Toward a Human Blood Serum Proteome

ANALYSIS BY MULTIDIMENSIONAL SEPARATION COUPLED WITH MASS SPECTROMETRY*[S]

Joshua N. Adkins‡, Susan M. Varnum‡, Kenneth J. Auberry§, Ronald J. Moore§, Nicolas H. Angell§¶, Richard D. Smith§, David L. Springer‡, and Joel G. Pounds‡‖

Blood serum is a complex body fluid that contains various proteins ranging in concentration over at least 9 orders of magnitude. Using a combination of mass spectrometry technologies with improvements in sample preparation, we have performed a proteomic analysis with submilliliter quantities of serum and increased the measurable concentration range for proteins in blood serum beyond previous reports. We have detected 490 proteins in serum by on-line reversed-phase microcapillary liquid chromatography coupled with ion trap mass spectrometry. To perform this analysis, immunoglobulins were removed from serum using protein A/G, and the remaining proteins were digested with trypsin. Resulting peptides were separated by strong cation exchange chromatography into distinct fractions prior to analysis. This separation resulted in a 3–5-fold increase in the number of proteins detected in an individual serum sample. With this increase in the number of proteins identified we have detected some lower abundance serum proteins (ng/ml range) including human growth hormone, interleukin-12, and prostate-specific antigen. We also used SEQUEST to compare different protein databases with and without filtering. This comparison is plotted to allow for a quick visual assessment of different databases as a subjective measure of analytical quality. With this study, we have performed the most extensive analysis of serum proteins to date and laid the foundation for future refinements in the identification of novel protein biomarkers of disease. ***Molecular & Cellular Proteomics 1:947–955, 2002.***

Serum, derived from plasma with clotting factors removed, contains 60–80 mg of protein/ml in addition to various small molecules including salts, lipids, amino acids, and sugars (1). The major protein constituents of serum include albumin, immunoglobulins, transferrin, haptoglobin, and lipoproteins (1, 2). In addition to these major constituents, serum also contains many other proteins that are synthesized and secreted, shed, or lost from cells and tissues throughout the body (3, 4). It is estimated that up to 10,000 proteins may be commonly present in serum, most of which would be present at very low relative abundances (5).

Historically, two-dimensional PAGE has been the primary method of separation and comparison for complex protein mixtures. This method has been critical in developing our understanding of the complexity and variety of proteins contained in cells and bodily fluids. Two-dimensional PAGE has been used to analyze serum and plasma (the unclotted parent fluid of serum) (6–13). Although impressive improvements in two-dimensional PAGE technologies have occurred in recent years, limitations remain. Two-dimensional PAGE is labor-intensive, requires relatively large sample quantities, is poorly reproducible, has a limited dynamic range for protein detection, and has difficulties in detecting proteins with extremes in molecular mass and isoelectric point (14). To address these limitations several types of mass spectrometry, in conjunction with various separation and analysis methods, are increasingly being adopted for proteomic measurements (15–22).

One of the driving forces in proteomics is the discovery of biomarkers, proteins that change in concentration or state in associations with a specific biological process or disease. Determination of concentration changes, relative or absolute, is fundamental to the discovery of valid biomarkers. The presence of higher abundance proteins (greater than mg/ml in serum) interferes with the identification and quantification of lower abundance proteins (lower than ng/ml in serum). Other methods such as two-dimensional PAGE have been used to demonstrate that the removal or separation of high abundance proteins enables greatly improved detection of lower abundance proteins (10, 11, 17, 23). The necessity of this removal or separation is also illustrated by noting that many proteins found useful as biomarkers for malignant and non-malignant disease (*e.g.* C-reactive protein, osteopontin, and prostate-specific antigen) are below 10 ng/ml, a value that is at least 7–8 orders of magnitude less than the most abundant serum proteins (1). Thus, the dynamic range typified by traditional proteomic methods are inadequate to allow for detection of these lower abundance serum proteins, or biomarkers, without effective removal or separation of the high abundance proteins.

One problem associated with any protein separation technique is that low abundance proteins may be removed along with the abundant species (24). Albumin is a protein of very high abundance in serum (35–50 mg/ml) that would be a prime candidate for complete selective removal prior to per-

From the ‡Biological Sciences Department and the §Environmental and Molecular Sciences Laboratory, Pacific Northwest National Laboratory, Richland, Washington 99352

Received, October 2, 2002, and in revised form, November 13, 2002

Published, MCP Papers in Press, November 15, 2002, DOI 10.1074/mcp.M200066-MCP200

forming a proteomic analysis of lower abundance proteins. However, albumin is a transport protein in blood serum that binds a large variety of compounds including hormones, lipoproteins, and amino acids (1, 25, 26). Thus, removal of albumin from serum may also result in the specific removal of low abundance cytokines, peptide hormones, and lipoproteins of interest.

Immunoglobulins, or antibodies, are also abundant proteins in serum that function by recognizing "foreign" antigens in blood and initiating their destruction. To recognize this enormous variety of antigens present in blood, immunoglobulins contain variable regions (1, 25, 27). These variable regions are a source of random peptide sequence in serum that can complicate protein identifications from peptide sequences. Therefore, with immunoglobulins binding foreign materials and the random nature of sequences from their variable regions, removal of immunoglobulins is important for a proteomic analysis of serum.

The purpose of this investigation was to establish new preparative methods to remove or separate high abundance serum proteins and to apply new proteomic approaches that increase the dynamic range available for the identification and characterization of serum proteins. These methods include the use of protein A/G covalently bound to acrylamide beads to selectively remove immunoglobulins, described earlier as a significant source of sequence variability found in serum. Further, these methods include the separation of trypsin-digested peptides prior to mass spectrometric analysis using both strong cation exchange (SCX)[1] chromatography and capillary gradient reversed-phase liquid chromatography. This investigation identifies a large number of proteins (490) from a single (submilliliter) serum sample and further provides the foundation for future studies with clinically important disease states.

