JOURNAL OF PROTEOMICS 75 (2012) 1303–1317

Available online at www.sciencedirect.com

SciVerse ScienceDirect

www.elsevier.com/locate/jprot

Identification and quantification of peptides and proteins secreted from prostate epithelial cells by unbiased liquid chromatography tandem mass spectrometry using goodness of fit and analysis of variance

*Angelica K. Florentinus, Peter Bowden, Girish Sardana, Eleftherios P. Diamandis, John G. Marshall**

Department of Chemistry and Biology, Ryerson University, Toronto, Canada

ARTICLE INFO

Article history:
Received 2 July 2011
Accepted 5 November 2011
Available online 15 November 2011

Keywords:
Prostate cancer
Epithelial cell line
Secreted proteins
LC-ESI-MS/MS
Chromatography optimization
Relative quantification of measured ion intensity

ABSTRACT

The proteins secreted by prostate cancer cells (PC3(AR)6) were separated by strong anion exchange chromatography, digested with trypsin and analyzed by unbiased liquid chromatography tandem mass spectrometry with an ion trap. The spectra were matched to peptides within proteins using a goodness of fit algorithm that showed a low false positive rate. The parent ions for MS/MS were randomly and independently sampled from a log-normal population and therefore could be analyzed by ANOVA. Normal distribution analysis confirmed that the parent and fragment ion intensity distributions were sampled over 99.9% of their range that was above the background noise. Arranging the ion intensity data with the identified peptide and protein sequences in structured query language (SQL) permitted the quantification of ion intensity across treatments, proteins and peptides. The intensity of 101,905 fragment ions from 1421 peptide precursors of 583 peptides from 233 proteins separated over 11 sample treatments were computed together in one ANOVA model using the statistical analysis system (SAS) prior to Tukey–Kramer honestly significant difference (HSD) testing. Thus complex mixtures of proteins were identified and quantified with a high degree of confidence using an ion trap without isotopic labels, multivariate analysis or comparing chromatographic retention times.

1. Introduction

Two challenges in the application of unbiased LC-ESI-MS/MS are the estimation of type I error rate and the large scale quantification of proteins between samples [1–6]. Empirical, heuristic or pragmatic approaches such as the so called false discovery rate (FDR) have been used for estimating the false positive rate of protein identification [7]. The problem of protein identification has been addressed by the use of the goodness of fit tests [8–12]. Spectral counting [13], isotopic labels [14] or multivariate statistics [1,15–19] have been utilized to quantitatively separate sample treatments. However, where unbiased LC-ESI-MS/MS is randomly and independently sampled from a normal probability distribution, the data may be

Abbreviations: ACN, acetonitrile; C18, reversed phase chromatography; CID, collision induced dissociation; ESI, electrospray ionization; FWER, family-wise error rate; HPLC, high pressure liquid chromatography; GPM, global proteomic machine; LC, liquid chromatography; MCP, multiple comparison procedure; (*.mgf), Mascot Generic Format; MS/MS, tandem mass spectrometry; P63104, 14-3-3 zeta; SAS, statistical analysis system; SAX, strong anion exchange; SQL, structured query language; XML, extensible mark up language.

* *Corresponding author*.
E-mail address: 4marshal@ryerson.ca (J.G. Marshall).

1874-3919/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.jprot.2011.11.002

analyzed by ANOVA and multiple comparison procedures (MCP). Chromatographic separations can be analyzed by the ion current intensity of tryptic peptides [20] from unbiased LC-ESI-MS/MS using the goodness of fit of MS/MS spectra to identify peptides and proteins [9,11,12] with ANOVA of peptide and fragment intensity values [21,22]. Here the large-scale database and statistical analysis tools SQL and SAS were used to analyze the results of LC-ESI-MS/MS by Chi square, general linear models such as ANOVA, and the Tukey–Kramer Honestly Significant Difference (HSD) test using only the mgf file provided by the mass spectrometer and the results of a correlation algorithm such as X!TANDEM.

1.1. *Type I and type II error of peptide and protein identification*

The basis of classical statistics is the expectation value of the fit of the data with respect to random chance. Peptide identification using algorithms such as X!TANDEM may be considered a correlation or goodness of fit problem of the many fragment ion *m/z* values to the predicted peptide fragments from a protein database. The expectation values of a type I error (false positive) can be generated by the goodness of fit test in X! TANDEM. The products of the replicate peptide goodness of fit tests for each peptide sequence and the different peptides within each protein combine to yield the confidence in the protein identification [10]. There are protections against type I error built into the X!TANDEM correlation algorithm to prevent false positive identification [11,12]. It is possible to confirm the type I error rate of protein identification by Chi square comparison of peptide-to-protein distribution to that of random results [11,12]. The Chi square comparison to random expectation confirmed that error rates were low and showed good agreement when the results of different experiments, instruments and correlation algorithms were compared [10–12,23,24]. The Chi square test agreed with the X!TANDEM and PARAGON algorithms for fitting spectra that type I error rates of protein identification were low and differed from those of the so called FDR test by many orders of magnitude [10–12,23,25]. The false discovery rate (FDR) is a multiple comparison protocol (MCP) designed to correct the significance of multiple means tests between many controls versus treatments over many parameters by holding a competition for significance [26]. In contrast, matching MS/MS spectra to peptide sequences within the mass tolerance of the instrument is not a multiple means comparison problem but a goodness of between the MS/MS spectra to the candidate peptides. The application of the so called FDR test to correcting the goodness of fit of MS/MS spectra resulted in a large type II error (false negative) and therefore an unacceptable total error [11,12]. The best fit of MS/MS spectra to a peptide sequence has already won a competition for significance between the candidate peptides within the specified parent mass tolerance.

1.2. *ANOVA and multiple comparison procedure (MCP)*

Comparing the mean intensity values of many peptides or proteins between treatments is a valid use of a multiple comparison procedure (MCP). When multiple means are compared between controls and treatments, the significance value accepted must be adjusted to compensate for the number of mean comparisons performed. Depending on the design and aims of the experiment, and whether the error in the measurements is known, MCP can take the form of a correction for family-wise error rate (FWER), such as the Tukey–Kramer HSD test, or the form of a competition for significance such as a False Discovery Rate [26]. When there is only a single category of data (i.e. chromatography fraction) and when the measurement error about the mean can be calculated and is equal between the treatments, the use of ANOVA with a multiple comparison procedure based on family wise error rate is appropriate [22,27]. One of the most rigorous MCP tests is the Tukey–Kramer Honestly Significant Difference test that is available in SAS [22]. Previously, the analysis of spectral peaks from MALDI MS was accomplished by using ANOVA by breaking the *m/z* scale into 5 *m/z* windows prior to analysis [21]. Here the application of ANOVA to unbiased LC-ESI-MS/MS is taken to its logical conclusion using the many intensity values from MS and MS/MS spectra matched to peptides and proteins.

