise enabled us to have a more global view. Figure 3 shows histograms of the RSD for the matched peak lists of datasets containing 10 analyses of separately prepared digested serum samples for the cap-LC-MS (blue trace) and the chip-LC-MS (red trace) systems.

Evaluating only the chip-LC-MS system on its own with 5 repeated injections of the same sample (green trace) confirms that most of this variability is due to the sample preparation steps and not to the LC-MS part. It appears from these histograms that cap-LC-MS has a slightly higher maximum (mode)[**] RSD (24.29±0.35%) than chip-LC-MS (22.38±0.46%) for the data sets, including variability due to sample preparation. Variability in the data set obtained with chip-LC-MS, including only the variability of the LC-MS system, was again lower (8.33±0.35%). These results confirm the data obtained with the analysis of the 6 peptides and are in agreement with the reproducibility values obtained on a prototype of the chip-LC-MS microfluidic device [17].

The presented data processing method leads to loss of the fine structure of the mass dimension and thus to a lower overall resolution of the analytical platform because of data reduction from 0.1 to 1 m/z and two-dimensional Gaussian filtering. It is envisioned that processing without data reduction, using profile spectra combined with a more accurate peak picking algorithm with higher overlapping peak resolving capabilities, will increase the precision of the information extracted from the raw data. This is particularly relevant for low-abundance peaks and for peaks eluting in highly crowded areas. Work along these lines is underway.

[**] The maximum was obtained based on the fit of 3 Gaussian curves on the histogram excluding the plateau region of the data. Among the 3 Gaussian curves the maximum of the highest curve was taken. The presented standard deviation is the value obtained by curve fitting.

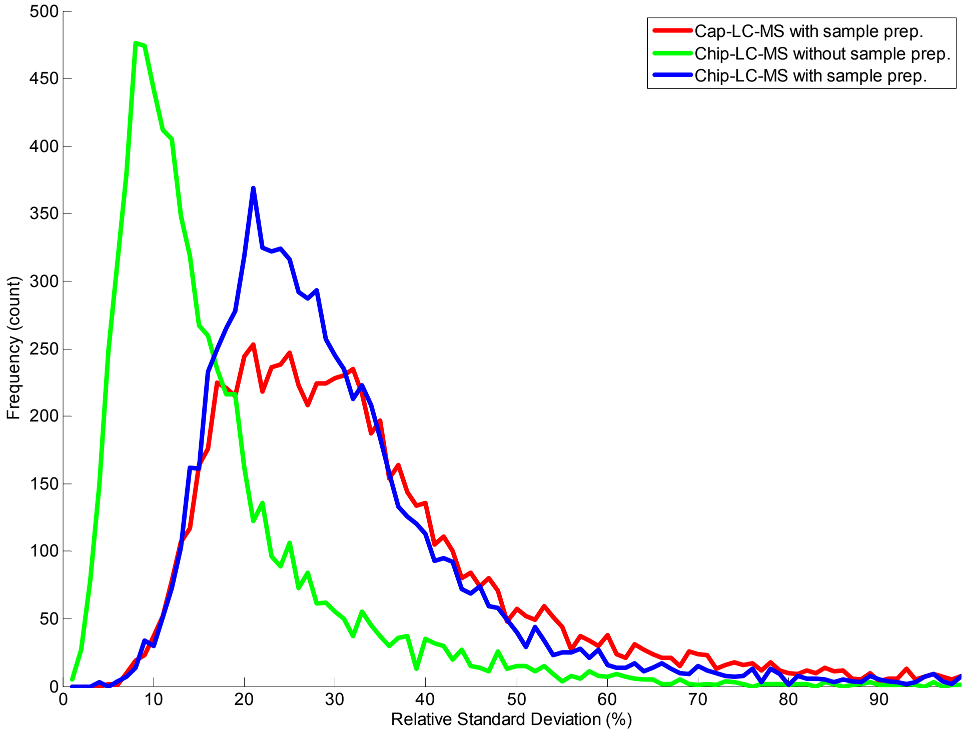

Figure 3. Relative Standard Deviation of peaks obtained using an M-N rule with a higher threshold (M = 3, N = 8) for cap-LC-MS (red) and chip-LC-MS (blue) obtained from 10 serum samples from the same patient including the sample preparation steps and for chip-LC-MS (green) obtained from 5 repetitive injections, representing thus only the variation of the LC-MS system.

3.2.2. Multivariate statistical analysis

In order to compare the ability of cap-LC-MS and chip-LC-MS to find discriminating features, 5 serum samples were spiked with 21 pmol of horse heart Cytochrome C for 2 μL serum. Due to losses during immunoaffinity depletion of high-abundance proteins, the actual amount of horse heart Cytochrome C after depletion was 4.2 pmol [16], corresponding to 3% of the total protein content. The obtained raw data were subjected to data processing as described above, followed by supervised classification and selection of discriminating features using the Nearest Shrunken Centroid (NSC) algorithm [25, 26]. The shrinkage parameter was optimized through “leave one out” cross validation (LOOCV). Initially a low threshold (M = 2, N = 5) was used for peak picking in order to see the ability of the classifier to find discriminating features in the presence of noise in the peak list. Figure 4 shows the LOOCV error and the number of selected variables plotted against the shrinkage for chip-LC-MS and cap-LC-MS.

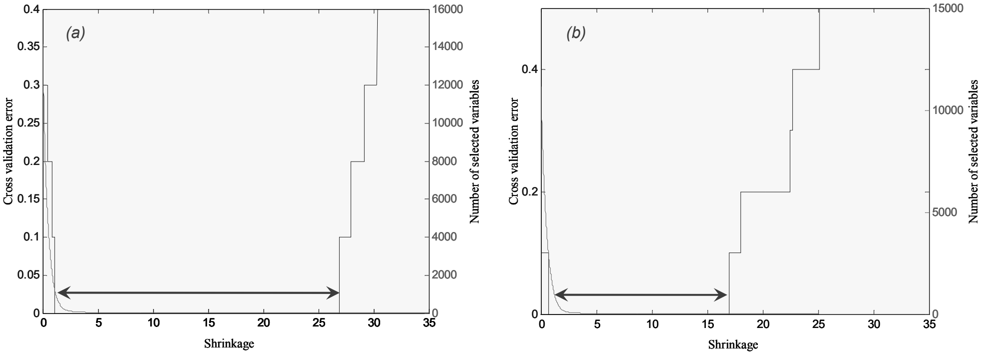

Figure 4. Representation of the LOOCV error and the number of selected variables as function of the shrinkage obtained using an NSC-based variable selection and classification algorithm on the matched peak list after peak detection (M-N rule with M = 2, N = 5) for chip-LC-MS *(a)* and cap-LC-MS *(b)*. The selected variables, where the shrinkage domain has no cross-validation error, are highlighted with red arrows (0.90-29.51 shrinkage and 2540 to 2 selected variables for chip-LC-MS; 0.61-16.80 shrinkage and 3496 to 2 selected variables for cap-LC-MS). For these domains, the selected variables enabled a perfect separation of the two classes of spiked and non-spiked samples.

There is a large domain of shrinkage where the cross validation error was 0 in both cases (0.90-29.51 for chip and 0.61-16.80 for cap-LC-MS), and where selecting variables at those shrinkage values resulted in a perfect classification. The most discriminating features are found at higher shrinkage values. Evaluating the 16 most discriminating features selected at shrinkages of 10 and 8.5 for chip- and cap-LC-MS, which corresponded to 6 different peptide peaks, showed that all of these 6 peptides for the chip-LC-MS and 5 for the cap-LC-MS peptides corresponded to *in-silico* predicted tryptic peptides of horse heart Cytochrome C (Table 1).

The remaining features are most probably due to non-specific cleavages, because they show clear absence and presence in the corresponding spiked or non-spiked chromatograms. These features are also present in the standard tryptic digest of horse heart Cytochrome C. Figure 5 shows that correct discrimination between spiked and non-spiked serum is easily possible, based on the selected peaks (Figure 5c and d).

Visualization of the EICs of some of the selected peaks (Figure 6) confirms that these peaks differ clearly between chromatograms obtained from spiked and non-spiked samples.

Table 1. Main characteristics of the 16 most discriminating peaks selected by NSC at a shrinkage of 10 and 8.5 for chip-LC-MS and cap-LC-MS, respectively (see Figure 3). The peaks in bold contributed most to the discrimination between the spiked and non-spiked serum samples.

Platform	m/z	Retention time [min]	Charge state	Miss cleavage	Location	*In-silico* Peptide sequence
Chip-LC-MS	605.5	26.29	1	0	56-60	GITWK
	584.5	43.00	2	0	28-38	TGPNLHGLFGR
	390.5	39.93	2	0	80-86	MIFAGIK
	391.5	39.93	2	0	80-86	MIFAGIK
	482.5	47.00	2	0	92-99	EDLIAYLK
	483.5	46.98	2	0	92-99	EDLIAYLK
	484.5	**46.98**	**2**	**0**	**92-99**	**EDLIAYLK**
	747.5	53.06	2	0	61-72	EETLMEYLENPK
	748.5	53.09	2	0	61-72	EETLMEYLENPK
	963.5	47.06	1	0	92-99	EDLIAYLK
	964.5	47.05	1	0	92-99	EDLIAYLK
	965.5	46.97	1	0	92-99	EDLIAYLK
	734.5	32.00	2	0	40-53	TGQAPGFTYTDANK
	735.5	32.00	2	0	40-53	TGQAPGFTYTDANK
	736.5	32.00	2	0	40-53	TGQAPGFTYTDANK
	737.5	32.06	2	0	40-53	TGQAPGFTYTDANK
Cap–LC-MS	557.5	104.90	3	2	23-38	GGKHKTGPNLHGLFGR
	558.5	104.92	3	2	23-38	GGKHKTGPNLHGLFGR
	559.5	**104.89**	**3**	**2**	**23-38**	**GGKHKTGPNLHGLFGR**
	577.5	111.36				
	578.5	111.35				
	735.5	87.29	2	0	40-53	TGQAPGFTYTDANK
	736.5	87.31	2	0	40-53	TGQAPGFTYTDANK
	737.5	87.27	2	0	40-53	TGQAPGFTYTDANK
	747.5	111.36	2	0	61-72	EETLMEYLENPK
	748.5	111.39	2	0	61-72	EETLMEYLENPK
	749.5	111.43	2	0	61-72	EETLMEYLENPK
	750.5	111.35	2	0	61-72	EETLMEYLENPK
	965.5	104.94	1	0	92-99	EDLIAYLK
	451.5	96.84	3	1	89-99	TEREDLIAYLK
	576.5	111.42				
	964.5	104.93	1	0	92-99	EDLIAYLK

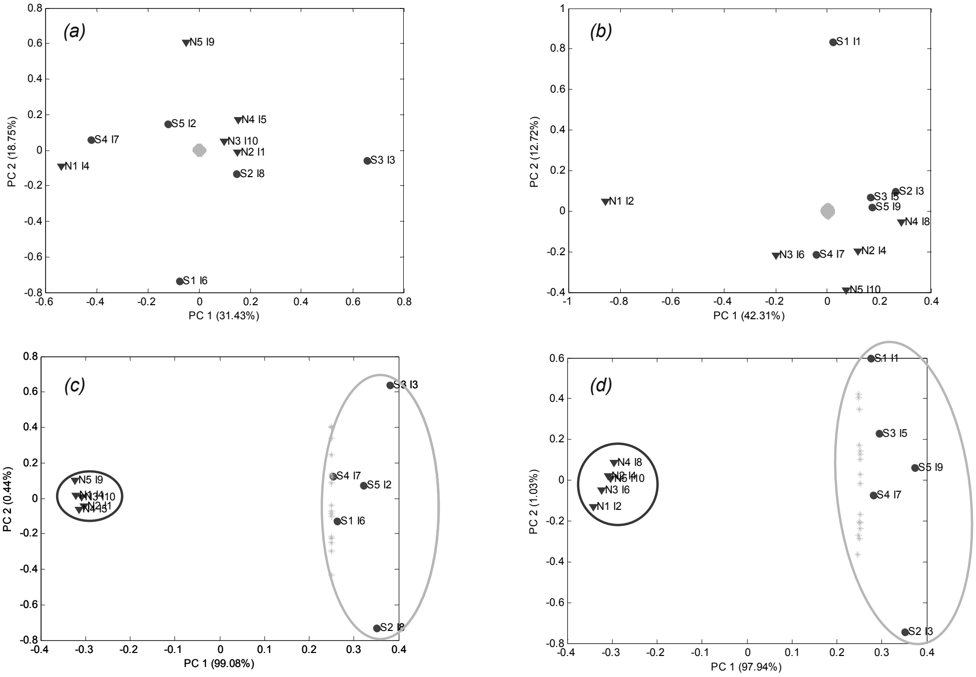

Figure 5. PCA plots using all peaks obtained with M = 2, N = 5 for chip-LC-MS *(a)* and cap-LC-MS *(b)*, (14091 for chip-LC-MS and 11256 for cap-LC-MS) and PCA plots of 16 features selected by NSC from datasets obtained with chip-LC-MS (shrinkage value of 10.00) (c) and cap-LC-MS (shrinkage value of 8.50) (d).

Among these variables, the peak at 484.5 m/z and 46.97 min (sequence EDLIAYLK) contributed most to the discrimination for chip-LC-MS and the peak at 559.5 m/z and 104.89 min (sequence GGKHKTGPNLHGLFGR) for cap-LC-MS. PCA analysis of the selected peaks (Figures 5c and d) shows that almost all variability in the data can be explained by Principal Component 1 (99% for chip- and 98% for the cap-LC), however, chip-LC-MS used 28-times less sample (the analyzed sample amount was 140 pmol for cap- and 5 pmol for chip-LC). Distribution of the selected variables in the PCA biplots also shows that all selected variables have higher intensities in the spiked samples as expected.

The necessity to apply NSC for variable selection prior to the final centroid classification is evident when using centroid classification with all variables, which did not result in any discrimination between spiked and non-spiked samples along PC1 and PC2 of the PCA plot using the complete peak list obtained by the M-N rule (Figures 5a and 5b).

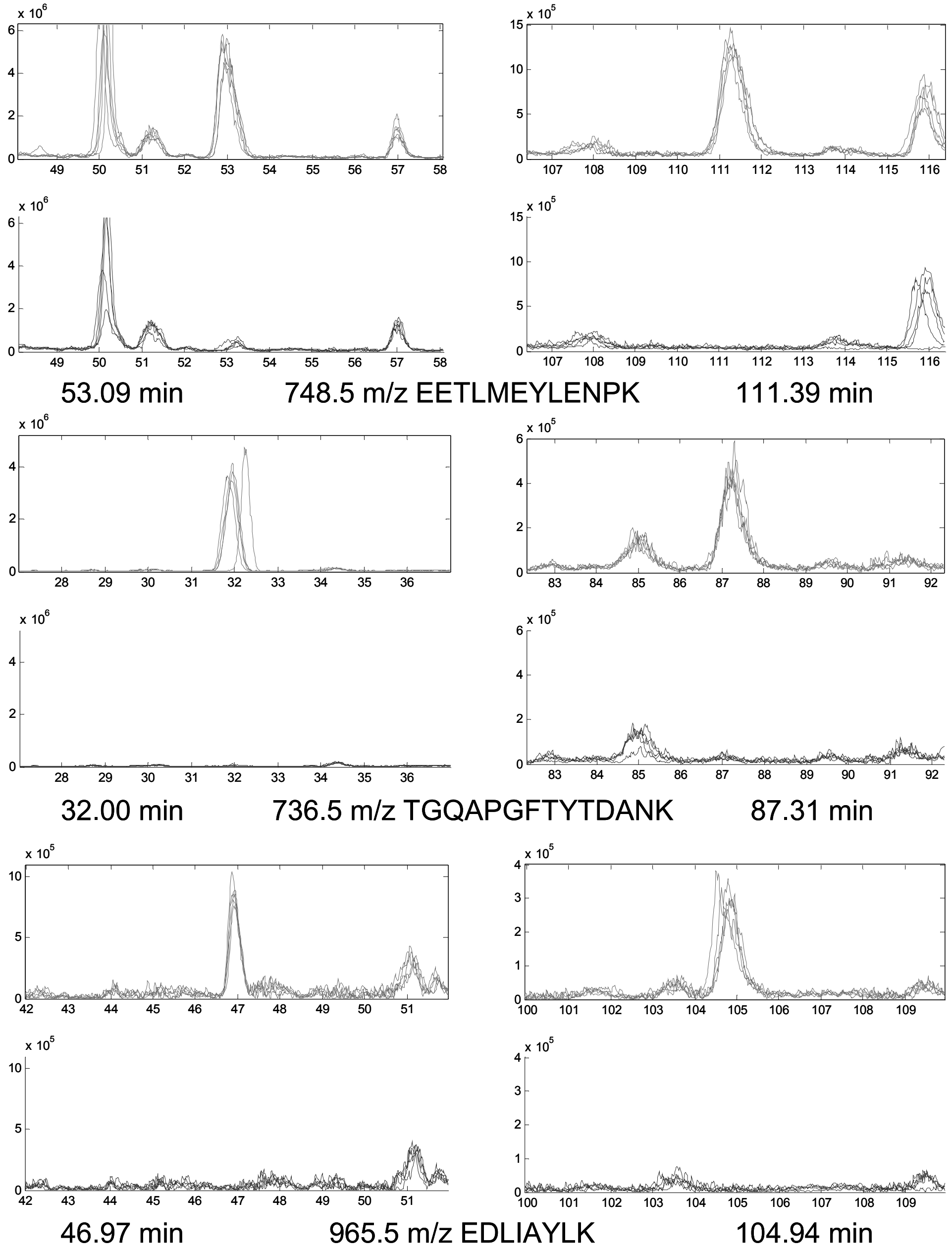

Figure 6. Example of EICs of NSC-selected peaks corresponding to tryptic fragments of horse heart Cytochrome C from datasets obtained with chip-LC-MS (left) and cap-LC-MS (right). The upper traces were obtained from spiked, the lower traces were obtained from non spiked samples.

4. Conclusions and future perspectives

We show that integration of a nano-HPLC (column ID=75 μm) system on a microfluidic chip directly connected to a mass spectrometer via an ESI interface results in repeatable analyses of highly complex samples such as trypsin-digested human serum. Indeed, overall performance was equal or superior to a previously used cap-LC-MS system while requiring ~30-times less sample. The higher chromatographic resolution resulted in an increased peak capacity and an overall lower charge state was observed, which resulted in improved use of the available time × m/z space (see Figures 1c-d). The data processing methodology presented is able to remove background ions efficiently and the visualization of the local baseline gives an overview over the intensity, distribution and average charge state of the non resolved peaks of each system, which is a valuable tool for the evaluation and further optimization of the analytical system. The presented data processing and classification methodology was able to discriminate between spiked and non-spiked samples at 3% of the total protein content based on the acquired data, with slightly better performance for the chip-LC-MS system.

The present analyses were performed with a quadrupole ion-trap mass analyzer and it is evident that use of the chip-LC interface with high resolution mass analyzers such as Time-Of-Flight, orbitrap- or Fourier Transform-Ion Cyclotron Resonance systems would further enhance the overall resolution. Replacing the particle-packed column on the chip with monolithic materials would, on the other hand, decrease the pressure and permit increased analysis speed. Faster analyses would in turn allow the analysis of prefractionated samples (e.g. after cation-exchange HPLC) with a higher throughput and thus to increase the dynamic range of the overall analysis. As the maximum pressure limit of the present chip is 150 bar, the development of a chip able to work at 300-1000 bar would allow the use of smaller diameter particles, which would in turn result in an increase in the resolution of the LC dimension. Further improvements along these lines should render microfluidics-based nano-LC-MS a highly useful tool for biomarker discovery.

Abbreviations

ACN	Acetonitrile
BSA	Bovine Serum Albumin
cap-LC-MS	HPLC-MS system using capillary column
chip-LC-MS	nanoHPLC-MS system integrated on a microfluidic device
COW	Correlation Optimized Warping
EICs	Extracted ion chromatograms
ESI	Electrospray Ionisation
FA	Formic Acid
FWHM	Full Width at Half Maximum
HPLC	High Performance Liquid Chromatography
ICAT	Isotope Coded Affinity Tag
LOOCV	Leave One Out Cross Validation
NSC	Nearest Shrunken Centroids
PCA	Principal Component Analysis
PCX	Principal Component number X
RSD	Relative Standard Deviation
TIC	Total Ion Current

References

[1] Horvatovich P, Govorukhina NI, Bischoff R. Biomarker discovery by proteomics: challenges not only for the analytical chemist. Analyst 2006;131:1-6.

[2] Anderson NL, Anderson NG. The human plasma proteome: history, character, and diagnostic prospects. Mol. Cell. Proteomics 2002;1(11):845-867.

[3] Petricoin EF, Liotta LA. SELDI-TOF-based serum proteomic pattern diagnostics for early detection of cancer. Curr. Opin. Biotechnol. 2004;15:24-30.

[4] Diamandis EP. OvaCheck: Doubts Voiced Soon After Publication. Nature 2004;430(7000):611.

[5] Diamandis EP. Proteomic Patterns to Identify Ovarian Cancer: 3 Years on. Expert. Rev. Mol Diagn. 2004;4:575-577.

[6] Ong SE, Kratchmarova I, Mann M. Properties of 13C-Substituted Arginine in Stable-Isotope Labeling by Amino Acids in Cell Culture (SILAC). J. Proteome Res. 2003;2(2):173-181.

[7] Ong SE, Foster LJ, Mann M. Mass spectrometric-based approaches in quantitative proteomics. Methods 2003;29(2):124-130.

[8] Gygi SP, Rist B, Gerber SA, Turecek F., et al., Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. Nat. Biotechnol. 1999;17(10):994-999.

[9] Ross PL, Huang YN, Marchese JN, Williamson B., et al., Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. Mol Cell Proteomics 2004;3(12):1154-1169.

[10] Leitner A, Lindner W. Current chemical tagging strategies for proteome analysis by mass spectrometry. J. Chromatogr. B. 2004;813(1-2):1-26.

[11] Higgs ER, Knierman DM, Gelfanova V, Butler PJ, Hale EJ. Comprehensive Label-Free Method for the Relative Quantification of Proteins from Biological Samples. J. Proteome Res. 2005;4(4):1442-1450.

[12] Washburn MP, Wolters D, Yates JR. Large-Scale Analysis of the Yeast Proteome by Multidimensional Protein Identification Technology. Nat. Biotechnol. 2001;9:242-247.

[13] Wolters DA, Washburn MP, Yates JR. An Automated Multidimensional Protein Identification Technology for Shotgun Proteomics. Anal. Chem. 2001;73:5683-5690.

[14] Meiring HD, van der Heeft E, ten Hove GJ, de Jong APJM. Maintaining fixed band spacing when changing column dimensions in gradient elution. J. Sep. Sci. 2002;25(9):557-568.

[15] Dolan JW, Snyder LR. Maintaining fixed band spacing when changing column dimensions in gradient elution. J. Chrom. A. 1998;799(1-2):21-34.

[16] Govorukhina NI, Reijmers TH, Nyangoma SO, van der Zee AGJ., et al., Analysis of human serum by liquid chromatography–mass spectrometry: Improved sample preparation and data analysis. J. Chromatogr. A. 2006;1120:142-150.

[17] Fortier MH, Bonneil E, Goodley P, Thibault P. Integrated Microfluidic Device for Mass Spectrometry-Based Proteomics and Its Application to Biomarker Discovery Programs. Anal. Chem. 2005;77:1631-1640.

[18] Vollmer M, Hörth P, Rozing G, Cout Y., et al., Multi-dimensional HPLC/MS of the nucleolar proteome using HPLC-chip/MS. J. Sep. Sci. 2006;29:499-509.

[19] Martosella J, Zolotarjova N, Liu H, Nicol G, Boyes BE. Reversed-Phase High-Performance Liquid Chromatographic Prefractionation of Immunodepleted Human Serum Proteins to Enhance Mass Spectrometry Identification of Lower-Abundant Proteins. J. Prot. Res. 2005;4(5):1522-1537.

[20] Dekker LJ, Bosman J, Burgers PC, van Rijswijk A., et al., Depletion of high-abundance proteins from serum by immunoaffinity chromatography: a MALDI-FTMS study. J. Chrom. B. 2007;847:65-69.

[21] Hao Ch, March ER. A survey of recent research activity in quadrupole ion-trap mass spectrometry. Int. J. Mass Spectrum. 2001;212:337–357.

[22] Wells JM, Plass WR, Patterson GE, Ouyang Zh. Chemical Mass Shifts in Ion-trap Mass Spectrometry: Experiments and Simulations. Anal. Chem. 1999;71:3405-3415.

[23] Tomasi G, van den Berg F, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic Data. J. Chemometrics 2004;18:231-241.

[24] Radulovic D, Jelveh S, Ryu S, Hamilton TG., et al., Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. Mol. Cell. Proteomics 2004;3:984-997.

[25] Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. USA. 2002;99:6567-6572.

[26] Tibshirani R, Hastie T, Narasimhan B, Soltys S., et al., Sample classification from protein mass spectrometry, by 'peak probability contrasts'. Bioinformatics 2004;20:3034-3044.

[27] Listgarten J, Emili A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. Mol. Cell. Proteomics 2005;4:419-434.

[28] Wagner M, Naik D, Pothen A. Protocols for disease classification from mass spectrometry data. Proteomics 2003;3:1692-1698.

[29] Hilario M, Kalousis A, Muller M, Pellegrini C. Machine learning approaches to lung cancer prediction from mass spectra. Proteomics 2003;3:1716-1719.

[30] Schlosser A, Volkmer-Engert RJ. Volatile Polydimethylcyclosiloxanes in the Ambient Laboratory Air Identified As Source of Extreme Background Signals in Nanoelectrospray Mass Spectrometry. Mass Spectrom. 2003;38:523-525.

[31] Spiteller M, Spiteller G. Massenspektrensammlung von Lösungsmitteln, Verunreinigungen Säulenbelegmaterialien und einfachen aliphatischen Verbindungen, Springer Verlag, Wien, New York 1973.

[32] Guo X, Bruins PA, Covey RT. Characterization of typical chemical background interferences in atmospheric pressure ionization liquid chromatography-mass spectrometry. Rapid Commun. Mass Spectrom. 2006;20:3145–3150.

[33] Karas M, Bahr U, Dülcks T, Fresenius. Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine. J. Anal. Chem. 2000;366:669–676.

[34] Schmidt A, Karas M, Dülcks T. Effect of Different Solution Flow Rates on Analyte Ion Signals in Nano-ESI MS, or: When does ESI turn into Nano-ESI? J. Am. Soc. Mass Spectrom. 2003;14:492-500.

[35] Juraschek R, Dülcks T, Karas M J. Nanoelectrospray—More Than Just a Minimized-Flow Electrospray Ionization Source. Am. Soc. Mass Spectrom. 1999;10:300–308.

[36] Bahr U, Pfenninger A, Karas M, Stahl B. High-Sensitivity Analysis of Neutral Underivatized Oligosaccharides by Nanoelectrospray Mass Spectrometry. Anal. Chem. 1997;69:4530-4535.

[37] Goodley CP, Fischer MS, Gourley LD. Self generating ion device for mass spectrometry of liquids. US patent 5559326A. 1996.

[38] Ebeling DD, Westphall SM, Scalf M, Smith ML. Corona Discharge in Charge Reduction Electrospray Mass Spectrometry. Anal. Chem. 2000;72:5158-5161.

Chapter IV.

Influence of clotting time on the protein composition of serum samples based on LC-MS data

N.I. Govorukhina, M. de Vries, T.H. Reijmers, P. Horvatovich, A.G.J. van der Zee, and R. Bischoff

Manuscript in preparation

Abstract

Many large banks of serum samples were acquired prior to the widespread use of proteomics in biomarker research. An important parameter that is often not exactly known is clotting time. We therefore investigated the influence of clotting time on the protein and peptide composition of serum by label-free and stable-isotope labeling techniques. The label-free analysis of trypsin-digested serum showed that the overall pattern of LC-MS data is not affected by clotting times varying from 2 to 8h. However, univariate and multivariate statistical analyses revealed that proteins that are directly involved in blood clot formation, such as the clotting-derived fibrinopeptides, change significantly. This is most easily detected in the supernatant of acid-precipitated, immunodepleted serum. Stable-isotope labeling techniques show that truncated or phosphorylated forms of Fibrinopeptides A and B increase or decrease depending on clotting time. These patterns can be easily recognized and should be taken into consideration when using serum sample collections of which the clotting time is not known. Leucine-rich alpha-2-glycoprotein (P02750) was down-regulated in all samples compared to the sample for which the clotting time was one hour based on iTRAQ labeling data.

1. Introduction

The discovery and validation of biomarkers for early diagnosis of disease at a stage where successful therapy is still possible is an important goal of modern biomedical research. To achieve this goal, high-resolution analytical techniques are applied to complex clinical samples, mostly body fluids. Serum is a body fluid that is representative of the composition of soluble proteins and peptides in blood and is thus a suitable starting material for biomarker discovery studies. Moreover, many existing large sample collections at major hospitals consist of serum that is stored frozen at -80°C. Since these collections may well contain important information about the health status of the corresponding patients and controls, especially when they have been followed over time, it is critical to evaluate under which conditions it is possible to compare samples from these collections with modern proteomics approaches.

The generation of serum requires that blood be coagulated and that the cellular components as well as the blood clot be removed by centrifugation or filtration. It has notably been argued that the time and conditions under which blood is allowed to clot (clotting time) are important parameters that must be controlled and kept constant in order to compare protein and peptide profiles [1-5]. However, most existing sample collections have not been obtained with subsequent proteomics analyses in mind and clotting time has often not been rigorously controlled. Many of the studies evaluating the influence of pre-analytical parameters on serum protein composition have been performed by Surface-Enhanced Laser Desorption Ionization Time-Of-Flight Mass Spectrometry (SELDI-TOF-MS) [4], a method that suffers from rather poor concentration sensitivity and that may be prone to mass spectrometric artifacts [6]. More sensitive approaches using enrichment of proteins and peptides on magnetic bead separators or by Liquid Chromatography (LC) followed by Matrix-Assisted Laser Desorption Ionization Time-Of-Flight Mass Spectrometry (MALDI-TOF-MS) have also indicated that conditions of sample handling and preparation are critical [2,3,5,7,8]. Our previous studies and those of others have shown that the combination of LC with Electrospray-Ionization Mass Spectrometry (ESI-MS), abbreviated LC-MS, is suitable to analyze body fluids such as serum or urine [9,10]. The increasing number of applications of LC-MS and LC-MS/MS for the profiling of body fluids or the targeted detection of individual proteins underscores furthermore that this method is capable of achieving concentration sensitivities in the ng-pg/mL range [11-18]. In return, LC-MS provides highly complex data sets when used in the profiling mode (measurement of all detectable compounds in a sample) and it is thus not easy to assess the effect of a given pre-analytical parameter on the overall result.

We describe here an approach to assess the effect of clotting time on LC-MS profiles of serum obtained from a healthy volunteer by univariate and multivariate statistical analysis after data processing. In order to render serum samples suitable to high-performance LC-MS analysis, proteins were digested with trypsin. Additionally, we investigated the supernatant of acid-precipitated serum samples which are highly enriched in low-molecular weight proteins and peptides (the so-called Peptidome) [3,19-24]. For comparison of samples we used label-free as well as stable-isotope labeling (iTRAQ^{TM}) [25] approaches.

2. Materials and Methods

2.1. Description of samples

Serum samples were prepared at the Department of Gynecological Oncology (University Medical Center Groningen, Groningen, The Netherlands) and stored at -80ºC in aliquots until analysis. All intermediate fractions that were obtained during sample preparation were stored at -20ºC. Glass tubes (Becton Dickinson, #367953), with a separation gel and micronized silica to accelerate clotting, were used for blood collection. Serum was obtained from a single healthy female volunteer, who consented to this study, after different clotting times (1, 2, 4, 6 and 8 h). Serum was prepared by letting the freshly collected blood coagulate at room temperature for 1, 2, 4, 6 or 8 h followed by centrifugation at room temperature for 10 min at 3000 rpm.