EXPERIMENTAL PROCEDURES

Human Blood Serum—The human blood serum was acquired from a healthy anonymous female donor (Donor No. M99869) (Golden West Biologicals, Temecula, CA). Immediately after collection, plasma was isolated from whole blood without anti-coagulants by centrifugation. The plasma supernatant was allowed to clot overnight at room temperature, and the clotted material was removed by centrifugation under sterile conditions. Upon receipt at our laboratory, the serum was aliquoted into 1-ml units and stored at −80 °C. In subsequent preparation steps, proteins were detected, and concentrations were estimated, where appropriate, using denaturing (SDS) polyacrylamide gel electrophoresis with GELCODE blue staining (Pierce catalog no. 24590), absorbance at 280 nm, and/or with a Bradford protein assay using bovine serum albumin (BSA) as a protein standard (24, 28).

Depletion of Serum Immunoglobulins and Trypsin Digestion—The immunoglobulins (Igs) were depleted by affinity adsorption chromatography using protein A/G. 500 μl of serum was diluted with an equal amount of 20 mM sodium phosphate, pH 8.0 and added to UltraLink Immobilized protein A/G beads (2:1, v/v) (Pierce) that had been equilibrated with 20 mM sodium phosphate, pH 8.0. This mixture was incubated with gentle rocking for 2 h at 25 °C. Immunoglobulin-depleted serum was separated from the protein A/G beads by centrifugation. The beads were washed three times with 5 volumes of PBS (150 mM NaCl, 10 mM sodium phosphate, pH 7.3), and the washes were pooled with the immunoglobulin-depleted serum. The diluted immunoglobulin-depleted serum sample was then dialyzed into 10 mM HCO_3NH_4, 5% acetonitrile, pH 7.5, digested with trypsin 1:50 (w/w) ratio (Promega, Madison, WI) for 2 h at 37 °C, and lyophilized.

Strong Cation Exchange Separation of Immunoglobulin-depleted Serum Peptides—Lyophilized, immunoglobulin-depleted serum peptides were resuspended in 2 ml of 75% 10 mM ammonium formate, 25% acetonitrile, pH 3.0 with formic acid. The sample was centrifuged to remove insoluble debris and then separated using an LC gradient ion exchange system consisting of a quaternary gradient pump (ThermoSeparations P4000, San Jose, CA) equipped with a polysulfoethyl A column (5 μm, 300 Å, PolyLC, Columbia, MD). Mobile phase A consisted of 75% 10 mM ammonium formate, 25% acetonitrile, pH 3.0 with formic acid, and mobile phase B was 75% 200 mM ammonium formate, 25% acetonitrile, pH 8.0. The column was initially loaded (2-ml injection loop) and equilibrated for 5 min with 0% B. Peptides were eluted using a linear gradient of 0–100% B over 30 min, and the column was subsequently washed at 100% B for an additional 25 min all at a flow rate of 4 ml/min. The column effluent was monitored at 280 nm with a Linear 200 UV detector (Micro-Tech Scientific, Sunnyvale, CA), and a total of 120 fractions were collected at 30-s intervals using a FRAC-100 (Amersham Biosciences). Collected fractions were lyophilized and stored at −80 °C for reversed-phase LC/MS/MS analysis.

Reversed-phase Separation and LCQ Ion Trap Analysis—Reversed-phase separation was performed with an Agilent 1100 capillary high pressure liquid chromatography system with a 60-cm capillary column (150-μm inner diameter × 360-μm outer diameter, Polymicro Technologies, Phoenix, AZ) packed with 5-μm Jupiter C_{18} particles (Phenomenex, Torrance, CA). Mobile phase A consisted of water and 0.1% formic acid, and mobile phase B consisted of acetonitrile and 0.1% formic acid. SCX fractions were dissolved in 50 μl of water, 0.1% formic acid. Peptides were injected on the column in 8 μl at a flow rate of 1.8 μl/min, and the column was re-equilibrated with 5% B for 20 min. Peptides were eluted with a linear gradient from 5 to 70% B over 80 min. The capillary column was interfaced to an LCQ Deca XP ion trap mass spectrometer (ThermoFinnigan, San Jose, CA) using electrospray ionization.

The mass spectrometer was configured to optimize the duty cycle length with the quality of data acquired by alternating between a single full MS scan followed by three MS/MS scans on the three most intense precursor masses (as determined by Xcaliber mass spectrometer software in real time) from the single parent full scan. Dynamic mass exclusion windows were used and varied from 3 to 9 min. In addition, MS spectra for all samples were measured with an overall mass/charge (*m/z*) range of 400–2000. Fractions 21, 34, 39, 46, and 53, which contained high peptide concentrations, were re-analyzed three times using overlapping *m/z* ranges of 500–1050, 1000–1550, and 1500–2000, respectively. These segmented mass range analyses also utilized static mass exclusion lists that removed *m/z* precursors corresponding to the 20 most abundant peptides that were observed in the initial unsegmented analysis.

SEQUEST Analysis of Peptides—Tandem mass spectra were analyzed by SEQUEST (Bioworks 2.0, ThermoFinnigan) (16, 29–32), which performs its analyses by cross-correlating experimentally ac-

[1] The abbreviations used are: SCX, strong cation exchange; HUPO, Human Proteome Organization; LC, liquid chromatography; MS, mass spectrometry; MS/MS, tandem mass spectrometry; NCBI, National Center for Biotechnology Information.

quired mass spectra with theoretical idealized mass spectra generated from a database of protein sequences. These idealized spectra are weighted largely with *b* and *y* fragment ions, *i.e.* fragments resulting from the amide linkage bond from the N and C termini, respectively. For these analyses, no enzyme rule restrictions were applied to the possible cleavage points available for peptide generation from the initial proteins, allowing identifications resulting from non-tryptic cleavage to be observed as well. The peptide mass tolerance was 3.0, and the fragment ion tolerance was 0.0.