1.3. *Univariate ANOVA verses multivariate ANOVA*

ANOVA can be univariate (e.g.. comparing the intensity of one peptide *m/z* value) or multivariate (e.g. including the intensity values of many different peptide *m/z* intensity values in the model). The unguarded use of overly powerful multivariate statistics led to the erroneous conclusion that pattern recognition methods could diagnose cancer with complete accuracy [17–19,28]. In contrast, the use of ANOVA to confirm the significance of individual parent ions intensity data prior to entry into a multivariate model (i.e. no garbage in) provided a quantitative means to compare normal versus disease samples [21]. ANOVA of individual ion intensity values followed by multiple ANOVA models were compared with linear discriminant analysis to ensure the significance of the results [21]. The approach of ANOVA is in keeping with traditional statistical theory and practice [29] while the approach of unguarded multivariate analysis was not the appropriate strategy for diagnostic experiments [30]. Essentially the data entering the multivariate model should be highly significant in itself by ANOVA and if so, then increasing the dimensionality of the data provides little additional benefit [21]. In agreement with these results, it has been subsequently concluded that ANOVA performs best [31]. The prerequisites to using ANOVA are that the data has been randomly and independently sampled from a normal population.

1.4. *Random and independent*

Biological experiments made under controlled laboratory conditions are usually analyzed by frequency-based statistics using the null random model and normal distribution followed by Chi-square, ANOVA or Student t-tests. ANOVA models of the results of complex liquid chromatography experiments of intact proteins followed by LC-ESI-MS/MS of the tryptic digests may be used to reveal which fractions are best for the identification of a certain protein of interest. In order to use ANOVA the data must be randomly and independently sampled from a normal population. It has been previously shown that different LC-ESI-MS/MS experiments agree on the

identity of the proteins in the sample but often detect different peptides from the same proteins [23]. Moreover repeating the same chromatography experiment and identifying the proteins by LC-ESI-MS/MS result in different sets of peptides and proteins identified in each run: Hence unbiased LC-ESI-MS/MS apparently makes a random sampling of the peptides generated in each experiment [24]. The peptides are progressively eluted from the HPLC column in order of solubility and are intended to be sampled without replacement in each recording especially if a rotating exclusion list is used. Critically, the results of one experiment must not be contaminated by the results from others experiments in order to achieve independence. Previously we achieved independence by creating a separate, disposable micro-chromatography column, or organic extract, for each intact polypeptide sample and then collecting the tryptic or endogenous peptides using another disposable micro-chromatography column prior to analysis by HPLC-ESI-MS/MS with cleaning between each sub-fraction [24,25,32,33]. Thus under our conditions, the sampling of the ions with respect to the start of each chromatographic run is random and independent where the sampling of an ion from one run, has no effect on the sampling of that ion from another run. Under these conditions LC-ESI-MS/MS data may be considered a random and independent sample of the ions eluting from the end of an HPLC column via an electrospray source over the course of an elution gradient from 95% water to about 65% acetonitrile (ACN).

1.5. *Log normal distribution*

Log transformation has been used to obtain a normal distribution and homogenize variance prior to statistical analysis in a number of applications including GC–MS and LC-ESI-MS/MS of small molecules, isotopic or isobaric tag ratios, spectral counts or MALDI spectra [13,34–37]. The log intensity distribution of ions in LC-ESI-MS/MS spectra from serum, protein standards and noise have been related to the normal distribution and log transformation has resulted in more homogenous variation [27,38]. In the present study, completely filtering out background noise of E3 or less for the randomly and independently sampled parent peptides, and filtering out noise in the MS/MS spectra of E2 or less for fragment ions, followed by log transformation lead to a normal distribution for the measured intensity that was linear over 2 orders of magnitude and thus suitable for analysis by ANOVA. Hence, the results of complex sets of LC-ESI-MS/MS experiments were analyzed in a complete ANOVA model by SQL and SAS. The ANOVA model can account for sources of error attributable to the different peptide sequences within proteins and the distribution of proteins over multiple chromatography fractions.

1.6. *Statistical strategy*

For the purpose of the statistical demonstration, we used the previously published results from the chromatographic separation of the intact, secreted proteins of the CaP cell line that were separated over strong anion exchange (quaternary amine) into a void volume and ten salt fractions. The secreted proteins from the human prostate cancer (CaP) epithelial cell line PC3(AR)6 were thus separated into eleven fractions by strong anion exchange (SAX) chromatography prior to digestion of each fraction with trypsin for peptide analysis by liquid chromatography and tandem mass spectrometry (LC-ESI-MS/MS). The many resulting LC-ESI-MS/MS experiments were distilled into a set of MASCOT generic format (mgf) files that contained the *m*/*z* and intensity values for the parent peptides and resulting CID fragment sets. The fragment *m*/*z* values from the mgf file were fit to human tryptic peptides by the goodness of fit algorithm X!TANDEM [9]. The peptide and protein results of the X!TANDEM algorithm may be stored in an SQL database and examined by a generic Statistical Analysis System (SAS) [10]. The peptide to protein frequency may be used to calculate the probability that the results are the same as random spectra by the Chi square test [11,12]. The parent peptide ions and the resulting fragment ion intensity values from many LC-ESI-MS/MS experiments were linked to the peptide sequence and protein names supplied by the X!TANDEM algorithm to create complete ANOVA models of each protein at the level of the different peptide sequences and their many fragment intensity values. The intensity of ions was compared between peptides, proteins or fractions by ANOVA followed by the Tukey–Kramer HSD test to establish differences between treatments [26]. The capacity to statistically compare the fragment intensity values of peptides from proteins might be applied to quantifying differences in chromatography fractions. However, there are many proteins in each fraction and each protein has many different peptide sequences each with its own unique chemical composition, ionization and fragmentation characteristics. Thus, the ANOVA model is an appropriate approach since the data can be blocked by the nominal variables of peptide sequence as well as parent protein and chromatography fractions to control for all sources of error [22].

