Published in final edited form as:

J Proteomics. 2011 December 10; 75(1): 202–210. doi:10.1016/j.jprot.2011.06.008.

Bioinformatic Strategies for Unambiguous Identification of Prostate Specific Antigen in Clinical Samples

Ákos Végvári¹, Melinda Rezeli¹, Jari Häkkinen², Carina Sihlbom³, Elisabet Carlsohn³, Johan Malm⁴, Hans Lilja⁵, Thomas Laurell^{1,6}, and György Marko-Varga¹

¹Clinical Protein Science & Imaging, Dept. of Measurement Technology and Industrial Electrical Engineering, Lund University, BMC C13, SE-221 84 Lund, Sweden

²Dept. of Oncology, Clinical Sciences, Lund University and Skåne University Hospital, Barngatan 2B, SE-221 85 Lund, Sweden

³Proteomics Core Facility, Göteborg University, Medicinaregatan 7, SE-413 90 Göteborg, Sweden

⁴Dept. of Laboratory Medicine, Section for Clinical Chemistry, Lund University, Skåne University Hospital in Malmö, SE-205 02 Malmö, Sweden

⁵Depts. of Clinical Laboratories, Surgery (Urology), and Medicine (GU-Oncology), Memorial Sloan-Kettering Cancer Center, New York, New York 10021, USA

⁶Dept. Biomedical Engineering, Dongguk University, Seoul, Korea

Abstract

Prostate specific antigen (PSA), as a widely used clinical biomarker in prostate cancer diagnostics, exists in multiple molecular forms. However, all of these forms might not be recognized in a given sample by the standard immunoassays. Therefore, we have investigated PSA isoforms separated by size using mass spectrometric analyses. The objective of these developments was to identify and specify the various forms of PSA. To optimize successful identification of different PSA forms, we have developed a bioinformatic strategy, consisting of high resolution MALDI-MS PMF and sequencing MS/MS data searches. To improve sequence-based identification, the recently introduced Proteios software environment was employed, allowing the combination of multiple database search engines in an automated manner.

We could unambiguously identify PSA in clinical samples by all detectable tryptic peptides, which were found to be common in several isoforms.

© 2011 Elsevier B.V. All rights reserved.

Corresponding author: Ákos Végvári Clinical Protein Science & Imaging, Dept. of Measurement Technology and Industrial Electrical Engineering, Lund University, BMC C13, SE-221 84 Lund, Sweden Phone: +46-46-222 3721 Fax: +46-46-222 4527 akos.vegvári@elmat.lth.se.

Conflict of interest

Dr. Hans Lilja holds patents for free PSA and kallikrein-2 (hK2) assays. All other authors declare no conflict of interest.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Keywords

prostate specific antigen isoforms; MALDI LTQ Orbitrap XL; ESI-LTQ FT-ICR; Proteios software environment; kallikrein-2

1. Introduction

Prostate cancer is commonly diagnosed by highly specific immunoassays, based on the detection of prostate specific antigen (PSA) as the clinical biomarker of disease. Today, PSA is the most widely used tumor marker despite a rather low specificity. Our previous studies have identified a number of PSA molecular forms that indicate complex disease biology. These PSA isoforms could play a significant role in the overall diagnostic readout of patients with prostate diseases [1].

There is an increasing knowledge and reporting on the metabolism and diversity of PSA and other kallikrein family members. PSA is a glycoprotein, containing approx. 8% carbohydrate [2]. The carbohydrate structure is thought to be a biantennary *N*-linked oligosaccharide of the *N*-acetylglucosamine type with a sialic acid group at the end of the two branches [3]. PSA in seminal fluid is heterogeneous mainly because of a variable degree of sialylation [4]. It has been shown that glycan structures on PSA differ greatly not only in seminal fluid but also in plasma of cancer patients [5]. There are also a number of scientific reports that describes both sequence variants [6], glyco-modified forms [7–9], as well as the variation in enzyme activity [4, 10–12].

Mass spectrometry has been intensively employed for proteomics characterization of clinical samples, including qualitative analysis of PSA in human seminal fluid [13] and blood serum [14]. Absolute quantitative analysis of PSA was implemented by using isotope dilution mass spectrometry that achieved a LOD of 0.04 and 0.16 $\mu\text{mol/L}$ in buffer and serum, respectively [15]. In recognition of unmet needs to improve prostate cancer diagnosis a combination of 2-D electrophoresis and MS analysis was proposed in a study of PSA subforms [16]. However, there is no systematic study on the relationship between PSA forms and clinical test readouts published to the best of our knowledge.