Protein Databases—SEQUEST analysis was performed using a modified version of the human FASTA protein database provided with SEQUEST (ThermoFinnigan). Database modifications included the removal of viral proteins and the removal of some redundant protein entries as well as minimizing the number of entries for abundant serum proteins (13). Additional analyses were conducted using the National Center for Biotechnology Information (NCBI) human protein database[2] and the Unigene human database[3] to determine whether important abundant serum proteins were missing from our modified database. Use of the additional various human databases did not alter the vast majority of SEQUEST peptide identifications. The use of the larger databases did result in an expected decrease in magnitude of the SEQUEST DelCN score in a fraction of peptide identifications. Most peptides not found in the smaller supplied database did not pass subsequent filters including visual inspection of fragmentation spectra (data not shown), and in the case of the Unigene database analysis required up to 2 weeks to finish on a modern PC. Currently no complete human protein database has been compiled, and one is not likely to exist for a number of years (35). Thus, the modified database was considered to be an adequate resource for this initial blood serum proteome analysis after comparisons to the NCBI and Unigene databases.[2,3]

Of concern with a shotgun proteomic approach is whether assumptions made for simple cases continue to apply with higher levels of complexity. To address the question for database choice, we sought to analyze LC/MS/MS results using a smaller database containing very few peptides with sequence identity to human proteins but still retaining the level of complexity observed in a complete genome. A locally available *Deinococcus radiodurans* FASTA database derived from the open reading frames of a completely sequenced genome (15) was used to generate SEQUEST analyses to compare against the human database-derived results. Five SCX fractions (fractions 21, 34, 39, 46, and 53) that contained the greatest number of fully tryptic peptides were analyzed against the *D. radiodurans* database for this comparison.

Filters for SEQUEST Results—SEQUEST results were filtered (Table I) with criteria similar to those developed by Yates and co-workers (31, 36). Serum proteins in circulation are frequently found cleaved by chymotrypsin and elastase (37). Thus, while trypsin was used to digest the serum proteins, the SEQUEST data filter was modified to allow for identification of peptides resulting from both chymotrypsin and elastase cleavage sites. The chymotrypsin and elastase filter levels were derived by comparing the SEQUEST-identified tryptic peptides to the identified non-tryptic albumin peptides. The high abundance and globular nature of albumin represented a useful reference for defining non-tryptic filter parameters. The resulting filters were those that resulted in four or more hits for any non-tryptic albumin peptide. These filters further resulted in 33 non-tryptic cleavage sites of the 133 total albumin cleavage sites.

The final filter parameters used to determine cross-correlation (Xcorr) cut-off values took into account both the charge state of the peptide and the proteolytic cleavage rules as shown in Table I. Additionally, a minimum value of 0.1 was used for DelCN, indicating that SEQUEST was readily able to distinguish between its first and second choices for identification (32). When three or fewer peptides for an individual protein passed the criteria shown in Table I, the mass spectra for those peptides were inspected manually. Manual inspection was performed using four criteria generally accepted as means for assessment of spectral quality (16, 36). First, the spectrum quality must be acceptable with the peaks to be used in the determination clearly above the noise base line. Second, some continuity must be present among the *b* or *y* fragments, *i.e.* fragments for three or more adjacent amino acids. Third, if proline is predicted to be present, then the corresponding *y* fragment should give an intense peak. Last, unidentified intense peaks should be verified as being either doubly charged or simply the mass of the precursor with one or two of the terminal amino acids removed.

TABLE I

Conservative filter parameters for SEQUEST results

The spectra for proteins with three or fewer unique peptide hits that met these criteria were manually inspected before inclusion to the protein list. Each protein with three or fewer passing peptide identifications had an average of 33 identifications that did not pass the above criteria but scored better than a 1.5 Xcorr and had a DelCN of at least 0.05.

Charge	Xcorr	Peptide type
+1	≥1.9	Fully tryptic
+1	≥2.1	Chymotryptic and/or elastic
+1	≥2.2	Partially tryptic, chymotryptic, and/or elastic
+2	≥2.2	Fully tryptic
+2	≥2.4	Partially tryptic, chymotryptic, and/or elastic
+2	≥3.0	No protease rules
+3	≥3.75	Tryptic, chymotryptic, and/or elastic only

RESULTS

Protein A/G for Immunoglobulin Depletion—We found that protein A/G affinity adsorption chromatography depleted essentially all of the immunoglobulins from serum as assessed by SDS-polyacrylamide electrophoresis (Fig. 1). Analysis of serum by MS is complicated by the fact that abundant proteins impede measurement of less abundant proteins. In addition, the abundant serum immunoglobulins have regions of high sequence variability that may complicate an MS-based sequence analysis of serum-derived peptides. Thus, to increase the dynamic concentration range and confidence of determination it is critical to remove the immunoglobulins from the serum sample. The heavy and light chain portions of the immunoglobulins were removed when visualized with GelCode Blue Stain (Fig. 1, *Lane 3*). Albumin is also slightly depleted by the same procedure (Fig. 1, *Lane 4*). This depletion is unexpected in that during the production of the chimeric protein A/G the albumin binding site from protein G was removed (38).

Multidimensional Peptide Separation—Albumin and other abundant non-immunoglobulin proteins may also present problems for an MS analysis. Many published methods of albumin separation have resulted either in poor depletion or

[2] NCBI, Hs GenBank™ Protein Databases ftp.ncbi.nlm.nih.gov/genomes/H_sapiens/protein/.

[3] NCBI, Hs Unigene Contig Databases ftp.ncbi.nlm.nih.gov/repository/UniGene/.

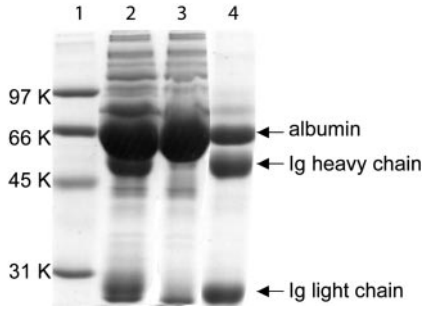

FIG. 1. **Igs, both heavy and light chain, are visibly depleted by protein A/G affinity chromatography as shown by SDS-PAGE.** Protein A/G was specific for immunoglobulins, but some cross-specificity for albumin was also present. *Lane 1*, molecular weight standards; *Lane 2*, unprocessed serum; *Lane 3*, serum after Ig depletion with protein A/G; *Lane 4*, proteins eluted from protein A/G.