1.7. *14-3-3 proteins*

To illustrate the power of the ANOVA approach to automatically account for peptide and fragment intensity values from the different peptides the arbitrarily selected example of the 14-3-3 proteins is presented. The 14-3-3 proteins found in all Eukaryotic cells to date are key regulators of cell division, signalling and apoptosis that function by facilitating the interaction of proteins [39]. The molecule scaffold 14-3-3 proteins are suspected biomarkers and potential therapeutic targets [40–42]. The detection of different 14-3-3 proteins and peptides over the course of a chromatographic separation experiment were analyzed by ANOVA to factor parent and fragment intensity values as well as peptide sequence and chromatographic fractions in the comparison of proteins. To illustrate the automatic calculations performed by SAS, the average intensity of the 14-3-3 proteins in the chromatography fractions were compared and the peptides of 14-3-3 zeta examined at the level of peptides and peptides nested within proteins.

1.8. *Data summary*

A complete ANOVA model summarizing the entire set of LC-ESI-MS/MS experiments was constructed for the measured intensity values of 101,905 fragment ions from 1421 different parent ions matched to 583 peptides from 233 proteins over 11 column fractions as calculated by SQL and SAS. The parent

and fragment m/z and intensity values were matched to the peptide sequences provided by the X!TANDEM algorithm in a Structured Query Language (SQL) database. The parent and fragment intensity data was transformed to log normal and tested by the log probability plot with >99.9% of the sampled distribution above the background noise. The intensity values for every parent peptide and its fragments from each chemically-distinct peptide sequence may be considered separately in the ANOVA model of the proteins over chromatography fractions. There is a need for a classical statistical analysis system that can be used to estimate type I error rate of protein identification and to compare log normal relative quantification using only the widely available and commercially standard SQL and SAS data systems.

2. Materials and methods

2.1. Cell culture

Human prostate cancer (CaP) epithelial cells lines $PC_3(AR)_6$ were cultured in roller flasks as previously described [43]. Briefly, the cells were grown in RPMI supplemented with 8% fetal calf serum (FCS) for two days, washed twice with phosphate-buffered saline and then cultured in chemically defined Chinese hamster ovary (CHO) medium supplemented with glutamine for 14 days. Afterwards, serum-free CHO media was collected. During roller bottle culture the levels of total protein increased linearly. Levels of the human kallikreins hK5 and hK6 were measured and showed steadily increasing amounts over the 14 day culture which was due to cell secretion and not cell death. We confirmed this by analyzing the cell pellet by Western blot where hK5 and hK6 was not present in significant amounts compared to the conditioned media (loading the same amount of total protein). In addition, we also monitored the confluency of the cells in the roller bottle to ensure they did not grow over ~80% and ensured there were not necrotic cells or cell debris by direct inspection with light microscopy. We previously showed agreement between the mass spectrometer and the ELISA assay on human Kallikrein 5 and 6 (HK5 and HK6) and Mac-2-binding protein from this study [43].

2.2. Sample preparation

The secreted proteins from the CaP human prostate epithelial cell line were collected from the growth medium prior to dialysis against 20 mM diethanolamine pH 8.9. Proteins from the cell media were pre-separated by partition chromatography using a SAX column as previously described [43]. Briefly, the proteins were applied to a strong anion exchange chromatography column under low pressure and resolved into 10 fractions at 1 ml per minute over a 10-minute gradient to 1 M NaCl in 20 mM diethanolamine as previously described [43]. An aliquot of each fraction (and the void volume fraction 0) containing ~100 μg of protein was digested with 1 μg trypsin in 200 mM Urea, 50 mM Tris pH 8.8, reduced with 1 mM DTT at 50 °C for 30 min and re-digested.

2.3. LC-ESI-MS/MS

The tryptic peptides were collected over preparative C18 in 5% formic acid and eluted in 2 μl 65% ACN, 5% formic acid before dilution into 0.1% formic acid for immediate injection and separation by micro HPLC for electrospray ionization and tandem mass spectrometry. HPLC grade water and solvents were used for all steps. The proteins were digested with trypsin and the peptides were separated over a 300 μm ID, 15 cm C18 reversed phase column with an Agilent 1100 HPLC pump. The LC-ESI-MS/MS analysis was recorded with Esquire 3000 ion trap (Bruker Daltonics, Bellerica, MA, USA) as previously described [44]. A federated library of ~135,000 human proteins predicted from cDNA and genomic sequences from the NCBI, Swiss Prot, Ensembl, Trembl and other sources was assembled in 2009 and rendered distinct with SQL prior to output in a FASTA format for correlation analysis [10].

2.4. MGF and X!TANDEM parser

The peptide and protein expectation values from X!TANDEM were parsed into an SQL database as previously described [10]. The MS and MS/MS spectra from parent ions greater than E3 in intensity were converted to .mgf files. The MS/MS spectra were correlated against the tryptic peptides of the federated human library by X!TANDEM within −3 to +3 Da for parent ions and within 0.5 Da for the +1 b and +1 y fragment ions and with no modifications considered [11,12]. Correlation of the≥E3 parent intensity values from blank runs into .mgf files for correlation yielded no protein identifications by X!TANDEM and thus noise made no contribution to the ion intensity data. The control of the blank runs with the LC-ESI-MS/MS system indicated that the measured ion intensity of noise and contamination spectra at the base line or from blank runs was≤E3 parent signal intensity and so no spectra from noise entered the data set. The parent and fragment data information in the mgf files were matched to the peptides and proteins from X! TANDEM using the file import and spectra numbers [10]. The MS/MS spectra were manually examined to ensure a goodness of fit.

2.5. Statistical analysis

The parent and fragment *m*/z and intensity data from the mgf files matched to the corresponding peptide identifications from the X!TANDEM correlation algorithm in SQL were analyzed by SAS. The *m*/z values in the spectra were treated as continuous variables and thus compared using a random spectra generator to reflect the true degrees of freedom in *m*/z values from tandem mass spectra. The peptides per protein counted in ordinal bins from 1 up to 133 peptides per protein. The parent and fragment intensity values were treated as continuous variables. The peptide and protein sequences were treated as nominal variables. The nominal and continuous variables were declared using the SAS JMP graphical interface prior automatic calculation of ANOVA models. The tables and graphical results from the "model" and "fit y to x" automatic SAS reports were converted to metafiles for inclusion in figures. The ANOVA results were converted to rich text formats for inclusion as tables.

3. Results

3.1. False positive identification rate

The peptides sequences were identified by X!TANDEM that matched the MS/MS spectra to amino acid sequences using a goodness of fit test. The parent ions with intensity≥E3 counts were matched to proteins that often showed multiple independent peptides correlations. In contrast, the random and false positive spectra showed mostly proteins with one peptide [11,12]. After calculating the expected peptide to protein distribution based on random frequency estimates [11,12], a Chi-square value of ≥750 was obtained showing a low probability that the data set is random or false positive identifications ($p<0.0001$) (Table 1). The results were in agreement with the previous use of Chi square to estimate error rates [10–12,23,45].