2.2. Preparation of serum samples

20 µL of serum were mixed with 80 µL of buffer A (Agilent, Santa Clara, California, USA) of which 80 µL were injected on a Multiple Affinity Removal column (Agilent, 4.6 × 50 mm, Part # 5185-5984) after filtration through a 0.22 µm spin filter (Part # 5185-5990) at 13000 g and 4°C for 10 min to remove particulates. Removal of abundant proteins was performed on a LaChrom HPLC System (Merck Hitachi, www.merck.com) with detection at 280 nm using the following timetable: 0-9 min, 100% buffer A (0.25 mL/min); 9.0-9.1 min, linear gradient 0-100 B % (1 mL/min), 9.1-12.5 min, 100% buffer B (1 mL/min); 12.5-12.6 min, linear gradient 100-0% buffer B (1 mL/min); 12.6-20 min, 100% buffer A (1 mL/min). The flow-through fraction (depleted serum collected between 2-6 min) of a total volume of appr. 1 mL was collected [9]. Each serum sample obtained after different clotting times (1, 2, 4, 6, 8 h) was depleted in duplicate.

Protein concentrations were determined with the Micro BCA^{TM} Protein assay reagent kit (www.piercenet.com) and calculated for an average protein molecular weight of 50 kDa. BSA was used as the calibration standard. Depleted serum samples were digested with trypsin (sequencing grade modified trypsin, Promega, cat # V5111, USA) at an enzyme-to-substrate ratio of 1 : 20 overnight at 37°C with shaking at 400 rpm (Eppendorf Thermomixer) and 4% of the digest were subjected to capillary LC-MS analysis.

2.3. Standard addition of horse heart Cytochrome C

Serum samples were spiked with 21 – 50 pmol of horse heart Cytochrome C (Sigma, www.sigmaaldrich.com, cat. # 9007-43-6) prior to immunodepletion (21 pmol (+) or 50 pmol (++)), of which 10% were subjected to the LC–MS analysis after trypsin digestion. Cytochrome C was alternatively digested with trypsin and added in the same amounts to depleted and trypsin-digested serum (1 h clotting time) prior to LC–MS (21 pmol (n+) or 50 pmol (n++)).

2.4. Cap-LC-MS

All LC-MS analyses were performed on an Agilent 1100 capillary HPLC system coupled on-line to an SL ion-trap mass spectrometer (www.home.agilent.com) equipped with an Atlantis™ dC 18 (1.0 × 150 mm, 3 µm) column that was protected by an Atlantis™ dC 18 in-line trap column (3 µm, 2.1 mm × 20 mm guard column). 40 µL of the pretreated (depleted and digested) fractions corresponding to ~8 µg or 160 pmol of total protein digest (calculated based on a 50 kDa protein) were injected. The autosampler (cat. # G1367A) was equipped with a 100 µL injection loop and a temperature-controlled cooler (cat. # G1330A) maintaining the samples at 4°C. The HPLC system had the following additional components: capillary pump (cat. #, G1376A), solvent degasser (cat. #, G1379A), UV detector (cat. # G1314A) and column holder (cat. #, G1316A). The sample was injected and washed in the back-flush mode for 30 min (0.1% aq. formic acid (FA) and 3% acetonitrile (AcN) at a flow rate of 50 µL/min). Peptides were eluted in a linear gradient from 0 to 70% (0.5%/min) AcN containing 0.1% FA at a flow-rate of 20 µL/min. After each injection, the in-line trap and the analytical column were equilibrated with eluent A (H_2O/AcN/FA; 950:50:1) for 20 min prior to the next injection.

The following settings were used for mass spectrometry during LC-MS. Nebulizer gas: 16.0 psi N2, drying gas: 6.0 L/min N2, skimmer: 40.0 V, ionisation voltage: 3500 V, cap. exit: 158.5 V, Oct. 1: 12.0V, Oct. 2: 2.48 V, Oct. RF: 150 Vpp (Voltage, Peak Power Point), Lens 1: -5.0 V, Lens 2: -60.0 V, Trap drive: 53.3, T: 325°C, Scan resolution: enhanced (5500 m/z per second scan speed). Target mass: 600. Scan range: 100-1500 m/z. Spectra were saved in centroid mode. LC-MS chromatographic data were analyzed with Bruker Data Analysis software, version 2.1 (Build 37) [9].

2.5. TCA precipitation of serum samples and MALDI-TOF-MS analysis

TCA, dissolved in 40 µL ice-cold water, was added to 20 µL of the original serum samples to reach a final concentration of 5%. After 30 min on ice, samples were filtered through 0.22 µm spin filters (Part # 5185-5990, Agilent) at 13000 g at 4°C for 10 min to remove particulates. Filtrates were used for further analysis.

For MALDI-TOF-MS analysis, 2 μL of the filtrates of the TCA precipitate were purified using C18 Stage tips (Proxeon, Odense, DK) according to the manufacturer's instructions. Peptides were eluted in 2 μL of 5 mg/mL α-Cyano-4-hydroxycinnamic acid (CHCA) in 50% AcN/0.1%TFA and directly spotted on a stainless-steel MALDI target. Analysis was performed in the positive ionization mode using a Voyager DE Pro instrument (Applied Biosystems, Foster City, California, USA). Spectra were acquired in reflectron mode with delayed extraction. Mono-isotopic molecular masses were considered in the further analysis and the instrument was calibrated using singly-charged BSA tryptic fragments with the m/z values of 927.49 Da and 2045.03 Da.

2.6. iTRAQ labeling of depleted serum

Changing of buffer and concentration of samples for iTRAQ labeling of depleted serum were done by ultrafiltration (Concentrators, Spin 5K MWCO, 4 mL, Part no 51855991, Agilent) using 5 mL of 10% AcN with 0.1% TFA in water. Samples were evaporated to dryness in a CentriVapConcentrator (LABCONCO, Kansas City, Missouri, USA) before labeling. iTRAQ labeling was performed according to the manufacturer's protocol [Applied Biosystems, iTRAQ™ Reagents Application Kit-Plasma (Amine-Modifying Labeling Reagents for Plasma Sample Applications)] with modification of the trypsin-to-protein ratio (1:14 instead 1:5.75 w/w). Sequencing grade modified trypsin was from Promega.

2.6.1. Strong Cation Exchange (SCE) fractionation

In order to remove excess iTRAQ reagent and to simplify the ensuing reversed-phase nanoLC-MS-MS analysis, the peptide mixture was washed and fractionated using a strong-cation exchange column [PolyLCinc 4.6 × 200mm column, column volume: 3.3 mL (Columbia, Maryland, USA)] operated at 0.2 mL/min (AKTA Purifier 10 with frac-900 fraction collector, GE Healthcare Bio-Sciences AB, Uppsala, Sweden). The mobile phase was comprised of two buffers: A: 5 mM KH_2PO_4/H_3PO_4 pH 3, 25% AcN and B: 5 mM KH_2PO_4/H_3PO_4 pH 3, 25% AcN, 1.0 M KCl. The KCl concentration was varied in three segments: 15%B [12 column volumes (CVs)], 50%B (3CVs), 100%B (5CVs), KCl (10 mM/min). The resulting 50 fractions (0.2 mL each) were pooled to obtain 20 fractions based on intensity and dried in a CentriVapConcentrator prior to nanoLC-MS-MS analysis.

2.6.2. Reversed-phase nanoLC-MS-MS

Derivatized peptides (pooled SCE fractions) were dissolved in 50 μL 2% AcN/0.1% FA and analyzed by nanoLC-MS/MS on a hybrid quadrupole time-of flight mass spectrometer (QSTAR® XL, Applied Biosystems) connected to an 1100 nanoHPLC system (Agilent). 10 μL of the SCE fractions were loaded onto a 0.3 mm × 0.5 cm C18-PepMap trapping column (Dionex, Sunnyvale, California, USA) at a flow rate of 10 µL/min (2% AcN/ 0.1% FA). After 30 minutes of washing, the trap column was switched and peptides were separated on a 75 µm × 15 cm C18 PepMap column (Dionex). Peptides were eluted at a flow rate of 300 nL/min using a 105 min gradient ranging from 95%A to 50%A [A: H_2O /AcN/FA (950:50:1), B: H_2O /AcN/FA (50/950/1)]. The outlet of the column was connected to an on-line standard coated silica picotip with a 10 µm ID (NewObjective, Woburn, Massachusetts, USA). The typical ionspray voltage was 2200V. Data was acquired using an independent data acquisition (IDA) protocol where, for each cycle, the 3 most abundant multiply-charged peptides (2 to 4 charges) in the MS scan with m/z values between 350 and 1500 amu were selected for MS/MS. For precursor ion selection, a threshold of 30 counts was applied. Each precursor ion was selected twice and then dynamically excluded for the following 60 seconds. Protein identification and quantification was carried out using ProQuant software v1.1 (Applied Biosystems). The search was performed against the Uniprot/SwissProt knowledge database (V49, downloaded May 2005). The search parameters allowed a tolerance up to 0.15 Da for precursor ion selection and the obtained MS/MS fragments, one missed cleavage (trypsin), oxidation of methionine (variable modification) and cysteine modification with iodoacetamide (fixed modification). ProGroup Viewer software v1.0.2 from Applied Biosystem was used to identify proteins with at least 95% confidence. The results obtained from ProGroup Viewer were exported to Microsoft Excel for further analysis.

2.7. iTRAQ labeling of TCA-precipitated serum supernatant

20 µL of TCA-precipitated serum supernatant were evaporated to dryness in a CentriVapConcentrator (LABCONCO) before labeling. iTRAQ labeling was performed according to the manufacturer's protocol with stable-isotopes [iTRAQ reagents with 114 (2 h) and 117 (6 h) isobaric tags] (see also section 2.6 for details). The samples were analyzed by LC-MS-MS with collision parameters optimized for observing the reporter mass tags.

2.8. NanoLC-MS-MS analysis of TCA-precipitated serum

NanoLC-MS/MS was performed as described under 2.6.2 with some modifications. The supernatant of TCA-precipitated serum was diluted ten fold with 2% AcN/ 0.1%FA, and 8 µL of the diluted sample were loaded onto a 0.3 mm × 0.5 cm C18-PepMap trapping column (Dionex) at a flow rate of 10 µL/min (2%AcN/0.1% FA). After 3 min of washing, the trap column was switched towards nanoLC and peptides were separated on a 75 µm × 15 cm C18 PepMap column (Dionex). Peptides were eluted using a 57 min gradient ranging from 98%A to 50%A (A: H_2O/AcN/FA; 950:50:1, B AcN/ H_2O /FA; 950:50:1). The column was connected to a spray capillary (10 µm tip) (NewObjective). The ion spray voltage applied was 2500 V. Data was acquired using an independent data acquisition (IDA) protocol where, for each cycle, the most abundant, multiply charged peptides (2 to 4 charges) in the MS scan with m/z between 350 and 1500 were selected for MS/MS. A threshold of 30 counts was applied. Each peptide was selected twice and then dynamically excluded for 15 seconds.

The resulting data was processed using Analyst QS 1.1/BioAnalyst software (Applied Biosystem) with MASCOT (Matrix Science, London, UK) used for protein/peptide identification : MS/MS search parameters: enzyme: none; variable modifications [deamidation (NQ), oxidation(M), Phospho (ST), pyroGlu (N-term E/Q), mono-isotopic, tolerance 0.2 Da]. The search was performed against the Uniprot/SwissProt knowledge database (V49, downloaded May 2005).

3. Data analysis

3.1. Computational environment

For (pre-)processing and multivariate statistical analysis the original Bruker Daltonics label-free LC-MS data files were converted into ASCII-format with the Bruker Data Analysis software. For all other type of data and statistical analyses (e.g. one-way ANOVA), and visualization Matlab (version 7.2.0.232 (R2006a), Mathworks, Natick, Massachusetts, USA) was used. Principal Component Analysis was performed using the PLS toolbox (version 3.5.2, Eigenvector Research Inc., Wenatchee, Washington, USA) under Matlab environment. All data preprocessing work was done on a personal computer equipped with a +3600 MHz AMD processor and with 4 GB of RAM.

3.2. Pre-processing and statistical analysis of label-free LC-MS data

Centroid data were smoothed and reduced using a normalized two-dimensional Gaussian filter, by rounding the nominal m/z ratios to 1 m/z (the original data had a resolution of 0.1 m/z). In the retention time dimension no data reduction was performed. This meshing procedure reduced the number of available data points by roughly a factor 10 and corrected for shifting m/z values as a result of different loadings of the ion-trap during elution of abundant peptides, a phenomenon that is common for ion-trap mass analyzers [26,27]. After meshing the data files, all chromatograms were time-aligned (warped) to a reference data file using Correlation Optimized Warping (COW) [28] based on TICs constructed from signals in the range 100-1500 m/z.

A modified M-N rule was applied for peak detection by first calculating a local median baseline using the sliding window technique separately for each m/z trace. A median window size of 1200 data points, corresponding to 20.84 min, was used with a moving rate of 10 points and the minimum median value was set to 200 counts. According to the M-N rule, a threshold of M-times the local baseline was used and a peak was assigned if, within one m/z trace, the signal exceeded this threshold for at least N consecutive points [26]. For each detected peak the m/z value, the mean retention times of the three highest measured intensities within the same peak reduced by the local baseline value were stored in a peak list created for every chromatogram. We used a similar approach as Radulovic et al. [26] to obtain optimal settings for M and N. Different values for M (1.5-4) and N (4-8) were applied to two blank LC-MS runs and to two LC-MS runs of depleted, trypsin-digested serum samples. Settings were used at which the ratio between the number of peaks (between 60-155 min) in the samples relative to the blank chromatograms was highest and at which a minimal number of peaks was extracted from the noise in the blank chromatogram (M = 2 and N = 5 in our case).

In order to combine the peak lists from different samples, one-dimensional peak matching was applied using the sliding window technique, in which the same m/z traces were evaluated for peaks that are proximate in time (step size 0.1 min; search window 1.0 min; maximal accepted standard deviation for all retention times within a group of matched peaks was 0.75 min). Missing peak allocation was performed by extracting the background subtracted local signal of the given m/z trace at the given retention time. The generated peak matrix, created from the peak lists of the individual samples, consisted of a peak(row)-sample(column)-intensity(value) matrix. This peak matrix was used for multivariate statistical analysis.

3.3. Analysis of iTRAQ data

The ProQuant software was used to calculate the intensity of 3 reporter ions (m/z: 115, 116 and 117) and to divide them by the intensity of the 4th reporter ion (m/z: 114) for each measured compound. Systematic experimental error occurring during sample preparation and measurement were corrected by subtracting the median of the corresponding ion ratio series (115/114, 116/114 and 117/114) for all compounds and adding one.

The data from two experimental series were analyzed. The first series (Series 1) contained samples with 1 and 2 h of clotting time with each sample measured in duplicate. The second series (Series 2) contained samples with 2, 4, 6 and 8 h of clotting time each of them measured once. Data from Series 1 were normalized to one sample with 1 h clotting time and data in Series 2 were normalized to the 2 h clotting time sample, which was used as denominator in the calculation of ion ratios. The number of identified proteins (95% confidence; see section 2.6.2) was 129 for Series 1 and 96 for Series 2, respectively.

Even after correction for experimental bias between different series of measurements as described above, there is fluctuation of the ion ratios within a single series. The aim of the further statistical analysis is to determine whether a difference in ion ratio is statistically significant relative to the measurement errors (expressed in terms of standard deviations). To this end, all ratios were transformed into decimal logarithms so that all ratios obtained from a single sample (excluding clotting time-related abundance differences) will have normal distribution. We selected "reference series" to calculate the standard deviation (SD) characterizing a particular set of measurements. In this study the reference series for Series 1 (1 and 2h clotting time measured in duplicate) were the duplicate samples at 1 and 2h, since differences between them cannot be due to clotting time. For Series 2 (2, 4, 6 and 8h clotting time) we based our reference series on the assumption that there are few, if any, differences between 6 and 8h clotting time and that differences between these time points are largely due to experimental error. Ion ratios were plotted against the number of identified proteins (supporting information Fig. S1), Gaussian curves were fitted on the smoothed histograms (histogram between -1 and 1 with 200 steps, smoothing using a Savitzky-Golay algorithm) and the standard deviations (SD) were determined. Proteins with log-transformed ion ratios differing by at least 3xSD (99.7% confidence) were considered significantly different from the random fluctuation calculated for the corresponding reference series. resulted in only one compound significantly different from the reference series. For Series 1 the mean of the 2 samples with 2 h of clotting time was used to select proteins that differed significantly different from the reference series. For Series 2 all proteins were included that showed a significant difference for at least one clotting time point. The log-transformed ratios of the selected proteins were used to visualize the clotting time-related differences (see Figure 7).

3.4. Statistical Analysis

For Principal Component Analysis (PCA) the processed data was mean-centered and normalized with respect to the standard deviation. One-way ANOVA was used for univariate statistics.

4. Results and discussion

The preparation of serum is a complex biochemical process and may be difficult to control, since cellular metabolism and the activity of extracellular enzymes continues for many hours after the collection of blood [6,7]. Some authors have noted that clotting time affects the resulting serum proteome, and that these effects are most pronounced in the low-molecular-weight portion, the so-called peptidome [1,3].

It is likely that changes in serum occur due to varying clotting times but it is not clear to which extent such changes affect subsequent proteomics analyses. This is dependent on the employed analytical approach (e.g. shotgun proteomics versus peptidomics) as well as on the use of label-free versus stable-isotope labeling techniques. Finally this depends also on the concentration sensitivity of the employed methodology, since high-abundance proteins may be less prone to significant relative change than those of lower abundance. Since biomarker discovery generally uses a comparative study design, it will also depend on how large and reproducible group-specific differences in serum proteomes are relative to variation induced by pre-analytical factors such as the clotting time.

In order to study the effect of clotting time on both the serum proteome and the serum peptidome, we have used two complementary methodologies. The first was based on tryptic digestion of serum after depletion of the six most abundant serum proteins followed by LC-MS (proteome), and the second on acid precipitation of serum and analysis of the supernatant (peptidome). The effect of clotting time was assessed with label-free and stable-isotope labeling techniques.

4.1. Label-free serum analysis after depletion and tryptic digestion

Serum was prepared and analyzed by capillary-LC ion-trap mass spectrometry (LC-MS) as previously described [9]. The six most abundant proteins were depleted by immunoaffinity chromatography [29]. In order to compare the effect of clotting time relative to a known change in the serum proteome, horse heart Cytochrome C was added to one group of samples at a concentration of 8 or 20µM prior to depletion (the concentration after depletion is 5-times lower due to co-depletion of Cytochrome C). This method was previously shown to have a concentration sensitivity of appr. 0.5µM for the added Cytochrome C. Clotting times of 1, 2, 4, 6 and 8h were compared, in duplicate, by both univariate (ANOVA) and multivariate (PCA) statistical analyses after data processing.

Using in-house-developed algorithms [10], a matched peak matrix was generated from the LC-MS data. This peak matrix contained 12659 aligned features that are defined in terms of retention time, m/z value and intensity. The LC-MS traces (Total Ion Chromatograms; TICs) of depleted, trypsin-digested serum samples showed little if any visible differences (Figure 1).

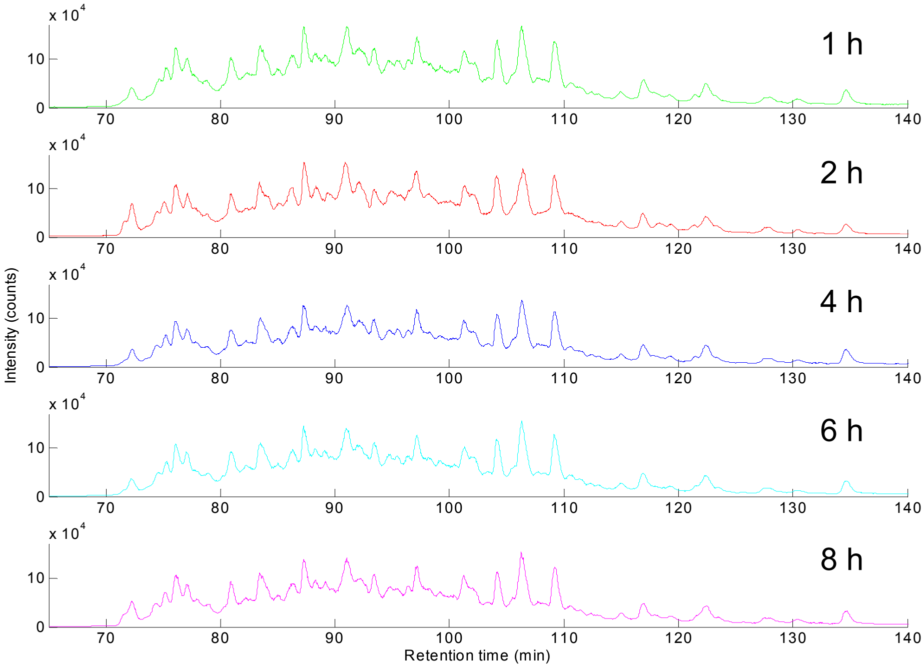

Figure 1. LC-MS analysis of immunodepleted, trypsin-digested human serum. Comparison of Total Ion Chromatograms (TICs) of serum samples obtained after 1, 2, 4, 6 and 8 h of clotting time.

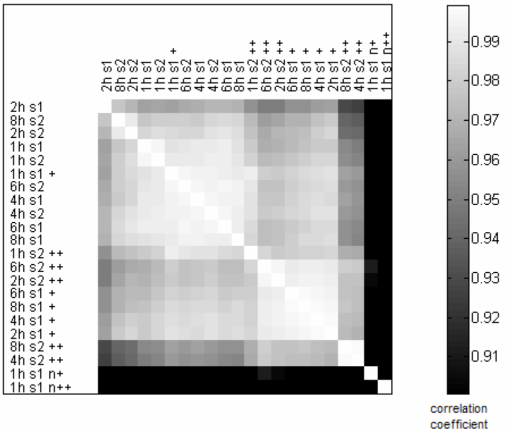

Figure 2. Correlation map of peak lists from all LC-MS analyses of depleted, trypsin-digested serum obtained after different clotting times (1, 2, 4, 6 and 8 h) and spiked with horse heart Cytochrome C (21 pmol (+), 50 pmol (++)) before the depletion step or with a tryptic digest of Horse heart Cytochrome C (21 pmol (n+) and 50 pmol (n++)) before LC-MS analysis. Highest correlation (1.0) = white, lowest correlation (0.9) = black. All non-spiked samples were analyzed in duplicate (s1 and s2) while single measurements were performed on the spiked samples. Samples are grouped by similarity (note that spiking of Cytochrome C results in a clearly decreased correlation coefficient, which is further decreased when a tryptic digest of Cytochrome C is added directly prior to LC-MS analysis).

A correlation map, comparing the peak lists of all LC-MS data files related to samples obtained after different clotting times (highest correlation (1.0) = white, lowest correlation (0.9) = black) confirmed this result by showing an overall high correlation (Figure 2). These results indicate the absence of major differences between serum samples with clotting times ranging from 1 to 8h analyzed by the described method. Contrary to this, the correlation map clearly separated the spiked (before depletion 21pmol (+) or 50pmol (++) and after depletion 21pmol (n+) and 50pmol (n++)) from the non-spiked samples, showing that the effect of spiking of Cytochrome C outweighs the effect of clotting time.

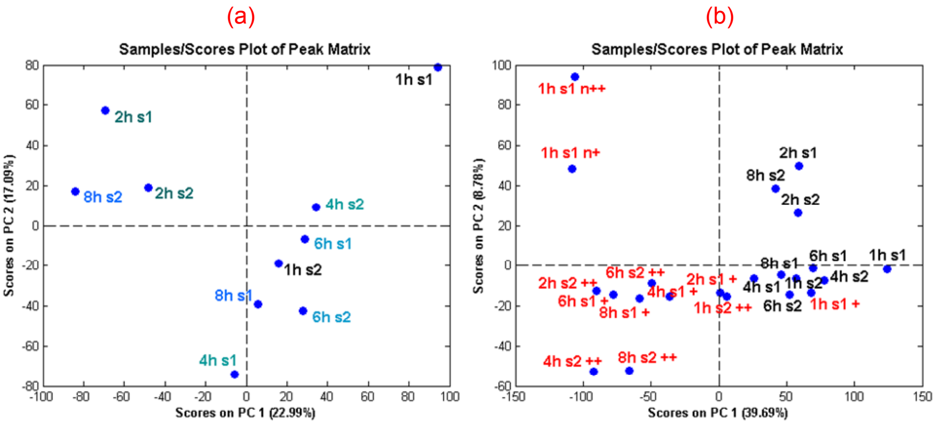

Figure 3. Principal Component Analysis (PCA) score plots of aligned peak matrixes obtained from LC-MS analysis of depleted, trypsin-digested serum after clotting times of 1, 2, 4, 6 and 8h. All analyses were performed in duplicate on different days (denominated as s1 and s2). Samples containing added horse heart Cytochrome C are labeled + (8 µM) or ++ (20 µM). In samples labeled n+ (8 µM) and n++ (20 µM) a tryptic digest of Cytochrome C was added directly prior to the LC-MS analyses. a) PCA scores plot for non-spiked samples; b) PCA scores plot for spiked (red) and non-spiked (black) samples.

The generated aligned peak matrix was subjected to Principal Component Analysis (PCA) and the calculated scores of PC1 and PC2 were visualized (Figure 3). Fig. 3a shows that there is no clear discrimination between samples with respect to clotting time. Differences between duplicate measurements (s1 and s2) appear to be similar when compared to differences between samples obtained after different clotting times. This indicates that varying the clotting time between 1 and 8h does not lead to detectable differences with the described analytical method. On the contrary, Fig. 3b shows that adding 8µM of Cytochrome C (appr. 1.6µM after depletion) leads to a clear discrimination between "spiked" (red) and "non-spiked" (black) samples. These data indicate that variable clotting times do not prevent detection of a single protein at a concentration of about 1.5 µM or 50µg/mL (for a 50kDa protein) using the described methodology. Proteins that are present at this concentration are, however, still considered "classical serum or plasma proteins" [30]. To find smaller local differences between samples that are the result of different clotting times supervised classification methods should be used instead of unsupervised multi-variate statistical methods.

The obtained peak lists were further analyzed by univariate statistical analysis (ANOVA). In this way a p-value is calculated for each peak expressing the magnitude of the effect of a certain factor (clotting time or spiking with Cytochrome C). Table 1 lists several peaks showing significant differences (p-values < 1.6×10^{-4}) between the samples obtained at different clotting times and Figure 4 shows box plots for several of the listed peaks. While there is a fairly even distribution of p-values when considering clotting time only, there are more peaks with very low p-values in the case of added Cytochrome C (p-values < 2.2×10^{-11}; Table 2) indicating that addition of Cytochrome C clearly outweighs the effect of clotting time.

The peaks shown in Table 1 were subjected to LC-MS-MS analysis and identified as corresponding to fragments of Fibrinogen alpha chain (FIBA_HUMAN (P02671); m/z = 733 and 809) or to Fibrinogen beta chain (FIBB_HUMAN (P02675); m/z = 1325) with a pyroglutamic acid at the N-terminus (see Fig. S2-S4, supporting information for further details), while all peaks in Table 2 were derived from Cytochrome C. Extracted Ion Chromatograms (EICs) of discriminative peaks corresponding to fragments of Fibrinogen alpha chain (m/z: 733, 91.22 min and m/z: 809, 93.42 min) at 1, 2, 4, 6 and 8 h of clotting times are presented in Figure S5 (supporting information).

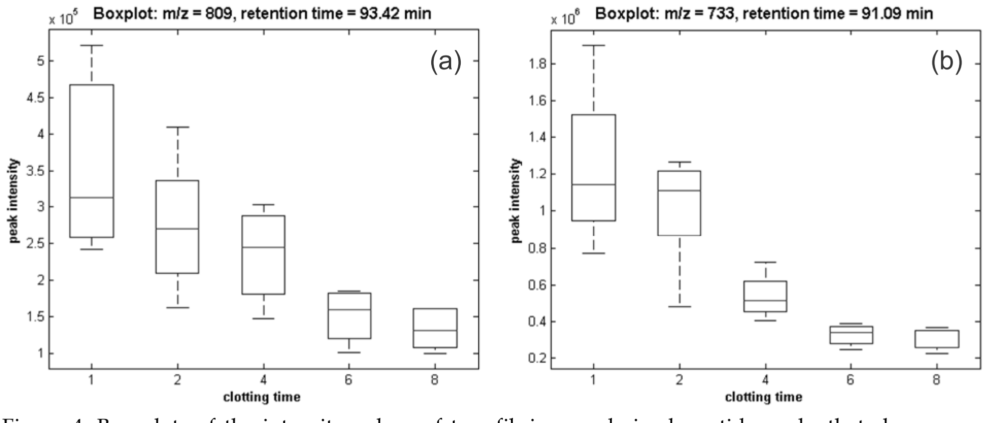

Figure 4. Box plots of the intensity values of two fibrinogen-derived peptide peaks that change significantly with clotting time (see Table 1). Peptides were identified as fragments of Fibrinogen alpha chain [(a) DSGEGDFLAEGGGVR; (b) ADpSGEGDFLAEGGGVR); FIBA_HUMAN (P02671)] (see Fig. S2 and S3, supporting information).

The peaks which correspond to p-values below 1.6 × 10^{-4} with respect to clotting time are listed in Table 1 and those with respect to Cytochrome C addition (p-values below 2.2 × 10^{-11}) in Table 2.

Table 1. Peaks detected in serum that are significantly affected by clotting times ranging from 1 to 8h. Calculations are based on a comparison of intensities by one-way ANOVA.

p-value	m/z	Retention time (min)	Peptide sequence
2.6 10^{-7}	733 [1]	91.09	DSGEGDFLAEGGGVR
1.1 10^{-4}	1325 [2]	103.87	QGVNDNEEGFFS
1.6 10^{-4}	809	93.42	ADSGEGDFLAEGGGVR

[1] isotopic peaks at m/z 732 (p = 1.4 x 10^{-4}) and 734 (p = 6.2 x 10^{-8}) show also significant differences.

[2] the isotopic peak at m/z 1324 shows also a significant difference (p = 2.8 x 10^{-4}).

Table 2. Peaks detected in serum that are significantly affected by the addition of horse heart Cytochrome C (8µM prior to depletion) independent of clotting times ranging from 1 to 8h. Calculations are based on a comparison of intensities by one-way ANOVA.

m/z measured[1]	m/z expected	charge state	retention time (min)	p-value	Cytochrome C sequence
634	634,4	+1	76.14	<1.00e-015	IFVQK
453	454,3	+2	92.18	1.74e-014	MIFAGIKK
966	964,5	+1	98.67	9.16e-014	EDLIAYLK
559	559,3	+3	99.22	9.51e-014	GGKHKTGPNLHGLFGR
452	450,9	+3	92.22	2.30e-013	TEREDLIAYLK
749	748,3	+2	104.36	4.57e-013	EETLMEYLENPK
782	781,4	+2	95.24	4.61e-013	HKTGPNLHGLFGRK
798	799,9	+2	103.19	7.82e-012	KTGQAPGFTYTDANK
1169	1168,6	+1	91.82	1.60e-011	TGPNLHGLFGR

[1] after meshing to integer values

4.2. Detailed analysis of Fibrinopeptides after TCA precipitation

The analysis of depleted and trypsin-digested serum indicated that fibrinogen-related peptides were affected by clotting time, as might be expected. We therefore examined the effect of clotting time on the overall serum peptidome in more detail. TCA precipitation after 2 and 6 hours clotting time was used to enrich the low molecular weight fraction and the resulting supernatant was analyzed by MALDI-TOF-MS to gain an overview of the peptide patterns (Figure 5).