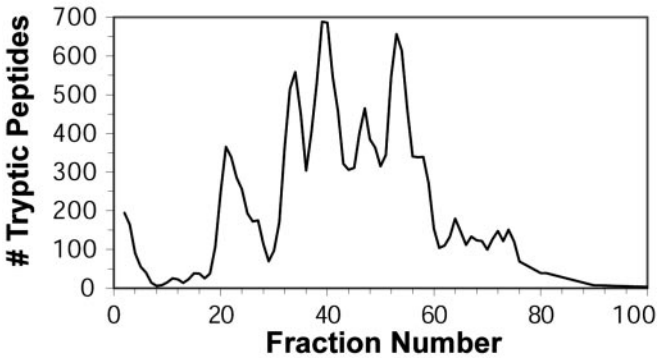

FIG. 2. **Strong cation exchange chromatography of human serum following depletion of immunoglobulins with protein A/G resulted in an excellent distribution of peptides over about 60 fractions (approximately fractions 19–79).** The number of fully tryptic peptides is as assessed by SEQUEST.

potential loss of specific low abundance proteins of interest in plasma (23) or in hemofiltrate (a plasma-derived fluid from dialysis patients) (17, 37). Rather than remove albumin from the serum, the strategy used here was to fractionate trypsin-derived peptides by SCX and then perform a second dimension separation with reversed-phase LC. The SCX chromatography resulted in good fractionation with the richest peptide samples eluted over about 60 fractions (fractions 19–79, Fig. 2).

The SCX fractionation illustrates the power of further analyzing specific fractions to increase the number of proteins determined by an LC/MS/MS analysis. Fractions 21, 34, 39, 46, and 53 were reanalyzed by LC/MS/MS using a static exclusion list for the 20 most commonly found peptides from the previous 400–2000 *m/z* MS analysis. In addition, each fraction was analyzed three times by limiting the *m/z* window to 500–1050, 1000–1550, or 1500–2000 for each run (illustrated in Fig. 3). The *m/z* segmentation resulted in approximately the same number of peptides passing SEQUEST data filters and manual inspection as the unsegmented analysis (Table II) but resulted in more proteins identified by multiple peptides and fewer numbers of serum albumin identifications. This increase in non-albumin identification is attributable to the MS analysis focusing on novel peptides rather than high abundance albumin peptides previously analyzed (Fig. 3). In addition, multidimensional separations allowed for important increases in dynamic range and decreases in individual analysis complexity. Here we show that some fractions may be complex enough to warrant further steps to simplify the MS analysis.

Proteins Identified in Serum—Using immunoglobulin depletion, SCX, and microcapillary reversed-phase high performance LC followed by data analysis with SEQUEST we have identified 490 proteins in serum. These proteins include those illustrated in Table III. Proteins found in this analysis also cover a large concentration range (as assessed from clinical reference normal values) from 85% coverage with 111 unique peptides from albumin (serum concentration 35–50 mg/ml), 31% coverage with 28 unique peptides from complement factor H (serum concentration 35 μg/ml); 29% coverage with 14 unique peptides from angiotensinogen (serum concentration 2.5–0.15 ng/ml) (39), 12% coverage with one peptide from prostate-specific antigen (serum concentration less than 1.0 pg/ml in a healthy female) (1). Our analysis identifies most serum proteins previously reported as well as a large number of proteins newly identified in serum (8–12, 37, 40).

Method for Visualizing and Accessing the Relative Quality of a Global SEQUEST Analysis—SEQUEST analysis results are typically scored using a combination of Xcorr and DelCN. Xcorr, in short, is the value of the best resulting correlation between a predicted peptide spectrum and an experimental spectrum. A higher Xcorr value provides better confidence of peptide identification. An Xcorr value greater than 2 is typically considered significant for peptide identifications. DelCN is the normalized difference in magnitude between the peptide fit with the highest Xcorr and the peptide fit with the second best Xcorr. A minimum acceptable value for DelCN is typically 0.1. More confidence is placed in protein identifications when multiple peptides occurring from the same protein that have Xcorr values greater than 2.0 and DelCN values greater than or equal to 0.1 (8, 16, 36).

To qualitatively evaluate the global results from a SEQUEST analysis, we compared the human peptides analyzed by MS/MS and *m/z* segmentation using SEQUEST with two different databases. The databases compared with SEQUEST analysis were an unrelated bacterial database (*D. radiodurans*) and a human protein database. The plot of DelCN *versus* Xcorr from a SEQUEST analysis with the *D. radiodurans* database generally defines a region of data that is composed of low confidence peptide identifications (Fig. 4*A*). A similar plot for a SEQUEST analysis using a human database identifies a second population of peptides with higher quality peptide identifications (Fig. 4*B*). The overlap between the poor quality and high quality populations contains many real peptide identifications. After filtering (see Table I), the SEQUEST analysis

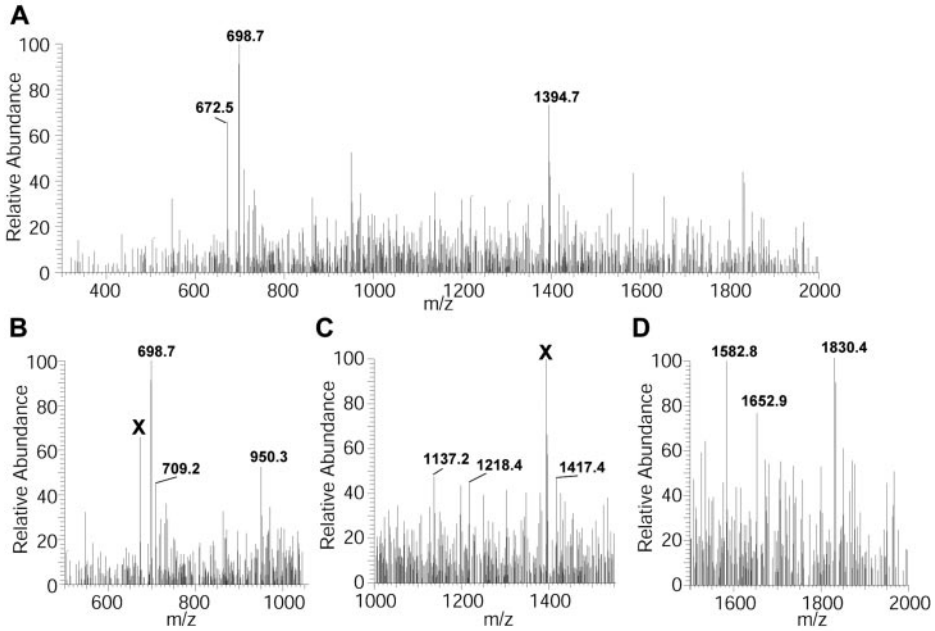

FIG. 3. **MS *m/z* segmentation illustration of the method used to measure peaks that may not have been missed during the initial analysis of SCX fractions 21, 34, 39, 46, and 53.** *A*, full scan with an *m/z* window of 300–2000; masses subsequently trapped and measured by tandem MS are labeled with *mass numbers*. *B*, *C*, and *D*, *m/z* segmented scans with *m/z* windows of 500–1050, 1000–1550, and 1500–2000, respectively. Masses analyzed by tandem MS are labeled with *mass numbers*, and masses disregarded due to static mass exclusion lists are labeled with an *X*.