Based on our previous observation that PSA exists in more molecular forms [1] than the currently used dissociation-enhanced lanthanide fluorescent immunoassay (DELFI) test of prostate cancer [17] can recognize, we have developed a proteomic strategy to monitor PSA isoforms. To verify the various molecular forms of PSA, possibly leading to novel biological functions, we utilized a combination of several mass spectrometry platforms, and bioinformatic tools and principles, to provide evidence for the characterization of PSA isoforms in clinical samples.

2. Material and Methods

2.1. Sample preparation

Seminal plasma samples were prepared by centrifugation from seminal fluid obtained from young men ($n=34$) as Figure 1 illustrates. The proteins of seminal plasma were separated by both SDS-PAGE and zymogram gel electrophoresis. The major difference in separation conditions was that the samples were boiled, reduced and alkylated before loading gels in the SDS-PAGE gels, while they were run under reducing condition in zymogram gels containing 0.1% gelatin without boiling. Enzymes retrieved their activity upon removal of the detergent following completion of protein separation in zymogram gels. Staining was performed with Coomassie blue R350 (GE Healthcare, Uppsala, Sweden) for 1 h according

to the manufacturer's recommendations. Protein bands within the molecular size range of 20–40 kDa were cut and digested with sequencing grade trypsin (Promega, Madison, WI) before subjected to mass spectrometric analysis (Figure 1).

2.2. Mass spectrometry

The analysis of trypsin digested gel pieces was performed on a MALDI LTQ Orbitrap XL instrument (Thermo Scientific, Bremen, Germany). Full mass spectra were obtained in positive mode within a mass range of 600–4,000 Da using the FT mass analyzer (Orbitrap) at 60,000 resolution. Twenty FT mass scans (2 microscans/scan) were collected from each spot, while the laser (operated at 10 μ J) was moved automatically after 2 microscans. Spectra were processed by Xcalibur software v 2.5.5, generating lists of extracted monoisotopic masses from averaged spectra. All analyses were performed in triplicates.

Sequence data collection MALDI-MS/MS was performed on the top 25 peaks, selected from a list of masses derived by theoretical (*in silico*) digestion of 35 PSA variants by Peptide-Mass (available at the ExPASy Proteomics Server web site, <http://expasy.org/sprot/>), choosing iodoacetamide as alkylation agent, oxidation on methionine and allowing 1 miss-cleavage. The minimal signal required for MS/MS data acquisition was 500 counts. Normalized collision energy was 50% during an activation time of 30 ms and activation Q of 0.250.

NanoLC-MS/MS was performed on a fused silica column (20 cm x 50 μ m i.d.) packed in-house with ReproSil-Pur C18-AQ 3 μ m porous particles (Dr. Maisch GmbH) and connected to a hybrid linear ion trap-Fourier transform ion cyclotron mass spectrometer (LTQ-FTICR-MS, Thermo Electron), equipped with a 7 T magnet. Sample injections (2 μ l) were performed with an HTC-PAL auto sampler (CTC Analytics AG), and the tryptic peptides were trapped on a C18-precursor column (4.5 cm x 100 μ m id) and separated in the analytical column. Following 6 min isocratic run at 5% ACN/0.2% formic acid, a linear gradient to 50% ACN in 34 min was started at constant flow rate of 200 nL/min (Agilent 1100 binary pump), while the eluent was electrosprayed (+1.4 kV) from the emitter tip (id 20 μ m, od 150 μ m; Polymicro Technologies). The mass spectrometer was operated in the data dependent mode to automatically switch between MS and MS/MS acquisitions. Survey MS spectra (m/z 400–2000) were acquired with the FT-ICR, and the five most abundant doubly, triply or quadruply protonated ions in each FT-scan were selected for isolation, fragmentation and detection in the linear ion trap. The exclusion time for isobaric precursor ions was set to 20 s. The same inclusion list of known PSA peptides, used in MALDI-MS/MS analysis, was combined with an exclusion list with possible peptides from co-eluting proteins in the gel. The minimal signal required for MS/MS data acquisition was set to 1000 counts. Normalized collision energy was 25% during an activation time of 30 ms and activation Q of 0.250.

2.3. Data analysis

Peptide mass fingerprint search of extracted mass lists was performed on the Mascot server (<http://www.matrixscience.com>), using the NCBI database, limited to Homo sapiens (release 20091024: 224,815 out of 9,937,670 sequences). Side chain modifications were set to fixed carbamidomethylation at cysteine and variable oxidation at methionine residues, respectively. Peptide tolerance was kept at 10 ppm. The samples were considered positive PSA identities if at least 2 specific peptide were identified in at least two of the triplicates.