A small set of prominent peptide peaks was observed. The spectrum was dominated by a peak at m/z 1465.8, which was identified as an N-terminally truncated form of Fibrinopeptide A (FPA) (DSGEGDFLAEGGGVR, doubly-charged ion of m/z 733.8 (supporting information, Fig. S2, Table S1)) by LC-MS/MS. This peptide was also shown to be most significantly affected by clotting time based on the univariate statistical analysis of the LC-MS data (see Table 1). A minor signal for intact FPA (ADSGEGDFLAEGGGVR) at m/z 1536.6 was also observed. Database-search of the LC-MS/MS data showed that the majority of the observed signals using this approach were derived from FPA (Table S1). Interestingly, two post-translational modifications of FPA were seen in the MALDI spectrum. Peaks at m/z 1545.7 and 1616.8 correspond to the phosphorylated forms DpSGEGDFLAEGGGVR (doubly-charged ion m/z 773.4) and ADpSGEGDFLAEGGGVR (doubly-charged ion m/z 808.8) as confirmed by LC-MS/MS (supporting information, Fig. S3,S4)). The MALDI spectrum showed also two less well-resolved peaks at m/z 1524,5 and m/z 1453 (Figure 5C). These peaks represented metastable ions due to the formation of dehydro-alanine (loss of phosphoric acid from the phosphorylated serine during flight to the detector). Besides FPA-related peptides, partial sequences of Fibrinopeptide B (FPB) were also identified.

In order to gain further insight into the ratios of FPA- and FPB-related peptides at 2 and 6h clotting time, supernatants after TCA precipitation were labeled with stable-isotopes (iTRAQ reagents with 114 (2h) and 117 (6h) isobaric tags). The samples were analyzed by LC-MS-MS with collision parameters optimized for observing the reporter mass tags. Database search of the MS/MS data sets allowed the identification of a greater variety of peptides than in the original, targeted analyses. Fragments of FPA and FPB were readily identified but quantification was problematic, due to the fact that the software (proQuantTM 1.1) is designed for tryptic peptides and not for naturally occurring peptides with non-tryptic cleavage sites. Extracted ion traces related to Fibrinopeptides from the data base search were therefore inspected manually. As an examples, the MS/MS spectra of the iTRAQ-labeled peptides ADpSGEGDFLAEGGGVR and FLAEGGGV are shown in Figure 6. In the insert the iTRAQ reporter ions indicate that the amount of this peptide is reduced by a factor 1.75 between 2 and 6h clotting time (Figure 6A), whereas another iTRAQ-labeled FPA peptide (FLAEGGGV; m/z 447 doubly-charged ion) showed a 2.8-fold increase over this time period (Figure 6B). An overview of all identified peptides is given in Table S2 (supporting information).

These experiments show that a number of peptides derived from Fibrinogen changed in abundance with respect to clotting time and that these peptides can be easily detected in the so-called peptidome due to their relatively high-abundance.

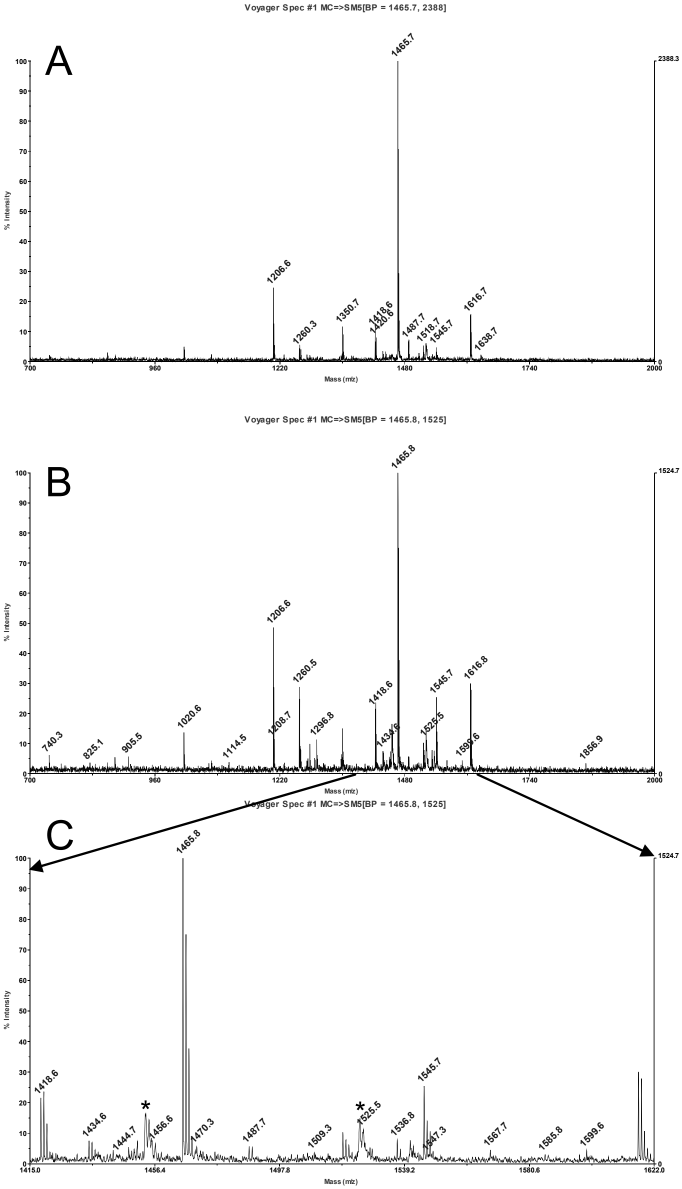

Figure 5. MALDI-TOF-MS spectrum of the supernatant of TCA-precipitated serum after 2 (A) and 6h (B) clotting time. The major peak at 1465.8 corresponds to an N-terminally truncated form of Fibrinopeptide A (FPA) and the minor peak at 1536.8 (Fig. 5C) corresponds to full-length FPA. These forms of FPA were also observed as serine-phosphorylated forms (m/z 1545.7 and m/z 1616.8). (C) Metastable ions at m/z 1453 and 1524,5 (labeled with * in panel (C)) indicate the presence of dehydro-alanine due to decay of the metastable phosphoserine residue during MALDI-TOF-MS analysis.

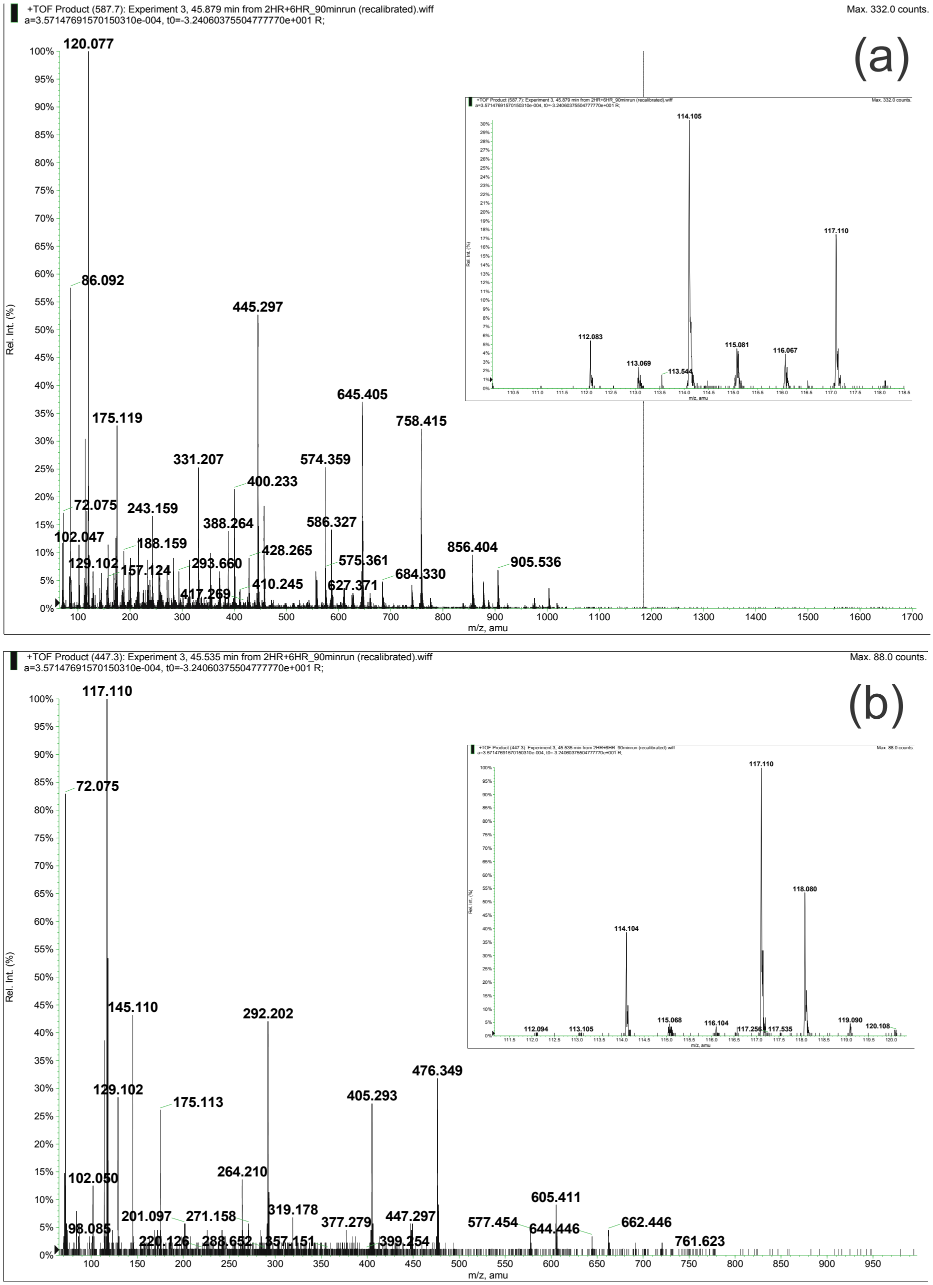

Figure 6. MS/MS spectrum of the iTRAQ-labeled, phosphorylated Fibrinopeptide A (FPA) (ADpSGEGDFLAEGGGVR) (a). The insert shows the relative abundance of this peptide at 2h [114 mass tag] and 6h [117 mass tag] clotting time (ratio 2h/6h = 1.75). MS/MS spectrum of iTRAQ-labeled truncated FPA (FLAEGGGV; m/z 447.3 doubly-charged ion) which increases by 2.8-fold between 2 [mass tag 114] and 6h [mass tag 117] clotting time (b).

4.3. Multiplexed analysis of trypsin-digested serum

Since the concentration sensitivity of the label-free analytical method described in section 4.1 reaches only about 0.5µM, it was of interest to apply a two-dimensional HPLC method with stable-isotope labeling to obtain a more comprehensive view of the effect of clotting time on the serum proteome. The initial experimental design consisted of varying clotting time between 2, 4, 6 and 8 h followed by immunodepletion of the six most abundant proteins. Proteins were furthermore reduced and alkylated and digested with trypsin prior to derivatization with a different isobaric tag for each clotting time [114 (2h), 115 (4h), 116 (6h) and 117 (8h)]. The combined samples (50 µg each) were fractionated by Strong Cation-Exchange HPLC (SCX; 20 fractions) prior to reversed-phase nanoLC-MS-MS analysis. The LC-MS-MS data were used for relative quantification and identification of proteins. 96 proteins were identified with a confidence higher than 95% and quantified relative to each other using the reporter mass tags in the MS/MS spectra of selected multiply-charged ions. Examples of spectra produced after iTRAQ labeling are shown (supporting information, Figure S6).

All the data files were analyzed separately using the ProQuant software. This software package is able to quantify the reporter ions and to calculate the ratio relatively to a preset denominator, in this case the 2h clotting time point (mass tag 114). A bias (multiplication factor) was applied in order to correct for systematic differences in the iTRAQ ratios due to experimental errors related to sample preparation (e.g. combining the different reaction mixtures; small differences in dilutions). The software calculates this bias and corrects the ratios assuming that the majority of proteins does not differ in response to clotting time and that the median iTRAQ ratio should be 1. In the experiment used the mean bias applied was 1 for the 114 (2h) signal (was set to 1), 1.09 for 115 (4h), 1.34 for 116 (6h) and 1.14 (8h) for 117.

Visualization of the obtained ratios for the 96 identified proteins (Figure 7a) showed that 10 proteins have significant differences in ratios relative to the 2h clotting time point (Figure 7a). Three of them (Complement component C8 alpha chain precursor (P07357), Hepatocyte growth factor activator precursor (HGF activator) (Q04756), Ig kappa chain C region (P01834)) were increased and seven (Insulin-like growth factor II precursor (IGF-II) (Somatomedin A) (P01344), Leucine-rich alpha-2-glycoprotein precursor (LRG) (P02750), Complement factor H-related protein 1 precursor (FHR-1) (H factor-like protein 1) (H-factor-like 1) (Q03591), Serum amyloid P-component precursor (SAP) (9.5S alpha-1-glycoprotein) ((P02743), Coagulation factor IX (P00740), Complement factor H-related protein 2 precursor (P36980) and Apolipoprotein C-IV precursor (P55056)) were decreased. All mentioned proteins were increased or decreased relative to the reference series (see Materials and Methods; section 3.3) with a confidence above 99.7%. However, none of the mentioned proteins was more than +/- 50% increased or decreased.

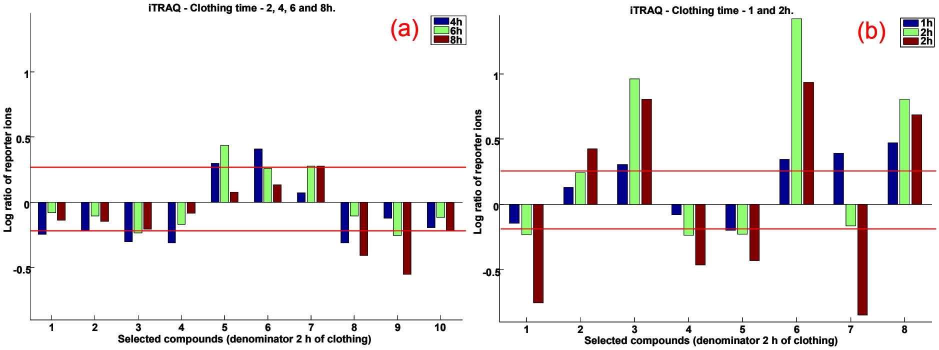

Figure 7. Compound showing significant difference in sample having clotting time of 2, 4, 6 and 8 hours (a) and clotting time 1 and 2 hours (in duplicate) (b) using iTRAQ quantification. The number on the figures are corresponding to the following proteins: (a): 1: SAMP_HUMAN (P02743) 2: FA9_HUMAN (P00740), 3: FHR2_HUMAN (P36980), 4: APOC4_HUMAN (P55056), 5: CO8A_HUMAN (P07357), 6: HGFA_HUMAN (Q04756), 7: KAC_HUMAN (P01834), 8: IGF2_HUMAN (P01344), 9: A2GL_HUMAN (P02750), 10: FHR1_HUMAN (Q03591); (b): 1: APOC1_HUMAN (P02654), 2: FIBA_HUMAN (P02671), 3: SC1_HUMAN (Q08554), 4: APOF_HUMAN (Q13790), 5: A2GL_HUMAN (P02750), 6: ALBU_HUMAN (P02768), 7: EPC2_HUMAN (Q52LR7), 8: HV3P_HUMAN (P01777).

Since it was not possible, at the time, to include more then 4 different clotting time points in one iTRAQ experiment, we compared the effects of shorter clotting times [1h (mass tags 114, 115) and 2 h (mass tags 116, 117)] in separate duplicate experiments. The analysis resulted in the identification of 129 proteins at the 95% confidence level. Based on an analogous analysis of the obtained data, eight proteins were shown to differ between serum samples with 1h as compared to 2h clotting time (Fig. 7b). Human serum albumin precursor (P02768), Fibrinogen alpha chain (P02671), Desmocollin-1 (Q08554) and Ig heavy chain V-III region TEI (P01777) were increased in concentration (up to +70%) and Apolipoprotein C-I (P02654), Apolipoprotein F (Q13790), Leucine-rich alpha-2-glycoprotein (P02750) and Enhancer of polycomb homolog 2 (Q52LR7) were decreased by at least 100%. Overall it appears that changes to the serum proteome are more drastic between 1 and 2 h clotting time and that the serum proteome “stabilizes” after 2 h.

Fibrinogen alpha chain (P02671), from which Fibrinopeptide A is cleaved off during blood clotting, was up-regulated after 2h clotting time compared to 1h but this difference was not significant at a confidence level of 99.7%. Fibrinogen alpha chain is involved in blood clotting and Fibrinopeptide A is the product of Fibrinogen cleavage during coagulation. Since we did not observed further changes in the relative abundance of this peptide after 2h clotting time, it is likely that coagulation was completed (Figure 7b).

Coagulation factor IX (P00740), consisting of Coagulation factor IXa light chain and Coagulation factor IXa heavy chain, was down-regulated after two hours clotting compare to 4, 6 and 8 hours but only in -0.1/-0.2 times which fall into the the normal measurement fluctuation. Leucine-rich alpha-2-glycoprotein (P02750) was down-regulated in all samples compared to one hour clotting sample. Leucine-rich alpha-2-glycoprotein is known as marker for granulocytic differentiation [31] which is involved in formation of platelets. Platelets are components of blood coagulation and activate the final step in the process of blood coagulation, thus Leucine-rich alpha-2-glycoprotein related to blood coagulation cascade.

Our studies with immunodepleted, trypsin-digested serum indicate that the serum proteome is rather stable once clotting has proceeded for at least 2h. Univariate statistical comparisons of the processed LC-MS data showed that fibrinogen-derived peptides change significantly with clotting time (p-values 1.6 x 10^{-4} to 7 × 10^{-8}) but that most other peptides remain stable in their intensities (+/- 25%). A more detailed analysis of these peptides revealed that they form a family of truncated forms, some of which contain post-translational modifications such as phospho-serine or pyroglutamic acid. Stable-isotope labeling combined with a two-dimensional chromatographic separation showed that 10 out of 96 identified proteins showed statistically significant differences between 2, 4, 6 and 8h clotting time. Our results indicate that serum stabilizes after 2h clotting under the described conditions, while shorter clotting times need to be more tightly controlled. This opens the possibility of using disease-relevant serum samples for biomarker discovery that would otherwise be very difficult to generate de novo.

Other authors have observed that clotting time needs to be tightly controlled when focusing on the so-called peptidome, the low-molecular weight complement of the proteome [3,5]. This is confirmed by our results, which show that especially fibrinogen-related peptides vary greatly with clotting time. These fibrinopeptides form a family of related molecules that are N- or C-terminally truncated as well as phosphorylated or containing N-terminal pyro-glutamic acid residues (see Table S2, supporting information). An increasing number of authors report that proteolytic degradation products of high-abundance proteins may be related to cancer and may thus be of interest as biomarkers [32,33]. This interesting observation was based on a rigorously controlled sample handling and preparation protocol, which may be difficult to implement in routine clinical diagnostic laboratories [8]. Some of these biomarker candidates were forms of Fibrinopeptides, which are also affected by clotting time. It will be interesting to follow the development of this field of biomarker research, since cancer development is often associated with changes in the balance between the proteolytic systems of the coagulation cascade, the fibrinolytic and the complement system. Villanueva et al. [33] assume that these major proteolytic systems provide the "founder peptides" that are subsequently further degraded by cancer-specific proteases.

Our study sheds additional light on the effect of clotting time on the composition of the serum proteome and peptidome. Biomarker discovery is relying on changes in protein or peptide abundance related to disease relative to unavoidable changes due to biological and analytical variations. Abundance ratios of some 100 identified proteins changed no more than 10 between 2 and 8h clotting time while changes were 8 between 1 and 2h. Most analyzed proteins were not significantly affected by clotting time. While the expected changes in putative, disease-relevant biomarkers can not be predicted, it is possible to make a calculation as to how large the relative difference should be in order to be statistically significant. The kinetics of blood coagulation are affected by multiple factors related to the blood collection tubes (e.g. some tubes contain clotting activators) as well as to the medication a patient might be taking (e.g. use of anti-coagulants, chemotherapy). We did not study these parameters but it is clear that they must be taken into account in comparative biomarker discovery studies. The detected fibrinopeptides might be good indicators to assess the clotting conditions a posteriori and thus to define inclusion/exclusion criteria for serum samples in existing biobanks. The question whether serum or plasma should be the preferred blood derivative for a given study cannot be answered definitively. Both types of samples have their advantages and disadvantages and plasma has also been shown to be dependent on a number of parameters, notably the anti-coagulant [3].

References

[1]. Drake RR, Cazares LH, Corica A, Malik G, Schwegler EE, Libby AE, Wright GL, Semmes OJ, Adam BL. Qualit control, preparation and protein stability issues for blood serum and plasma used in biomarker discovery and proteomic profiling assays. Bioprocessing Journal 2004;43-49.

[2]. Hsieh SY, Chen RK, Pan YH, Lee HL. Systematical Evaluation of the Effects of Sample Collection Procedures on Low-Molecular-Weight Serum/Plasma Proteome Profiling. Proteomics 2006;6(10):3189-3198.

[3]. Tammen H, Schulte I, Hess R, Menzel C, Kellmann M, Mohring T, Schulz-Knappe P. Peptidomic Analysis of Human Blood Specimens: Comparison Between Plasma Specimens and Serum by Differential Peptide Display. Proteomics 2005;5:3414-3422.

[4]. Timms JF, Arslan-Low E, Gentry-Maharaj A, Luo Z, T'jampens D, Podust VN, Ford J, Fung ET, Gammerman A, Jacobs I, Menon U. Preanalytic Influence of Sample Handling on SELDI-TOF Serum Protein Profiles. Clin. Chem. 2007;53(4):645-656.

[5]. Villanueva J, Philip J, Chaparro CA, Li Y, Toledo-Crow R, DeNoyer L, Fleisher M, Robbins RJ, Tempst P. Correcting Common Errors in Identifying Cancer-Specific Serum Peptide Signatures. J Proteome. Res. 2005;4:1060-1072.

[6]. Ekblad L, Baldetorp B, Ferno M, Olsson H, Bratt C. In-Source Decay Causes Artifacts in SELDI-TOF MS Spectra. J. Proteome Res. 2007;6(4):1609-1614.

[7]. Aristoteli LP, Molloy MP, Baker MS. Evaluation of Endogenous Plasma Peptide Extraction Methods for Mass Spectrometric Biomarker Discovery. J. Proteome Res. 2007;6:571-581.

[8]. Villanueva J, Philip J, Entenberg D, Chaparro CA, Tanwar MK, Holland EC, Tempst P. Serum Peptide Profiling by Magnetic Particle-Assisted, Automated Sample Processing and MALDI-TOF Mass Spectrometry. Anal. Chem. 2004;76:1560-1570.

[9]. Govorukhina NI, Reijmers TH, Nyangoma SO, van der Zee AGJ, Jansen RC, Bischoff R. Analysis of Human Serum by Liquid Chromatography-Mass Spectrometry: Improved Sample Preparation and Data Analysis. Journal of Chromatography A 2006;1120:142-150.

[10]. Kemperman RF, Horvatovich PL, Hoekman B, Reijmers TH, Muskiet FA, Bischoff R. Comparative Urine Analysis by Liquid Chromatography-Mass Spectrometry and Multivariate Statistics: Method Development, Evaluation, and Application to Proteinuria. J. Proteome Res. 2007;6:194-206.

[11]. Aguiar M, Masse R Gibbs BF. Mass Spectrometric Quantitation of C-Reactive Protein Using Labeled Tryptic Peptides. Analytical Biochemistry 2006;354:175-181.

[12]. Berna MJ, Zhen Y, Watson DE, Hale JE, Ackermann BL. Strategic Use of Immunoprecipitation and LC/MS/MS for Trace-Level Protein Quantification: Myosin Light Chain 1, a Biomarker of Cardiac Necrosis. Anal. Chem. 2007 Jun 1;79(11):4199-205.

[13]. Bondar OP, Barnidge DR, Klee EW, Davis BJ, Klee GG. LC-MS/MS Quantification of Zn-{Alpha}2 Glycoprotein: A Potential Serum Biomarker for Prostate Cancer. Clinical Chemistry;2007;53(4):673-678.

[14]. Lin S, Shaler TA, Becker CH. Quantification of Intermediate-Abundance Proteins in Serum by Multiple Reaction Monitoring Mass Spectrometry in a Single-Quadrupole Ion-trap. Anal. Chem. 2006;78:5762-5767.

[15]. Liu T, Qian WJ, Gritsenko MA, Xiao W, Moldawer LL, Kaushal A, Monroe ME, Vamum SM, Moore, R. J.; Purvine, S. O.; Maier, R. V.; Davis, R. W.; Tompkins, R. G.; Camp, D. G.; Smith RD. High Dynamic Range Characterization of the Trauma Patient Plasma Proteome. Mol. Cell Proteomics. 2006;5(11):2167-2174.

[16]. Qian WJ, Jacobs JM, Liu T, Camp DG, Smith RD. Advances and Challenges in Liquid Chromatography-Mass Spectrometry Based Proteomic Profiling for Clinical Applications. Mol. Cell Proteomics. 2006;5(10):1727-1744.

[17]. Shen Z, Want EJ, Chen W, Keating W, Nussbaumer W, Moore R, Gentle TM, Siuzdak G. Sepsis Plasma Protein Profiling With Immunodepletion, Three-Dimensional Liquid Chromatography Tandem Mass Spectrometry, and Spectrum Counting. J. Proteome. Res. 2006;5:3154-3160.

[18]. Wang G, Wu WW, Zeng W, Chou CL, Shen RF. Label-Free Protein Quantification Using LC-Coupled Ion-trap or FT Mass Spectrometry: Reproducibility, Linearity, and Application With Complex Proteomes. J. Proteome. Res. 2006;5,1214-1223.

[19]. Petricoin EF, Belluco C, Araujo RP, Liotta LA. The Blood Peptidome: a Higher Dimesion of Information Content for Cancer Biomarker Discovery. Nat. Rev. Cancer. 2006;12:961-967.

[20]. Raida M, Schulz-Knappe P, Heine G, Forssmann WG. Liquid Chromatography and Electrospray Mass Spectrometric Mapping of Peptides From Human Plasma Filtrate. Journal of the American Society for Mass Spectrometry 1999;10:45-54.

[21]. Schrader M, Schulz-Knappe P. Peptidomics Technologies for Human Body Fluids. Trends in Biotechnology 2001;19:S55-60.

[22]. Schulz-Knappe P, Schrader M. Peptidomics in Biomarker and Drug Discovery. Current Drug Discovery 2003;21-24.

[23]. Schulz Knappe P, Raida M, Meyer M, Quellhorst EA, Forssmann WG. Systematic Isolation of Circulating Human Peptides: the Concept of Peptide Trapping. Eur J Med Res 1996;1,223-36.

[24]. Schulz Knappe P, Schrader M, Standker L, Richter R, Hess R, Jurgens M, Forssmann WG. Peptide Bank Generated by Large-Scale Preparation of Circulating Human Peptides. J Chromatogr A 1997;776:125-132.

[25]. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S, Purkayastha S, Juhasz P, Martin S, Bartlet-Jones M, He F, Jacobson A, Pappin DJ. Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents. Mol Cell Proteomics 2004;3,1154-1169.

[26]. Radulovic D, Jelveh S, Ryu S, Hamilton TG, Foss E, Mao Y, Emili A. Informatics Platform for Global Proteomic Profiling and Biomarker Discovery Using Liquid Chromatography-Tandem Mass Spectrometry. Mol. Cell Proteomics. 2004;3,984-997.

[27]. Hao C, March RE. A Survey of Recent Research Activity in Quadrupole Ion-trap Mass Spectrometry. International Journal of Mass Spectrometry 2001;212,337-357.

[28]. Tomasi GF. v. d. BCA. Correlation Optimized Warping and Dynamic Time Warping As Preprocessing Methods for Chromatographic Data. Journal of Chemometrics 2004;18:231-241.

[29]. Dekker LJ, Bosman J, Burgers PC, van Rijswijk A, Freije R, Luider T, Bischoff R. Depletion of High-Abundance Proteins From Serum by Immunoaffinity Chromatography: A MALDI-FT-MS Study. Journal of Chromatography B 2007;847:65-69.

[30]. Anderson NL, Anderson NG. The Human Plasma Proteome: History, Character, and Diagnostic Prospects. Mol Cell Proteomics 2002;1:845-867.

[31]. O'Donnell LC, Druhan LJ, Avalos BR. Molecular characterization and expression analysis of leucinerich α-2-glycoprotein, a novel marker of granulocytic differentiation. Journal of Leukocyte Biology 2002;72:478-485.

[32]. Overall CM, Dean RA. Degradomics: Systems Biology of the Protease Web. Pleiotropic Roles of MMPs in Cancer. Cancer Metastasis Rev. 2000;25:69-75.

[33]. Villanueva J, Shaffer DR, Philip J, Chaparro CA, Erdjument-Bromage H, Olshen AB, Fleisher M, Lilja H, Brogi E, Boyd J, Sanchez-Carbayo M, Holland EC, Cordon-Cardo C, Scher HI, Tempst P. Differential Exoprotease Activities Confer Tumor-Specific Serum Peptidome Patterns. J. Clin. Invest. 2006;116,271-284.

Supplementary material

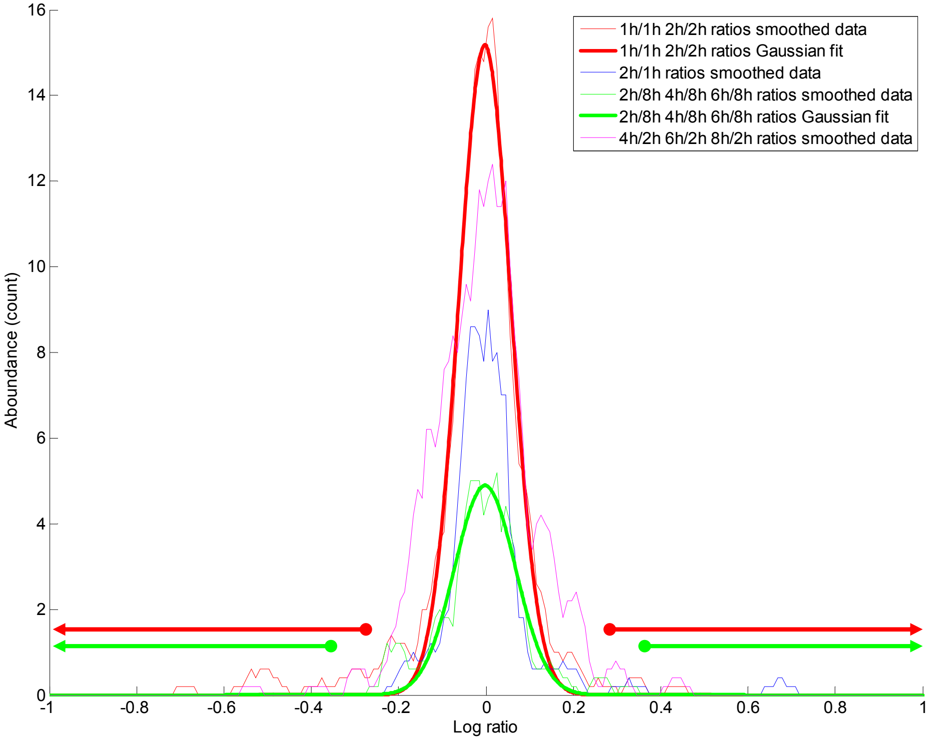

Figure S1. Evaluation of iTRAQ data. The thin red and green are the smoothed logtransformed ratio of the references sets after correction with the median, and the bold corresponding lines are the Gaussian fit. The thin blue and purple curve is the ratios corresponding to set 1 and 2 in which significance peak ratios has to be determined. The arrows are indicating the log ratio of area considered as significant 3 SD (99.7%).