TABLE II
Proteins found comparing a single MS/MS analysis to m/z window segmentation

Comparison of peptide information from a single analysis 400–2000 full MS scan of the five peak fractions (fractions 21, 34, 39, 46, and 53) (Fig. 2) and three *m/z* segmentations of the same five peak fractions with static exclusion list.

	400–2000 *m/z* full MS scan	500–1050, 1000–1550, and 1500–2000 *m/z* full MS scans
Total peptide hits	2014	2179
Proteins determined by ≥3 peptides[a]	26	58
Proteins determined by 1–2 peptides[a]	25	83

[a] Passing data filters in Table I.

of peptides using the *D. radiodurans* database eliminates all but 1% or 76 of the original low confidence peptide identifications (Fig. 4*C*). In contrast, after filtering (Table I), 20% or 2179 of the peptides from the human database remain (Fig. 4*D*). The filtering method results in more qualitative confidence for the peptide identifications using the human protein database at a global scale. While it is expected that most of the peptides identified from the *D. radiodurans* database that passed the data filters do not appear as proteins serum, some of these peptides may, by chance or evolutionary conservation, be legitimately found using the *D. radiodurans* database.

DISCUSSION

Blood plasma, like cells, has many high abundance proteins that perform various housekeeping functions. Blood plasma contains numerous secreted or shed low abundance proteins that are critical for signaling cascades and regulatory events. During necrosis, apoptosis, and hemolysis, contents of cells may be released into the plasma. The presence of these components in blood reinforces the benefits of using a proteomic approach for identifying biomarkers for disease states. In this study, we report an analysis of serum identifying 490 proteins (Table III and supplemental data table), at least a 3–5-fold increase in the number of identified proteins from a blood-derived fluid found in previous reports.

Previous proteomic characterizations of human plasma have used two-dimensional PAGE. These studies such as the seminal work of Anderson and co-workers (10, 41) have been summarized by the ExPASy on-line human plasma two-dimensional PAGE database (ca.expasy.org/ch2d/). These previous investigations have focused on plasma and thus are not directly comparable to the serum results reported here. However, of the 58 named proteins identified in this on-line human plasma protein database, we identified 51 in our serum analysis. There are several possible explanations for not identifying these seven proteins, including fibrinogen B, fibrinogen γ, C-reactive protein, and actin. First, plasma but not serum samples contain the clotting factors fibrinogen B and fibrinogen γ. Second, our serum was obtained from a single healthy female. The concentration of certain blood proteins may make detection difficult for our single source sample *versus* a general population; an example is C-reactive protein, which is typically at subnanogram per milliliter concentrations in a healthy female (1). Finally, the sample preparation and analytical methods used by these previous investigators differ significantly from those reported here. The lack of detection for the other proteins, such as actin, may be due to differing methods of sample collection, processing, and analysis. Overall our approach is superior for global identification since the two-dimensional PAGE database is made up of nine published reports but identified only 58 proteins, while we found

TABLE III

Selected 134 categorized proteins from the 490 total proteins detected

MAP, mitogen-activated protein; ERK, extracellular signal-regulated kinase.

Common circulating blood proteins	Albumin, haptoglobin, hemopexin, fibrinogen A, α_1-microglobulin, β_2-microglobulin, α_2-glycoprotein(Zn), α_2-HS-glycoprotein, serum amyloid proteins (A2-β and A), vitronectin, apolipoproteins (A-I, A-II, A-IV, B, C-I, C-II, C-III, D, E, F, and L), gelsolin, histidine-rich glycoprotein, leucine-rich α_2-glycoprotein, low density lipoprotein-related proteins (1 and 2), α_1-acid glycoprotein 1 (orsomucoid 1), α_1-acid glycoprotein 2 (orsomucoid 2), clusterin, Kell blood group protein, perlecan (heparan sulfate proteoglycan), ferroxidase
Coagulation and complement factors	Complement factors (B, C1R, C2, C3, C4A, C4B, C5, C6, C7, C8α, C8β, C8γ, C9, H, and I), coagulation factors (II, V, VIII, XII, and XIIIb)
Blood transport and binding proteins	Transferrin, transthryretin, retinol-binding protein, vitamin D-binding protein, insulin-like growth factor-binding proteins (5 and 7), calcium-binding protein P22, complement C4-binding protein α, hemoglobins (A and B), high density lipoprotein-binding protein, histidine-rich calcium-binding protein, hyaluronan-binding protein 2, latent transforming growth factor-β-binding protein, S100 calcium-binding protein A2, thyroglobulin, corticosteroid-binding globulin, selenoprotein P
Protease inhibitors	α_2-Antiplasmin inhibitor, complement C1 inhibitor, heparin cofactor II (protease inhibitor leuserpin 2), inter-α-(globulin) inhibitor H4 (plasma kallikrein-sensitive glycoprotein), inter-α-trypsin inhibitor, plasminogen activator inhibitor, protease inhibitor 4 (kallistatin), α_1-antichymotrypsin, α_1-antitrypsin
Proteases	Kallikrein, angiotensinogen, plasminogen, α-thrombin, carboxypeptidase N
Other enzymes	Antioxidant protein 1, arginosuccinase, hexokinase 3, folate hydrolase 1 (prostate-specific membrane antigen), nicotinamide nucleotide transhydrogenase, paraoxonase/arylesterase, phosphodiesterase 5A, phosphoglycerate kinase 1, squalene monooxygenase, triacylglycerol lipase, methylmalonyl coenzyme A mutase, thioredoxin-dependent peroxide reductase
Cytokines and hormones	Atrial natriuretic factor, human growth hormone, inhibin, interleukin-12a, interferon (α-inducible protein 27), fibroblast growth factor-12, prostate-specific antigen, growth/differentiation factor 5, pigment epithelium-derived factor
Channel and receptor-derived peptides	ATP-sensitive inward rectifier K^+ channel 11, chemokine (CX_3C) receptor 1, G protein-coupled receptor 1, γ-aminobutyric acid receptor B, prostaglandin E receptor (subtype EP3), protein tyrosine phosphatase (receptor type, f polypeptide), solute carrier family 5 (sodium iodide symporter), T-cell receptor α chain VJ region, tumor necrosis factor receptor-associated factor 5, interleukin-2 receptor γ chain, integrin α (4, 8, and E)
Miscellaneous (structural, nuclear, etc.)	Keratins (1, 2, and 9), microtubule-associated protein, microtubule-vesicle linker clip-170, plectin, sytaxin, elastin, MAP/ERK kinase kinase 5, bullous pemphigoid antigen, centromere protein f, collagens (IV and XI), titans, elongation factor tu, epidermal growth factor receptor pathway substrate 8

the 490 proteins, including those that would be expected to be common between studies.