The MS/MS search was performed on the Proteios Software Environment (ProSE) [18]. ProSE has built-in support to interface several protein identification engines such as Mascot and X!Tandem, and combines search results from multiple search engines [19]. Three

different search engines were used: Mascot, X!Tandem with native scoring and X!Tandem with k-score scoring. All searches were performed with the IPI Human database ver3.71 containing 86,745 protein sequences [20]. The search settings included side chain modifications with fixed carbamidomethylation at cysteine and variable oxidation at methionine residues. Precursor tolerance was 10 ppm and the fragment mass tolerance was 0.5 Da. Proteins were considered correctly identified when at least two different peptides were present with a false discovery rate (FDR) less than 0.01 [19].

Measured peptide masses by FT-ICR and their CID spectra were analyzed using Proteome Discoverer version 1.1 (Thermo Fisher Scientific), which incorporates Mascot database search software version 2.2.0 (Matrix Science). The Swiss-Prot database limited to *Homo sapiens* (from March 2010) was used for protein identification. Database interrogation was carried out using monoisotopic masses at 10 ppm and 0.5 Da precursor-ion and fragment-ion mass tolerance window, respectively. Carbamidomethylation of cysteine and oxidation of methionine residues were used as static and variable modifications, respectively. The maximum number of allowed missed cleavages was set to two.

3. Results and Discussion

We have analyzed 34 biological samples (seminal plasma prepared from semen of young men) by MALDI-Orbitrap MS and MALDI-MS/MS, resulting in 417 tryptic digests (analyzed in triplicates), resulting in a total of 2502 mass spectrometric analyses. Additionally, 14 of the seminal plasma samples were run in a gel-based activity assay from where 84 tryptic digests were analyzed by LC-MS/MS.

The high number of sample outputs from the study and the extensive data generated, put requirements on a sound bioinformatic strategy in regards to software preferences, search engine outputs as well as the algorithms. The strategy should also be compatible with the data originating from the different MS platforms used.

3.1. Outline of mass spectrometric identification strategy

The molecular forms of PSA were identified by high resolving MALDI- and FT-ICR-MS, as well as peptide sequencing utilizing the linear ion trap of the hybrid mass spectrometers (see Figure 2). For unambiguous identification of PSA, firstly the accurate masses, obtained by the Orbitrap mass analyzer, were utilized in a PMF database search. Secondly, our MS/MS sequencing data collection strategy was based on *in silico* predictions of tryptic peptides of 35 PSA sequences from protein databases and applied as inclusion mass list for precursor ion selection. Thirdly, in order to link the protein expression to the respective functionality, nanoLC-MS/MS runs were performed on enzymatically active gel bands utilizing the high resolving power of the FT-ICR mass spectrometer for precursor ion selection. In order to improve the fragmentation efficiency of PSA peptides in the presence of gelatin, an exclusion mass list of tryptic background signals was also employed. The inclusion of this step was essential to be able to assign identities to a highly complex matrix background of these samples.

3.2. Outline of bioinformatic strategy

The database search was tailored to achieve optimal identification rate with high confidence. As Figure 2 illustrates, peptide mass fingerprint and sequencing database search was performed on MALDI FT-MS and IT-MS/MS scans, respectively.

Furthermore, we present for the first time clinical sequence data whereby the MALDI-MS/MS platform is interfaced with the ProSE that provides a combined readout from three different search algorithms (Mascot, X!Tandem with native and k-scores). The ProSE

bioinformatic system gives an automated data processing ability with these search engine queries, delivering high speed repository build of large protein sequence data sets at high speed.

In the third part, an automated informatic processing was employed on the high resolution mass spectrometric data, where the precursor ions were selected in a data dependent mode, followed by automated database search.

As a final result, the PMF and sequence identifications were combined at the protein level efficiently annotating PSA isoforms in the clinical samples. Further structural data was collected by identifying enzymatically active forms of PSA.

3.3. High resolution MALDI FT-MS (Orbitrap) and peptide mass fingerprint

For the investigation of expressed PSA profiles, the trypsinated samples were analyzed by collecting both high resolution MALDI MS full scans, as well as MALDI MS/MS sequencing data. These data were generated by sequential analysis in two separate runs, with resulting outcomes in two types of raw data file sets. Firstly, the full mass scans specified typically 100–300 monoisotopic signals with a total output of as high as 218,265 extracted monoisotopic masses. These data were used directly as input queries for database searches without any additional filtering.

One of the objectives of this study was to identify any isoform speciation of PSA. For this reason we selected the NCBI database in combination with the Mascot search engine. The advantage with choosing the NCBI database is the high number of annotated proteins. The ever-increasing new variants and sequence information from scientists around the world are being presented within updated versions of NCBI on regular and frequent basis. This was also true for the molecular variants of PSA in relation to other databases such as Swiss-Prot, MSDB and IPI that all are interfaced with Mascot.