3.2. Distribution of peptides and fragments

A total of 103,326 intensity values (1421 parent and 101,905 fragment ions) from the .mgf files were matched to the proteins and peptides from the X!TANDEM .xml files in an SQL database [10]. The 1421 parent peptide intensity values appeared symmetrical after log transformation and approximated the log normal distribution (Fig. 1A). Most of the log transformed data from many ions fell within the diagnostic plot for the log normal distribution (Fig. 1B). With a column loading of 1–5 μg, setting the parent fragment intensity filter to E3, essentially captures about 99.9% of the sampled normal distribution (Fig. 1B). The 101,905 fragment ions intensity values approximated a normal distribution after log transformation (Fig. 1C). The normal diagnostic plot showed that most of the fragment ion intensity values were not far from the predicted normal and that more than 99.9% of the signal intensity distribution was above the background (Fig. 1D). The population of fragment ion intensity values from individual peptides showed a distribution that was often indistinguishable from normal examined on a peptide by peptide basis (not shown). The fragment intensity data were numerous and showed acceptable normality. Thus, the normal distribution analysis and baseline measurements were in agreement that parent ions of less than E3 and fragment ions of less than E2 were similar to noise and not required for the analysis.

Table 1 – The redundant peptide to protein frequency distribution of CaP proteins versus three different control distributions from noise, random spectra and a reversed human library. All results were calculated with the X! TANDEM algorithm with the identical parameters. The expected distributions were re-calculated using the frequency values of Zhu et al. [11,12] such that the total number of protein identified in each category matches the present study. For this analysis noise spectra were correlated from parent ions with intensity values≤E3 from blank runs.

Count	Noise	Random	CaP [a]	Reversed
>18	9	0	410	2
18	0	0	14	0
17	0	0	11	2
16	0	0	15	0
15	0	0	14	0
14	0	0	21	0
13	0	0	24	2
12	2	0	24	1
11	15	0	29	2
10	9	0	25	0
9	11	0	35	1
8	49	0	39	5
7	11	2	44	7
6	13	1	52	9
5	45	1	61	9
4	54	5	76	13
3	52	29	94	34
2	215	152	109	212
1	937	1230	324	1122

[a] The probability that the CaP peptide to protein distribution was the same as that of any of the controls was less than 1/10,000 or $p(x)<0.0001$ by Chi square analysis.

3.3. Range of parent and fragment ion intensity values

The range of identified peptide intensity values sampled by the ion trap were found to span about four orders of magnitude from E3 to about E7 (Fig. 2A). The range of fragment intensity values also covered five orders of magnitude from about E2 to about E7 arbitrary count values (Fig. 2B). In both cases, the intensity values showed a linear increase over about 2 orders of magnitude from the inflections. Hence, the potential exists to directly quantify the measured ion intensity values from the unbiased LC-ESI-MS/MS experiment of CaP proteins with a linear range of about two orders of magnitude in this experiment.

3.4. Parent and fragment ion m/z distribution

The parent m/z values appeared to show a consistent increase from about 500 to 800 m/z with the highest value of about 1100 m/z with relatively few peptides sampled at extreme m/z values (Fig. 2C). The parent ion [M+H] values from the MS spectra of peptides ranged from about 800 to 2200 Da (Fig. 2D). The fragment m/z values showed a linear range from 200 to about 1100 m/z with a sharp inflection to a maximal value of 1900 m/z (Fig. 2E). Most or all of the identified peptides were apparently [M+2H] ions (Fig. 2F). Hence the best fit of parent ions of≥E3 intensity by X!TANDEM was entirely restricted to 2^+ ions.

3.5. The effect of sample treatment, protein, peptide and fragments

Whole models of log-transformed parent (1421) and fragment (101,905) intensity values were fit at the level of the sample treatments, proteins, and the many peptide sequences associated with each protein. The many replicated parent peptides each with the many fragment intensity values have the cumulative effect of building confidence in the results of the ANOVA. The results of such a whole model analysis show that there is a significant probability of real variation between sample treatments, proteins and peptides (Table 2). There were significant differences between proteins even after the effect of the

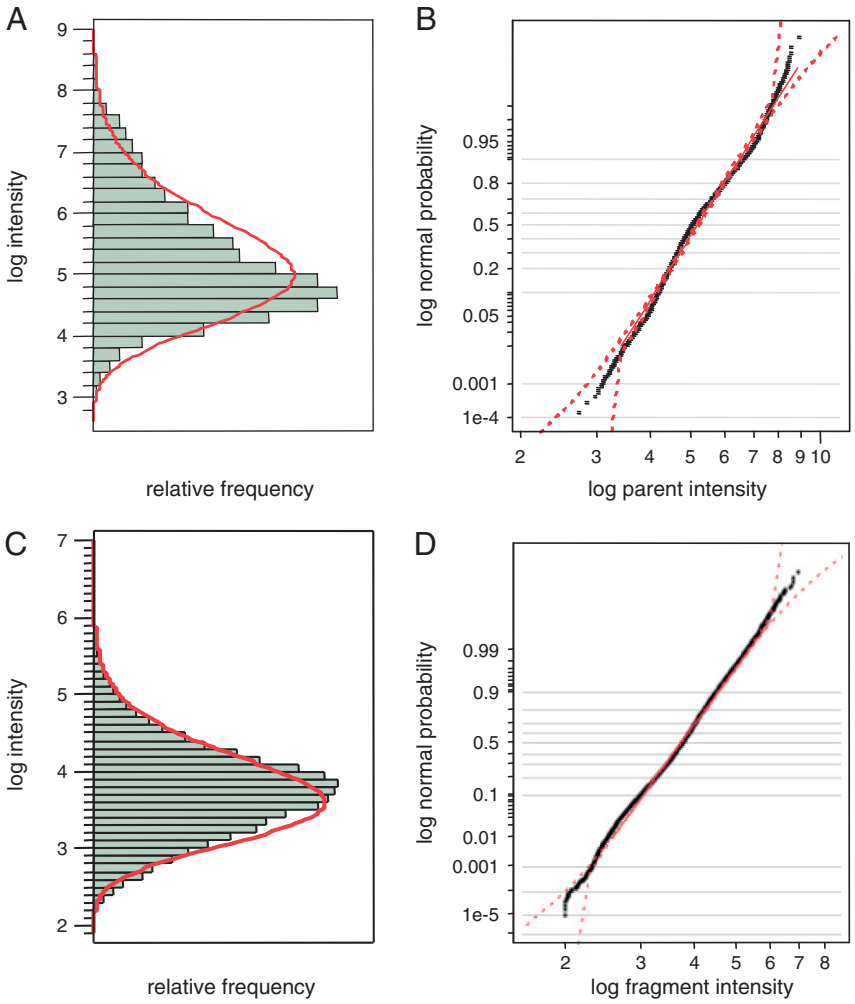

Fig. 1 – Fit of the normal function to parent and fragment intensity values (arbitrary counts) with Log transformation. Panels: A, the normal function fitted to the log transformed peptide intensity distribution; B, the distribution of the log transformed peptide intensity data plotted along the normal function diagnostic plot; C, the normal function fitted to the distribution of the log fragment intensity values; D, the distribution of the log fragment intensity data plotted along the normal function diagnostic plot. The distributions from 1421 peptide and 101,905 fragments are shown.

different peptide chemical compositions have been taken into account ($p \leq 0.0001$) (Table 2) and so it was possible to compare average protein intensity values summed over all peptides.