Another family of serum/plasma studies for comparison is the characterization of rat serum by Gianazza and co-workers (6–8, 11, 12). These studies identified 34 proteins with human homologues and characterized the changes in protein abundance with disease states or chemical exposures associated with inflammatory disease. These rat serum studies concluded that even abundant proteins could be markers for disease states. Our study identified the human homologues of 31 of the 34 identified rat proteins. We did not find the human equivalent of thyroxine-binding globulin, thiostatin, or C-reactive protein. Many of the same reasons for a lack of complete overlap with the ExPASy plasma two-dimensional PAGE database apply here. In addition, species-specific differences may explain differing proteins and expression levels.

Serum is a complex biological fluid with many functions, and the presence, absence, or concentration of a specific protein may be non-intuitive until the serum proteome is fully understood. In an analysis of this complexity, it is important to note that expectations often differ from results for many proteins. Examples of unexpected results are hemoglobin and actin, which are both ubiquitous in the red blood cells. Therefore between high quantity and rapid turnover of red blood cells it may be expected that hemoglobin and actin should be readily detectable in serum (42). In contrast to our expectations, few hemoglobin-derived peptides and no actin-derived peptides were identified. In fact, both hemoglobin and actin are actively sequestered and cleared from the serum via the abundant serum proteins haptoglobin and vitamin D-binding protein, respectively (42–45). Another example of unexpected results are the identification of immunoglobulin-derived peptides, although depletion was complete when evaluated by SDS-PAGE. It is unclear whether these peptides originated from incomplete depletion of immunoglobulins *in vitro* or from proteolyzed immunoglobulins circulating in blood.

As global proteomic approaches become more common, there is an increasing need to evaluate and visualize large data sets with improvements in individual scoring methods (46–48). Often proteomic studies are less concerned with individual peptide identifications than with globally studying changes. In fact, a recent study using a global approach to profile proteins only by masses using surface-enhanced laser desorption-ionization MS with blood serum has been shown

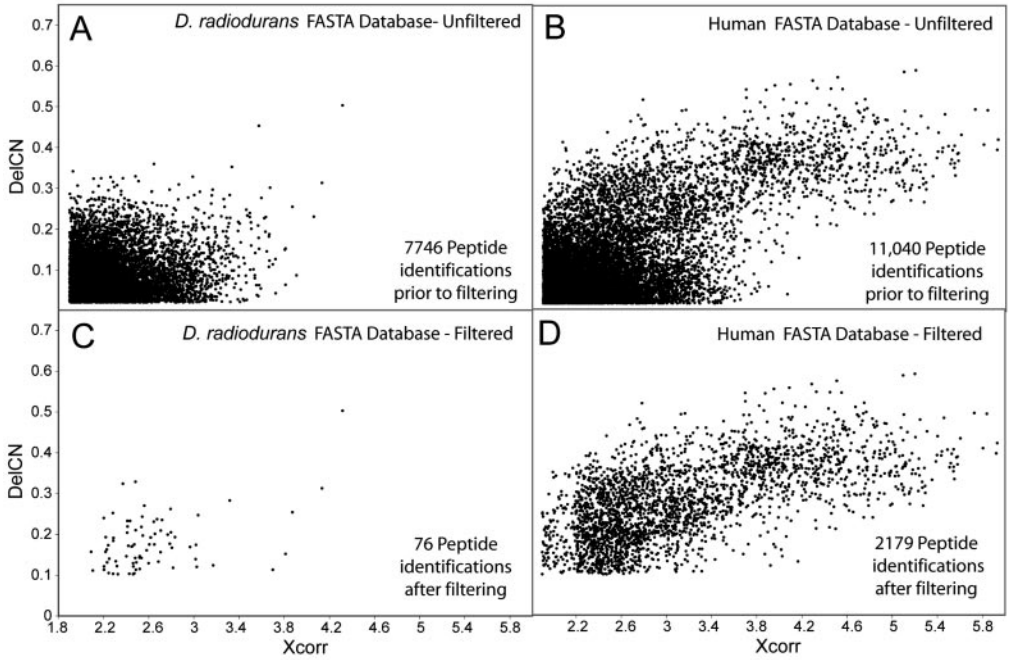

FIG. 4. **Global affects of SEQUEST peptide identification filters.** Shown are populations using a "random sequence" database with genome level complexity and a human protein database. The *D. radiodurans* database was used here as the random sequence database with results unfiltered (*A*) and filtered (*C*). A database derived from NCBI human protein sequence data was used as the human protein database with results unfiltered (*B*) and filtered (*D*).

to have predictive value in ovarian cancer (33). One of the difficulties related to the use of SEQUEST for peptide identifications is the lack of methods to globally evaluate the quality of data and the lack of methods to access global changes created by filtering schemes and/or database changes. Here, by comparing our SEQUEST results to multiple databases, we have illustrated an intuitive and easily adopted method for analyzing LC-MS/MS experiments in global terms (Fig. 4).