PSA was identified in 248 runs (out of 1251) by using the PMF function of Mascot, thereby fulfilling our search criteria (see Materials and Methods). An additional number of 3 single PSA identifications were excluded based on one of the criteria defined as at least 2 IDs in triplicate runs were required for a positive annotation.

Totally, 16 tryptic peptides (m/z values) were associated with a PSA identity, as given in Table 1. More than half of the peptides were identified with at least 60% frequency rate. All possible tryptic peptide sequences of the major PSA form (UniProt ID: P07288) could be recognized with masses within the detection range (600–4000 Da) at least in one sample. We have also observed that three methionine-containing peptides could be identified by two monoisotopic masses, as well as oxidized and non-oxidized form (see Table 1). Furthermore, only one single peptide with a missed cleavage site was identified at a low frequency of 4% as shown in Table 1.

As a prerequisite for reliable PMF identification, the mass accuracy of the Orbitrap was outstanding with an error typically kept below 3 ppm as presented in Figure 3. The data in Figure 3 also shows the characteristic error distribution of the Orbitrap along the mass range. The error values were typically positive between 1000 and 3000 Da and negative below and above this range (depicted in Figure 3). The mean error of all of the detected PSA peptides was calculated and given in Table 1.

In addition, the PMF analysis provides an interesting observation. If a peptide was found in only one of the replicates, the error value was higher than the average, *i.e.*, above 3 ppm in most cases. This finding strongly supports our identification criterion that requires an

identity of ≥ 2 in series of triplicate analysis runs. As a resulting outcome, this observation could also be proposed as a filtering parameter, in order to exclude those peptides with error rates higher than 3 ppm.

3.4. Protein annotations by PMF

The data output from our Orbitrap analyzer provided results where PSA was presented as the first significant protein candidate in the Mascot searches by PMF alternative annotations with 8 different molecular variants (see Table 2). For a positive protein identity, a MOWSE score of 66 was required using the NCBI protein database, in comparison to Swiss-Prot that requires a scoring of 62. The experience from large scale query is that the scoring was found to be 35% of all PSA annotations in the 248 runs fulfilling the NCBI significance level, as shown in Figure 4A.

As outlined in the histogram representation (Figure 4B), the sequence coverage was found to be at least 10%, whereas the mean value was found to be 32%. In samples with strong protein bands in gel, the sequence coverage reached as high as 79%, which equals to 100% of measurable peptides of PSA.

The Mascot search algorithm ranks hits by the MOWSE score that relates to the number of peptides associated with a given protein and also to the length of the protein sequence/sequence coverage. Consequently, in cases when a low peptide number was associated with a given protein, we found that the PSA isoform ranked as first hit was the one with the shortest sequence. The search engine used this preference in order to generate the highest protein score. These are explanations that form the basis of the 8 identified PSA isoforms presented in Table 2. Throughout all samples, we were unable to identify unique, isoform-specific peptides, which leaves the gene-coded predictions open for interpretations.

3.5. MS/MS sequencing and Proteios Software Environment

PSA expression was also investigated using sequencing data generated by collision induced dissociation (CID) fragmentation in the linear ion trap. The ion trap has high sensitivity and is suitable for MS/MS spectra collection, despite that the resolution of this mass analyzer is significantly lower than that of the Orbitrap. In addition, ProSE is known to possess improved performing characteristics with CID than higher-energy C-trap dissociation (HCD) fragmentation data [18]. ProSE has a built-in support function for several proteomics workflows (see Figure 2). The three search engines (Mascot, X!Tandem with native and k-scores), being integrated, automatically combined the annotated sequence results with the accompanying FDR. The prerequisite of this function is an identical database used for all search engines, including an identical decoy part for all scoring algorithms that enables an estimation of hits in the decoy part with more than one search engine.

The triplicate analytical runs of the 417 protein digests resulted in 434 data files with MS/MS spectra, which were then used as input for the ProSE searches. The standard database hits were found to be 2332 and 1847 after filtering with FRD 0.01; whereas the random (decoy) database hits were 570 before and 18 after filtering.

By 5 fragmentation experiments with MS/MS spectra in each analysis, throughout the 197 runs we could clearly detect PSA sequences, which originated from 65 gel bands. Thirteen of the PSA peptides were sequenced and associated with the samples as listed in Table 3. Interestingly, only one of the tryptic peptides (FLRPGDDSSHDLMLLR), containing methionine, was identified in both the oxidized as well as the non-oxidized form. We also identified two methionine-containing PSA peptides as their respective oxidized (HSLFHPEDTGQVFQVSHSFPHPLYDM*SLLK) or non-oxidized (FMLCAGR) form only.