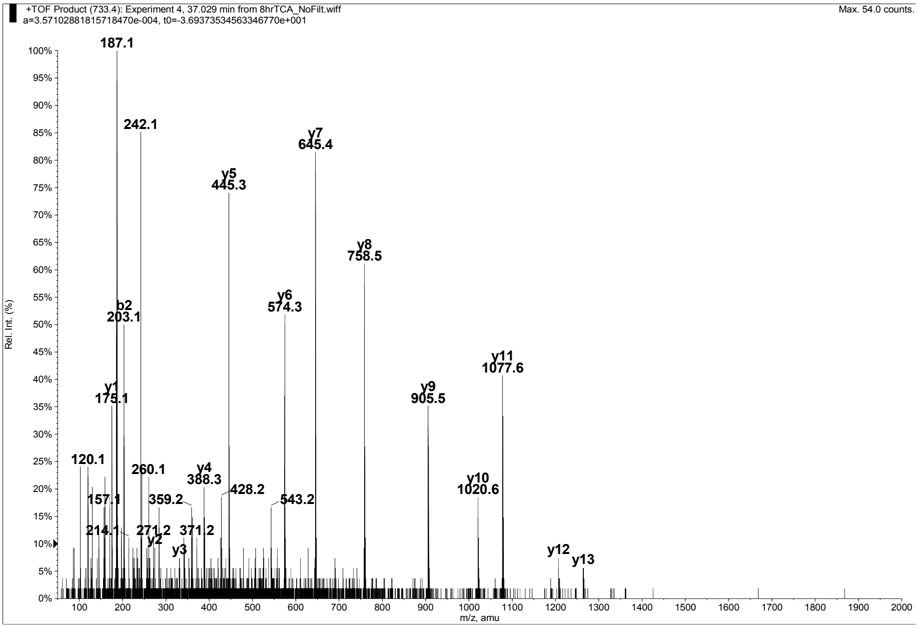

Figure S2. MS-MS spectrum of a FPA fragment (DSGEGDFAEGGGGVR; m/z 733.4 , doubly charged)

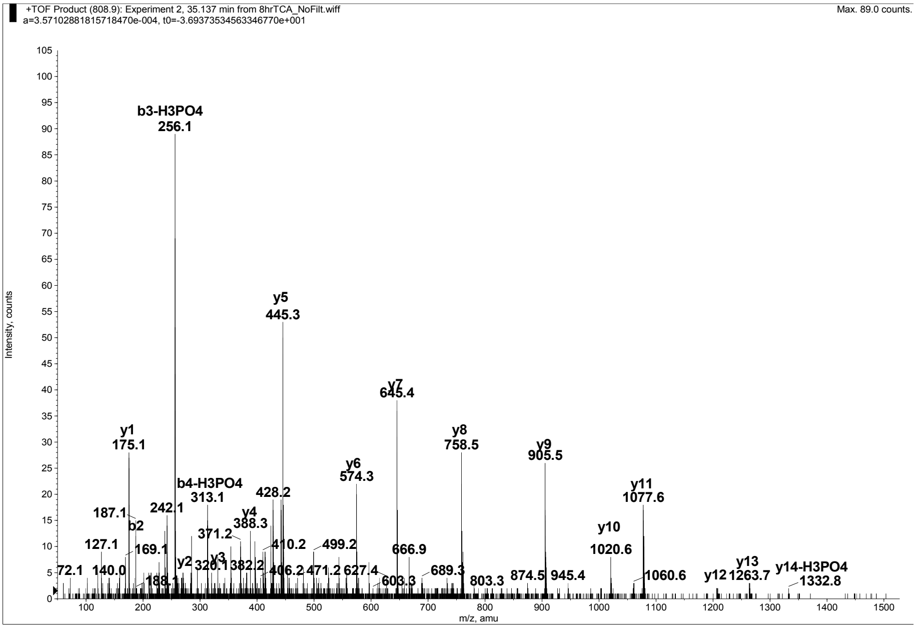

Figure S3. MS/MS spectrum of a phosphorylated FPA fragment (ADpSGEGDFLAEGGGVR; m/z 808.9, doubly-charged).

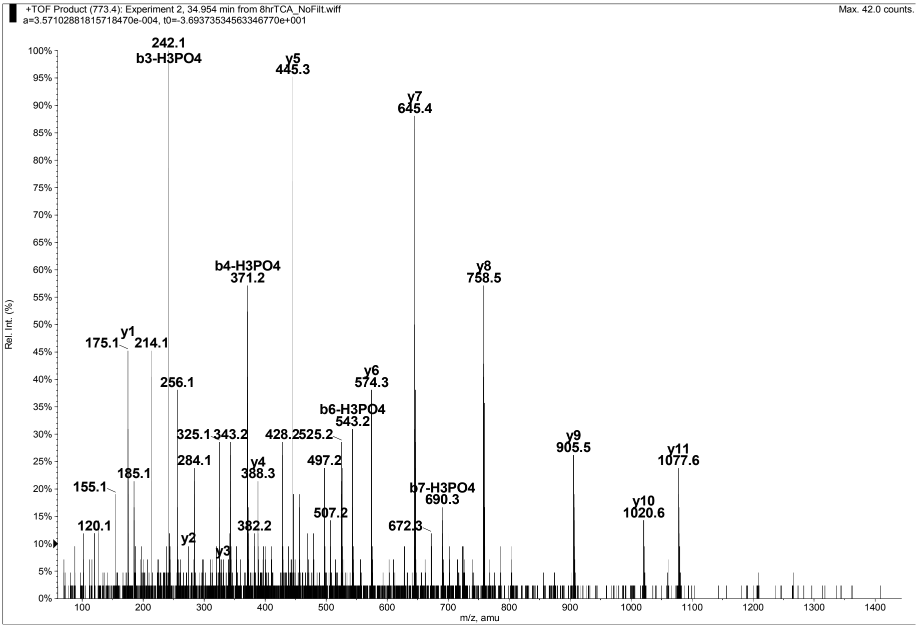

Figure S4. MS-MS spectrum of a phosphorylated FPA fragment m/z 773.4 (MM1464.8, DpSGEGDFLAEGGGVR, doubly charged)

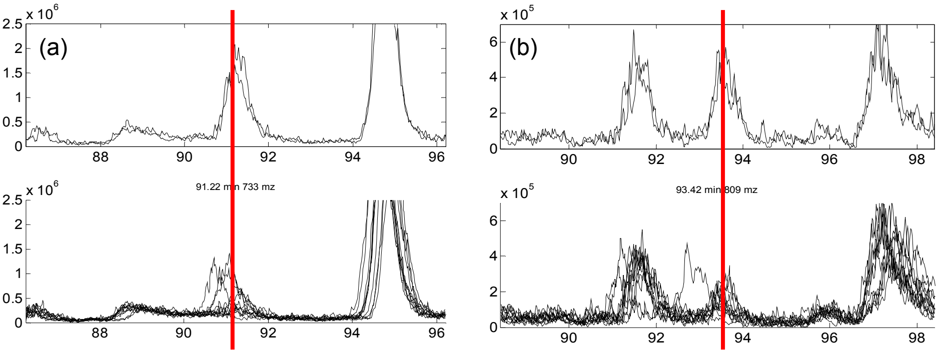

Figure S5. Discriminative peaks [733 m/z, 91.22 min (a) and 809 mz and 93.42 min (b)] between 1 (upper traces) and 2 (lower traces), 4, 6 and 8 h (lower traces) of clothing time.

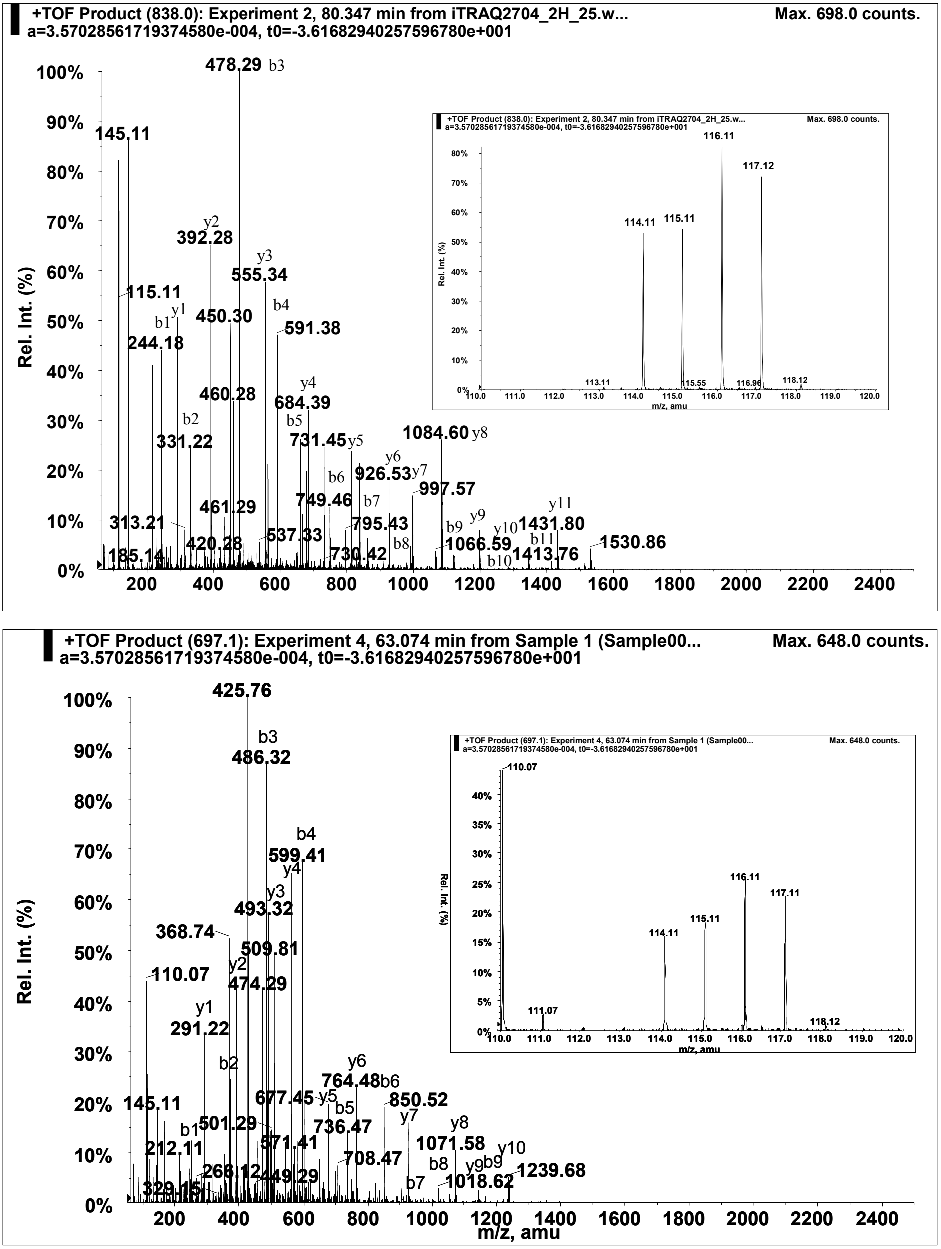

Figure S6. Examples of iTRAQ data for 2, 4, 6 and 8h clotting time is given. ApoA1 (a) and Complement C3 precursor (b) identified. In inserts 114, 115, 116 and 117 ions showed.

Mascot Paramaters; Database Swissprot,
Type of search : MS/MS Ion Search

Enzyme : None
Variable modifications : Q->p-E(N-term Q), E->p-E(N-term E), Oxidation(M), Phos (ST), Phos(Y)
Mass values : Monoisotopic
Protein Mass : Unrestricted
Peptide Mass Tolerance : ± 0.2 Da
Fragment Mass Tolerance: ± 0.2 Da
Max Missed Cleavages : 1
Instrument type : Default
Data File Name : D:\mascot mgf files\8hrsTCA.mgf Number of queries : 145

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Rank	Peptide
277	432.75	863.49	863.40	0.08	0	47	1	DFLAEGGGV
305	510.81	1019.60	1019.50	0.10	0	36	2	DFLAEGGGVR
307	525.79	1049.56	1049.47	0.10	0	23	1	EGDFLAEGGGV
310	539.32	1076.63	1076.52	0.10	0	47	1	GDFLAEGGGVR
314	554.30	1106.58	1106.49	0.09	0	28	1	GEGDFLAEGGGV
323	597.82	1193.63	1193.52	0.11	0	28	1	SGEGDFLAEGGGV
324	603.85	1205.68	1205.57	0.11	0	51	1	EGDFLAEGGGVR
340	655.34	1308.67	1308.55	0.12	0	24	1	DSGEGDFLAEGGGV
348	675.88	1349.74	1349.62	0.12	0	62	1	SGEGDFLAEGGGVR
366	730.85	1459.68	1459.55	0.13	0	20	3	ADSGEGDFLAEGGGV+Phos.(ST)
368	733.40	1464.78	1464.65	0.13	0	78	1	DSGEGDFLAEGGGVR
383	773.38	1544.75	1544.61	0.14	0	(72)	1	DSGEGDFLAEGGGVR+Phos.(ST)
389	808.90	1615.79	1615.65	0.14	0	74	1	ADSGEGDFLAEGGGVR+Phos.(ST)

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Rank	Peptide
311	546.27	1090.52	1090.42	0.10	0	18	2	QGVNDNEEGF+Q->p-E(N-term Q)
328	619.81	1237.61	1237.49	0.12	0	23	2	QGVNDNEEGFF+Q->p-E(N-term Q)
341	663.33	1324.64	1324.52	0.12	0	20	4	QGVNDNEEGFFS+Q->p-E(N-term Q)
357	698.85	1395.69	1395.56	0.14	0	23	1	QGVNDNEEGFFSA+Q->p-E(N-term Q)

Table 1. A list of identified peptides using TCA precipitation and LC-MS/MS by Mascot is shown. Top 2 from mascot hits.

1. FIBA_HUMAN Mass: 94914 Total score: 1044 Peptides matched: 29, (P02671) Fibrinogen alpha chain precursor [Contains: Fibrinopeptide A]

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Rank	Peptide
23	433.29	432.28	432.25	0.04	0	9	3	GGGV+iTRAQ4p(N-term)
307	447.30	892.58	892.48	0.10	0	42	1	FLAEGGGV+iTRAQ4p(N-term)
477	504.82	1007.62	1007.50	0.11	0	53	1	DFLAEGGGV+iTRAQ4p(N-term)
508	519.81	1037.60	1037.43	0.17	0	43	1	SGEGDFLAEGG
517	525.36	1048.70	1048.58	0.12	0	39	1	FLAEGGGVR+iTRAQ4p(N-term)
528	533.33	1064.65	1064.53	0.12	0	91	1	GDFLAEGGGV+iTRAQ4p(N-term)
551	544.81	1087.60	1087.53	0.07	0	35	1	TGKTFPGFFS
556	548.33	1094.64	1094.50	0.14	0	42	1	EGDFLAEGGG+iTRAQ4p(N-term)
594	582.89	1163.76	1163.61	0.15	0	34	1	DFLAEGGGVR+iTRAQ4p(N-term)
600	594.86	1187.70	1187.56	0.15	0	27	1	EGDFLAEGGGVR+E->p-E(N-term E)
608	597.88	1193.74	1193.57	0.17	0	63	1	EGDFLAEGGGV+iTRAQ4p(N-term)
624	611.40	1220.78	1220.63	0.16	0	32	1	GDFLAEGGGVR+iTRAQ4p(N-term)
659	620.86	1239.70	1239.54	0.16	0	59	1	DSGEGDFLAEG+iTRAQ4p(N-term)
667	626.39	1250.76	1250.59	0.17	0	68	1	GEGDFLAEGGGV+iTRAQ4p(N-term)
702	649.38	1296.74	1296.56	0.18	0	54	1	DSGEGDFLAEGG+iTRAQ4p(N-term)
727	669.91	1337.80	1337.62	0.18	0	88	1	SGEGDFLAEGGGV+iTRAQ4p(N-term)
738	675.93	1349.84	1349.62	0.22	0	82	1	SGEGDFLAEGGGVR
782	469.96	1406.85	1406.69	0.16	0	47	1	GEGDFLAEGGGVR+iTRAQ4p(N-term)
802	727.43	1452.85	1452.65	0.20	0	73	1	DSGEGDFLAEGGGV+iTRAQ4p(N-term)
808	730.88	1459.75	1459.55	0.20	0	57	2	ADSGEGDFLAEGGGV+Phos (ST)
825	498.97	1493.90	1493.72	0.18	0	(59)	1	SGEGDFLAEGGGVR+iTRAQ4p(N-term)
832	767.42	1532.83	1532.62	0.21	0	(20)	2	DSGEGDFLAEGGGV+iTRAQ4p(N-term); Phos (ST)
834	773.42	1544.83	1544.61	0.22	0	(37)	1	DSGEGDFLAEGGGVR+Phos(ST)
846	533.16	1596.47	1596.59	-0.12	0	8	9	GSSGTGGTATWKPGSS+2 Phos(ST)
852	537.32	1608.95	1608.75	0.20	0	73	1	DSGEGDFLAEGGGVR+iTRAQ4p(N-term)
867	808.95	1615.88	1615.65	0.23	0	48	1	ADSGEGDFLAEGGGVR+Phos (ST)
918	561.01	1680.00	1679.79	0.21	0	(47)	1	ADSGEGDFLAEGGGVR+iTRAQ4p(N-term)
922	563.99	1688.94	1688.72	0.22	0	(32)	1	DSGEGDFLAEGGGVR+iTRAQ4p(N-term); Phos(ST)
960	587.67	1759.98	1759.75	0.23	0	(35)	1	ADSGEGDFLAEGGGVR+iTRAQ4p(N-term); Phos(ST)

2. FIBB_HUMAN Mass: 55892 Total score: 466 Peptides matched: 13, (P02675) Fibrinogen beta chain precursor [Contains: Fibrinopeptide B]

Query	Observed	Mr(expt)	Mr(calc)	Delta	Miss	Score	Rank	Peptide
551	544.81	1087.60	1087.46	0.14	0	35	1	DNEEGFFS+iTRAQ4p(N-term)
587	580.33	1158.64	1158.50	0.15	0	24	1	DNEEGFFSA+iTRAQ4p(N-term)
611	601.83	1201.65	1201.50	0.15	0	43	1	NDNEEGFFS+iTRAQ4p(N-term)
618	607.86	1213.70	1213.54	0.16	0	46	1	VNDNEEGFF+iTRAQ4p(N-term)
653	619.83	1237.65	1237.49	0.16	0	43	1	QGVNDNEEGFF+Q->p-E(N-term Q)
675	636.37	1270.72	1270.56	0.16	0	53	1	GVNDNEEGFF+iTRAQ4p(N-term)
678	637.36	1272.71	1272.54	0.17	0	56	1	NDNEEGFFSA+iTRAQ4p(N-term)
720	663.35	1324.69	1324.52	0.16	0	58	1	QGVNDNEEGFFS+Q->p-E(N-term Q)
743	679.90	1357.78	1357.59	0.19	0	45	1	GVNDNEEGFFS+iTRAQ4p(N-term)
751	686.91	1371.80	1371.61	0.19	0	27	1	VNDNEEGFFSA+iTRAQ4p(N-term)
765	698.88	1395.75	1395.56	0.20	0	45	1	QGVNDNEEGFFSA+Q->p-E(N-term Q)
767	698.89	1395.77	1395.56	0.21	0	(30)	1	QGVNDNEEGFFSA+Q->p-E(N-term Q)
787	715.42	1428.82	1428.63	0.20	0	50	1	GVNDNEEGFFSA+iTRAQ4p(N-term)

Table 2. Peptides in the supernatant identified after precipitation and itraq labeling by LCMSMS and subsequent SwissProt database search by the MASCOT algorithm.

MASCOT parameters: Type of search: MS/MS Ion Search.
Enzyme : None;Variable modifications: Q->E(N-term Q), E->p-E(N-term E), iTRAQ4p(N-term), iTRAQ4p(Y), Oxidation(M), Phos(ST), Phos(Y); Mass values: Monoisotopic; Protein Mass: Unrestricted; Peptide Mass Tolerance: ±0.3 Da; Fragment Mass Tolerance: ±0.2 Da).

MASCOT Search Parameters
Type of search: MS/MS Ion Search
Enzyme: None
Variable modifications: Q->p-E (N-term Q), E->p-E (N-term E), iTRAQ4p(N-term), iTRAQ4p (Y), Oxidation (M), Phospho (ST), Phospho (Y)
Mass Monoisotopic
Protein Mass: Unrestricted
Peptide Mass Tolerance: ± 0.3 Da
Fragment Mass Tolerance: ± 0.2 Da
Max Missed Cleavages: 1
Instrument type: Default
Data File Name: D:\mascot mgf files\itraq_2_6hrs sup.mgf

Chapter V.

Factorial design of serum protein profiling by LC-MS

P. Horvatovich, N.I. Govorukhina, F. Suits, I.M. Westra, T.H. Reijmers, A.G.J. van der Zee, and R. Bischoff

Manuscript in preparation

1. Introduction

Biomarker discovery requires complex procedures that involve close collaboration between medical, analytical and computational sciences (Figure 1). Biofluids that are easily accessible, such as serum, are generally used for biomarker discovery, but they are extremely complex. During the process of biomarker discovery doctors working in medical sciences are responsible for diagnosis and patient classification, sample collection and storage as well as documentation. Scientists working in analytical science are responsible for sample preparation and analysis (e.g. by LC-MS), and must strive to minimize the variation introduced by these analytical steps. Scientists working in computational science process and analyze the generated data by extracting information related to the abundance of the various proteins and peptides in different samples and by using sophisticated statistical methods to find significant abundance differences between preclassified samples. Due to the complexity of the overall process, it is possible that certain experimental parameters may have an effect on the abundance of the measured protein and peptides thus affecting the outcome of the statistical analysis with the risk of creating false positives with respect to biomarker candidates or of missing potentially relevant candidates. It is thus important to determine the various factors, which affect the measured peptide profiles, in the case of trypsin-digested serum. This will allow the experimentalist to keep close control over the most important factors and to relax the stringency of the experimental protocol for those factors, which have little effect on the outcome of the analyses. There are preanalytical factors that cannot be changed retrospectively, for example, when working with already existing biobanks (e.g. the hemolysis level) but that may well affect the protein composition. These factors may be used as entry criteria and can help to remove samples that do not meet these criteria prior to analysis with the added benefit that non–class-specific variance will be reduced.

To assess the importance of a selected number of factors, we used a Factorial Design approach on serum samples that were depleted of the 6 most abundant proteins. As our analytical procedure by cap-LC-MS is rather time-consuming, we prioritized seven factors that were analyzed at two levels each. The analyzed factors were:

- type of blood collection tube (BD 367784, BD 368430, abrv.: Blo)
- hemolysis level (low and high, abrv.: Hem)
- clotting time (2 and 6 hours, abrv.: Clo)
- number of freeze-thaw cycles (1 and 3 cycles, abrv.: Fre)
- trypsin to protein ratio (1:20 and 1:100, abrv.: Try)
- stopping of trypsin digestion with acid (yes and no, abrv.: Sto)
- residence time in the autosampler (0 and 30 days, abrv.: Sta) at 4°C

A full factorial design would requires 2^7=128 analyses and would give information on the main effect of each factor and on all interactions between different factors from the second till the seventh order. In order to reduce the total analysis time, we have chosen to perform a two-level fractional factorial design with resolution VI (a so-called 2_{VI}^{7-3} design). This design requires only 16 analyses while resolution VI means that only the main effect of each factor can be evaluated directly, since these effects are not confounded by other effects or by two-factor interactions. Two-factor interactions are confounded in triads and thus in this design it is not possible to distinguish which of them is responsible for the measured effect. Higher order interactions between factors are generally negligible and thus confounding them with main and two-factor interactions will note affect the overall evaluation of our design. Table 1 contains the main effect and two-factor confounding pattern of the 2_{VI}^{7-3} design as used in this study. In order to estimate the data variances not related to the studied factors we have performed 3 repetitions of one analysis leading to total of 19 analyses [1, 2].

Automatic processing of ion-trap LC-MS data resulted in 8 000 - 10 000 aligned features or measured variables. The evaluation of this multivariate data set was done by eliminating one factor at a time to produce the average highest variance in the dataset.

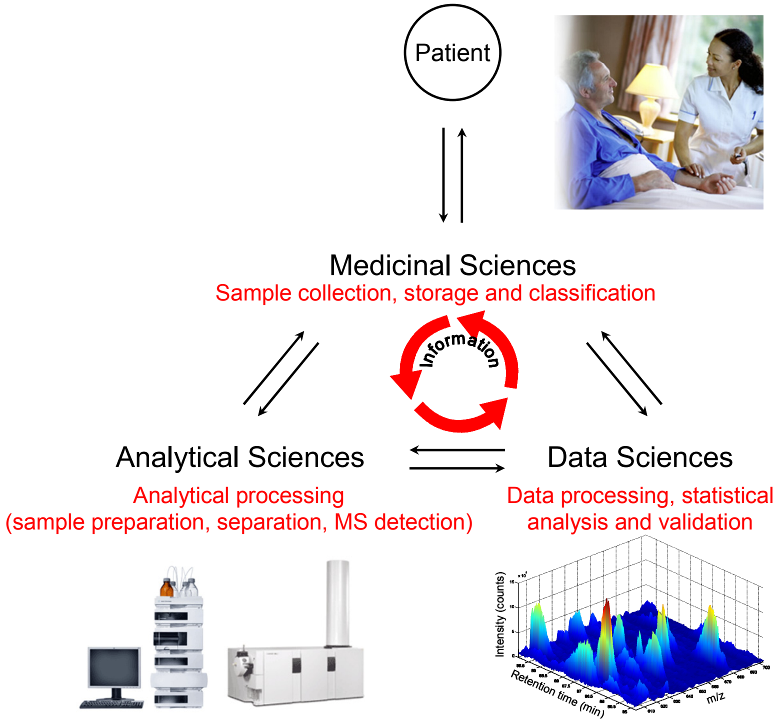

Figure 1. Schematic view of the collaborative environment of a biomarker discovery project. The final result may be affected by variable factors, as exemplified in red.

Table 1. Main effect and two-factors confounding pattern of a 2_{VI}^{7-3} fractional factorial design as used in this study.

Term	Confounded with	Confounded with
Blo		
Hem		
Clo		
Fre		
Try		
Sto		
Sta		
Blo*Hem	Clo*Try	Sto*Sta
Blo*Clo	Hem*Try	Fre*Sta
Blo*Fre	Clo*Sta	Try*Sto
Blo*Try	Hem*Clo	Fre*Sto
Blo*Sto	Hem*Sta	Fre*Try
Blo*Sta	Hem*Sto	Clo*Fre
Hem*Fre	Clo*Sto	Try*Sta

2. Methods

2.1. Factorial design

Different handling of serum samples can influence the resulting LC-MS data reflecting the peptide/protein composition of the samples. Based on previous experience with the analysis of serum samples we prioritized the following seven factors and defined the levels to be investigated: tubes for blood collection, high/low hemolysis level, different clotting times (2 or 6h), freeze-thaw cycles (1 or 3), digestion with trypsin at different enzyme to substrate ratios (1:20 or 1:100), stopping of trypsin digestion with acid or not, stability of the samples at 4^{o}C prior to LC-MS analysis. Samples were analyzed according to the factorial design scheme given in Table 1.

Table 2. 2_{VI}^{7-3} Fractional design pattern.

Experiment name	Run order	Blood collection tube	Hemolysis	Clotting time	Freeze-thaw cycles	Trypsin digestion	Stopping trypsin	Stability sample
N1	11	BD368430	Low	2 hours	1 cycle	1:20	Yes	0 days
N2	3	BD367784	Low	2 hours	1 cycle	1:100	Yes	30 days
N3	9	BD368430	High	2 hours	1 cycle	1:100	No	0 days
N4	17	BD367784	High	2 hours	1 cycle	1:20	No	30 days
N5	18	BD368430	Low	6 hours	1 cycle	1:100	No	30 days
N6	2	BD367784	Low	6 hours	1 cycle	1:20	No	0 days
N7	10	BD368430	High	6 hours	1 cycle	1:20	Yes	30 days
N8	16	BD367784	High	6 hours	1 cycle	1:100	Yes	0 days
N9	8	BD368430	Low	2 hours	3 cycles	1:20	No	30 days
N10	15	BD367784	Low	2 hours	3 cycles	1:100	No	0 days
N11	13	BD368430	High	2 hours	3 cycles	1:100	Yes	30 days
N12	7	BD367784	High	2 hours	3 cycles	1:20	Yes	0 days
N13	12	BD368430	Low	6 hours	3 cycles	1:100	Yes	0 days
N14	6	BD367784	Low	6 hours	3 cycles	1:20	Yes	30 days
N15	5	BD368430	High	6 hours	3 cycles	1:20	No	0 days
N16	19	BD367784	High	6 hours	3 cycles	1:100	No	30 days
N17	14	BD368430	Low	2 hours	1 cycle	1:20	Yes	0 days
N18	1	BD368430	Low	2 hours	1 cycle	1:20	Yes	0 days
N19	4	BD368430	Low	2 hours	1 cycle	1:20	Yes	0 days

The samples were injected into the LC-MS according to the order shown in column 2 "Run order".

2.1.1. Description of samples

Serum samples were obtained from 2 healthy volunteers (Fact. Design #1, female and Fact. Design #2, male) and provided by the Department of Gynecological Oncology (University Medical Center Groningen, The Netherlands). They were stored at -80ºC in aliquots until analysis.

2.1.2. Blood collection tubes

Two kinds of tubes were used for the collection of blood: BD368430 (a "red stopper clotting tube", which is a glass tube with a siliconized inner wall used to avoid retention of red blood cell on the walls of the tube) and BD367784 (a "gel tube", which is a glass tube with a separation gel and micronized silica to accelerate clotting). During centrifugation, the polymer gel moves up the inner wall of this tube forming a barrier between the supernatant (serum) and sediment (blood clot and cells)). These were selected, since they were both used at the University Medical Center Groningen, The Netherlands to acquire the available serum biobank from cervical cancer patients.

2.1.3. Variation of clotting time

Serum samples were allowed to clot for 2 or 6 hours at room temperature before centrifugation to obtain serum.

2.1.4. Stability of trypsin-digested serum samples in the autosampler prior to LC-MS analysis

Stability of the trypsin-digested serum samples was evaluated by keeping them for 30 days at 4°C in the autosampler. Samples labeled "0 days samples" were injected directly after thawing (storage at -80°C).

2.1.5. Level of hemolysis

To simulate a high level of hemolysis, a lysate of red blood cells was added to the serum prior to depletion. Red blood cells were collected according to the following protocol: 0.5 mL lysis buffer (NH_4Cl 155 mmol/L, EDTA 0.1mmol/L) was added to 0.5 mL fresh blood and centrifuged for 20 minutes at 2000 rpm. 4 mL of lysis buffer was added to the pellet and incubated overnight at 4°C. The next day the lysate was filtered through spin filters (0.22µm; # 5185-5990, Agilent) at 13000 rpm. Aliquots of the filtrate were stored at -80°C.

The amount of lysed red blood cells that should be added to serum to mimic a high level of hemolysis was determined by the addition of different amounts (1, 3, 5, 7 and 10 µL) of red blood cell lysate to 20 µL serum that was immediately diluted with ice-cold water to a total volume of 60 µL followed by centrifugation at 13000 rpm for 30 min at 4°C. Another 15 µL of ice-cold water were added to the supernatant and the absorbance was measured at 340, 380, 415 and 450 nm (Biowave S2100 UV/Vis Diode Array Spectrophotometer (Biochrom Ltd, Cambridge, UK)). A calibration line with respect to hemoglobin (Hb) was obtained using the formula: Hb [g/L] = $(167.2 * A_{415} – 83.6 * A_{340/380} – 83.6 * A_{450}) / 1000$ [4]. The absorbance of a serum sample from the serum bank that appeared to be very red, indicating a high level of hemolysis, was measured and the amount (in µL) of red blood cells to reach the corresponding Hb level was calculated from the calibration curve. To determine the concentration of hemoglobin that corresponded to a given volume of lysed red blood cells, a calibration curve of hemoglobin (Hemoglobin human, Sigma, #9008-02-0) was made. It was calculated that 4 µL red blood cell (containing 6.68 µg hemoglobin) should be added before depletion to the serum to simulate a high level of hemolysis as observed in clinical samples from the biobank.

2.1.6. Preparation of serum samples

80 µL (80% of the total amount (20 µL of crude serum mixed with 80 µL of buffer A (Agilent)) of diluted crude serum was injected on a Multiple Affinity Removal column (Agilent, 4.6 x 50 mm, # 5185-5984) after filtration through 0.22 µm spin filters (# 5185-5990) at 13000g and 4˚C for 10 min to remove particulates. The removal of abundant proteins was performed on a LaChrom HPLC System (Merck Hitachi, www.merck.com) with detection at 280 nm using the following timetable: 0-9 min, 100% buffer A (0.25 mL/min); 9.0-9.1 min, linear gradient 0-100 B % (1 mL/min), 9.1-12.5 min, 100% buffer B (1 mL/min); 12.5-12.6 min, linear gradient 100-0% buffer B (1 mL/min); 12.6-20 min, 100% buffer A (1 mL/min). The flow-through fraction (depleted serum collected between 2-6 min) of a total volume of appr. 1 mL was collected [3].