3.6. *General analysis of secreted proteins and peptides*

The general description of the ion trap data made with SAS showed that most of the calculated peptide mass values closely corresponded to the predicted [M+H] of the correlated peptide (Fig. 3A). Peptide [M+H] values from about 1200 to 2000 Da showed the lowest expectation values (best identifications) but much of the significant data was still collected in the range of 1000 to 1700 Da (Fig. 3B). Most of the peptide correlation results were obtained from peptides of about 7 to 16 amino acids with the best scores observed for peptides of 12 to 14 residues (Fig. 3C). Fragment intensity values showed a mean of about E4 intensity and there was little trend in intensity versus [M+H] indicating the measured intensity is not much affected by the peptide mass (Fig. 3D). The individual peptide expectation values were as low a E-12 with more than 1200 peptides showing expectation values of false positive identification of E-2 (1/100) or less (Fig. 3E). The product of the log peptide expectation values yield the total protein expectation of type I error (false positive) as low as E-346 with some 900 protein identifications of about E-10 or less (Fig. 3F). Most of the protein expectation scores were derived from polypeptides of 300 amino acids in length or less and there was no trend towards the better identification of longer proteins (Fig. 3G). Examining the distribution of peptide and protein expectations in a scatter plot showed many identifications where the peptide and/or protein expectations values of type 1 error were less that E-4 (1/10,000) as calculated by X!TANDEM (Fig. 3H). The calculation of type I error rates by the goodness of fit of many proteins by X!TANDEM seems acceptable and is not necessarily far different

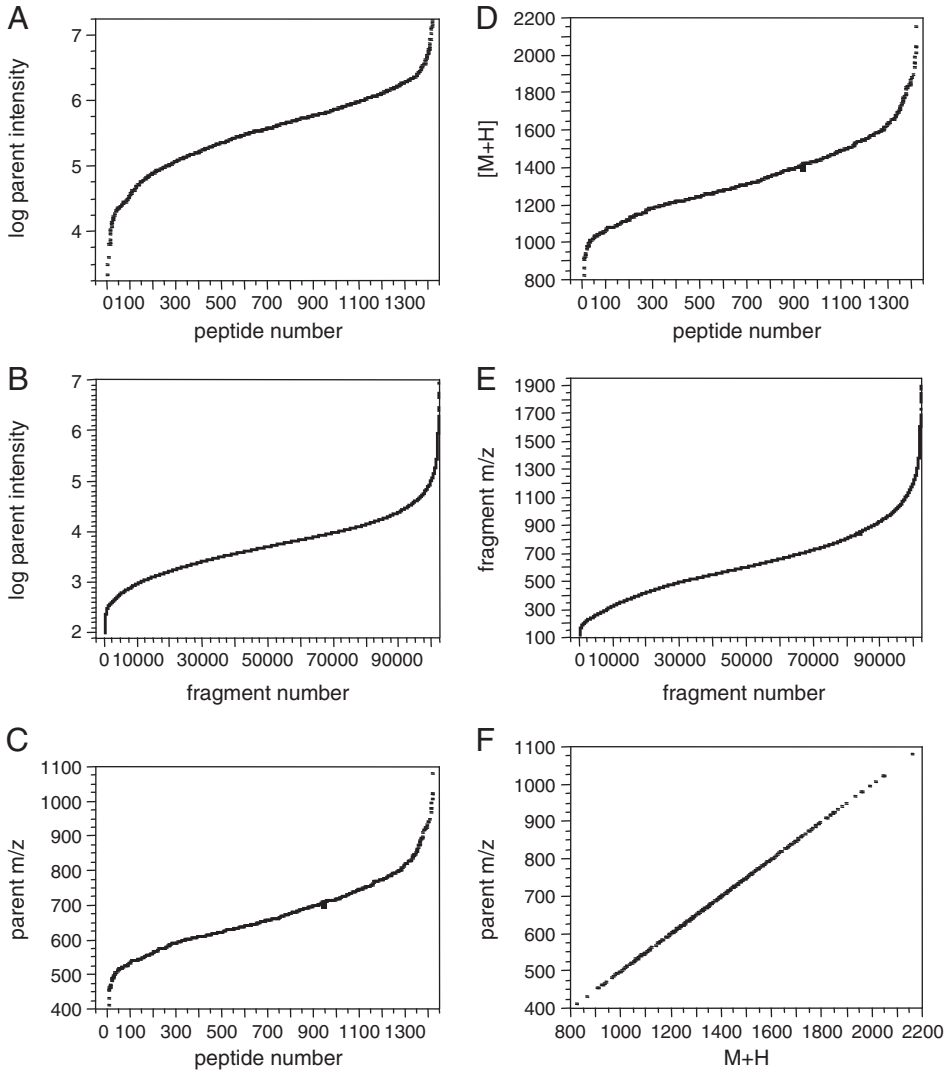

Fig. 2 – The distribution of intensity values (arbitrary counts) of parent peptides and fragments. Panels: A, the log parent ion intensity distribution versus protein number; B, the log fragment ion intensity distribution versus fragment number; C, the parent peptide m/z values; D, the parent ion [M+H] values; E, the fragment ion m/z values; F, the relationship between m/z and [M+H]. The results of 1421 parent and 101,905 fragment ions are shown.

from the 1/10,000 E-4 or $p<0.0001$ for the entire experiment calculated by Chi square.