The combination of several search engines in general leads to a significant increase in the number of identified peptides, compared to a single search engine, given the FDR is the same. Comparing the results of these databases, Mascot generated the highest number of PSA sequences, almost double as much as any of the X!Tandem algorithms (see Figure 5). Furthermore, a large portion of the X!Tandem database search results was not overlapping, favoring the k-scoring algorithm against the native score, which were found to be able to annotate 1.25-fold more spectra.

3.6. Protein identification by ProSE

Due to the fact that MS/MS data acquisition was based on an inclusion mass list of all possible PSA sequences listed in databases, more than 99% of the MS/MS spectra identified peptide sequences pointing to PSA. ProSE has identified PSA by 7 IPI entries that were associated with 10 PSA sequences in the UniProt database (Table 4). Interestingly, the most frequent PSA forms were not reviewed in UniProt KB. The UniProt ID: Q15098 (APS protein with 234 amino acids) and the UniProt ID: Q8IXI4 (PSA isoform CRA_h with 227 amino acids) are both shorter than the UniProt ID: P07288 confirmed PSA sequence (261 amino acids) and identified at transcript level only (see Figure 6 for PSA sequences). Interestingly, this latter P07288 form contains both the signaling peptide and the propeptide sequences, which are missing in the other two forms and characteristic of the active PSA. Additional isoforms were found at high frequency rates (Table 4) but identified by the same peptides that could be found in the other PSA sequences (*i.e.*, isoform specific, unique peptides were not detected).

Furthermore, another important serine protease, kallikrein-2 (hK2), was also clearly identified, by a peptide sequence (IVGGWECEK) common with PSA, in 4 different IPI entries (6 UniProt accession codes as shown in Table 4). Interestingly, only the Mascot scoring algorithm could identify kallikrein-2 by another sequence (VVHYR), common to both PSA and hK2. All these indications that hK2 could be identified in clinical samples at expression level must be confirmed experimentally where other tryptic peptide of hK2 are sequenced as well.

3.7. Comparison of PMF and sequencing identifications

In a comparative analysis, we found that PSA was recognized in 35 gel bands by PMF with significant MOWSE scores, whereas sequence based identification could clearly associate PSA in 65 gel bands. The overlap between these two annotation strategies was higher than 91%, as only one protein band was identified by PMF without sequencing data. Based on our experiences with these large number of PSA spectral data, we were able to accept those protein identifications by PMF with scores lower than 66, when sequencing data supported the presence of PSA in the sample. This calculation resulted in the final read out value of total 65 unambiguous PSA identifications combining the PMF and ProSE sequencing outputs, of which 61 were supported by both approaches.

3.8. Enzymatic activity monitored by FT-ICR

In order to obtain additional structural information about PSA isoforms in clinical samples, the enzymatic activity was monitored. In these experiments, the selection of precursor masses by FT-ICR during LC-MS/MS data collection provided a novel source of protein functionality, which is then linked to structural information reporting on the intactness of the active site in PSA molecules.

Positive identification of enzymatically active PSA forms was accomplished in 25 protein bands by the peptide (LSEPAELTDAVK). Due to the high chemical background and the otherwise strict criteria applied on database search by Mascot, these identifications were

judged satisfactory. In all cases, a single PSA form was annotated, pointing to the major isoform with entire protein sequence (UniProt ID: P07288 with 261 amino acids). As Figure 6 presents, only 4 out of the 16 sequences of identified PSA isoforms have intact catalytic triad (H65, D120, S213).

4. Concluding Remarks

We have developed a bioinformatic strategy with multiple database searches in order to investigate the microheterogeneity of PSA in clinical samples. Obviously, the currently available databases are not prepared to select among molecular forms and sequence variants of proteins. Therefore, the combination of database search results based on both high resolution MS and MS/MS has great significance. Our results suggest that PMF based on high resolution MALDI-MS analyses is a powerful identification tool due to the excellent mass accuracy of the Orbitrap. On the other hand, sequencing MS/MS data should be superior when the Proteios software was used allowing a combination of database search algorithms. Since the isoform profiles of PSA in clinical samples may have importance in elucidation of diseased linked biomarkers, the presented PSA isoform data as well as the combined multiplex MS platform and database strategy might be a useful tool in future studies to decipher a possible diagnostic value of different PSA isoforms.