Protein concentrations were determined with the Micro BCA™ Protein assay reagent kit (www.piercenet.com) and calculated for an average protein molecular weight of 50 kDa. BSA was used as the calibration standard.

2.1.7. Digestion of serum samples

Trypsin (sequencing grade modified trypsin, Promega, # V5111, USA) in ratios 1:20 or 1:100 wt/wt (enzyme to total protein in depleted serum) was used for digestion at 37°C at 450rpm (Eppendorf thermomixer, overnight).

2.1.8. Stopping of trypsin digestion (according to Table 1)

To stop the reaction with trypsin, formic acid was added after overnight digestion to reach a final concentration of 0.5% (v/v).

2.1.9. Freeze-thaw cycles

One or three freeze-thaw cycles (-80°C/room temperature) were include for comparative analysis (according to Table 1).

2.2. Data processing

For processing and multivariate statistical analysis of the original Bruker Daltonics HPLC-MS data, the files were converted into ASCII-format with the Bruker data analysis software LC/MSD Trap, version 3.3 (build 146) (Bruker Daltonics, Bremen, Germany) and saved in centroid mode. The time alignment algorithm was written in C++ using Microsoft Visual Studio [ver. 8.0.50727.762 (SP.050-727-7600), Redmont, WA, USA]. For further data analysis, Matlab [version 7.4.0.287 (R2007a), Mathworks, Natick, Massachusetts, USA] was used.

Centroid data were smoothed and reduced using a normalized two-dimensional Gaussian filter, with rounding the nominal m/z ratios to 1 m/z (the original data had a resolution of 0.1 m/z). In the retention time dimension no data reduction was performed. This meshing procedure reduced the number of available data points by roughly a factor 10 and corrected for shifting m/z values as a result of different loadings of the ion-trap during elution of abundant peptides, a phenomenon that is common for ion-trap mass spectrometers [5,6]. After meshing the data files of all chromatograms, they were time-aligned (warped) to a reference data file using Correlation Optimized Warping (COW) [7] based on peak lists obtained from the chromatograms. This time alignment algorithm optimizes time shifts using the overlapping between the extracted peaks using 2-dimensional extent of the peaks (retention time and m/z value) in the two chromatograms. Thus all chromatograms were aligned to each other in a pair-wise manner starting with randomly selected chromatogram as reference. The accuracy of correcting retention time shifts by time alignment was manually checked by visualization the 4 most intensive peaks for 10 equal time segments between 60-155 min before and after time alignment.

A modified M-N rule was applied for peak detection on meshed data with data reduction 1:10 in m/z resulting rounded integer m/z in the mass spectra by first calculating a median local baseline using the sliding window technique separately for each rounded m/z trace. A median window size of 1200 data points, corresponding to 20.17 min, was used with a moving rate of 10 points and a minimum median value of 200 counts. According to the M-N rule, a threshold of M-times the local baseline was used and a peak was assigned if, within one m/z trace, the signal exceeded this threshold for at least N consecutive points [8]. For each detected peak the m/z value, the mean retention times of the three highest measured intensities within the same peak reduced by the local baseline were stored in a peak list created for every chromatogram.

We used a similar approach as Radulovic *et al.* [8] to obtain optimal settings for M and N. Different values for M (1.5-4) and N (4-8) were applied to two blank LC-MS runs and two LC-MS runs of depleted, trypsin-digested serum samples. Settings were used at which the ratio between the number of peaks between 60 and 155 min in the samples relative to the blank chromatograms was highest and at which a minimal number of peaks was extracted from the noise in the blank chromatogram (M = 2 and N = 5 in our case).

In order to combine the peak lists from different samples into a common peak matrix, one-dimensional peak matching was performed by using the sliding window technique, in which the same m/z traces were evaluated for peaks that are proximate in time (step size 0.1 min; search window 1.0 min; maximal accepted standard deviation for all retention times within a group of matched peaks 0.75 min). Missing peak allocation was performed by extracting the background subtracted local signal of the given m/z trace at the average retention time corresponded to the chromatograms where that peak was present. The generated peak matrix, created from the peak lists of the individual samples, consisted of a peak(row)-sample(column)-intensity(value) matrix. This peak matrix was used for analysis. All data preprocessing work was done on a personal computer equipped with a dual core +3800 MHz AMD 64 X2 processor equipped with 4 GB of RAM.

2.3. Statistical analysis

Multiple Linear Regression was performed using the MODDE software (ver 7.0.0.1) from Umetrics (Kinnelon, NJ, USA) (Umetrics INC, USA, ver 7.0.0.4). The evaluation of aligned peak matrices was performed using a stepwise method. First the aligned peak matrix was mean centered and normalized with respect to the standard variation for each peak (row) and thus relative standard deviation was obtained (RSD). Because of the small sample size and the high number of peaks (variables), occurrence of small standard deviations were avoided by adding 0.001. At each step the main factor effects were removed from the peak's intensity by subtracting the corresponding mean factor-level . At a given step the factor resulting in the highest decrease in overall sum of the RSD (SRSD) was retained and eliminated from the remaining iterations. For the remaining steps the peak matrix corrected with the selected factor was used. Pearson correlation plots produced in Matlab were used to visualize the results.

3. Results

Specific effect of the selected seven preanalytical parameters on LC-MS data was evaluated by integrating the peak areas of peptides belonging to high-abundant proteins (Apolipoprotein A, α-2-Macroglobulin precursor, human Complement C3 precursor with masses of 756.7, 909.6, 682.5, 753.5, 619.2, 694.4). These protein-derived peptides have been shown by us to be rather stable in quantity in serum samples [3] and may thus be considered to reflect major effects of preanalytical factors, if affected.

In Figure 2 peak areas for these peptides are shown for serum from female (2a) and male blood (2b). The deviations of the mean retention time (RT) of endogenous standards were less than 0.8% confirming reproducibility of the LC-MS analyses. Variation in peak area ranged from approximately 40% a the peak derived from Apolipoprotein A (doubly charged peptide DLATVYVDVDVLK, m/z 619.2, accession number P02647) up to 130% for a peak derived from Alpha-2-macroglobulin precursor (doubly charged peptide AAQVTIQSSGTFSSK, accession number P01023); (m/z 753.5) for factorial design #1 (female serum) (Figure 2a) showing that the chosen preanalytical factors do have a significant effect even on peptides derived from high-abundance proteins.

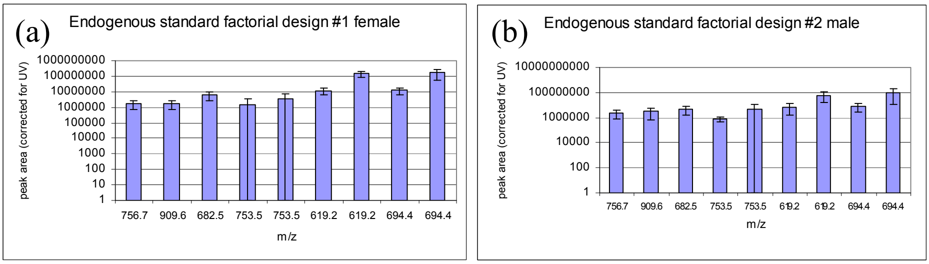

Figure 2. Six m/z values derived from high-abundance proteins remaining after depletion were analyzed to assess the overall effect of the selected seven preanalytical factors (see Table 2 for the factorial design) based on changes in peak areas of extracted ion chromatograms. a) female serum; b) male serum. Peptides with the same m/z values: 619,2; 694,4; 753,5 were eluted twice at different retention time. Standard deviations are calculated based on the peak area's of the Selected Ion Chromatogram of ±0.5 m/z values.

In male samples (factorial design #2) the deviation of the mean RT of the same peptides was also less than 0.8%. Standard deviations of peak areas varied from approximately 75% for a peak m/z 756.7 up to 135% for the peak with m/z 753.5 from Alpha-2-macroglobulin precursor.

In order to study the effects of the seven selected factors on these peptides, we applied a multiple linear regression model (Figure 3). The model showed that only one peptide was significantly affected by the studied factors in female serum (m/z 694.4, shown in red).

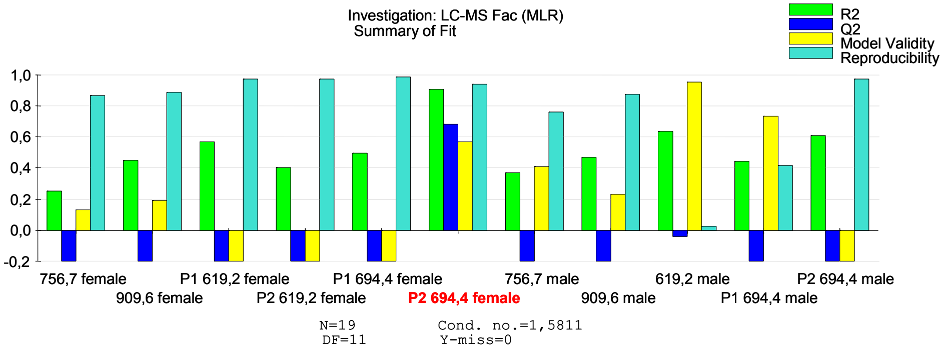

Figure 3. Peak areas of peptides derived from high-abundance proteins (see Table 3) were analyzed using a Multiple Linear Regression model. Data were obtained from two fractional factorial design studies (one in male and one in female serum). In the figure R2 refer to Residual Standard Deviation and is the fraction of variation of the response explained by model. Q2 refer to the predictive power of the model (Prediction Residual Sum of Squares) and is the fraction of variation of the response that can be predicted by the model.

The peak area of the peptide derived from Apolipoprotein A (doubly charged ion of peptide VSFLSALEEYTK, accession number P02647, m/z 694.4) in female serum showed a linear correlation between the expected (as predicted by the model) and observed peak areas (Figure 4a). Analysis of the contribution of different factors showed that clotting time and the trypsin-to-substrate ratio affected the measured peak area significantly (significance level of 95%) (Figure 4b). In male dataset the peak areas of the same peptide were less well predicted by the model, although a similar pattern was observed (data not shown).

In order to perform a global analysis of all detected features and how they respond to the variation of a given factor, it is necessary to generate a common peak matrix from all 19 LC-MS runs. This requires correction of retention time shifts between the different LC-MS analyses. However, Correlation Optimized Warping using the Total Ion Chromatogram (COW-TIC) was not successful, due to the major differences in chromatographic patterns when changing certain factors (Figure 5a). Poor time alignment of chromatograms lead to mismatched peaks and resulted in misinterpretations. The failure of the COW-TIC algorithm to align the chromatograms is due to the fact that some factors have a large effect on the measured peak profiles leading to very high analytical variability and to rather diverse TICs in spite of the fact that the serum sample was taken from the same patient (no biological variance included in the fractional factorial design experiment). Applying the COW algorithm not to the TIC but to a 2-dimensional extent of all detected peaks in the retention time and m/z dimensions (2D-COW), thus taking the 3-dimensional structure of the data into account, resulted in proper alignment of most peaks (see Figure 5b for an example).

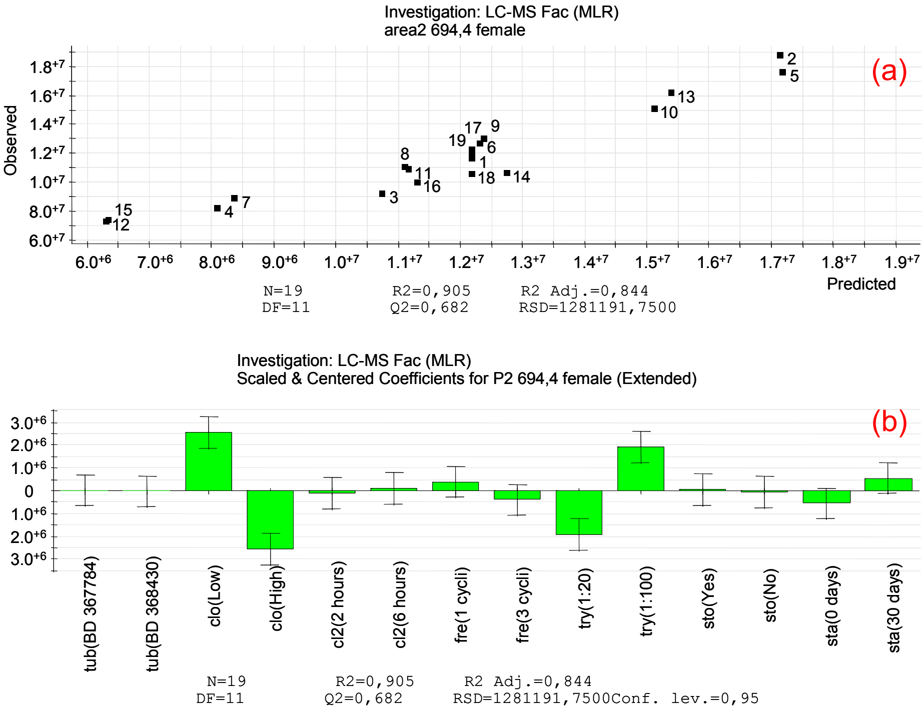

Figure 4. Multiple Linear Regression model of the peptide derived from Apolipoprotein A (doubly charged ion of peptide VSFLSALEEYTK, accession number P02647, m/z 694.4). A) Correlation of the predicted with the observed peak area (R^2 = 0.905; N=19 experiments), b) coefficient of factors obtained in the MLR model.

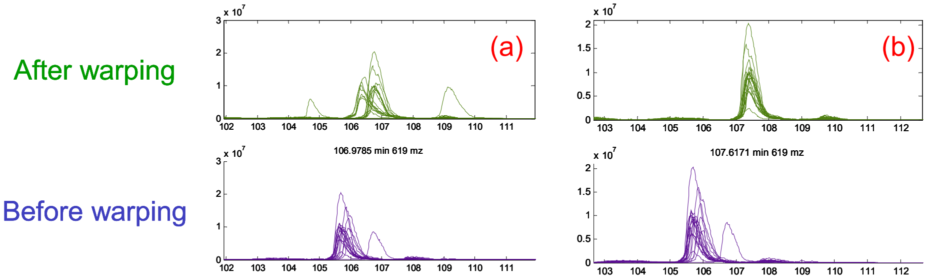

Fig. 5. Time alignment of the 19 chromatograms of the fractional factorial design study (the example of one peptide peak at appr. 107 min retention time (m/z 619) is shown) using the COW-TIC algorithm (a) or COW in conjunction with a newly developed algorithm that aligns peaks in 2-dimensions (retention time and m/z) (b).

This allowed automatic processing of thousands of peaks in the datasets and the subsequent evaluation of the effect of the selected factors on the overall pattern. Correlation plots with hierarchical clustering of the aligned peak matrixes from the study of the different factors showed that the level of hemolysis is the most important factor affecting the overall correlation between LC-MS datasets (Figure 6).

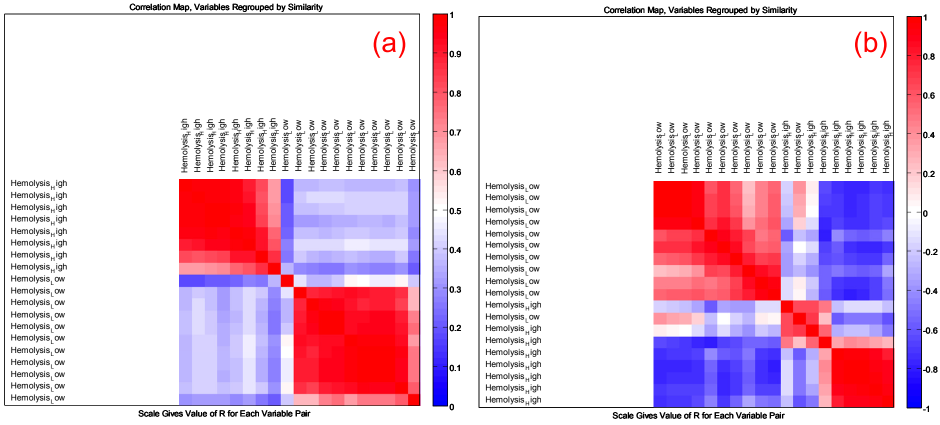

Figure. 6 Correlation plots with hierarchical clustering of two fractional factorial design studies (see Table 2) for (a) male serum, and (b) female serum. Note that samples with low or high hemolysis levels, respectively, cluster together.

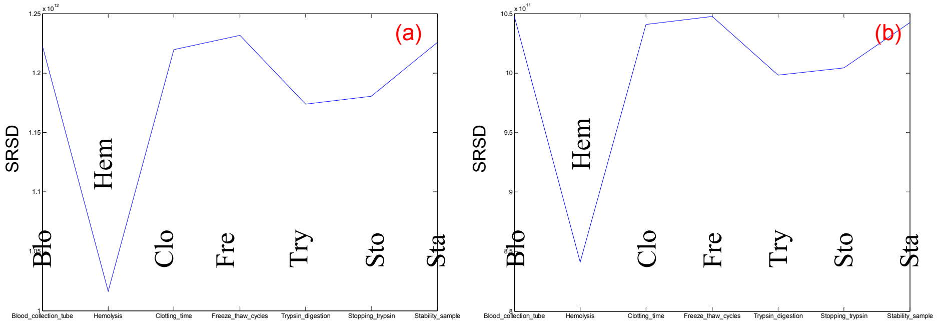

Figure 7. Sum of relative standard deviation (SRSD) after correction with the mean of corresponding factors for each level for all 7 factors (a, male) and (b, female).

Further analysis of the contribution of each factor to the overall sum of relative standard deviation (SRSD) between the datasets was carried out by correcting the contribution of each factor by the mean of each level. The factor showing the lowest SRSD has the greatest influence on peak intensity for largest number of peaks, and may thus be considered to be the most important preanalytical factor in the design with respect to the analysis of depleted from the 6 most abundant protein and trypsin-digested serum samples (Figure 7). Again the level of hemolysis stands out as being highly relevant. In order to assess the relative importance of the remaining factors, the most important factor was discarded from the list of factors and the aligned peak matrix corrected with the most important factor was retained to determine the second most important factor. These iterations resulted in a list of factors ordered by their importance (Figure 8). The most important factors were: hemolysis level > trypsin-to-protein ratio > stopping the trypsin digestion with acid. Factors such as clotting time, type of blood collection tube, stability of sample in the autosampler or freez-thaw cycles had rather small effects on the resulting LC-MS data. The results concerning clotting time corroborate our findings described in **Chapter IV**.

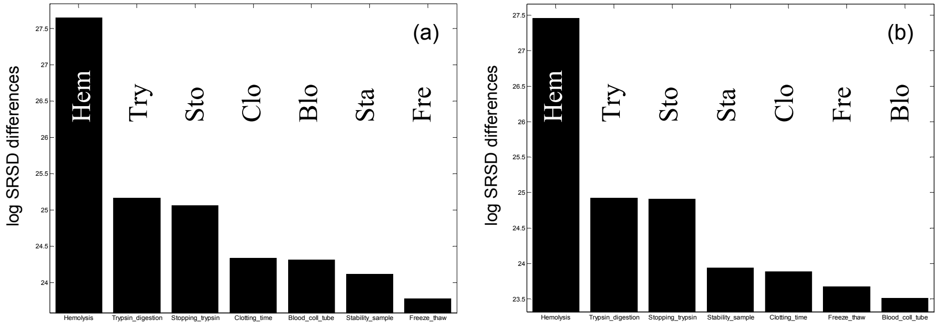

Figure 8. Differences of the SRSD (log scale; log SRSD) after recursive correction with the mean of factor effect having the highest influence on SRSD using all peptide peaks of the 2D-COW aligned-peak matrices for male (a) or female serum (b).

4. Conclusions

The main result of this study is that the hemolysis level, which is a parameter that needs to be controlled during blood collection (responsibility of medical sciences; see Figure 1), makes the largest contribution to variability in the datasets. This implies that sera with high hemolysis levels have to be discarded from further analysis because this level cannot be set a priori but is largely influential on the expected differences in concentrations of potential biomarkers. Determining this relationship will be a matter of future studies. It is, however, wise to take special care to avoid hemolysis during blood collection and the preparation of serum. The other two most important factors, trypsin-to-protein ratio and stopping digestion with acid, can be easily controlled by the analytical chemist.

Our results describe only a partial evaluation of the multivariate data obtained from the fractional factorial design study of depleted and trypsin-digested serum samples. We only evaluated the main effect of each factor on the final data and more detailed analyses of the data need to apply multivariate statistical methods (e.g. Partial Least Squares methods) to obtain additional information about possible interactions between two factors. Beside the factor and interaction order we would like to assess the significance level of the main factors and multiple factor interactions. As the design of the experiments has a resolution of IV, meaning that at least 3 out of two factor-interactions are confounded, a full factorial design, of the 3 main factors should be performed. This work is presently ongoing.

References

[1]. Box GEP, Hunter JS, Hunter WG. Statistics for Experimenters. ed. Wiley, 2005.

[2]. Mason LR, Gunst, FR, Hess LJ. Statistical design and analysis of experiments. ed. Wiley, 2003.

[3]. Govorukhina NI, Reijmers TH, Nyangoma SO, Zee AGJ van der, Jansen RC, Bischoff R. Analysis of human serum by LC-MS: improved sample preparation and data analysis. J. Chromatogr. A. 2006,110:142-150.

[4]. Cookson P, Sutherland J, Cardigan R. A simple spectrophotometric method for the quantification of residual haemoglobin in platelet concentrates. Vox Sanguines. 2004;87:264-271

[5]. Hao Ch, March R. A survey of recent research activity in quadrupole ion-trap mass spectrometry Int. J. Mass Spectrum. 2001;212:337–357.

[6]. Wells JM, Plass WR, Patterson GE, Ouyang Zh. Chemical Mass Shifts in Ion-trap Mass Spectrometry. Anal. Chem. 1999;71:3405-3415.

[7]. Tomasi G, van den Berg F, Andersson C. J. Correlation optimized warping and dynamic time warping as preprocessing methods forchromatographic Data. Chemometrics 2004;18:231-241.

[8]. Radulovic D, Jelveh S, Ryu S, Hamilton TG, et al. Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry, Mol. Cell. Proteomics 2004;3:984-997.

Chapter VI.

Comparative analysis of cervical cancer patients (long-term survivors) before and after treatment

N.I. Govorukhina, A. Hesseling, P. Horvatovich, M. de Vries, F. Suits, K.A. ten Hoor, A.G.J. van der Zee, R. Bischoff

1. Introduction

Cervical cancer is one of the most common malignant diseases of women worldwide [1]. It is more common in developing countries, where 83% of cases occur and where cervical cancer accounts for 15% of newly diagnosed female cancers while in developed countries it accounts for 3,6 % of all new cancer cases [2]. It was shown that 99% of cervical squamous cell carcinomas are linked to Human Papilomavirus (HPV) infection [3]. Especially "high-risk" strains such as 16 and 18 [4] are associated with the development of high-grade intraepithelial squamous lesions (HSIL) and may eventually lead to cervical cancer [5].

Squamous cell carcinoma antigen (SCC-ag) is a tumor marker in serum for patients with squamous cell carcinoma of the cervix. Concentration of SCC-ag in the serum is the good marker for monitoring response to therapy. SCC-ag levels return to normal upon complete remission after treatment (threshold below 1.9 µg/L) while SCC-ag levels raise again in patients with recurrent disease. Measuring SCC-ag levels has, however, no predictive value, does not contribute to better survival [6] but provides useful information for management of the disease [7]. SCC-ag is also not very disease-specific, since it was shown that squamous cell carcinoma antigen 1 (SCC-ag 1) was up-regulated not only in the uterine cervix cancer but also in lung cancer, cancer of the esophagus and skin cancer [8-11] as well as in cancer of the tongue [12]. It was shown previously that TGFβ1 (transforming growth factor β1) is produced by cervical cancer cells inducing PAI-1 expression (plasminogen activator inhibitor-1) [13]. PAI-1 expression has been correlated with decreased survival in breast cancer and associated with poor prognosis in cervical cancer [13]. Thus there is a continuing need for better biomarkers to assist in early diagnosis of cervical cancer, in staging of the disease and in following up on therapeutic efficacy.

In this study we compared serum from patients with cervical cancer obtained before and after treatment that resulted in complete/partial remission or stable disease. Efficient depletion of high-abundance proteins (6 or 20) was combined with two complementary analytical approaches: gel electrophoresis (1D or 2D) or Liquid Chromatography-Mass Spectrometry (LC-MS). Dedicated algorithms for data pre-processing were developed and applied followed by multivariate statistical analysis (Principal Components Analysis, PCA).

2. Methods

2.1. Description of samples

Serum samples from cervical cancer patients were obtained at the Department of Gynecological Oncology (University Medical Centre Groningen, The Netherlands) and stored at – 80 °C in aliquots. Most of the patients (Table 1) showed no recurrence of disease after therapy (except patient 22: partial remission and patient 23: stable disease: tumor remains without progression). The diagnosis was done by histological analysis and gynecological examination: inspection and palpation of the genitalia and SCC-ag test. Patients with remission have no complains and normal SCC-ag level. All patients used in this study had advanced disease (stage III or IV) according to the International Federation of Gynecology and Obstetrics (FIGO) classification [14] and belonged to a group of long-term survivors. The level of the Squamous cell carcinoma antigen-1 (SCCag) was determined by ELISA [6].

2.2. Preparation of serum samples

Depletion of the 6 most abundant serum proteins (on a Multiple Affinity Removal column (4.6 x 50 mm, # 5185-5984, www.chem.agilent.com), digestion of the remaining proteins in depleted serum with trypsin, capillary LC-MS analysis (Atlantis dC 18 (1.0 x 150 mm, 3 μm) with Atlantis dC 18 in-line trap column for cap-LC-MS (Waters, Milford, Massachusetts, USA, www.waters.com) were performed as previously described (15).

Depletion of the 20 most abundant proteins was performed according to the manufacturer's instructions (ProteoPrep 20 Plasma Immunodepletion Kit (PROT20), www.sigma-aldrich.com). 8μL of serum were depleted on a spin column. The depletion procedure was repeated 5-times (40μL of original serum) and followed by final (second) depletion of samples after first depletion to rise the level of depletion to >99% (www.sigma-aldrich.com). The depleted sera were pre-fractionated by high-temperature reversed-phase HPLC (see below) followed by 1D- or 2D-gel electrophoresis. Protein concentrations were determined with the Micro BCA™ Protein assay reagent kit (www.piercenet.com) and calculated for an average protein molecular weight of 50 kDa. BSA was used as the calibration standard.

2.2.1. Pre-fractionation of depleted serum by high-temperature reversed-phase HPLC

0.48 g of urea (#084K0063, Sigma, www.sigmaaldrich.com) and 13µL of glacial acetic acid (#1.00063.1000, www.darmstadt.merk.de), were added to ~300µg (about 300µL) of depleted serum according to the manufacturer's instructions (www.agilent.com/chem/bioreagents). Solvent A (97% H_2O/0.1% TFA (# 28902, www.pierce.com)) was added further to a final volume of 1mL and the sample was injected onto a Macroporous Reversed-Phase mRP-C18 column (Agilent, 4.6 x 50mm, # 5188-5231). Prefractionation was performed at 80 °C and pH <4.5 at a flow rate of 0.75mL/min with UV detection at 280nm using a gradient from 3 to 30% of solvent B (97% AcN/0.1% TFA) in 6 min, to 55% solvent B in 40 min and finally up to 100% B in 53 min. The volume of the collect fractions was 0.75mL (1min fractions). The proteins in these fractions were further analyzed by SDS-PAGE in a pair-wise manner with the same fraction from a given patient before and after treatment next to each other.

2.2.3. SDS-PAGE

SDS PAGE was performed in a Mini-Protein III cell (Bio-Rad, www.biorad.com) using 12,5% gels with 0.1% SDS according to the manufacturer's instructions. All chemicals were from Bio-Rad. Samples were boiled in sample buffer containing 0.02M DTT for 1 min, cooled down and applied directly to the gel. PageRuler™ Prestained Protein Ladder (#SM0671, www.fermentas.com) were used. Proteins were stained with Coomassie Brilliant Blue R concentrate (Sigma, www.sigmaaldrich.com) diluted and used as prescribed by the manufacturer.

2.2.4. In-gel digestion

Performed according to protocol of Shevchenko et all. [16]. Trypsin (sequencing grade modified trypsin, # V5111, www.promega.com) was used (10µg/mL) for digestion overnight at 37°C with shaking at 450 rpm.

2.5. 2D gel electrophoresis

Isoelectric focusing (IEF; first dimension) was performed by loading 250 µL (250µg) of serum after depletion of six most abundant proteins (www.chem.agilent.com) onto a ReadyStrip (IPG strip pH 4-7; Bio-Rad #163-2099, www.biorad.com) and rehydrated overnight [17]. The IPG strip was placed in the focusing tray and IEF was performed at 50µA/strip (150 V) for 30 min, 300 V for 1 hour, 600 V for 1 hour and 8000 V to reach a total of 35.000

V•hours) at 20^{o}C. For the second dimension electrophoresis (SDS-PAGE) the IPG strip was equilibrated in buffer containing 0.05 M Tris-HCl pH 8.8, 6 M urea, 30% w/v glycerol and 2% w/v SDS with 1% (w/v) Dithiothreitol (DTT) (Sigma #D0632, www.sigmaaldrich.com) for 15 min, followed by incubation with the same buffer containing 4% iodoacetamide (Sigma #11149) for another 15 minutes in the dark [18]. Equilibrated IPG strips were placed on top of a precast 12.5% polyacrylamide gel (Bio-Rad #345-0102, www.biorad.com) and fixed with a 0.5 % agarose solution containing 25 mM Tris, 192mM glycine, 0.1% SDS and a trace bromophenolblue. Reference marker PageRuler™ Prestained Protein Ladder (Fermentas, #SM0671, www.fermentas.com) was used. Electrophoresis was performed at room temperature, 100 V for 2.5 hours. After electrophoresis, the gel was fixed with a 40% methanol, 10% acetic acid solution for 30 min followed by Coomassie Brilliant Blue staining (Page Blue (Fermentas, #R0571, www.fermentas.com) at room temperature for 3 hours and subsequent overnight destaining in milliQ water. Gels were analyzed and scanned using the PD*Quest* Image analysis program (Bio-Rad, www.biorad.com). Differentially expressed spots were excised from the gel for in-gel digestion (see above) and mass spectrometric analysis. To be able to detect low-abundance proteins, the gels were destained and stained using a mass-spec-compatible silver staining method [19]. In-gel digestion performed as in 2.2.4.

2.3. Matrix-Assisted Laser Desorption Ionization (MALDI) Mass Spectrometry (MS) analysis

MALDI-MS analysis of in-gel digested protein bands (1D SDS-PAGE) or spots (2D gel electrophoresis) was performed in the positive ionization mode using a Voyager DE Pro instrument (Applied Biosystems, www.appliedbiosystems.com). Spectra were acquired in reflectron mode with delayed extraction considering only mono-isotopic molecular masses for protein identification. The spectra were internally calibrated using the tryptic auto-digestion peaks at m/z 842.5099 and 2211.1046 amu. For identification 0.5 µL of the digestion mixture were directly spotted on a stainless-steel MALDI target plate and 0.5 µL of a 5 mg/mL α-cyano-4-hydroxycinnamic acid (CHCA) in 50% acetonitrile (AcN)/0.1%TFA were added on top of the deposit. If this yielded no useful spectra, 5 µL of the remaining tryptic digests were purified using C18 Stage tips (Proxeon, www.proxeon.com) according to the manufacturer's instructions. Peptides were eluted in 2 µL of a 5 mg/mL α-cyano-4-hydroxycinnamic acid (CHCA) in 50% AcN/0.1%TFA and directly spotted on the stainless-steel target plate. Expasy's aldente was used for identification (http://www.expasy.org/cgibin/aldente/help.pl) Type of search: MS/MS Ion Search; Enzyme: Trypsin; Variable modifications: Oxidation (M), Mass values: Monoisotopic; Protein Mass: Unrestricted; Peptide

Mass Tolerance: ± 0.2 Da; Fragment Mass Tolerance: ± 0.2 Da; Max Missed Cleavages: 1.