Major technical issues complicate the routine characterization of the plasma/serum proteome. First, plasma/serum proteins, like tissue proteins, may be post-translationally modified, and many plasma proteins are glycosylated (13). Other important factors include modifications such as sulfation, phosphorylation, oxidation, glycation, lipidation, and γ-carboxyglutamylation. Currently there are no commercially available tools that can identify peptides with this variety and number of modifications. The serum proteins in this study (Table III) were identified from translationally unmodified peptides. Significant improvements to sample processing and informatics are needed to identify these protein modifications. Second, protease digestion further adds to the complexity of a proteomic analysis of serum (13). Here we filtered peptide identifications based on protease modifications to take *in situ* proteolysis (chymotrypsin and elastase) into account. Third, the concentration range of plasma/serum proteins encompasses at least 9 orders of magnitude. Thus, significant improvements in the sample processing and separation with improvement in the dynamic range, sensitivity, and ability to quantitate results from mass spectrometry are needed to elaborate the plasma/serum proteome beyond the 490 proteins identified in this report. Last, the immature status of human protein databases further complicates analysis because there are likely to be protein identifications even in this mid-abundance range that have not yet been added to any publicly available human protein database (35).

The Human Proteome Organization (HUPO) has been founded to consolidate and organize future efforts in human proteomics (34). Among the many of the stated goals of HUPO are the research goals of characterizing the human plasma/serum proteomes and the informatic goals of standardizing proteome data and annotations with the improvement of bioinformatic tools for proteome analysis (34). Here we report a large improvement for proteomic analysis of serum; this analysis identifies 490 proteins, about 10% toward a 5000 protein goal of HUPO. Further, we have presented a visualization method that can be used to evaluate the quality of a global SEQUEST proteomic analysis along with the ability to subjectively evaluate protein database quality for a SEQUEST analysis.

Acknowledgments—We gratefully acknowledge the insightful discussions of David Wunschel and Richard Zangar, the technical assistance of Deanna Auberry, and the encouragement and support of Robert Miller and David Koppenaal.

* This work was supported by the Biotechnology section of Core Technology, Battelle Memorial Institute. The costs of publication of this article were defrayed in part by the payment of page charges.

This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

S The on-line version of this article (available at http://www.mcponline.org) contains supplemental data: peptide identifications.

¶ Current address: Human Genome Sciences, 9410 Key West Ave., Rockville, MD 20850.

‖ To whom correspondence should be addressed: Biological Sciences Dept., Pacific Northwest National Laboratory, P.O. Box 999, MSIN: P7-58, Richland, WA 99352. Tel.: 509-376-1015; Fax: 509-376-9449; E-mail: joel.pounds@pnl.gov.