Acknowledgments

We would like to thank Kerstin Strupat, Martin Hornshaw and Egon Rosén, at Thermo Fisher Scientific, for mass spectrometry support. The authors are grateful for funding support from the Swedish Research Council, Vinnova and Foundation for Strategic Research – The Programme: Biomedical Engineering for Better Health grant no: 2006-7600 and grant no: K2009-54X-20095-04-3, Swedish Cancer Society (08-0345), Knut and Alice Wallenberg Foundation, Crafoord Foundation, Carl Trygger Foundation, Fundación Federico SA, Royal Physiographic Society, Sten Lexner Foundation, Sidney Kimmel Center for Prostate and Urologic Cancers, National Cancer Institute Specialized Programs of Research Excellence (P50-CA92629) and David H. Koch through the Prostate Cancer Foundation.

Non-standard abbreviations

PSA	prostate specific antigen
FDR	false discovery rate
DELFI	dissociation-enhanced lanthanide fluorescent immunoassay
ProSE	Proteios software environment
HCD	higher-energy C-trap dissociation
hK2	human kallikrein-2

References

1. Végvári Á, Rezeli M, Welinder C, Malm J, Lilja H, Marko-Varga G, et al. Identification of Prostate Specific Antigen (PSA) Isoforms in Complex Biological Samples Utilizing Complementary Platforms. *J Proteomics*. 2010; 73:1137–47. [PubMed: 20102753]
2. Armbruster DA. Prostate-specific antigen: biochemistry, analytical methods, and clinical application. *Clin Chem*. 1993; 39:181–95. [PubMed: 7679336]
3. Bélanger A, van Halbeek H, Graves HCB, Grandbois K, Stamey TA, Huang L, et al. Molecular mass and carbohydrate structure of prostate specific antigen: studies for establishment of an international PSA standard. *Prostate*. 1995; 27:187–97. [PubMed: 7479385]
4. Zhang W-M, Leinonen J, Kalkkinen N, Dowell B, Stenman U-H. Purification and characterization of different molecular forms of prostate-specific antigen in human seminal fluid. *Clin Chem*. 1995; 41:1567–73. [PubMed: 7586544]

5. Tabares G, Radcliffe CM, Barrabes S, Ramirez M, Aleixandre RN, Hoesel W, et al. Different glycan structures in prostate-specific antigen from prostate cancer sera in relation to seminal plasma PSA. *Glycobiology*. 2006; 16:132–45. [PubMed: 16177264]
6. Hilz H, Noldus J, Hammerer P, Buck F, Lück M, Huland H. Molecular heterogeneity of free PSA in sera of patients with benign and malignant prostate tumors. *Eur Urol*. 1999; 36:286–92. [PubMed: 10473986]
7. Isono T, Tanaka T, Kageyama S, Yoshiki T. Structural Diversity of Cancer-related and Non-Cancer-related Prostate-specific Antigen. *Clin Chem*. 2002; 48:2187–94. [PubMed: 12446475]
8. Okada T, Sato Y, Kobayashi N, Sumida K, Satomura S, Matsuura S, et al. Structural characteristics of the N-glycans of two isoforms of prostate-specific antigens purified from human seminal fluid. *Biochim Biophys Acta*. 2001; 1525:149–60. [PubMed: 11342264]
9. Prakash S, Robbins PW. Glycotyping of prostate specific antigen. *Glycobiology*. 2000; 10:173–6. [PubMed: 10642608]
10. Mikolajczyk SD, Grauer LS, Millar LS, Hill TM, Kumar A, Rittenhouse HG, et al. A precursor form of PSA (pPSA) is a component of the free PSA in prostate cancer serum. *Urology*. 1997; 50:710–4. [PubMed: 9372880]
11. Mikolajczyk SD, Millar LS, Wang TJ, Rittenhouse HG, Wolfert RL, Marks LS, et al. "BPSA," a specific molecular form of free prostate-specific antigen, is found predominantly in the transition zone of patients with nodular benign prostatic hyperplasia. *Urology*. 2000; 55:41–5. [PubMed: 10654892]
12. Noldus J, Chen Z, Stamey TA. Isolation and characterization of free form prostate specific antigen (f-PSA) in sera of men with prostate cancer. *J Urol*. 1997; 158:1606–9. [PubMed: 9302183]
13. Fung KYC, Glode LM, Green S, Duncan MW. A comprehensive characterization of the peptide and protein constituents of human seminal fluid. *Prostate*. 2004; 61:171–81. [PubMed: 15305340]
14. 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 Research*. 2001; 61:957–62. [PubMed: 11221890]
15. Barnidge DR, Goodmanson MK, Klee GG, Muddiman DC. Absolute quantification of the model biomarker prostate-specific antigen in serum by LC-MS/MS using protein cleavage and isotope dilution mass spectrometry. *J Proteome Res*. 2004; 3:644–52. [PubMed: 15253448]
16. Jung K, Reiche J, Boehme A, Stephan C, Loening SA, Schnorr D, et al. Analysis of subforms of free prostate-specific antigen in serum by two-dimensional gel electrophoresis: Potential to improve diagnosis of prostate cancer. *Clin Chem*. 2004; 50:2292–301. [PubMed: 15472031]
17. Dickson EF, Pollak A, Diamandis EP. Ultrasensitive Bioanalytical Assays Using Time-Resolved Fluorescence Detection. *Pharmacol Therapeut*. 1995; 66:207–35.
18. Häkkinen J, Vincic G, Månsson O, Wårell K, Levander F. The Proteios Software Environment: An Extensible Multiuser Platform for Management and Analysis of Proteomics Data. *J Proteome Res*. 2009; 8:3037–43. [PubMed: 19354269]
19. Levander F, Krogh M, Wårell K, Gärdén P, James P, Häkkinen J. Automated reporting from gel-based proteomics experiments using the open source Proteios database application. *Proteomics*. 2007; 7:668–74. [PubMed: 17295359]
20. Kersey PJ, Duarte J, Williams A, Karavidopoulou Y, Birney E, Apweiler R. The International Protein Index: An integrated database for proteomics experiments. *Proteomics*. 2004; 4:1985–8. [PubMed: 15221759]