Identification of proteins after in-gel digestion was done also on nano-LC-MS system (Agilent) equipped with a microfluidics (chip-cube) interface (cat. n° G4240A) including a chip (cat. n° G4240-62001) with a 40 nL trap column (75 µm × 11 mm) and a 75 µm × 43 mm analytical column both containing C-18SB-ZX 5 µm chromatographic material. The interface contains a nanoelectrospray tip (2 mm length with conical shape: 100 µm OD × 8 µm ID) and was coupled on-line to an MSD-Trap-SL ion-trap mass spectrometer. Injections (2µL were performed with an autosampler (Agilent, cat. n° G1389A) equipped with an injection loop of 8 µL (this includes the dead volume up to the trapping column) and a thermostated cooler maintaining the samples in the autosampler at 4°C during the analysis. All parameters for LC-MS analysis are described in Chapter III. MS spectra were deconvoluted with respect to charge state and isotopes. The resulting spectra were saved in mascot (Matrix Science, London, U.K.) generic file format and submitted to a Web based version of the Phenyx36 search engine (v2.1) for UniProt_Sprot (r. 48.8 of 10-Jan-2006) queries.

2.4. Stable-isotope labeling (iTRAQ)

Changing of buffer and concentration of samples for iTRAQ labeling of depleted serum were done by ultrafiltration (Concentrators, Spin 5K MWCO, 4 mL, # 51855991, www.chem.agilent.com) using 5 mL of 10% AcN with 0.1% TFA in water. Samples were evaporated to dryness in a CentriVapConcentrator (www.labconco.com), before labeling. Samples were resolved in 25µL of 1M Triethylammonium bicarbonate. iTRAQ labeling was performed according to the manufacturer's protocol (Applied Biosystems, iTRAQ Reagents Application Kit-Plasma (Amine-Modifying Labeling Reagents for Plasma Sample Applications)) with modification of the trypsin-to-protein ratio (1:14 instead 1:5.75 w/w enzyme-to-substrate). Sequencing grade modified trypsin was from Promega (# V5111, www.promega.com). iTRAQ labeling was performed twice for patients 30 and 25 (both with complete remission after treatment; see Table 1) by derivatizing the peptide mixture with the isobaric tags 114 (patient 30, pre-therapy), 115 (patient 25, pre-therapy), 116 (patient 30, post-therapy) and 117 (patient 25, post-therapy).

2.5. Pre-fractionation by Strong-Cation Exchange HPLC

In order to remove excess, unbound iTRAQ reagent and to simplify the mixture before nanoLC-MS-MS analysis, the peptide mixture was washed and fractionated off-line using a strong cation exchange (PolySULFOETHYL A) column (4.6x 200mm column (www.doi.wiley.com) operated at 0.2 mL/min flow rate using an AKTA Explorer Purifier 10 with a frac-900 fraction collector. The mobile phase was comprised of two buffers: A: 5mM KH_2PO_4/H_3PO_4 pH

3, 25% Acetonitrile [AcN] and B: 5mM KH_2PO_4/H_3PO_4 pH 3, 25% can, 1.0 M KCl. The salt gradient varied in three segments: 15%B (12 column volumes (CVs)), 50%B (3CVs), 100%B (5CVs), KCl (10mM/min). The resulting 50 fractions were pooled in 20 fractions based on intensity and dried in a CentriVapConcentrator (LABCONCO) prior to mass spectrometric analysis.

2.6. Cap-LC-MS

All cap-LC-MS analyses were performed on an Agilent 1100 capillary HPLC system coupled on-line to an SL ion-trap mass spectrometer (www.home.agilent.com) equipped with an Atlantis™ dC 18 (1.0 × 150 mm, 3 µm) column that was protected by an Atlantis™ dC 18 in-line trap column (3 µm, 2.1 mm × 20 mm guard column). 40 µL of the pretreated (depleted and digested) fractions corresponding to ~8 µg or 160 pmol of total protein digest (calculated based on a 50kDa protein) were injected. The autosampler (# G1367A) was equipped with a 100 µL injection loop and a temperature-controlled cooler (# G1330A) maintaining the samples at 4°C. The HPLC system had the following additional components: capillary pump (#, G1376A), solvent degasser (#, G1379A), UV detector (# G1314A) and column holder (#, G1316A). The sample was injected and washed in the back-flush mode for 30 min (0.1% aq. formic acid (FA) and 3% AcN at a flow rate of 50 µL/min). Peptides were eluted in a linear gradient from 0 to 70% (0.5%/min) AcN containing 0.1% FA at a flow rate of 20 µL/min. After each injection the in-line trap and the analytical column were equilibrated with eluent A for 20 min prior to the next injection.

The following settings were used for mass spectrometry during LC-MS. Nebulizer gas: 16.0 psi N_2, drying gas: 6.0 L/min N_2, skimmer: 40.0 V, Ionisation voltage: 3500 V, cap. exit: 158.5 V, Oct. 1: 12.0V, Oct. 2: 2.48 V, Oct. RF: 150 Vpp (Voltage, Peak Power Point), Lens 1: -5.0 V, Lens 2: -60.0 V, Trap drive: 53.3, T: 325°C, Scan resolution: enhanced (5500 m/z per second scan speed). Target mass: 600. Scan range: 100-1500 m/z. Spectra were saved in centroid mode. LC-MS chromatographic data were analyzed with Bruker Data Analysis software, version 2.1 (Build 37), [15].

Label-free LC-MS data were first analyzed with the data analysis software provided by the instrument manufacturer, (version 3.3, build 146; Bruker Daltonics, Bremen, Germany). For the analysis label-free data (as well as of iTRAQ data analysis) Matlab (version 7.4.0.287 (R2007a), Mathworks, Natick, MA, USA) were used. For PCA calculation and visualization the PLS toolbox (version 3.5.2, Eigenvector Research Inc., Wenatchee, Washington, USA) were used under Matlab.

2.7. nanoLC-MS-MS analysis of iTRAQ-labeled samples

Dried SCX fractions were resuspended in 50 µL 2%AcN/0.1%formic acid (FA) and analyzed by nanoLC–MS-MS with electrospray ionization (ESI)

(QSTAR-XL hybrid quadrupole-time-of-flight mass spectrometer, Applied Biosystems, Foster City, USA). 10 µL of the resuspended SCX fractions were loaded onto a 0.3 mm x 0.5cm C18-Pepmap trapping column (www.dionex.com) at a flow rate of 10 µL/min (2%AcN/0.1%FA). After 30 min of washing, the trap column was switched into the Agilent 1100 nanoflow system and peptides were separated on a 75 µm x 15 cm C18 PepMap column (Dionex) at 300 nL/min. Peptides were eluted from the column using a 105 minute gradient from 95%A to 50%A (A: H_2O/AcN/FA (950:50:1), B: AcN/ H_2O /FA (950/50/1)). The column was connected to a nanospray emitter (360/20 µM ID, 10 µm tip; NewObjective, doi.wiley.com). The typical ion-spray voltage applied was 2200V. Data was acquired using an independent data acquisition (IDA) protocol where, for each cycle, the 3 most abundant multiply-charged peptides (2 to 4 charges) in the MS scan with m/z values between 350 and 1500 amu reaching a threshold of 30 counts were selected for MS/MS. The collision energy was adjusted to gain more information in the reporter ion region. Each peptide was selected twice, and then dynamically excluded for 60 seconds. Data were processed using Analyst QS 1.1/BioAnalyst software with protein identification and quantification (ProQUANT/ProGroup programs).

2.8. Label-free data analysis and multivariate statistics

For (pre-)processing and multivariate statistical analysis of the original Bruker Daltonics capillary LC-MS data, the files were converted into ASCII-format with the Bruker data analysis software and saved in centroid mode. Centroid data were smoothed and reduced using a normalized two-dimensional Gaussian filter, with rounding the nominal m/z ratios to 1 m/z unit (the original data had a resolution of 0.1 m/z unit). In the retention time dimension no data reduction was performed. This meshing procedure reduced the number of available data points by roughly a factor 10 and corrected for shifting m/z values as a result of different loadings of the ion-trap during elution of abundant peptides, a phenomenon that is common for ion-trap mass spectrometers [20,21]. After meshing the data files of all chromatograms, they were time-aligned (warped) to a reference data file using Correlation Optimized Warping (COW) [22] based on TICs constructed from signals in the range 100-1500 m/z. The accuracy of correcting retention time shifts in this manner was manually checked by visualization before and after time alignment of the 4 most intensive peaks for 10 equal segments between 60-155 min.

A modified M-N rule was applied for peak detection by first calculating a median local baseline using a sliding window technique that was applied separately to each m/z trace. A median window size of 1200 data points, corresponding to 20.17 min, was used with a moving rate of 10 points and a minimum median value of 200 counts. According to the M-N rule, a threshold of M-times the local baseline was used and a peak was assigned if within one m/z trace the signal exceeded this threshold for at least N consecutive points [23]. For each detected chromatographic peak the m/z value, the mean retention times of the three highest measured intensities within the mass spectrum of this peak reduced by the local baseline were stored in a peak list created for every chromatogram.

We used a similar approach as Radulovic et al. [23] to obtain optimal settings for M and N. Different values for M (1.5-4) and N (4-8) were applied to two blank LC-MS runs and two LC-MS runs of depleted, trypsin-digested serum samples. Settings were used at which the ratio between the number of peaks (between 60-155 min) in the samples relative to the blank chromatograms was highest and at which a minimal number of peaks was extracted from the noise in the blank chromatogram (M = 2 and N = 5 in our case).

In order to combine peak lists from different samples into one common matrix, one-dimensional peak matching was performed by using the sliding window technique, in which the same m/z traces were evaluated for peaks that are proximate in time (step size 0.1 min; search window 1.0 min; maximal accepted standard deviation for all retention times within a group of matched peaks 0.75 min). Missing peak allocation was performed by extracting the background subtracted local signal of the given m/z trace at the given retention time. The generated peak matrix, created from the peak lists of the individual samples, consisted of a peak(row)-sample(column)-intensity(value) matrix. This peak matrix was used for multivariate statistical analysis. All data preprocessing work was done on a personal computer equipped with a dual core +3800 MHz AMD 64 X2 processor equipped with 4 GB of RAM.

2.9. Classification and multivariate statistical methods

To select the most discriminating peaks between serum sampled before (A) and after treatment (B), the Nearest Shrunken Centroid (NSC) classification algorithm was applied [24,25]. NSC regularizes data whereby class-specific centroids are "shrunken" toward the overall (non-class-specific) centroid, which has the effect of eliminating the influence of the most weakly discriminating peaks, thereby reducing the risk to overfit the data [25]. This algorithm selects those peaks that are most relevant for the discrimination of the predefined classes. NSC is used in conjunction with leave-one-out cross-validation (LOOCV) to find the shrinkage value that gives the minimal LOOCV error [26]. In LOOCV one observation per class is iteratively omitted from the data set that is used to construct the classification model, which is then used to classify the omitted observation. The selected peaks at the highest shrinkage value giving the lowest LOOCV error were employed for construction of the final classification model. The selected peaks were then subjected to auto-scaling (all variables have zero mean and unit variance) and

Principal Components Analysis (PCA) and visualized using biplots of the first two principal components [27,28].

2.10. Analysis of iTRAQ data

The total number of identified proteins were 117 and 131. The reporter ion ratios of two samples obtained at one time point (between 25A-30A and 25B-30B named as reference set) and thus presenting differences mostly related to biological variance considered as noise were natural log-transformed and median-subtracted. Gaussian curves were fitted on the smoothed histograms (smoothing using Savitzky-Golay with 2 degree of polynomial) of the above-mentioned data and standard deviations (SD) were determined. Differences in the natural log-transformed and median-subtracted reporter ion ratios (in ion ratio corresponding time point A were used as denominator for each sample) were considered to be significant when their respective mean ratio of patient 25 and 30 exceeded 2-times the SD determined on the ratios reference set. 2SD resulted in a confidence level of 95.4%. Using that value 11 and 7 proteins were found to be significantly different in the reference set. When comparing samples before treatment with samples after treatment, 24 and 14 proteins were found to be significantly different, in the two replicats

3. Results

3.1. Label-free data analysis and multivariate statistics

20 serum samples obtained from 10 cervical cancer patients before (A) and after (B) therapy (Table 1) were analyzed.

Depletion of serum samples on multiple Affinity Removal column (Agilent) allowed to remove the 6 most abundant proteins (albumin, IgG, antitrypsin, IgA, transferrin and haptoglobin). The resulting depleted sera were digested and analyzed by capLC-MS (an equivalent of appr. 2µL of the original serum were injected). After processing the LC-MS data to obtain a common, aligned peak matrix (see **Chapter IV**), datasets obtained from samples at time points A and B, respectively, were compared using the Nearest Shrunken Centroid (NSC) classification algorithm [24,25]. The NSC algorithm selected 51 peaks (at a shrinkage of 2.88) contributing to the discrimination between pre- and post-treatment samples. The relative abundance (time point A is considered as 100%) of these peaks is shown in Figure 1.

Table 1. Cervical cancer patients (long-term survivors) before (A) and after (B) treatment used for comparative serum analysis. Samples B were collected 7-11 months after treatment.

Patient number	Serum number	SCC level (ELISA), µg/L	Stage of the cancer (FIGO), [14]	Follow-up
21A	911837	3,6	IIIb	
21B	921094	0,4		complete remission
22A	971468	26,5	IVa	
22B	981748	1,4		partial remission
23A	951089	44,9	IVa	
23B	961080	6,9		stable disease
24A	870921	6,2	IIIb	
24B	880986	1,3		complete remission
25A	890702	42,0	IIIb	
25B	900226	0,2		complete remission
26A	900694	6,6	IIIb	
26B	910122	0,6		complete remission
27A	921695	2,2	IIIb	
27B	930861	0,8		complete remission
28A	930095	6,5	IIIb	
28B	931482	1,1		complete remission
29A	931355	9,1	IIIb	
29B	940841	0,9		complete remission
30A	890552	12,6	IIIa/b	
30B	940841	0,9		complete remission

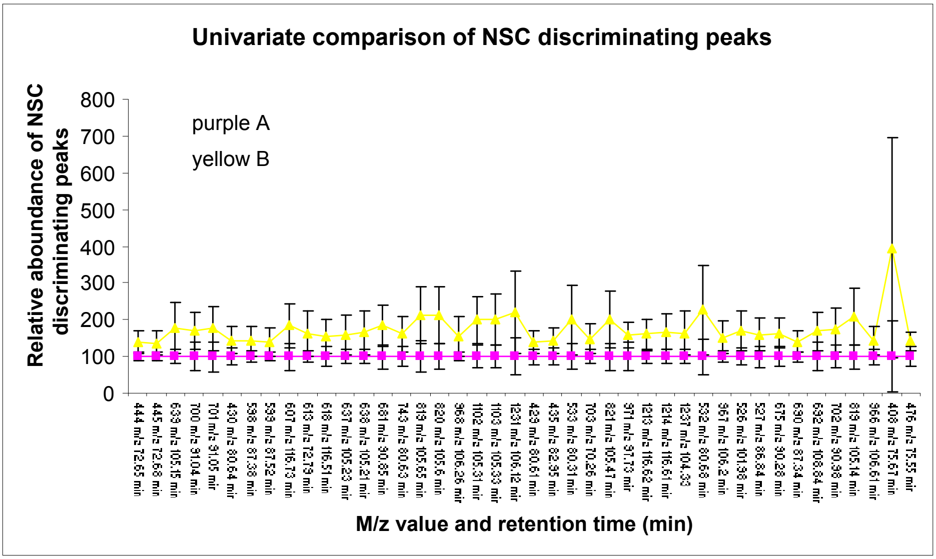

Figure 1. Relative abundance (all values obtained from time point A were set to 100%) of the 51 NSC-selected peaks (shrinkage 2.88) that contribute to discrimination between time points A (pre-treatment) and B (post-treatment) are shown with their relative standard deviations (10 patient samples in each dataset).

None of the observed differences reached statistical significance 95% when analyzed in a univariate manner. It is noteworthy that all intensities of peaks from time point B were higher than those at time point A, indicating that there is an underlying normalization problem that requires further attention. Although statistical significance was not reached, there were quantitative differences for certain peptides with extracted m/z values of: 606, 408, 532, 613, 675, 527, 1231, and 968 (Figure 2: upper traces, patients before treatment (A), lower traces patients after treatment (B)).

Principal component analysis (PCA biplot) of all 13531 peaks (Figure 3a) and of the 51 NSC-selected peaks (Figure 3b) resulted in some clustering in the case of the 51 selected peaks. Whether clustering is due to the above-mentioned normalization problem or to a real discrimination between the samples awaits further data processing and statistical analyses. A correlation map of the same 10 patient samples before (A) and after treatment (B) using all 13531 peaks (variables) showed high correlation coefficients for all samples (>0.7) indicating a high similarity between the analyzed samples (Figure 4).

3.2. Stable-isotope (iTRAQ) labeling

In order to evaluate whether more subtle differences could be detected between pre- and post-treatment serum samples, stable-isotope labeling using the commercially available iTRAQ reagents [29] were performed in duplicate on two patient sera (patients 30 and 25; see Table 1) obtained before and after treatment. Samples were depleted as described above and proteins (50 µg) were reduced, alkylated and digested with trypsin (see Methods). Peptides were subsequently derivatized with the isobaric tags 114 (patient 30, pre-therapy), 115 (patient 25, pre-therapy), 116 (patient 30, post-therapy) and 117 (patient 25, post-therapy). All samples were labeled in separate reactions and mixed together for further work-up, separation by strong cation-exchange HPLC (20 fractions) and nanoLC-MS-MS analysis. All 20 fractions were analyzed separately by LC-MS-MS. 131 or 117 proteins were identified with a confidence higher than 95% in experiments one and two, respectively, and quantified relative to each other using the reporter ions (m/z 114(30A)/115(25A)/116(30B) and 117(25B)) after MS-MS fragmentation of selected multiply-charged precursor ions.

10 proteins from 131 were up-regulated after medical treatment: these included Apolipoprotein B-100 (P04114), Apolipoprotein A-IV (P06727), Serum albumin (P02768), Apolipoprotein A-I (P02647), Lumican (P51884), C4b-binding protein alpha chain (P04003), Apolipoprotein A-II (P02652), Tetranectin (P05452), Properdin (P27918) and Fibrinogen alpha chain (P02671).

Down regulation of proteins after therapy were shown in both experiments for 4 proteins: Alpha-1-antichymotrypsin (P01011), Platelet basic protein (P02775), Leucine-rich alpha-2-glycoprotein (P02750), Ig heavy chain V-III region TUR (P01779). Figure 6 shows examples of LC-MS-MS spectra after iTRAQ labeling of up- (Figure 6A, Apolipoprotein II.) and down- regulated (Figures 6B and C, Platelet basic protein) proteins.

Serum albumin should be removed during depletion step, one of the reason of its identification could be due to partial fragmentation of protein.

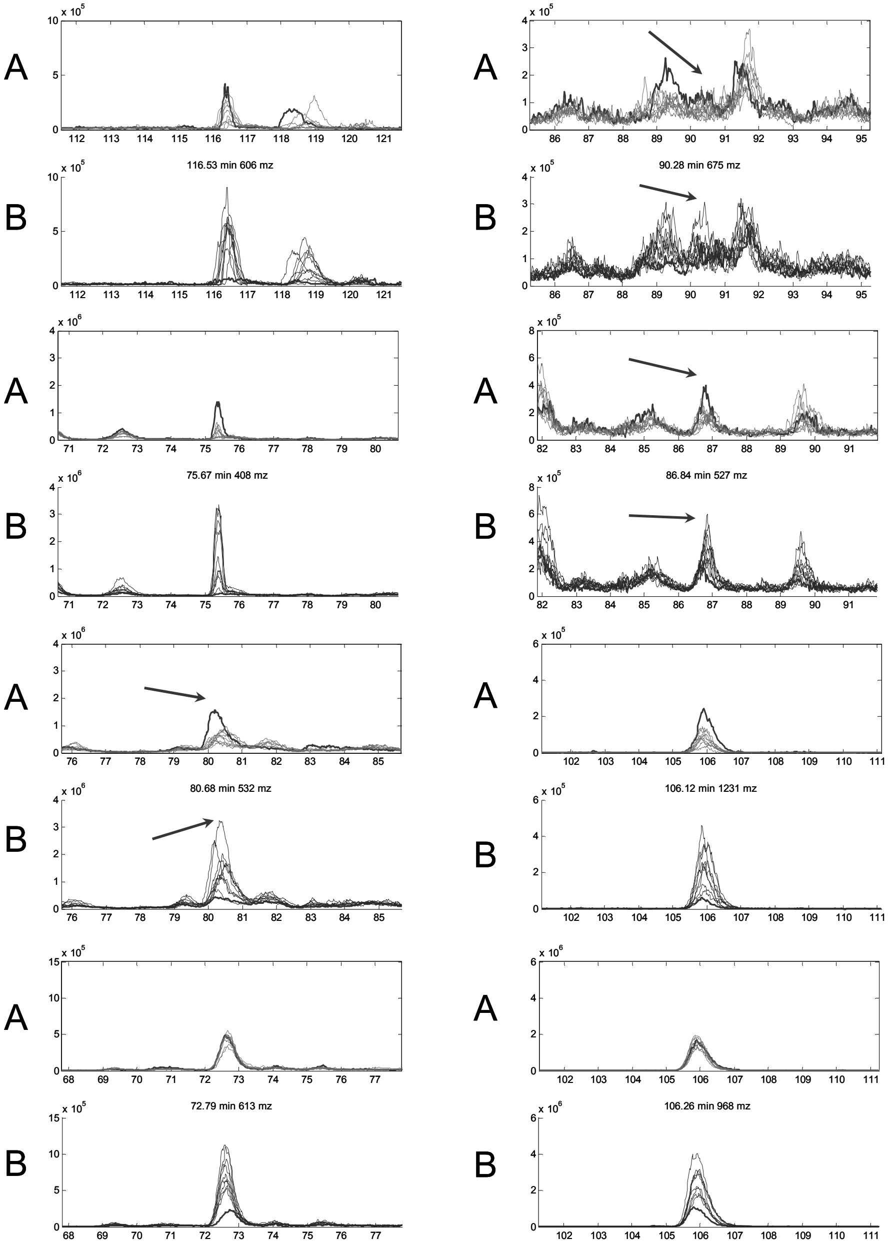

Figure 2. Extracted ion chromatograms of 8 signals from a total of 51 that were selected by the NSC algorithm to contribute to the discrimination between pre- (A) and post-treatment (B) serum samples from cervical cancer patients (late-stage survivors, see Table 1).

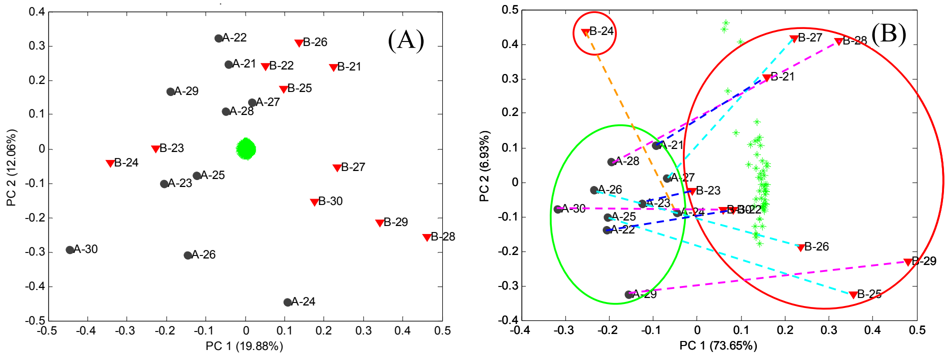

Figure 3. PCA biplot based on all (A, 13531 variables) and on the 51 NSC-selected variables (B, shrinkage 2.88) from 10 serum samples collected at the time of diagnosis of advanced stage of cervical cancer (●) and 7-11 months after treatment (▼). NSC-selected variables (✳).

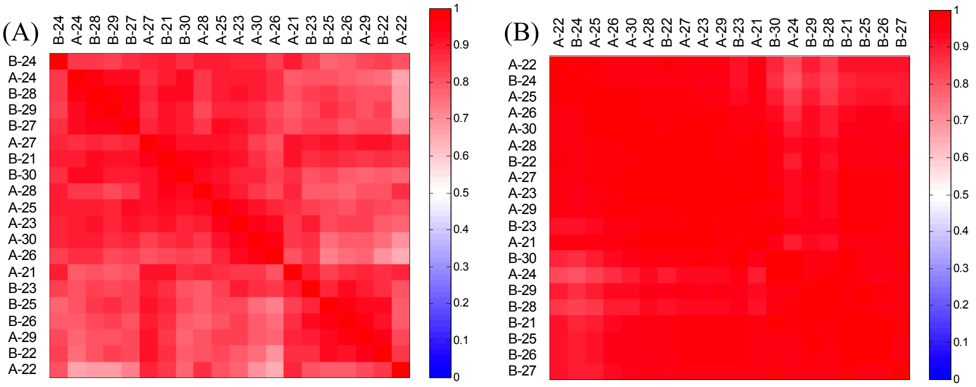

Figure 4. Correlation plots of serum samples from 10 patients collected at the time of diagnosis of advanced stage of cervical cancer (A) and 7-11 months after treatment (B) using all 13531 variables and using the NSC selected 51 variables that were obtained using an MN rule with M = 2, N = 5 for peak picking.

The LC-MS-MS data analyzed using the ProQuant software with quantification the reporter ions and calculation of the ratio relatively to a preset the patient 30, pre-therapy as denominator (114). The bias was applied for correction of the systematic error in the iTRAQ ratio due to experimental error. In experiment 1 the bias applied were: 1 for 114 signal, 1.22 for 115; 1,13 for 116 and 1.23 for 117 and 1; 1,23; 1,043; 0.92 respectively in experiment 2. Based on those ratios the bar graphs were created (Figure 5, exp. 1)

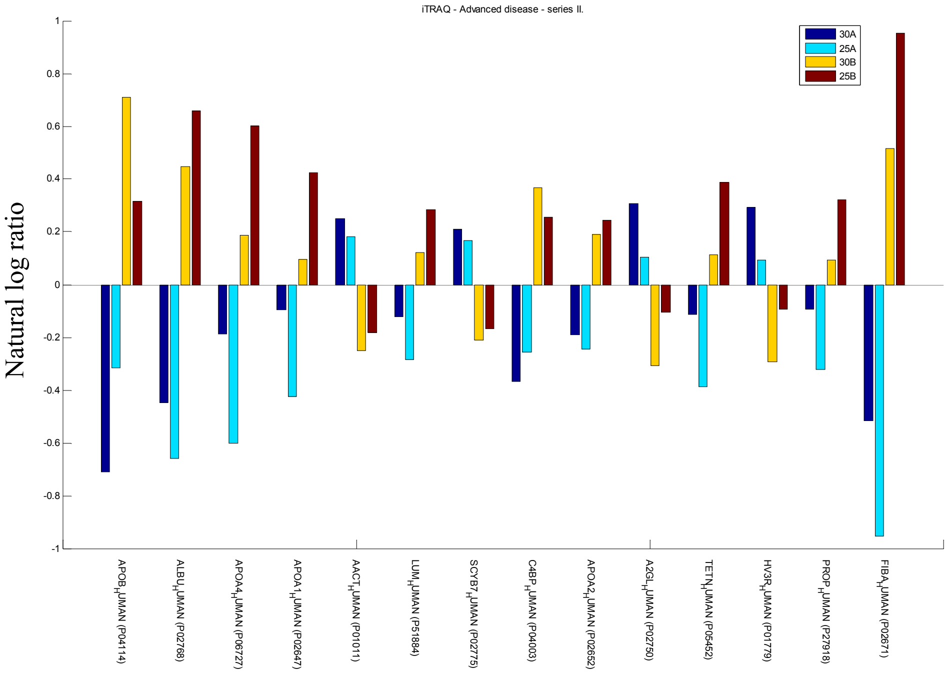

Figure 5. Proteins showing significant difference in patient sample before (30A, 25A) and after (30B, 25B) treatment using iTRAQ quantification.

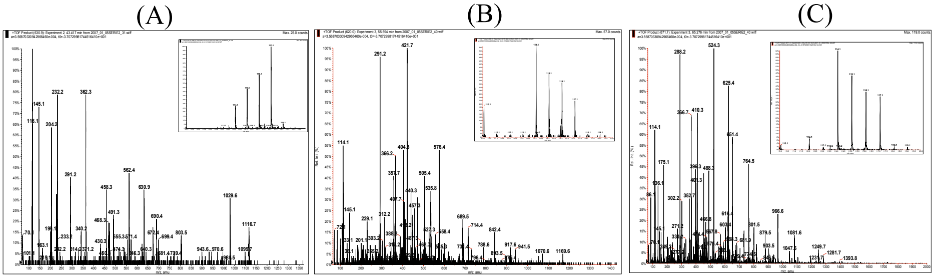

Figure 6. A: Up-regulated peptide (m/z 630,9, RT 43,417 min) derived from Apolipoprotein AII [SPELQAEAJ (P02652)]. B: Down-regulated peptide (m/z 671.7, RT 65,276 min) derived from Platelet basic protein [GJEESLDSDLYAELR (P02775)]. C: Another down-regulated peptide (m/z 620,0, RT 55,594 min) derived from Platelet basic protein [GTHCNQVEVIATLJ (P02775)]. J: abbreviation for iTRAQ-Lys.

3.3. Pre-fractionation of proteins in depleted serum by high-temperature reversed-phase HPLC on an mRP column followed by gel electrophoresis

In order to enhance the concentration sensitivity of our comparative proteomics method and since we did not get a clear-cut separation of groups before and after therapy (based on the analysis of trypsin-digested, depleted serum by cap-LC-MS), we decided to perform a more thorough pre-fractionation of depleted serum on a recently developed macroporous reversed-phase HPLC column at elevated temperature under denaturing conditions (mRP column; Agilent) followed by 1D (Patients 21, 24, 25 and 30) or 2D gel electrophoresis (Patient 30). Pre-fractionation (50 fractions) of depleted serum in a gradient of AcN containing 0.1% TFA allowed to reduce complexity significantly (Figure 7A,B).

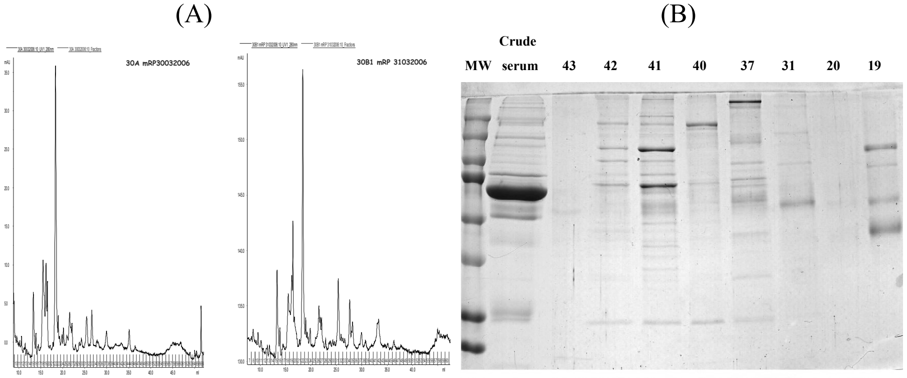

Figure 7. (A): Pre-fractionation of depleted serum from patient 30 (before treatment: left panel; after treatment: right panel) on an mRP column using an increasing gradient of AcN/0.1% TFA. (B): Example of 1D gel electrophoresis of some collected fractions (19-20, 31, 37, 40-43) demonstrate that complexity has been greatly reduced relative to original serum (lane 2).