The data indicate that successful correlations from the X! TANDEM were strongly non-random in agreement with the Chi Square result. Furthermore, of these 2^+ ions the best correlations scores were obtained from peptides of about 12 to 14 amino acids. Previously endogenous polypeptides of greater than ~6500 resisted CID fragmentation presumably because the momentum of the molecule is too great to be fragmented by N_2 gas [25]. In the present study of tryptic peptides with the lighter He_2 gas few peptides of greater than 2000 [M+H] were observed. Perhaps at a charge state of 2^+ peptides of much less that 12 amino acids may not provide enough b and y fragment ions for optimal statistical confidence by goodness of fit.

3.7. *Comparison of fragment intensity by column fraction*

It might be useful to compare the ion current intensity of peptides from proteins in chromatography fractions to develop separation protocols. The analysis of the LC-ESI-MS/MS results from the many SAX factions showed significant differences on average fragment ion intensity of proteins at the level of sample treatments by the Tukey–Kramer HSD test (Fig. 4A). About 10 different pair-wise difference ranges were identified among 11 column fractions as illustrated. As expected, the void volume (fraction 0) had low measured ion intensity values compared to those of ~200–300 mM NaCl that showed greater ion intensity values than the 600–700 mM NaCl fractions but thereafter declined to the final fraction (10) of ~1 M NaCl. The trend in measured intensity values closely

Table 2 – The whole model of log fragment intensity values (arbitrary counts) for a set of LC-ESI-MS/MS experiments at the level of sample fractions, proteins and peptides. A model total of 101,905 log transformed fragment ions intensity values from 1421 parent peptides were computed from 11 chromatography fractions with 583 different peptides from 233 proteins identified from the parent and fragment *m/z* values by X!TANDEM. The probability that the transformed LC-ESI-MS/MS data fails to show significant variation apparently approaches zero. The effect of each chemically distinct peptide sequence, parent proteins and chromatography columns was modeled by ANOVA. Note the differences between sample treatments, proteins and peptides achieved F values of 2561, 63, and 111 respectively indicating the approach shows great statistical power. Note that only parent ions with intensity values≥E3 that were successfully correlated by X!TANDEM were accepted and so no noise spectra entered the analysis.

Analysis of variance

Source	DF	Sum of squares	Mean square	F ratio
Model	623	20,309.137	32.5989	236.3603
Error	101,281	13,968.730	0.1379	**Prob>F**
C. total	101,904	34,277.867		0.0000*

Effect tests

Source	Nparm	DF	Sum of squares	F ratio	Prob>F
Sample treatment	11	11	3886.2862	2561.611	0.0000*
Protein accession	232	22	193.0022	63.6078	<.0001*
Peptide sequenc	582	380	5856.5406	111.7451	0.0000*

* The probability associated with the whole model or effect is shown.

matched the trend in protein assays of the SAX column (not shown).

3.8. *Comparison of protein fragment intensities*

It was possible to provide relative quantification of the ion currents from different peptides by MALDI mass spectrometry with ANOVA and Tukey–Kramer means testing [21]. Here, the parent (1421) and fragment (101,905) ion intensity values were organized under all of the 11 fractions, 233 protein and 583 peptide sequences to permit complete statistical modeling. The intensity values from any or all proteins of interest may be rapidly exported from SQL database and can be instantly queried by SAS or run as a batch file in total and stored. To illustrate the types of statistical analysis that might be instantly performed on any or all proteins, the example of the 14-3-3 proteins was selected. After specifying all 14-3-3 protein data or 14-3-3 zeta data from the database in SQL, the parent and fragment ion intensity data of the 14-3-3 proteins were examined in detail by SAS. Since there were significant differences between 14-3-3 proteins even after the effect of the different peptide chemical compositions have been taken into account ($p \leq 0.0001$), it is possible to compare average protein intensity values taking the effect of all peptide sequences into account separately in the ANOVA (Table 3).

3.9. *Comparison of 14-3-3 proteins*

The average peptide intensity of all proteins that contain the query term 14-3-3 in the descriptor field is shown in Fig. 4B. The analysis of the many LC-ESI-MS/MS results from the SAX columns fractions showed significant differences in the ion intensity of peptide fragments over the 14-3-3 proteins by the Tukey–Kramer HSD test (Fig. 4B). Six different 14-3-3 proteins were detected which could be resolved into four different pair-wise difference ranges. See overlapping circles for comparison of all means by Tukey–Kramer at $p \leq 0.05$ at the side of the figure where proteins with circles that resolved are significantly different (Fig. 4B). Similarly, the mean fragment ion intensity values for all the 233 proteins detected in this experiment are provided with significance limits in Supplemental Table 1. About 60 different pair-wise difference ranges were observed among the proteins identified by the Tukey–Kramer HSD Test (Supplemental Table 1).

3.10. *Comparison of 14-3-3 zeta over column fractions*

The mean fragment ion intensity from all the peptides of 14-3-3 zeta (P63104) were compared to the chromatography fractions and showed significant differences in fragment ion intensity. The greatest 14-3-3 zeta intensity values were observed in fraction 7 with much lower intensity observed in fractions 8 and 9 that followed (Fig. 4C). There was no 14-3-3 zeta identified in any of the other fractions. Similarly, many other proteins were discretely detected by multiple observations of peptides in only one, or a few adjacent, salt fractions consistent with successful chromatography. Hence it was possible to monitor and observe the efficacy of chromatographic separations using statistical analysis of the ion intensity values.

3.11. *Comparison of fragment intensity values for the peptides of 14-3-3 zeta*

To take into account the effect of the different chemical composition and arrangement of amino acids, each different sequence may be declared as a separate nominal variable in the ANOVA model. The fragment intensity values for 14-3-3 zeta peptides ranged from a maximum of ~E 4.2 for the peptide VLAEVAAGDDK to a minimum of ~E 3.5 for the peptide GIVDQSQQAYQEAFELSK. The average intensity of the six different 14-3-3 zeta peptides varied by about one order of magnitude and were resolved by the Tukey–Kramer HSD test into 4 different pair-wise comparison ranges (Fig. 4D). Example MS/MS spectra for the 14-3-3 zeta polypeptide are shown in Fig. 5 alongside a noise spectra typical of data collected from blank runs of less than E3 intensity. The spectra matched to 14-3-3 zeta show rich fragmentation patterns evenly spaced along the backbone of the peptide with few ions that are not accounted for by the predicted 1^+ b and y