Figure 1. Illustration of the sample preparation workflow used for the identification of PSA isoforms in seminal fluid.

Figure 2.

Overview of the bioinformatics strategy applied on the unambiguous identification of PSA in clinical samples.

Figure 3. Error distribution of the Orbitrap mass analyzer: represented by a single full mass analysis with high number of PSA peptides (for the mean error of each peptide identified at various frequency rates, see Table 1).

Figure 4A

Figure 4B

Figure 4. (A) Distribution of the MOWSE scores in the Mascot protein identification result and (B) distribution of the sequence coverage. Histfit (data) plots a histogram of the values in the vector data using the number of bins equal to the square root of the number of elements in data, and then superimposed a fitted normal distribution.

Figure 5. Venn diagram of the peptide identification by Proteios showing the hits by the separate scoring algorithms.

Sequence comparison of the PSA isoforms identified as first hit in Mascot search by PMF. Blue=signal peptide; Red=Propeptide; Green=active site; Orange underlined=unique sequence. A=NCBI ID: gi|166235452 and UniProt ID: Q15098; B= NCBI ID: gi|119592329; C= NCBI ID: gi|189530; D= NCBI ID: gi|71834855, E= NCBI ID: gi|119592331; F= NCBI ID: gi|119592330 and UniProt ID: Q8NCW4; G= NCBI ID: gi|5419891; H= NCBI ID: gi|71834857; 1= UniProt ID: Q15096; 2= UniProt ID: Q8IXI4; 3= UniProt ID: Q96A30; 4= UniProt ID: C9JXH3; 5= UniProt ID: P07288 and Q546G3; 6= UniProt ID: Q6LDS3.

Table 1

List of PSA peptides identified in high resolution MALDI-FT MS dataset by PMF (amino acid positions are based on the PSA precursor UniProt ID: P07288). PSA was identified in 251 FT-MS runs. Oxidized methionine is in bold type setting and marked with an asterisk. All cysteine residues were alkylated.

Calculated mass	Peptide	Position	No. of identification	Mean Δmass (ppm)
1076.4961	IVGWECEK	25–33	241 (96.0%)	0.755
1406.7419	HSQPWQVLVASR	34–45	243 (96.8%)	0.937
2343.2093	AVCGGVLVHPQWVLTAAHCIR	48–68	39 (15.5%)	2.210
756.4858	SVILLGR	71–77	230 (91.6%)	1.534
3492.6925	HSLFHPEDTGQVFQVSHSFPHPLYDMSLLK	78–107	5 (2.0%)	2.333
3508.6874	HSLFHPEDTGQVFQVSHSFPHPLYDM*SLLK	78–107	31 (12.4%)	1.250
1870.9360	FLRPGDDSSHDLMLLR	110–125	104 (41.4%)	1.280
1886.9309	FLRPGDDSSHDLM*LLR	110–125	176 (70.1%)	1.257
1271.6609	LSEPAELTDAVK	126–137	227 (90.4%)	0.656
3539.6476	VM*DLPTQEPALGTTCYASGWGSIPEEFLTPK	138–169	7 (2.8%)	1.500
2587.2999	KLQCVDLHVISNDVCAQVHPQK	146–167	92 (36.7%)	0.940
2459.2050	LQCVDLHVISNDVCAQVHPQK	171–191	96 (38.2%)	1.000
853.3939	FMLCAGR	195–201	176 (70.1%)	1.269
869.3888	FM*LCAGR	195–201	180 (71.7%)	1.193
1382.6588	FMLCAGR_WTGGK ^a	195–206	10 (4.0%)	1.857
672.3707	VVHYR	246–250	171 (68.1%)	2.315

^a One missed cleavage site.