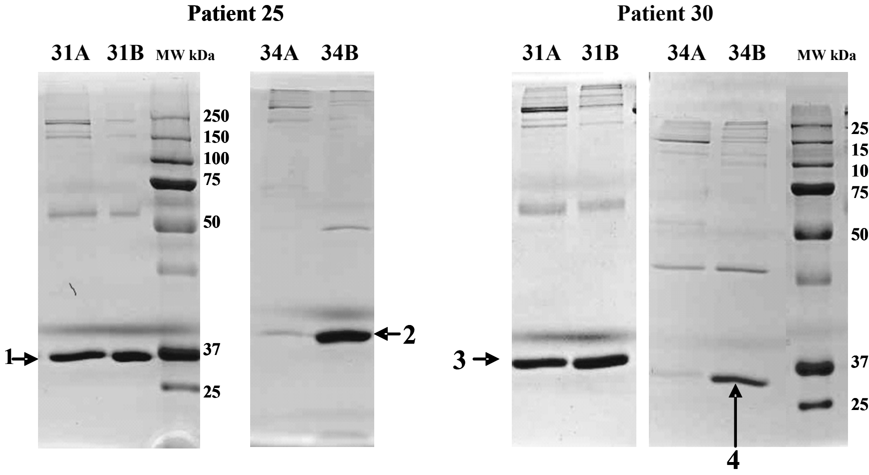

Figure 8. 1D gel electrophoresis of mRP column fractions (patients 25 and 30):
31A: A- patient before treatment, fraction #31 (mRP column)
31B: B- patient after treatment, fraction #31 (mRP column)
34A: A-patient before treatment, fraction #34 (mRP column)
34A: B-patient after treatment, fraction #34 (mRP column).
Arrows (1-4) indicate Apolipoprotein A1.

Due to the high reproducibility of the reversed-phase separation, fractions could be compared by SDS-PAGE in a pair-wise manner between sera before and after treatment obtained from the same patient (Figure 8). In most of the fractions patterns were the same, however, clear differences of protein patterns were observed in a few fractions (Figure 8).

Apolipoprotein A1 was identified in fractions 31 and 34 before and after treatment for all four examed patients (21, 24, 25 and 30; all complete remission). In all four cases the amount of ApoA1 was higher in samples after treatment. Interestingly, ApoA1 eluted in two different fractions that were well separated from each other, while it was absent in fractions 32 and 33. This is most likely due to modifications. To gain further insight into the microheterogenity of Apo A1, mRP fractions 31 and 34 (pre/post therapy) were compared by 2D gel electrophoresis (Figure 9).

After in-gel digestion of all spots and LC-MS-MS analysis, it was confirmed that all spots contained ApoA1. This confirmed that Apo A1 was indeed heterogenous with respect to its isoelectric point. The same number of ApoA1 spots were observed in fractions 31 before and after therapy (Figure 9A,B), while fractions 34 showed two additional spots after treatment (Figure 9C,D) and the abundance of ApoA1 in fraction 34 increased after treatment, confirming earlier results obtained by SDS-PAGE (see Figure 8). An example of the identification of ApoA1 in spot (*) (Figure 9D) by MALDI-TOF-MS is given as Figure S1 (supplementary material).

Another possibility to reduce complexity of serum samples is to remove more of the high-abundant proteins. The recently released ProteoPrep 20 Plasma Immunodepletion column (PROT20, Sigma-Aldrich), removes the 20 most abundant proteins from human serum (~97-99% of total protein content). This could allow to visualize further low-abundance proteins, which co-migrate during 1D gel electrophoresis and are thus masked by high-abundant proteins. Unfortunately only spin column are presently available on the market, which renders the depletion step time consuming. Serum (pre-/post-therapy) from patient 30 served again as initial test material for this immunodepletion step. Depleted serum (corresponding to 40μL of original serum) was pre-fractionated on the mRP column (Figure 10) followed by 1D gel electrophoresis. Not all 20 proteins (according description of column) were removed completely after depletion, since Haptoglobin precursor, Transferrin precursor, IgHG2_human, were identified after in-gel digestion of the corresponding bands.

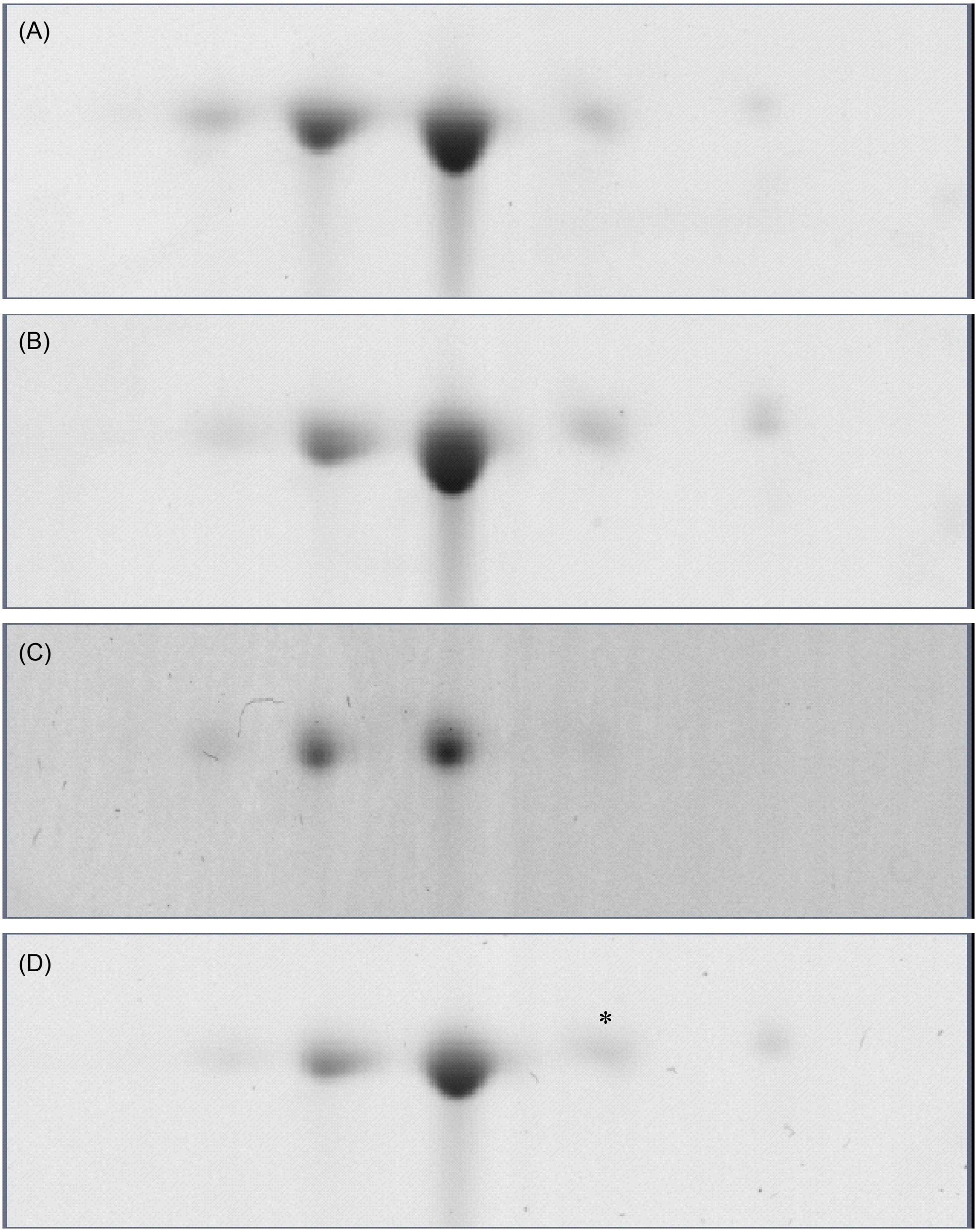

Figure 9. 2D gel electrophoresis (zoom-in) of fractions 31 before treatment (A), after treatment (B) and fractions 34 before treatment (C); after treatment (D) after pre-fractionation on mRP column. The same spots observed in fraction 31 (A,B), but in fraction 34 two more spots are present (D) after therapy.

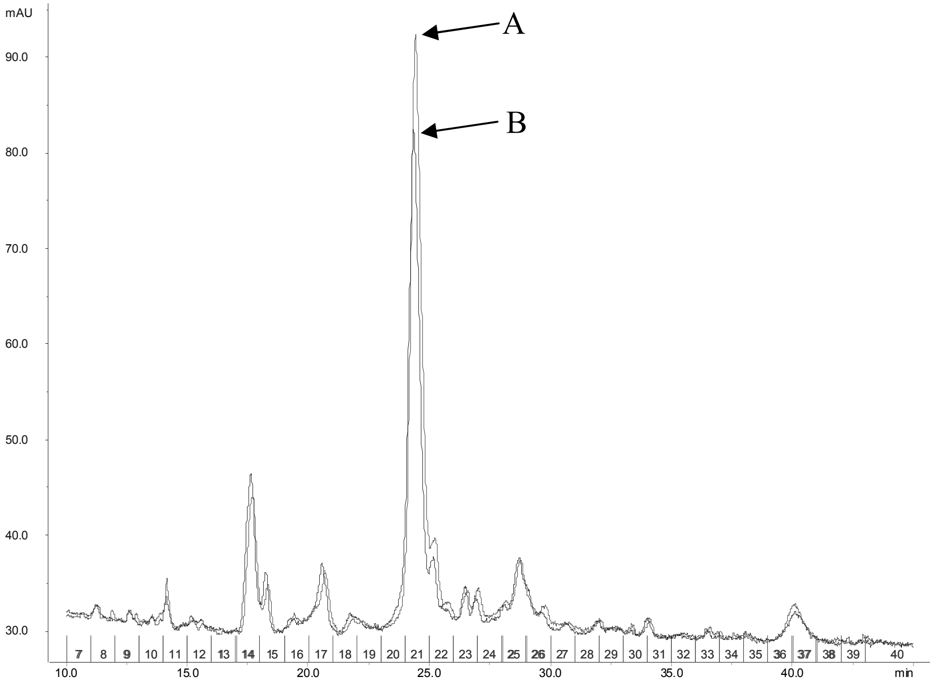

Figure 10. . Pre-fractionation of serum depleted of the 20 most abundant proteins on an mRP column using a gradient of increasing AcN/0.1% formic acid. Depleted serum of patient 30 before (A) and after (B) treatment.

While the overall chromatogram (Figure 10) does not show any major differences between pre- and post-treatment samples, gel electrophoretic analysis of the collected fractions showed that Vitronectin/Serum-spreading factor (P04004), Beta-1B-glycoprotein/Hemopexin (P02790), Complement factor H Precursor (P08603), Zinc-alpha-2-glycoprotein precursor (P25311) were present in serum prior to treatment and absent or down regulated after therapy (Figure 11).

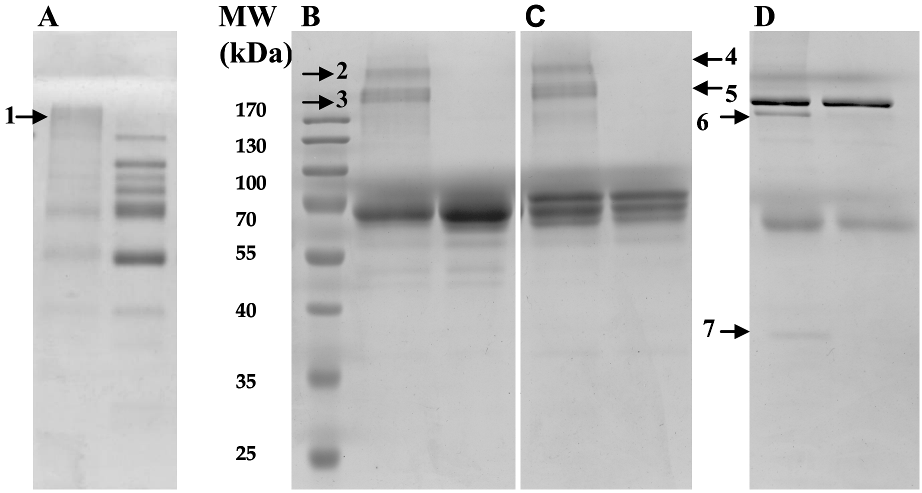

Figure 11. 1D gel electrophoresis of mRP column fractions (patient 30 before and after treatment; pair-wise analysis (from left to the right))
1: Zinc-alpha-2-glycoprotein (P25311) score 231(mascot), 102.05 (Phenyx), p=4.56E-41, fraction 17.
2, 4: Serum-spreading factor (P04004), fractions 21 (B) and 22 (C), score 7.55 (Phenyx), p= 7.68E-11 and 36.5, p=5.8E-14, respectively.
3, 5: Hemopexin (P02790), fractions 21 (B) and 22 (C), score 102,68 (Phenyx), p= 3,14E-28 and 103,45, 3=1,01E-32
6-7: Complement factor H: a and b, respectively (P08603), fraction 15, score 102,68 (Phenyx), p= 3,14E-28 and 103,45, 3=1,01E-32

4. Discussion

Pre-processing and multivariate statistics of LC-MS data obtained from advanced stage patients with cervical cancer did not allow us to detect significant differences in peak areas or intensities for pre- or post- therapy samples after depletion of the 6 most abundant serum proteins followed by trypsin-digestion and cap-LC-MS.

Stable-isotope labeling (iTRAQ) and additional prefractionation of the digest by strong cation-exchange HPLC detected four different Apolipoproteins that were found to be up- regulated in serum from 2 patients after treatment. They were: Apolipoprotein B-100 (P04114), Apolipoprotein A-IV (P06727), Apolipoprotein A-I (P02647) and Apolipoprotein A-II (P02652). Apo A-IV is a major component of HDL (high density lipoprotein) and chylomicrons [30]. Apo A-II stabilizes the structure of HDL [31,32]. Apo B-100 plays a role as a recognition signal for the cellular binding by the apoB/E receptor [33,34]. Apo-A-I is important in the reverse transport of cholesterol from tissues to the liver and also as a cofactor for lecithin cholesterol acyltransferase (LCAT) [35]. Up-regulation of Apo-A-I after treatment agreed well with results of 1D gel electrophoretic analysis after pre-fractionation on an mRP column (Figure 8, fraction 34) and with subsequent 2D gel analysis (Figure 10C, 10D). Apo A-I has been shown to inhibit herpes simplex virus-induced cell fusion [36], inhibits HIV-induced syncytium formation [37]. It has previously been shown that changes in the amounts of isoforms of Apolipoprotein A-1 could be a useful biomarker for HIV patients (diagnosis and effect of therapy) [38]. It is thus conceivable that increasing amounts of Apo A-I and changes in its isoforms, as detected by 2D gel electrophoresis in patient 30, could be a marker for patient recovery after medical treatment.

Fibrinogen (Fibrinogen alpha chain (P02671) forms monomers that polymerize into fibrin after cleavage by thrombin and that promote platelet aggregation [39]. Tetranectin is involved in packaging of molecules for exocytosis [40]. Properdin stabilizes the C3/C5 convertase enzyme complexes (regulation of the complement pathway) [41].

Lumican (P51884) is a small proteoglycan that regulates the fibrillar and suprafibrillar organization of collagen in the periodontal ligament [42]. C4b-binding protein alpha chain (P04003) controls activation of the complement pathway and interacts with the anticoagulant protein S and with a component of serum amyloid P [43].

Serum albumin (P02768) is the main protein in plasma and serum, regulating the colloidal osmotic pressure of blood, binding to Ca^{2+}, Na^{+}, K^{+}, fatty acids, hormones, bilirubin and other compounds [44,45]. C-reactive protein (CRP) was down-regulated in experiment 2. CRP is associated with host defense including agglutination, bacterial capsular swelling, phagocytosis and is a well-known "acute-phase" protein that is up-regulated upon inflammation [46]. It is thus expected to be down-regulated when a patient's health improves.

With respect to our search for potential biomarkers, down-regulated proteins after therapy are of special interest. Alpha-1-antichymotrypsin (P01011), Platelet basic protein (P02775), Leucine-rich alpha-2-glycoprotein (P02750) and Ig heavy chain V-III region TUR (P01779) were down-regulated in both iTRAQ labeling experiments. Leucine-rich alpha-2-glycoprotein is most probably a membrane-associated or membrane-derived protein, associated with binding properties [47,48], Alpha-1-antichymotrypsin inhibits neutrophil cathepsin G and mast cell chymase. Both proteins are involved in the conversion of angiotensin-1 (inactive) to the angiotensin-2 (active) [49]. Alpha-antichymotrypsin belongs to "acute-phase" proteins and likely being related to generally improved health after successful treatment. Platelet basic protein (contains antibacterial proteins TC-1 and TC-2) stimulates DNA synthesis, mitosis, glycolysis and the secretion of plasminogen [50,51]. Ig heavy chain V-III region TUR contains 1 Ig-like (immunoglobulin-like) domain and this chain was isolated from an IgA1 myeloma protein [52]. The variable region sequences of five human immunoglobulin heavy chains of the VH3 subgroup form a multifunctional locus that is needed both for the IgG and the IgA class.

It is hard to say why this protein should be increased in cervical cancer samples and decrease after successful therapy but one possibility is that it is connected to the immune response against viral infections, as was shown for other diseases caused by an underlying infection [53,54].

After depletion of 20 of the most abundant serum proteins, it was possible to load a bigger volume of the original serum (in our case 40μL) on the high-temperature reversed-phase pre-fractionation column in order to improve concentration sensitivity and to reach into a range, where known potential tumor markers have been identified (ng/mL range). 30-50 fold increase in the relative amount of low abundant proteins could be reached on ProteoPrep 20 Plasma Immunodepletion (PROT20) column, as example we were able to detect E1 protein: Human papillomavirus type 31 (W1WL31) after depletion of 20 proteins followed by mRP fractionation (fraction 31) in serum Patient 25 with confidence 95% (data are not shown).

Vitronectin/Serum-spreading factor (P04004), Beta-1B-glycoprotein/ Hemopexin (P02790), Complement factor H Precursor (P08603) and Zinc-alpha-2-glycoprotein precursor (P25311) were all present in serum samples prior to treatment and absent or down-regulated in post-treatment samples (Figure 12 and Figure S2 in supplementary materials). Human Zinc-alpha-2-glycoprotein was previously found in benign and malignant breast tissues [55]. It is quite likely that this protein is an indicator of tumor development.

Recently, the analysis of the plasma proteome from a mouse model of intestinal cancer has been described [56]. Interestingly, these authors found that the levels of some apolipoproteins, hemoglobins, fibronectin and immunoglobulins changed upon tumor progression. This indicates that some of our findings could reflect cancer development, rather than a more generic response to medical treatment.

In order to group the identified proteins that change in response to successful treatment according to functional classes, we used a Gene Ontology approach accessed through the WebGestalt server (http://bioinfo.vanderbilt.edu/webgestalt). It appeared that a group of proteins (Apolipoprotein A-II, Complement component 4 alpha chain binding protein (up-regulated after treatment) and Vitronectin, Platelet basic protein (chemokine (C-X-C motif) ligand 7), Complement factor H, Zink-alpha-2-glycoprotein, Hemopexin (down-regulated after treatment) were clustered in categories related to the immune system/response or inflammatory and response to injury. It is quite likely therefore to assume that the observed changes are part of a coordinated common function. Whether any of them is specific for cancer and cervical cancer, in particular, remains to be investigated.

This requires the analysis of additional patients preferably using the ProteoPrep 20 Plasma Immunodepletion column in the format of an HPLC rather than a spin column. We are also in the process to perform more extensive longitudinal comparisons of samples from the same patient.

References

[1]. Franco EL, Schlecht NF, Saslow D. The epidemiology of cervical cancer. The cancer journal 2003;9:348-359

[2]. Parkin DM, Bray F, Ferlay J, Pisani P. Global cancer statistics 2002. CA Cancer J Clin. 2005;55:74–108

[3]. Monsonego J. HPV infections and cervical cancer prevention. Priorities and new directions. Highlights of EUROGIN 2004 International Expert Meeting, Nice, France, October 21–23, 2004. Gynecologic Oncology 2005;96:830–839.

[4]. Cotton SC, Sharp L, Seth R, Masson LF, Little J, Cruickshank ME, Neal K, Waugh N. Lifestyle and socio-demographic factors associated with high-risk HPV infection in UK women. Br J Cancer. 2007;97(1):133-139.

[5]. Nguyen HH, Broker TR, Chow LT, Alvarez RD, Vu H.L, Andrasi J, Brewer LR, Jin G, Mestecky. J. Immnune responses to human papillomavirus in genital tract of women with cervical cancer. Gynecologic oncology 2005;96:452-461.

[6]. Esajas MD, Duk JM, de Bruin HW, Aalders JG, Willemse PH, Sluiter W, Pras B, ten Hoor K, Hollema H, van der Zee AGJ. J. Clin. Oncol. 2001;19:3960-3966

[7]. Hong JH, Tsai CS, Chang JT., et al., The prognostic significance of pre-and post-treatment SCC levels in patients with squamous cell carcinoma of the cervix treated by radiotherapy. Int. J. Radiat. Oncol. Biol. Phys. 1998;41:823-830.

[8]. Ho YJ, Hsieh JF, Tasi SC, Lee JK, Kao CH. Tissue polypeptide specific antigen and squamous cell carcinoma antigen for early prediction of recurrence in lung squamous cell carcinoma. Lung 2000;178:75–80.

[9]. Micke O, Bruns F, Schafer U, Horst E, Buntzel J, Willich N. The clinical value of squamous cell carcinoma antigen in patients irradiated for locally advanced cancer of the head and neck. Anticancer Res. 2003;23:907–911.

[10]. Hefler L, Obermair A, Tempfer C., et al., Serum concentrations of squamous cell carcinoma antigen in patients with vulvar intraepithelial neoplasia and vulvar cancer. Int J Cancer 1999;84:299–303.

[11]. Lin H, ChangChien CC, Huang EY, Tseng CW, Eng HL, Huang CC. The role of pretreatment squamous cell carcinoma antigen in predictingnodalmetastasis in early stage cervical cancer. Acta Obstet Gynecol Scand 2000;79:140–144.

[12]. Huang X, Wei Y, Li L, Wen Y, Yang J, Liu B, Song X, Zhao J. Serum proteomics study of the squamous cell carcinoma antigen 1 in tongue cancer. Oral Oncology 2006;42:26–31.

[13]. Hazelbag S, Kenter GG, Gorter A, Fleuren GJ. Prognostic relevance of TGF-beta1 and PAI-1 in cervical cancer.Int J Cancer. 2004 Dec 20;112(6):1020-1028.

[14]. Benedet JL, Bender H, Jones H, Ngan HYS, Pecorelli S. FIGO staging classifications and clinical practice guidelines in the management of gynecologic cancers. FIGO Committee on Gynecologic Oncology. Int. J. Gynec. Obst. 2000;70:207-212.

[15]. Govorukhina NI, Reijmers TH, Nyangoma SO, Zee AGJ van der, Jansen RC, Bischoff R. Analysis of human serum by LC-MS: improved sample preparation and data analysis. J. Chromatogr. A. 2006;110:142-150.

[16]. Shevchenko A, Wilm M, Vorm O, Mann M. Mass Spectrometric Sequencing of Proteins from Silver-Stained Polyacrylamide Gels. Anal. Chem. 1996;68:850-858.

[17]. Rabilloud T, Valette C, Lawrence JJ. Sample application by in-gel rehydration improves the resolution of two-dimensional electrophoresis with immobilized pH-gradients in the first dimension. Electrophoresis 1994;15:1552-1558.

[18]. Görg A, Obermaier C, Boguth G, Harder A, Scheibe B, Wildgruber R, et al., The current state of two-dimensional electrophoresis with immobilized pH gradients. Electrophoresis 2000;21:1037-1053.

[19]. Yan JX, Wait R, Berkelman T, Harry RA, Westbrook JA, Wheeler CH, et al., A modified silver staining protocol for visualization of proteins compatible with matrix-assisted laser desorption/ionization and electrospray ionization-mass spectrometry. Electrophoresis 2000;21:3666-3672.

[20]. Hao Ch., March ER. A survey of recent research activity in quadrupole ion-trap mass spectrometry, Int. J. Mass Spectrum. 2001;212:337–357.

[21]. Wells JM, Plass WR, Patterson GE, Ouyang Zh, Badman ER, Cooks RG. Chemical Mass Shifts in Ion-trap Mass Spectrometry: Experiments and Simulations. Anal. Chem. 1999;71:3405-3415.

[22]. Tomasi G, van den Berg F, Andersson C. Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic Data. J. of Chemometrics 2004;18:231-241.

[23]. Radulovic D, Jelveh S, Ryu S, Hamilton TG, Foss E, Mao YY, Emili A. Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. Mol. Cell. Proteomics 2004;3:984-997.

[24]. Tibshirani R, Hastie T, Narasimhan B, Chu G. Diagnosis of multiple cancer types by shrunken centroids of gene expression. Proc. Natl. Acad. Sci. USA. 2002;99:6567-6572.

[25]. Tibshirani R, Hastie T, Narasimhan B, Soltys S, Shi G, Koong A, Le QT. Sample classification from protein mass spectrometry by 'peak probability contrasts'. Bioinformatics 2004;20:3034-3044.

[26]. Listgarten J, Emili A. Statistical and computational methods for comparative proteomic profiling using liquid chromatography-tandem mass spectrometry. Mol. Cell. Proteomics 2005;4:419-434.

[27]. Wagner M, Naik D, Pothen A. Protocols for disease classification from mass spectrometry data. Proteomics 2003;1692-1698.

[28]. Hilario M, Kalousis A, Muller M, Pellegrini C. Machine learning approaches to lung cancer prediction from mass spectra. Proteomics 2003,3:1716-1719.

[29]. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K., et al., Multiplexed Protein Quantitation in Saccharomyces Cerevisiae Using Amine-Reactive Isobaric Tagging Reagents. Mol Cell Proteomics 2004;3:1154-1169.

[30]. Deeb SS, Nevin DN, Iwasaki L, Brunzell JD. Two novel apolipoprotein A-IV variants in individuals with familial combined hyperlipidemia and diminished levels of lipoprotein lipase activity. Hum. Mutat. 1996;8:319-325.

[31]. Lux SE, John KM, Ronan R, Brewer HB Jr. Isolation and characterization of the tryptic and cyanogen bromide peptides of apoLp-Gln-II (apoA-II), plasma high density apolipoprotein. J. Biol. Chem. 1972;247:7519-7527.

[32]. Kumar MS, Carson M, Hussain MM, Murthy HM. Structures of apolipoprotein A-II and a lipid-surrogate complex provide insights into apolipoprotein-lipid interactions.Biochemistry 2002;41:11681-11691.

[33]. Dashti N, Lee DM, Mok T. Apolipoprotein B is a calcium binding protein. Biochem. Biophys. Res. Commun. 1986;137:493-499.

[34]. Zhao Y, McCabe JB, Vance J, Berthiaume LG. Palmitoylation of apolipoprotein B is required for proper intracellular sorting and transport of cholesteroyl esters and triglycerides. Mol. Biol. Cell 2000;11:721-734.

[35]. Hoeg JM, Meng MS, Ronan R, Fairwell T, Brewer HB Jr. Human apolipoprotein A-I. Post-translational modification by fatty acid acylation. J. Biol. Chem. 1986;261:3911-3914.

[36]. Srinivas RV, Venkatachalapathi YV, Rui Z, Owens RJ, Gupta KB, Srinivas SK, Anantharamaiah GM, Segrest JP, Compans RW. Inhibition of virus-induced cell fusion by apolipoprotein A-I and its amphipathic peptide analogs. J Cell Biochem. 1991;45(2):224-37.

[37]. Owens BJ, Anantharamaiah GM, Kahlon JB, Srinivas RV, Compans RW, Segrest JP. Apolipoprotein A-I and its amphipathic helix peptide analogues inhibit human immunodeficiency virus-induced syncytium formation. J Clin Invest. 1990 Oct;86(4):1142-50.

[38]. Kim S, Kim M, Shin B, Na H, Choi J, Kee M, Chong S, Nam. M. Different isoforms of Apolipoprotein A1 present heterologous post-translational expression in HIV infected patients. J. of Proteome Research 2007;6:180-184.

[39]. Everse SJ, Spraggon G, Veerapandian L, Doolittle RF. Conformational changes in fragments D and double-D from human fibrin(ogen) upon binding the peptide ligand Gly-His-Arg-Pro-amide. Biochemistry 1999;38:2941-2946.

[40]. Fuhlendorff J, Clemmensen I, Magnusson S. Primary structure of tetranectin, a plasminogen kringle 4 binding plasma protein: homology with asialoglycoprotein receptors and cartilage proteoglycan core protein. Biochemistry 1987;26:6757-6764.

[41]. Hartmann S, Hofsteenge J. Properdin, the positive regulator of complement, is highly C-mannosylated. J. Biol. Chem. 2000;275:28569-28574.

[42]. Matheson S, Larjava H, Hakkinen L. Distinctive localization and function for lumican, fibromodulin and decorin to regulate collagen fibril organization in periodontal tissues. J Periodontal Res. 2005 Aug; 40(4):312-324.

[43]. Liu T, Qian W-J, Gritsenko M, Camp DG. II, Monroe ME, Moore RJ, Smith RD. Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry. J. Proteome Res. 2005;4:2070-2080.

[44]. Carter DC, He X.-M. Structure of human serum albumin. Science 1990;249:302-303.

[45]. He X-M, Carter DC. Atomic structure and chemistry of human serum albumin. Nature 1992;358:209-215.

[46]. Thompson D, Pepys MB, Wood SP. The physiological structure of human C-reactive protein and its complex with phosphocholine. Structure 1999;7:169-177.

[47]. Takahashi N, Takahashi Y, Putman FW. Periodicity of leucine and tandem repetition of a 24-amino acidsegment in the primary structure of leucine-rich α 2-glycoprotein of human serum. Biochemistry 1985;82:1906-1910.

[48]. Cummings C, Walder J, Treeful A, Jemmerson R. Serum leucine-rich alpha-2-glycoprotein-1 binds cytochrome c and inhibits antibody detection of this apoptotic marker in enzyme-linked immunosorbent assay. Apoptosis 2006;11:1121–1129.

[49]. Morii M, Travis J. Structural alterations in alpha 1-antichymotrypsin from normal and acute phase human plasma. Biochem. Biophys. Res. Commun. 1983;111:438-443.

[50]. Holt JC, Harris ME, Holt AM, Lange E, Henschen A, Niewiarowski S."Characterization of human platelet basic protein, a precursor form of low-affinity platelet factor 4 and beta-thromboglobulin. Biochemistry 1986;25:1988-1996.

[51]. Piccardoni P, Evangelista V, Piccoli A, de Gaetano G, Walz A, Cerletti C. Thrombin-activated human platelets release two NAP-2 variants that stimulate polymorphonuclear leukocytes. Thromb. Haemost. 1996;76:780-785.

[52]. Steiner LA, Pardo AG, and Margolied MN. Amino Acid Sequence of the Heavy-Chain Variable Region of the Crystallizable Human Myeloma Protein Dobt Biochemistry. 1979;18(19):4068-4080.

[53]. Chang Q, Abadi J, Alpert P, Pirofski L. A pneumococcal capsular polysaccharide vaccine induces a repertoire shift with increased VH3 expression in peripheral B cells from human immunodeficiency virus (HIV)-uninfected but not HIV-infected persons.J Infect Dis. 2000 Apr;181(4):1313-1321.

[54]. Bessudo A, Rassenti L, Havlir D, Richman D, Feigal E, Kipps TJ. Aberrant and unstable expression of immunoglobulin genes in persons infected with human immunodeficiency virus. Blood. 1998 Aug 15;92(4):1317-1323.

[55]. Freije JP, Fueyo A, Uria J, Lorez-Otin C. Human Zn-alpha-2-glycoprotein cDNA cloning and expression analysis in benign and malignant breast tissues. FEBS lett.1991;290:247-249.

[56]. Hung KE, Kho AT, Sarracino D, Richard LG, Krastins B, Forrester S, Haab BB, Kohane IS, Kucherlapati R. Mass spectrometry-based study of the plasma proteome in a mouse intestinal tumor model. J Proteome Res. 2006;5(8):1866-1878.