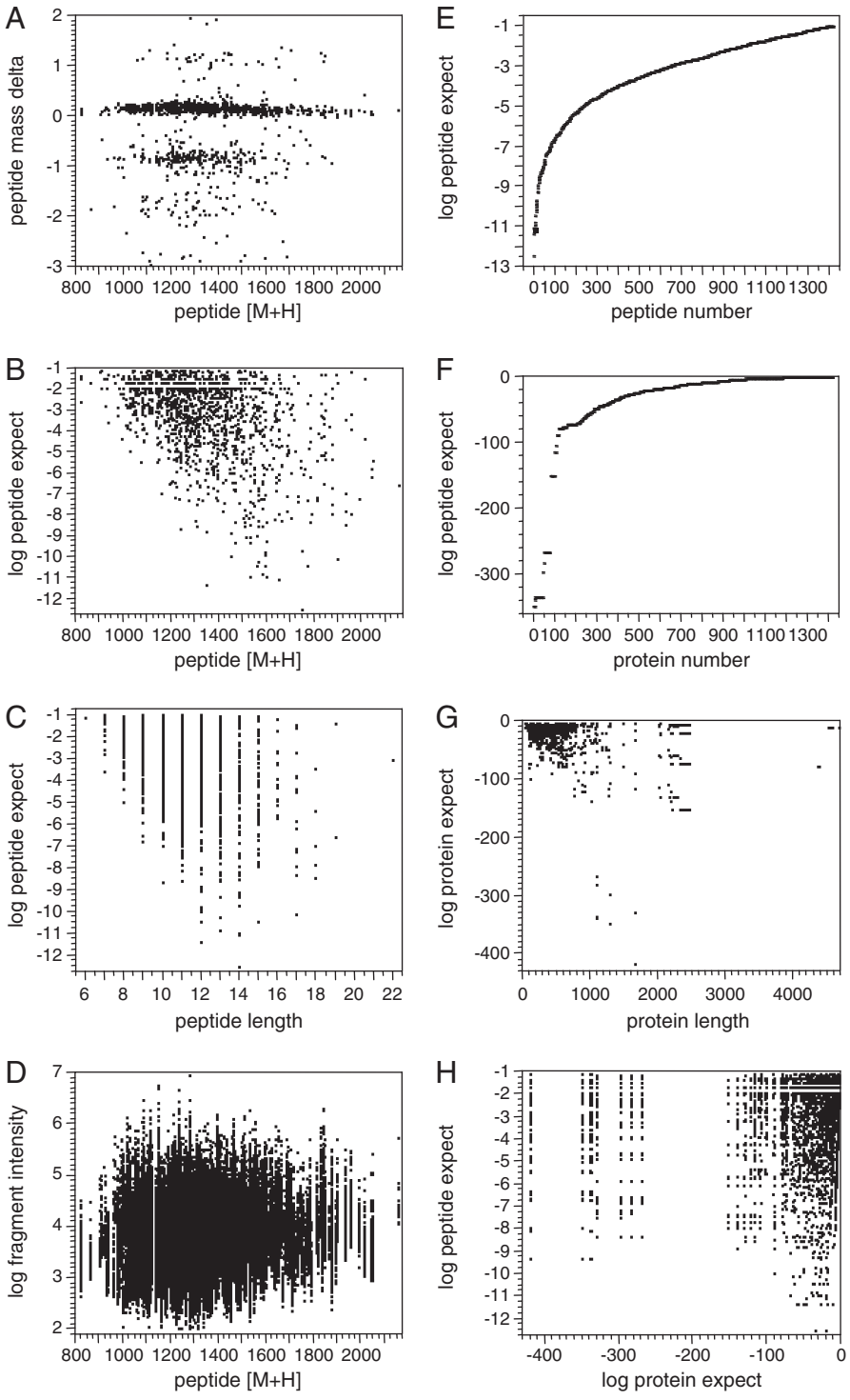

Fig. 3 – General statistical description of the secreted Cap protein dataset. Panels: A, delta mass versus [M+H]; B, Log peptide expectation value versus [M+H]; C, Log peptide expectation value versus peptide length in amino acid residues; D, log fragment intensity versus peptide [M+H]; E, log peptide expectation value versus peptide number; F, log protein expectation value versus protein number; G, log protein expectation number versus protein length; H, log peptide expectation number versus protein expectation number. The results of 101,905 fragment ions from 1421 peptide ions redundantly correlated to 583 distinct peptides from 233 proteins are shown.

fragment ions of the indicated peptides. Noise ions of less that E2 were filtered out of the MS/MS data set using SAS prior to statistical analysis. Similarly, it was possible to statistically analyze the fragment ion intensity values associated with every one of the 583 peptide sequences in the 11 chromatography fractions resulting in a large number of pair-wise