Table 2

List of PSA isoforms identified in Mascot database search by PMF.

NCBI ID	NCBI definition	No. of identification
gi 166235452	Chain P, Crystal Structure Of Human Prostate Specific Antigen Complexed With An Activating Antibody	157
gi 189530	Prostate specific antigen precursor [Homo sapiens]	41
gi 119592329	Kallikrein 3, (prostate specific antigen), isoform CRA_b [Homo sapiens]	28
gi 71834855	Prostate specific antigen isoform 4 preproprotein [Homo sapiens]	13
gi 119592330	Kallikrein 3, (prostate specific antigen), isoform CRA_c [Homo sapiens]	3
gi 5419891	Prostate specific antigen [Homo sapiens]	3
gi 119592331	Kallikrein 3, (prostate specific antigen), isoform CRA_d [Homo sapiens]	2
gi 71834857	Prostate specific antigen isoform 5 preproprotein [Homo sapiens]	1

Table 3

List of PSA sequences identified in MALDI-MS/MS dataset by ProSE combined database search using Mascot, X!Tandem and XiTandem with k-score (amino acid positions are based on the PSA precursor UniProt ID: P07288). Oxidized methionine is in bold type setting and marked with an asterisk. All cysteine residues were alkylated.

Expected mass	Peptide	Position	No. of identification	Mean Amass (ppm)
1076.496	IVGGWCEK	25–33	330	1.53
1406.7454	HSQPWQVLV ^{ASR}	34–45	406	1.41
2343.2093	AVCGGVLVHPQWVLTAAHCIR ^a	48–68	114	1.35
756.4858	SVILLGR	71–77	19	2.24
3508.6874	HSLFHPEDTGQVFQVSHSPHPPL ^{YDM} *SLLK	78–107	7	1.97
1870.9376	FLRPGDDSSHDLMLLR	110–125	153	1.62
1886.9314	FLRPGDDSSHDL ^M *LLR	110–125	322	1.57
1271.6608	LSEPAELTDAVK	126–137	111	1.34
2587.2999	KIQCVDLHVISNDVCAQVHPQK	146–167	141	1.44
2459.2050	LQCVDLHVISNDVCAQVHPQK	171–191	114	1.27
853.3939	FMLCAGR	195–201	33	2.38
672.3699	VVHYR	246–250	79	3.31
1044.5599	ALPERPSLY ^b	235–243	5	1.36

^aThis tryptic peptide was identified in 24 cases as truncated sequence (AVCGGVLVHPQWVLTAAH, m/z 1913.99) by both XiTandem and X!Tandem with k-score but as a complete sequence by Mascot.

^bThis non-tryptic peptide was clearly identified by both XiTandem and X!Tandem with k-score but not by Mascot.

Table 4

List of PSA isoforms identified in ProSE combined search by sequencing data. Reviewed proteins are in bold typeset, indicating evidence at protein level; whereas other proteins are non-reviewed and evidence at transcript level.

UniProt KB	Accession ID	Name	Length (AA)	No. of identification
IP100040297	Q15096	APS protein	234	1750 (94.7%)
IP100383517	Q8IXI4	Kallikrein 3 (PSA) isoform CRA_h	227	1750 (94.7%)
IP100386209	Q96A30, Q15098	Kallikrein 3 (PSA) isoform CRA_e, APS protein	180 176	1333 (72.2%)
IP100640523	-	Prostate Specific Antigen isoform 4 preproprotein	218	1332 (72.1%)
IP100386036	Q8NCW4	Prostate Specific Antigen	220	1247 (67.5%)
IP100941321	C9JXH3	Uncharacterized protein (KLK3 gene product)	69	850 (46.0%)
IP100010858	P07288, Q546G3, Q6LDS3	Prostate Specific Antigen, Kallikrein 3 (PSA) isoform CRA_m, APS protein	261 261 257	782 (42.3%)
IP100022227	P20151, B4DU93, B4DU77	Kallikrein-2, cDNA FLJ56711, highly similar to Kallikrein-2 cDNA FLJ56711, highly similar to Kallikrein-2	261 159 244	1462 (79.2%)
IP100941408	P20151	Kallikrein-2	261	330 (17.9%)
IP100434943	B4DUB0	cDNA FLJ56711, highly similar to Kallikrein-2	224	330 (17.9%)
IP100219231	P20151, Q6T775, B0AZU9	Kallikrein-2, Kallikrein 2 isoform 5 preproprotein, cDNA FLJ56711, highly similar to Kallikrein-2	261 180 62	330 (17.9%)