REFERENCES

1. Burtis, C. A., and Ashwood, E. R. (2001) *Tietz Fundamentals of Clinical Chemistry*, 5th Ed., W. B. Saunders Company, Philadelphia, PA
2. Turner, M. W., and Hulme, B. (1970) *The Plasma Proteins: An Introduction*, Pitman Medical & Scientific Publishing Co., Ltd., London
3. Schrader, M., and Schulz-Knappe, P. (2001) Peptidomics technologies for human body fluids. *Trends Biotechnol.* **19,** S55–S60
4. Kennedy, S. (2001) Proteomic profiling from human samples: the body fluid alternative. *Toxicol. Lett.* **120,** 379–384
5. Wrotnowski, C. (1998) The future of plasma proteins. *Genet. Eng. News* **18,** 14
6. Eberini, I., Agnello, D., Miller, I., Villa, P., Fratelli, M., Ghezzi, P., Gemeiner, M., Chan, J., Aebersold, R., and Gianazza, E. (2000) Proteins of rat serum V: adjuvant arthritis and its modulation by nonsteroidal anti-inflammatory drugs. *Electrophoresis* **21,** 2170–2179
7. Eberini, I., Miller, I., Zancan, V., Bolego, C., Puglisi, L., Gemeiner, M., and Gianazza, E. (1999) Proteins of rat serum IV. Time-course of acute-phase protein expression and its modulation by indomethacine. *Electrophoresis* **20,** 846–853
8. Haynes, P., Miller, I., Aebersold, R., Gemeiner, M., Eberini, I., Lovati, M. R., Manzoni, C., Vignati, M., and Gianazza, E. (1998) Proteins of rat serum: I. establishing a reference two-dimensional electrophoresis map by immunodetection and microbore high performance liquid chromatography-electrospray mass spectrometry. *Electrophoresis* **19,** 1484–1492
9. Edwards, J. J., Anderson, N. G., Nance, S. L., and Anderson, N. L. (1979) Red cell proteins. I. two-dimensional mapping of human erythrocyte lysate proteins. *Blood* **53,** 1121–1132
10. Anderson, L., and Anderson, N. G. (1977) High resolution two-dimensional electrophoresis of human plasma proteins. *Proc. Natl. Acad. Sci. U. S. A.* **74,** 5421–5425
11. Miller, I., Haynes, P., Gemeiner, M., Aebersold, R., Manzoni, C., Lovati, M. R., Vignati, M., Eberini, I., and Gianazza, E. (1998) Proteins of rat serum: II. influence of some biological parameters of the two-dimensional electrophoresis pattern. *Electrophoresis* **19,** 1493–1500
12. Miller, I., Haynes, P., Eberini, I., Gemeiner, M., Aebersold, R., and Gianazza, E. (1999) Proteins of rat serum: III. gender-related differences in protein concentration under baseline conditions and upon experimental inflammation as evaluated by two-dimensional electrophoresis. *Electrophoresis* **20,** 836–845
13. Peters, T., Jr. (1987) Intracellular precursor forms of plasma proteins: their functions and possible occurrence in plasma. *Clin. Chem.* **33,** 1317–1325
14. Rabilloud, T. (2002) Two-dimensional gel electrophoresis in proteomics: old, old fashioned, but it still climbs up the mountains. *Proteomics* **2,** 3–10
15. Conrads, T. P., Alving, K., Veenstra, T. D., Belov, M. E., Anderson, G. A., Anderson, D. J., Lipton, M. S., Pasa-Tolic, L., Udseth, H. R., Chrisler, W. B., Thrall, B. D., and Smith, R. D. (2001) Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and 15N-metabolic labeling. *Anal. Chem*. **73,** 2132–2139
16. Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M., and Yates, J. R., III (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17,** 676–682
17. Raida, M., Schulz-Knappe, P., Heine, G., and Forssmann, W. G. (1999) Liquid chromatography and electrospray mass spectrometric mapping of peptides from human plasma filtrate. *J. Am. Soc. Mass Spectrom.* **10,** 45–54
18. Liotta, L. A., Kohn, E. C., and Petricoin, E. F. (2001) Clinical proteomics: personalized molecular medicine. *J. Am. Med. Assoc.* **286,** 2211–2214
19. Smith, R. D. (2000) Evolution of ESI-mass spectrometry and Fourier transform ion cyclotron resonances for proteomics and other biological applications. *Int. J. Mass Spectrom.* **200,** 509–544
20. Yates, J. R., III (2000) Mass spectrometry. From genomics to proteomics. *Trends Genet*. **16,** 5–8
21. Wu, S.-L., Amato, H., Biringer, R., Choudhary, G., Shieh, P., and Hancock, W. S. (2002) Targeted proteomics of low-level proteins in human plasma by LC/MSn: using human growth hormone as a model system. *J. Proteome Res.* **1,** 459–465
22. Bergquist, J., Palmblad, M., Wetterhall, M., Hakansson, P., and Markides, K. E. (2002) Peptide mapping of proteins in human body fluids using electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry. *Mass Spectrom. Rev.* **21,** 2–15
23. Georgiou, H. M., Rice, G. E., and Baker, M. S. (2001) Proteomic analysis of human plasma: failure of centrifugal ultrafiltration to remove albumin and other high molecular weight proteins. *Proteomics* **1,** 1503–1506
24. Scopes, R. K. (1994) *Protein Purification: Principles and Practice*, 3rd Ed., Springer-Verlag, New York
25. Ritchie, R. F., and Navolotskaia, O. (eds) (1996) *Serum Proteins in Clinical Medicine*, 1st Ed., Vol. 1, Foundation for Blood Research, Scarborough, ME
26. Beutler, E., and Williams, W. J. (1995) *Williams Hematology*, 5th Ed., McGraw-Hill Inc. Health Professions Division, New York
27. Anderson, N. L., and Anderson, N. G. (2002) The human plasma proteome: history, character, and diagnostic prospects. *Mol. Cell. Proteomics* **1,** 845–867
28. Bradford, M. M. (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72,** 248–254
29. Yates, J. R., III, Carmack, E., Hays, L., Link, A. J., and Eng, J. K. (1999) Automated protein identification using microcolumn liquid chromatography-tandem mass spectrometry. *Methods Mol. Biol.* **112,** 553–569
30. Yates, J. R., III, McCormack, A. L., and Eng, J. K. (1996) Mining genomes with MS. *Anal. Chem.* **68,** 534–540
31. Washburn, M. P., Wolters, D., and Yates, J. R., III (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol*. **19,**242–247
32. Eng, J. K., Mccormack, A. L., and Yates, J. R. (1994) An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5,** 976–989
33. Petricoin, E. F., Ardenkani, A. A., Hitt, B. A., Levine, P. J., Fusaro, V. A., Steinberg, S. M., Mills, G. B., Simone, C., Fishman, D. A., Kohn, E. C., and Liotta, L. A. (2002) Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359,** 572–577
34. Hanash, S., and Celis, J. (2002) The human proteome organization: a mission to advance proteome knowledge. *Mol. Cell. Proteomics* **1,** 413–414
35. Harrison, P. M., Kumar, A., Lang, N., Snyder, M., and Gerstein, M. (2002) A question of size: the eukaryotic proteome and the problems in defining it. *Nucleic Acids Res.* **30,** 1083–1090
36. Wolters, D. A., Washburn, M. P., and Yates, J. R., III (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Anal. Chem.* **73,** 5683–5690
37. Richter, R., Schulz-Knappe, P., Schrader, M., Standker, L., Jurgens, M., Tammen, H., and Forssmann, W. G. (1999) Composition of the peptide fraction in human blood plasma: database of circulating human peptides. *J. Chromatogr. B Biomed. Sci. Appl.* **726,** 25–35
38. Pierce Endogen (1995) *Vol. 0497 Instructions: UltraLink Immobilzed Protein A/G*, pp. 1–4, Pierce Endogen, Rockford, IL
39. Vinck, W. J., Fagard, R. H., Vlietinck, R., and Lijnen, P. (2002) Heritability of plasma renin activity and plasma concentration of angiotensinogen and angiotensin-converting enzyme. *J. Hum. Hypertens.* **16,** 417–422
40. Eckerskorn, C., Strupat, K., Schleuder, D., Hochstrasser, D., Sanchez, J. C., Lottspeich, F., and Hillenkamp, F. (1997) Analysis of proteins by direct-scanning infrared-MALDI mass spectrometry after 2D-PAGE separation and electroblotting. *Anal. Chem.* **69,** 2888–2892
41. Hoogland, C., Sanchez, J. C., Tonella, L., Bairoch, A., Hochstrasser, D. F., and Appel, R. D. (1999) The SWISS-2DPAGE database: what has changed during the last year. *Nucleic Acids Res.* **27,** 289–291
42. Houmeida, A., Hanin, V., Constans, J., Benyamin, Y., and Roustan, C. (1992) Localization of a vitamin-D-binding protein interaction site in the COOH-terminal sequence of actin. *Eur. J. Biochem.* **203,** 499–503
43. Emerson, D. L., Galbraith, R. M., and Arnaud, P. (1984) Electrophoretic

demonstration of interactions between Gc (vitamin D-binding protein), actin and 25-hydroxycholecalciferol. *Electrophoresis* **5,** 22–26
44. Goldschmidt-Clermont, P. J., Van Baelen, H., Bouillon, R., Shook, T. E., Williams, M. H., Nel, A. E., and Galbraith, R. M. (1988) Role of group-specific component (vitamin D binding protein) in clearance of actin from the circulation in the rabbit. *J. Clin. Investig.* **81,** 1519–1527
45. Haddad, J. G., Hu, Y. Z., Kowalski, M. A., Laramore, C., Ray, K., Robzyk, P., and Cooke, N. E. (1992) Identification of the sterol- and actin-binding domains of plasma vitamin D binding protein (Gc-globulin). *Biochemistry* **31,** 7174–7181
46. Keller, A., Purvine, S., Nesvizhskii, A. I., Stolyar, S., Goodlett, D. R., and Kolker, E. (2002) Experimental protein mixture for validating tandem mass spectral analysis. *Omics* **6,** 207–212
47. MacCross, M. J., Wu, C. C., and Yates, J. R., III (2002) Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Anal. Chem.* **74,** 5593–5599
48. Field, H. I., Fenyo, D., and Beavis, R. C. (2002) RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. *Proteomics* **2,** 36–47