Supplementary materials

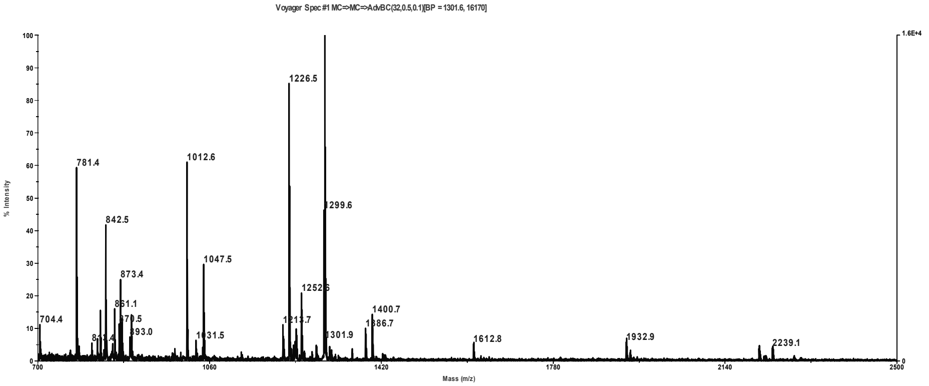

Figure S1. Example of a MALDI spectrum that served to identify Apolipoprotein A-I(1-242), P02647 in one spot after 2D gel electrophoresis as indicated with an asterisk (*) in Figure 9. M.W. 28kDa; Score: 90.36, Hits: 13, sequence coverage: 52%.

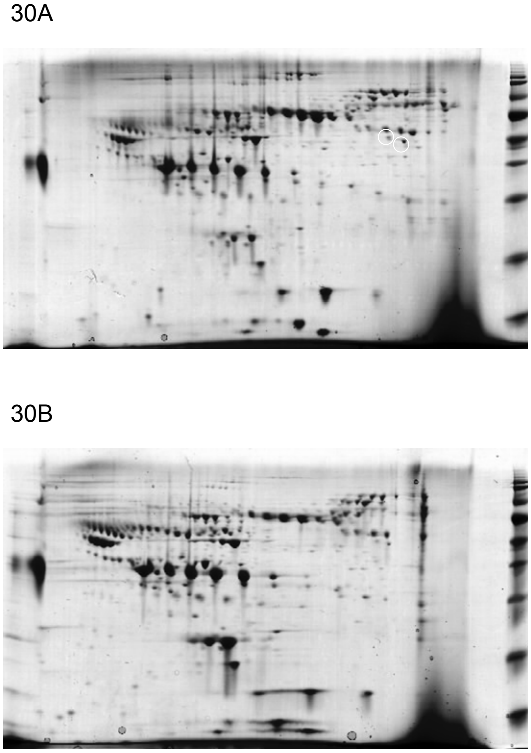

Figure S2. 2D gel electrophoresis of 6 protein depleted serum of patient 30 before (A, above) and after (B, below) medical treatment. White circles (above): Hemopexin _Human (P02790) Score: 34,17; 31% of coverage (left spot); Hemopexin _Human (P02790) Score: 13.68; 24% of coverage (right spot).

Chapter VII.

Summary and perspectives

This thesis describes the analysis of serum samples of cervical cancer patients before and after treatment using LC-MS of trypsin-digested serum followed by data processing and statistical data analysis. Paired samples from the same patients before treatment and after remission were selected in order to optimize the chance of discovery cancer-related changes in the proteome. While we did not discover major changes in the serum proteome, some changes in protein composition were found in samples taken before and after treatment. It is thus demonstrated that the described methods are applicable to highly complex body fluids such as serum and that further studies into the relevance of the discovered changes of the serum proteome are warranted.

At present, we are able to discriminate serum proteins in healthy (patients after successful treatment) and cancer patients at a concentration which is still considered to represent "classical serum or plasma proteins" [1]. Changes in those proteins reflect the condition of patients in general. We have to reach lower concentration ranges for disease-specific biomarkers (ng/mL range). For this purpose, protein enrichment will be very helpful (e.g. an enrichment strategy for less abundant serum glycoproteins).

Proteins such as cytokines, growth factors, cycling dependent kinases (CDKs) and inhibitors of cell cycling belong to the very low-abundance proteins in the cell. Although DNA microarray studies demonstrate that they are very important in the diagnosis of malignant disease [2], the staging of cancer and the characterization of the cell cycling in general, these proteins are not readily detectable in proteomics studies. Part of the problem originates from the source of the samples: cell cycle controlling proteins are localized in the nucleus. It may thus be of interest to select the most appropriate type of tissue and to perform subfractionation of cell organelles prior to MS analysis.

The problem of biological variance of protein concentrations that are unrelated to disease is very common and well recognized in cancer biomarker studies. In our case it is necessary to analyze additional patient samples and to perform more extensive longitudinal studies to discriminate between actual cancer markers and unrelated confounding effects.

The importance of genome-wide searches for disease markers in cancer has been proven with DNA microarray studies [3]. Microarrays are increasingly important tools for the diagnosis and classification of various cancers [4-6]. Expression profiles with microarrays in their present form are capable to detect nanogram quantities of mRNA, but say very little about the corresponding protein levels in normal and malignant tissues

The technology of oligonucleotide-based arrays is well developed and well understood, which is not the case for attempts to create similar protein microarrays [7]. In this respect, the LC-MS approach offers additional possibilities to the protein-chip concept. The performance of mass spectrometers has been steadily improving in recent years and present modern LC-MS systems are capable of analyzing complex protein/peptides samples generating thousands of peaks. Whether these analyses are comprehensive is still doubtful (limitation of proteomics methods with respect to concentration sensitivity) but it is fair to assume that a considerable part of the proteome can be covered. It has become increasingly clear that the successful integration of sample preparation, protein/peptide separation and data analysis is key to an overall successful biomarker discovery strategy.

Data pre-processing and statistical analysis of comprehensive LC-MS profiles is a rather new field. In collaboration with other groups, we are working on further improving our statistical analysis platform, since it is a crucial step in the discrimination of samples and the search for potential biomarkers.

By nature, cervical cancer is a localized disease at its early stage and will therefore quite often escape the general immune response mediated by natural killer cells and T-cells circulating in blood. Because of such a limited exchange between proteins of the cervical epithelium and blood, finding specific biomarkers seems more promising in comparative studies of cervical cancer and healthy tissue. Work along this line has been initiated using Laser Dissection Microscopy in conjunction with mass spectrometry.

Although an effective vaccine against Human Papillomavirus (HPV) will most probably decrease the incidence of cervical cancer significantly, early diagnosis still remains relevant. Moreover, finding biomarkers for cervical cancer can help in other cases where HPV virus seem to be involved. As an example, HPV 16 was found in lung carcinoma cases in Chile [8] and Korea [9].

It is thus quite likely that this virus can infect other kinds of epithelia than the cervix alone.

Finally, I would like to stress that only well-integrated, multi-disciplinary studies will have the potential to discover relevant and new biomarkers in cervical cancer, resulting in earlier diagnosis and thereby better treatment of patients with cervical neoplasia.

References

[1]. Anderson NL, Anderson NG. The Human Plasma Proteome: History, Character, and Diagnostic Prospects. Mol Cell Proteomics 2002;1:845-867.

[2]. Lehman NL, Tibshirani R, Hsu JY, Natkunam Y, Harris BT, West RB, Masek MA, Montgomery K, van de Rijn M, Jackson PK. Oncogenic regulators and substrates of the anaphase promoting complex/cyclosome are frequently overexpressed in malignant tumors. Am J Pathol. 2007;170(5):1793-1805.

[3]. Ewis AA, Zhelev Z, Bakalova R, Fukuoka S, Shinohara Y, Ishikawa M, Baba Y. A history of microarrays in biomedicine. Expert Rev Mol Diagn. 2005;5(3):315-328.

[4]. Miller LD, Liu ET. Expression genomics in breast cancer research: microarrays at the crossroads of biology and medicine. Breast Cancer Res. 2007;9(2):206.

[5]. Henrickson SE, Hartmann EM, Ott G, Rosenwald A. Gene expression profiling in malignant lymphomas. Adv Exp Med Biol. 2007;593:134-46.

[6]. Perez-Diez A, Morgun A, Shulzhenko N. Microarrays for cancer diagnosis and classification. Adv Exp Med Biol. 2007;593:74-85.

[7]. Whiteley GR. Proteomic patterns for cancer diagnosis: promise and challenges. Mol Biosyst. 2006 Aug;2(8):358-363.

[8]. Aguayo F, Castillo A, Koriyama C, Higashi M, Itoh T, Capetillo M, Shuyama K, Corvalan A, Eizuru Y, Akiba S. Human papillomavirus-16 is integrated in lung carcinomas: a study in Chile. Br J Cancer. 2007 Jul 2;97(1):85-91.

[9]. Park MS, Chang YS, Shin JH, Kim DJ, Chung KY, Shin DH, Moon JW, Kang SM, Hahn CH, Kim YS, Chang J, Kim SK, Kim SK. The prevalence of human papillomavirus infection in Korean non-small cell lung cancer patients. Yonsei Med J. 2007 Feb 28;48(1):69-77.

Samenvatting

In dit proefschrift wordt de analyse van serum monsters van patiënten met baarmoederhalskanker vóór en na behandeling beschreven. Hierbij werd gebruik gemaakt van LC-MS in combinatie met door trypsine behandeld serum waarna de gegevens werden verwerkt en vervolgens statistisch getoetst. Bij een aantal patiënten werden simultaan serummonsters van vóór behandeling en na herstel van de kanker vergeleken om een zo groot mogelijke kans te hebben kankergerelateerde veranderingen in het proteoom te vinden. Hoewel we geen grote veranderingen in het serumproteoom vonden, zagen we in de vergelijking van monsters genomen vóór en na de behandeling veranderingen van de eiwitsamenstelling. Zo is aangetoond dat de beschreven methodes toepasbaar zijn voor zeer ingewikkelde lichaamsvloeistoffen zoals serum en dat verdere studies naar de relevantie van de ontdekte veranderingen van het serum proteoom hierdoor nodig zijn.

Op dit moment zijn we in staat om serumeiwitten te onderscheiden van gezonde (patiënten na succesvolle behandeling) en kankerpatiënten, bij een concentratie waarin de "klassieke serum- of plasma eiwitten" nog vertegenwoordigd zijn [1]. Veranderingen in de samenstelling van deze eiwitten weerspiegelen de algemene conditie van patiënten. Voor ziektespecifieke biomarkers is een (detectie methode met een) hogere resolutie (ng/mL) nodig. Voor dit doel is het verrijken van eiwitten wenselijk bijvoorbeeld een verrijkingsstrategie voor de in lage concentratie aanwezige glycoproteines. Eiwitten zoals cytokines, groeifactoren, "cycling-dependent kinasen"(CDK's) en remmers van de celcyclus, horen bij de groep eiwitten die in zeer lage concentraties in een cel aanwezig zijn. Hoewel DNA microarray studies erg belangrijk zijn bij de diagnose van kwaadaardige ziekten [2], bij de identificatie van de stadia van kanker en het kenschetsen van de celcyclus in het algemeen, zijn in proteoomstudies deze eiwitten zijn niet makkelijk aantoonbaar. Een deel van dit probleem komt voort uit de aard van de monsters: celcyclus sturende eiwitten bevinden zich in de celkern. Het kan daarom van belang zijn om het meest geschikte weefsel te selecteren en een fractionering van celorganellen uit te voeren voor de MS analyse.

Het probleem van biologische variatie van eiwitconcentraties die niet gerelateerd zijn aan een ziekte is een algemeen en veel voorkomend fenomeen in kanker biomarker studies. In ons geval is het noodzakelijk om aanvullende patiëntmonsters te analyseren en om uitgebreider en breder onderzoek te doen naar het onderscheid tussen echte kanker merkers en gevonden effecten die niet ziektegerelateerd zijn.

Het belang van het genoom breed zoeken naar biomarkers in kanker is aangetoond met DNA microarray studies [3]. DNA microarrays worden in toenemende mate gebruikt als instrumenten voor de diagnose en classificatie van verschillende soorten van kanker [4-6]. Door de huidige expressie profilering zijn we in staat om nanogram hoeveelheden mRNA te detecteren maar dit zegt niet genoeg over het bijbehorende eiwitniveaus in normaal en in maligne weefsels.

De technologie van op oligonucleotiden gebaseerde DNA microarrays is goed ontwikkeld en begrepen, hetgeen nog lang niet het geval is bij een vergelijkbaar eiwit microarray systeem [7].

In dit kader levert de LC-MS benadering aanvullende mogelijkheden ten opzichte van het eiwitchip concept. De werking van massa spectrofotometers is de afgelopen jaren gestaag verbeterd en de huidige moderne LC-MS systemen zijn in staat complexe eiwit/peptide monsters te analyseren en duizenden pieken te genereren. Het valt te betwijfelen of deze analyses veelzeggend zijn (beperking van proteomics methoden wat betreft de concentratie gevoeligheid) maar het is redelijk aanneembaar dat een aanzienlijk deel van het proteoom bestreken kan worden. Het wordt in toenemende mate duidelijk dat een succesvolle integratie van monsterbereiding, eiwit/peptide scheiding en data analyse de sleutel is naar een algemeen succesvolle strategie voor het ontdekken van biomarkers.

Voorbewerking van data en statistische analyse van uitgebreide LC-MS profielen zijn een relatief nieuw onderzoeksgebied. In samenwerking met andere groepen werken we aan verdere verbetering van ons statistisch analytisch platform omdat dit een cruciale stap is in zowel het onderscheiden van monsters als de zoektocht naar potentiële biomarkers.

Van nature is baarmoederhalskanker in het eerste stadium een gelokaliseerde ziekte waardoor het vaak ontsnapt aan een algemene immunorespons die wordt verzorgd door natuurlijke “killer cellen”en T-cellen die in het bloed voorkomen. Door de beperkte uitwisseling tussen eiwitten van de epitheelcellen in de baarmoederhals en bloed, lijkt het vinden van specifieke biomarkers veelbelovend in vergelijkende studies van baarmoederhalskanker en gezond weefsel. Dit werk is geïnitieerd door de combinatie van Laser Dissection Microscopie en massa spectrometrie.

Hoewel een effectief vaccin tegen Humaan Papillomavirus (HPV) het voorkomen van baarmoederhalskanker waarschijnlijk significant zal verminderen, blijft een vroege diagnose nog steeds relevant. Bovendien kan het vinden van biomarkers voor baarmoederhalskanker helpen in andere gevallen waarbij het HPV virus schijnt te zijn betrokken.HPV 16 is bijvoorbeeld gevonden in gevallen van longkanker in Chili [8] en Korea [9]. Het is daarom aannemelijk dat dit virus andere epitheelweefselsoorten dan alleen die van de baarmoederhals kan infecteren.

Ten slotte wil ik benadrukken dat alleen goed geïntegreerde, multidisciplinaire studies potentie hebben om relevante en nieuwe biomarkers in baarmoederhalskanker te ontdekken die zouden kunnen resulteren in een vroegtijdige diagnose en daarmee betere behandeling van patiënten met neoplasia in de baarmoederhals.

References

[1]. Anderson NL, Anderson NG. The Human Plasma Proteome: History, Character, and Diagnostic Prospects. Mol Cell Proteomics 2002;1:845-867.

[2]. Lehman NL, Tibshirani R, Hsu JY, Natkunam Y, Harris BT, West RB, Masek MA, Montgomery K, van de Rijn M, Jackson PK. Oncogenic regulators and substrates of the anaphase promoting complex/cyclosome are frequently overexpressed in malignant tumors. Am J Pathol. 2007;170(5):1793-1805.

[3]. Ewis AA, Zhelev Z, Bakalova R, Fukuoka S, Shinohara Y, Ishikawa M, Baba Y. A history of microarrays in biomedicine. Expert Rev Mol Diagn. 2005;5(3):315-328.

[4]. Miller LD, Liu ET. Expression genomics in breast cancer research: microarrays at the crossroads of biology and medicine. Breast Cancer Res. 2007;9(2):206.

[5]. Henrickson SE, Hartmann EM, Ott G, Rosenwald A. Gene expression profiling in malignant lymphomas. Adv Exp Med Biol. 2007;593:134-46.

[6]. Perez-Diez A, Morgun A, Shulzhenko N. Microarrays for cancer diagnosis and classification. Adv Exp Med Biol. 2007;593:74-85.

[7]. Whiteley GR. Proteomic patterns for cancer diagnosis: promise and challenges. Mol Biosyst. 2006 Aug;2(8):358-363.

[8]. Aguayo F, Castillo A, Koriyama C, Higashi M, Itoh T, Capetillo M, Shuyama K, Corvalan A, Eizuru Y, Akiba S. Human papillomavirus-16 is integrated in lung carcinomas: a study in Chile. Br J Cancer. 2007 Jul 2;97(1):85-91.

[9]. Park MS, Chang YS, Shin JH, Kim DJ, Chung KY, Shin DH, Moon JW, Kang SM, Hahn CH, Kim YS, Chang J, Kim SK, Kim SK. The prevalence of human papillomavirus infection in Korean non-small cell lung cancer patients. Yonsei Med J. 2007 Feb 28;48(1):69-77.

Acknowledgments

First of all, I would like to thank my promoters Prof. Dr. Rainer P.H. Bischoff Prof. Dr. Ate G.J. van der Zee for giving me the opportunity to do this research and for the helpful discussions.

I would also like to thanks the reading committee , Prof. Dr. Sabeth Verpoorte (RuG), Prof. Dr. Christian Huber (Saarland University, Germany), Prof. Dr. R.J. Vonk (UMCG) for the reading, the important comments and for their time.

Special thanks to Dr. A.P. Bruins for his help at the very beginning of my work at Pharmacy and regular conversations. All the people at the MS center and especially Marcel de Vries for being able to answer all my questions concerning mass-spectrometry. It was a great help! Thanks to Dr. Begona Barroso for the introduction to Mass Spectrometry field.

It was my pleaser to collaborate with people from the group of Prof. Dr. Theo Luider: Centre for Biomics, Erasmus University of Rotterdam (Lennard Dekker, A.L.C.T. van Rijswijk, Hans Dalebout) as well as the long term collaboration with Dr. Theo Reijmers (University of Leiden, Leiden, The Netherlands) and the collaboration with the group of Molecular Genetics (RUG), Prof. Dr. Oscar P. Kuipers (especially my friend Anne Hesseling-Meinders for 2D gel analysis). I would also like to say thanks to Klaske A. ten Hoor and Harry G. Klip (Department of Gynecology, University Medical Centre, Groningen) for the always quick response and help.

Thanks to all members of the Analytical Biochemistry (in particular to Jolanda Meindertsma, Therese Rosenling, Nicolas Abello and Ramses Kemperman), Pharmaceutical Analysis groups and our ex-colleague Ineke Keizer Gunnink.

Thank you to Peter Horvatovich for the work we have done together and all the different ways he has helped and assisted me.

I would also like to thank people from the various companies: Rod B. Watson and Tristan Moore (Manager Protein Consumable Business Europe Applied Biosystems), for giving the trial first anti-body column and iTRAQ reagents; Erik Vaessen (Sales Development Biotechnology Sigma-Aldrich Chemie BV) for 20 depletion column, Ed van Dam and Otto Hering (Senior Service Engineer) from GE Healthcare for their support and special thanks to Dr. Jack Wenstrand, Gerard Rozing and Martin Haex (Agilent Technologies).

Thanks to Prof. Dr. A. Brisson and all group of Laboratory of Biophysical Chemistry, University of Groningen for being in my life. Not directly related to this thesis, but it has been really great to work together.

And a lot of thanks to my wonderful students: Suzanne Roelfsema (University of Amsterdam), Inge M. Westra (RUG), Lydia Terborg (University of Münster Institute of Inorganic and Analytical Chemistry, Germany) and

Valentina Moscato (Faculty of Pharmacy University of Calabria,Arcavacata di Rende(CS) Italy, The Leonardo da Vinci project).

Also a very warm thank you to M.Th.E.Bink-Boelkens (MD, PhD) and Dr. T.W. Waterbolk for all they have done.

Finally I want to say thanks to my family for love, support and giving solid backgrounds under my feet.

List of publications
Book chapters

[1]. **Natalia Govorukhina** and Rainer Bischoff: Sample preparation of body fluids for proteomics analysis; In: Proteomics of human body fluids. In: Principles, Methods, and Applications (Visith Thongboonkerd, Ed.), Humana Press (Totowa, New Jersey, USA). (2007);Ch. 2:31-71.

[2]. **Natalia Govorukhina**, Peter Horvatovich and Rainer Bischoff: Label-Free Proteomics of Serum (in press). In: Functional Proteomics, the Humana Press (Totowa, New Jersey, USA).

Publications

[3]. Horvatovich, P., **Govorukhina, N.,** Bischoff, R.: Current Methodological Advances in the Discovery of Protein and Peptide Disease Markers. Boletin de la SECyTA (Bulletin of the Spanish Chromatographic Society). (2005);26:3-11.

[4]. **Govorukhina, N.I.,** Keizer-Gunnink, A., van der Zee, A.G.J.,de Jong, S.. de Bruijn, H.W.A., Bischoff, R. Sample preparation of human serum for the analysis of tumormarkers: a comparision of different approches for albumin and γ-globulin-depletion. J.Chrom. A. (2003); 1009:171-178.

[5]. **Govorukhina, N.I.,** Reijmers, T.H., Nyangoma, S.O., Zee, A.G.J. van der, Jansen, R.C., Bischoff, R., Analysis of human serum by LC-MS: improved sample preparation and data analysis. J. Chromatogr. A. (2006); 110:42-150.

[6]. Horvatovich, P., **Govorukhina, N.,** Bischoff, R., Biomarker Discovery by Proteomics: challenges not only for the analytical chemist. The Analyst. (2006); 131:1193-1196.

[7]. Horvatovich, P., **Govorukhina, N. I.,** Reijmers, T. H., van der Zee, A. G. J., Bischoff, R., Evaluation of HPLC-chip/MS platform for label-free profiling for biomarker discovery. Electrophoresis 2007 (in press).

List of publications not related to the thesis

[1]. Loginova, N.V., **Govorukhina, N.I**., Trotsenko, Yu.A. Metabolism of obligate methylotroph *Methylophilus methanolovorus*. Mikrobiologya (1981); 50:305-310 (in Russian).

[2]. Loginova, N.V., **Govorukhina, N.I**., Trotsenko, Yu.A. Authotrophic methanol metabolism in *Blastobacter viscosus*. Mikrobiologya (1981); 50:591-597 (in Russian).

[3]. Loginova, N.V., **Govorukhina, N.I.**, Trotsenko, Yu.A. Enzymes of ammonium assimilation in bacteria with different assimilation pathways of C1 compounds. Mikrobiologya (1982);51:38-41 (in Russian).

[4]. Loginova, N.V., **Govorukhina, N.I.**, Trotsenko, Yu.A. *Blastobacter aminooxidans* – a new strain of bacterium which growth autotrophically on methylated amines. Mikrobiologya (1982);52:709-715 (in Russian).

[5]. **Govorukhina, N.I.**, Trotsenko, Yu.A. A procedure of isolation of glutamate dehydrogenase. Patent N1017730. 1982.

[6]. Doronina, N.V., **Govorukhina, N.I.**, Trotsenko, Yu.A. Properties of facultative methylotroph *Protaminobacter ruber*. Mikrobiologya (1985); 54:363-369 (in Russian).

[7]. Trotsenko, Yu.A., Dorinina, N.V., **Govorukhina, N.I.** Metabolism of non-motile obligately methylotrophic bacteria. FEMF Microbiology Letters (1986); 33:293-297.

[8]. **Govorukhina, N.I.**, Kletsova, L.V., Tsygankov, Yu.D., Trotsenko, Yu.A., Netrusov, A.I. Characteristics of a new obligate methylotroph. Mikrobiologya (1987);56:849-855 (in Russian).

[9]. Kletsova, L.V., **Govorukhina, N.I.**, Tsygankov, Yu.D., Trotsenko, Yu.A. Metabolism of an obligate methylotroph Methylobacterium flagellatum. Mikrobiologya (1987); 56:901-906 (in Russian).

[10]. **Govorukhina, N.I.**, Doronina, N.V., Trotsenko, Yu.A. Effect of multi-carbon compounds on growth and metabolism of obligate methylotroph *Methylophilus methanolovorus*. Mikrobiologya (1987); 56:564-569 (in Russian).

[11]. Doronina, N.V., **Govorukhina, N.I.** Current state of a taxonomy of methylotrophic bacteria (Yu.A. Trotsenko, ed.). 1987;85-94. Puschino (in Russian).

[12]. Vedenina, I.Ja., **Govorukhina, N.I.** Formation of methylotrophic denitrifying population in wastewater refinery from nitrates. Mikrobiologya (1988); 57:320-328 (in Russian).

[13]. Doronina, N.V., **Govorukhina, N.I.**, Lysenko, A.M., Trotsenko, Yu.A. Analysis of DNA-DNA homologies in obligate methylotrophic bacteria. Mikrobiologya (1988); 57:629-633 (in Russian).

[14]. Doronina, N.V., **Govorukhina, N.I.**, Trotsenko, Yu.A. New strains of restricted facultative methylotrophic bacteria. Mikrobiologya (1988); 57:828-834 (in Russian).

[15]. Sokolov, A.P., **Govorukhina, N.I.**, Trotsenko, Yu.A. Characterization of methanol dehydrogenases of restricted facultative and obligate methylotrophs. Biokhimia (1989); 54:811-816 (in Russian).

[16]. **Govorukhina, N.I.**, Doronina, N.V., Andreev, L.V., Trotsenko, Yu.A. *Methylomicrobium*, a new genus of facultative methylotrophic bacteria. Mikrobiologya (1989):58, 326-333 (in Russian).

[17]. Bulygina, E.S., Galchenko, V.F., **Govorukhina, N.I.**, Netrusov, A.I., Nikitin, D.I., Romanovskaya, V.A., Trotsenko, Yu,A., Chumakov, K.M. Classification of methylotrophic bacteria by sequence analysis of ribosomal 5S RNA. Molek.Genet.Mikrobiol.Virusol. (1989); 4:18-24 (in Russian).

[18]. Sysoev, O.V., **Govorukhina, N.I.**, Trotsenko, Yu.A. Role of glutathione in methylotrophic bacteria with different pathways of C1-assimilation. Mikrobiologya (1989); 58:549-553 (in Russian).

[19]. **Govorukhina, N.I.**, Trotsenko, Yu.A. Phospholipid composition of methylotrophic bacteria. Mikrobiologya (1989); 58:405-411(in Russian).

[20]. Skladnev, D.A., Tzygankov, Yu.D., Kuznetzov, E.V., Bayev, M.B., **Govorukhina, N.I.**, Trotsenko, Yu.A. Effect of various organic compounds on growth of methylotrophic bacterium Methylobacterium MB1. Biotekhnologya (1989); 5:559-564 (in Russian).

[21]. **Govorukhina, N.I.**, Trotsenko, Yu.A. Quantification of poly-β-hydroxybutirate in methylotrophic bacteria with different pathways of methanol assimilation. Prykladnaya Biokhimia i Mikrobiologya (1991); 27(1):98-101 (in Russian).

[22]. **Govorukhina, N.I.**, Chetina, E.V., Trotsenko, Yu.A. Effect of multi-carbon compounds on growth and metabolism of *Methylobacillus flagellatum*. Mikrobiologya (1990); 9:249-256.

[23]. **Govorukhina, N.I.**, Trotsenko, Yu.A. *Methylovorus* - a new genus of restricted facultative methylotrophic bacteria. Mikrobiologya (1990); 59:880-890 (in Russian).

[24]. Sysoev, O.V., **Govorukhina, N.I.**, Gruzman, M.B. Glutathion-S-transferase of methylotrophic bacteria: distribution and characterization. Prykladnaya Biokhimia i Mikrobiologya (1990); 26:456-461 (in Russian).

[25]. Arfman, N., Bystrykh, L.V., **Govorukhina, N.I.**, Dijkhuizen, L. 3-hexulose-6-phosphate synthase from the thermotolerant methylotroph Bacillus sp. C1. Methods in Enzymology, Hydrocarbons and methylotrophy (ed. Lidstrom, M.E.), (1990);188, 391-397.

[26]. Bulygina, E.S., Galchenko, V.F., **Govorukhina, N.I.**, Netrusov, A.I., Nikitin, D.I., Trotsenko, Yu,A., Chumakov, K.M. Taxonomic studies on methylotrophic bacteria by 5S ribosomal RNA sequencing. J.Gen.Microbiol. (1990); 136:441-446.

[27]. **Govorukhina, N.I.**, Trotsenko, Yu.A. Methylovorus, a new genus of restricted facultatively methylotrophic bacteria. Int. J. System. Bacteriol. (1991);41:158-162.

[28]. Bulygina, E.S., **Govorukhina, N.I.**, Netrusov, A.I., Trotsenko, Y.A., Chumakov, K.N. Comparative studies on 5S RNA sequences and DNA-DNA hybridization of obligately and restricted facultatively methylotrophic bacteria. System.Appl. Microbiol. (1993); 16:85-91.

[29]. Bayev, M.B., Kuznetzov, E.V., Skladnev, D.A., **Govorukhina, N.I.**, Sterkin, V.E., Tzygankov, Yu.D. Growth and enzymological characteristics of a pink-pigmented facultative methylotroph *Methylobacterium sp*. MB1. Folia Microbiol. (1992); 37:93-101.

[30]. Arfman, N., Dijkhuizen, L., Kirchhof, G., Ludwig, W., Schleifer, K.-H., Bulygina, E.S., Chumakov, K.M., **Govorukhina, N.I.**, Trotsenko, Yu.A., White, D., Sharp, R.J. *Bacillus methanolicus sp. nov*., a new species of thermotolerant, methanol-utilizing endospore-forming bacteria. Int.J.System.Bacteriol. (1992); 42:439-445.

[31]. Bystrykh, L.V., **Govorukhina, N.I.**, van Ophem, P.W., Hektor, H.J., Dijkhuizen, L, Duine, J.A. 1993, Formaldehyde dismutase activities in Gram-positive bacteria oxidizing methanol. J. Gen. Microbiol. (1993); 139:1979-1985.

[32]. Bystrykh, L.V., Vonck, J., van Bruggen, E.F.J., van Beeumen, J., Samyn, B., **Govorukhina, N.I.**, Arfman, N., Duine, J.A., Dijkhuizen, L. Electron microscopic analysis and structural characterization of novel NADP(H)-containing methanol: NDMA oxidoreductases from the Gram-positive methylotrophic bacteria *Amycolatopsis methanolica* and *Mycobacterium gastri* MB19. J.Bacteriol. (1993); 175(6):1814-1822.

[33]. Arfman, N., Hektor, H.J., Bystrykh, L.V., **Govorukhina, N.I.**, Dijkhuizen, L., Frank, J. Properties of an NAD(H) -containing methanol dehydrogenase and its activator protein from *Bacillus methanolicus*. Eur.J.Biochem. (1997); 244:426-433.

[34]. Bystrykh, L.V., **Govorukhina, N.I.**, Dijkhuizen, L., Duine, J.A. Tetrazolium- dye-linked alcohol dehydrogenase of the methylotrophic actinomycete *Amycolatopsis methanolica* is a three-component complex. Eur.J.Biochem. (1997); 247:280-287.

[35]. Reviakine, I., Bergsma-Schutter, W., Mazères-Dubut, Ch., **Govorukhina, N.**, Brisson, A. Surface topography of the p3 and p6 Annexin V crystal forms determined by atomic force microscopy. J.Struct.Biol. (2000); 131:234-239.

[36]. Oling, F., Sopkova-de Oliviera Santos, J., **Govorukhina, N.**, Mazères-Dubut, Bergsma-Schutter, W., Oostergetel, G., Keegstra, W., Lambert, O., Lewit-Bentley, A., Brisson, A. Structure of membrane-bound Annexin A5 trimers: a hybrid cryo-EM – X-ray crystallography study. J.Mol.Biol. (2000); 304:561-573.

[37]. **Govorukhina, N.**, Bergsma-Schutter, W., Mazeres-Dubut, Ch., Mazeres, S., Drakopoulu, E., Bystrykh, L., Oling, F., Reviakine, I., Mukhopadhyay, A., Brisson, A. Conditions of self-assembly of annexin A5 in membrane-bound trimers and 2D arrays of trimers. In "Annexins: Biological importans and annexin-related patologies. Landers Bioscience, Georgetown USA, Ed. J. Bandorowicz-Pikula (2003); 61-79.