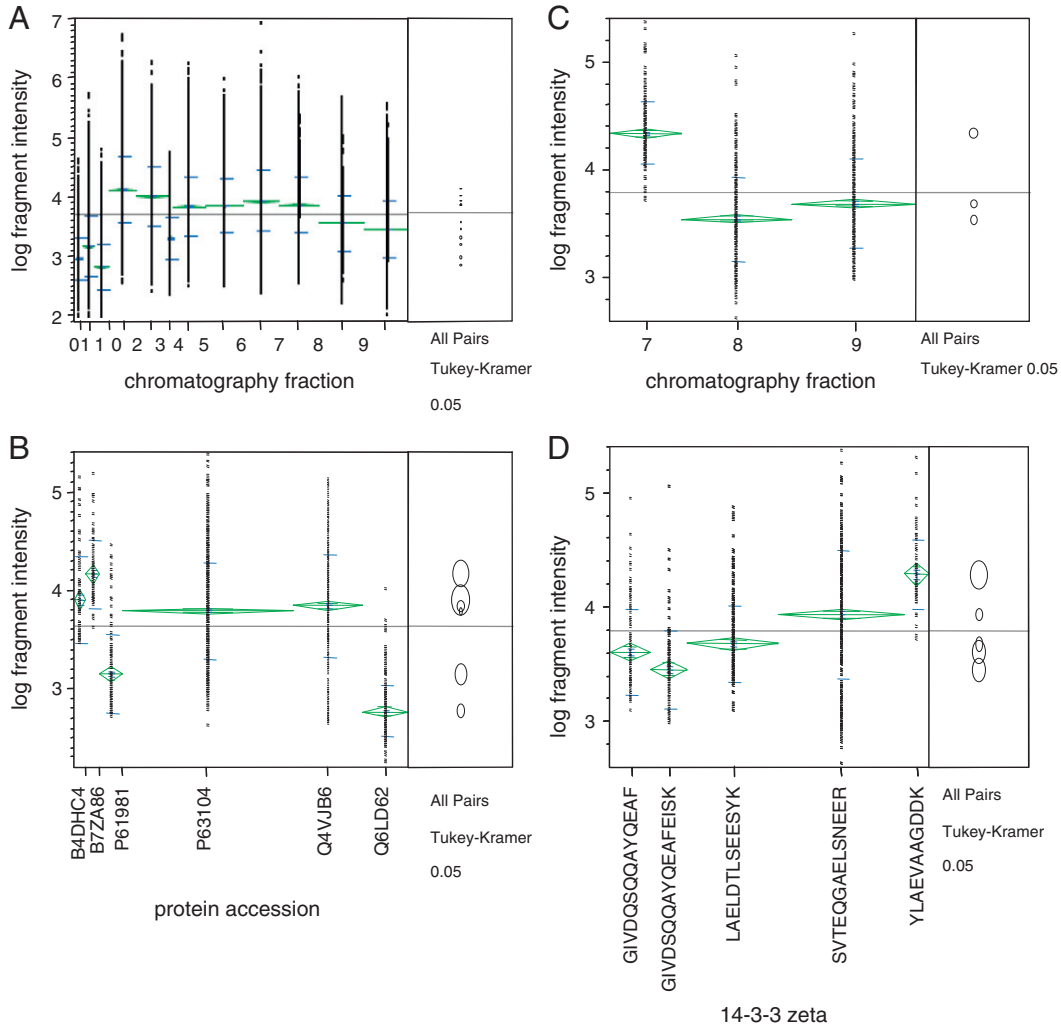

Fig. 4 – Comparison of mean intensity values at the level of chromatography fractions, proteins, peptides using the automatic calculation of statistical differences between sample treatments by SAS. Panels: A, the fractions 1 to 10 correspond with 0 to 1000 mM NaCl where fraction 0 is the void volume. Note there are two fractions of ~400 mM NaCl (fraction 4) as an injection failure was observed on the first attempt; B, comparison of all 14-3-3 proteins at the level of log fragment intensity values averaged over all sample fractions with testing of all pairs by the Tukey–Kramer HSD test; C, all 14-3-3 protein zeta fragment log intensity values compared between chromatography fractions with the Tukey–Kramer HSD test; D, all 14-3-3 zeta peptides compared at the level of fragment intensity values over all fractions with testing of all pairs by the Tukey–Kramer HSD test. ANOVA of the many fragment ion intensity values showed an expectation value of $p < 0.0001$ at the level of the sample fractions, proteins and peptides prior to means testing. The ANOVA tables are in the supplemental data. The separated circles indicate statistically different pairs at the level of $p \leq 0.05$ after correction for multiple comparisons.

significance ranges by the Tukey–Kramer test (Supplemental Table 2).

4. Discussion

4.1. *The analysis of unbiased LC-ESI-MS/MS experiments*

A standard means to convey the qualitative and quantitative parameters between proteomics experiments will be absolutely crucial to comparing and contrasting the results from different laboratories. The use of the industry standard SQL and SAS data systems may provide the means to compare the results of different studies; For example the similar analysis from different instruments and algorithms provided in Bowden et al. (2009), Williams et al., (2010) and Tucholska et al. (2010) and the present experiment provide a complete graphical description of the experimental, instrumental, and search parameters used in each study that include the degree peptide digestion, the accuracy of the parent and fragments, the typical charge state and the size of the peptides correlated by the algorithm and the statistical confidence in the results.

Table 3 – The ANOVA analysis of 14-3-3 zeta at the level of sample, peptide and ion type. The ANOVA table indicates significant effects at the level of samples, the chemical composition of the peptides sequence and parent versus fragment or ion type. The model takes into account the effect of the different chemical compositions of the peptides on ion intensity values (arbitrary counts). See Table 2 for other details.

Analysis of variance

Source	DF	Sum of squares	Mean square	F ratio
Model	7	150.29198	21.4703	157.2872
Error	1039	141.82735	0.1365	Prob>F
C. total	1046	292.11933		<.0001*

Tests

Source	Nparm	DF	Sum of squares	F ratio	Prob>F
Sample treatment	2	2	54.778774	200.6494	<0.0001*
Peptide sequence	4	4	5.241488	9.5995	<0.0001*
Peptide or fragment Ion	1	1	46.361702	339.6369	<0.0001*

All of these important parameters may be graphically presented on the common SQS and SAS system that does not require proteomic specific software [46]. Thus it will be necessary to routinely collect all this information into a relational database and to include these blocking variables in ANOVA analysis prior to means comparisons. Here, the peptide and protein results of the X!TANDEM algorithm were used to organize the peptide and fragment intensity values to examine the separation efficacy from a set of chromatography fractions. In this paper, we considered that parent and fragment ion intensity values were randomly and independently sampled without replacement as they eluted from the end of the reversed phase column in LC-ESI-MS/MS. Furthermore, this study shows that after transformation the intensity values of the identified peptides show a log normal distribution. Hence the parent and fragment ion intensity values were randomly and independently sampled from a normal population of ions that have been unambiguously assigned to peptides and proteins for ANOVA. A complete ANOVA model was constructed for the hundreds of thousands of measured intensity values from thousands of peptides from hundreds proteins over many sample treatments by classical ANOVA without having to take elution times into account. This observation has profound consequences for proteomics. It was possible to quantify and compare the measured ion intensity values from peptide and proteins over many chromatography fractions using only classical statistical methods exploiting the automatic properties of widely available SQL and SAS data system and without proteomic-specific software systems [46]. Fragment intensity values were shown to increase linearly from about E2.5 to E5 and so most of the data for the ANOVA analysis shown fell within the linear range. The normal distribution analysis indicated sufficient sample was loaded on the column to ensure that>99% of mean average signal intensity distribution was above the baseline intensity cut off at about E3 counts.

4.2. *Calculation of the false positive rate*

The mass spectrometry data in this study was previously shown to be in agreement with ELISA measurements of human Kallikrein 5 and 6 as well as Mac-2-binding protein. Numerous studies have now shown good agreement between biochemical methods and LC-ESI-MS/MS data [21,42,43,47–74]. Agreement of biochemical studies and mass spectrometry are consistent with previous studies controlled by noise or random spectra which showed that identified peptide ions with intensity values greater than E3 counts had low type I error rates [11,12]. The peptide to protein distribution is a key descriptive feature of a set of protein LC-ESI-MS/MS experiments where the confidence of identification increases with number of different peptides correlated to a protein [24,75,76]. The peptide to protein distributions can be compared to that of random expectation to yield an estimate of type I (false positive) error rate for a set of experiments. Frequency based statistical approaches, such as the Chi square test, may be utilized to estimate the probability of agreement between different experimental results [10–12,23,24,45]. Random number or noise generators were invented, and commonly used for, making customized models of random expectation for the experimental variable under measurement [77]. Random spectra are a source of false positive results that typically show about 89% of the proteins with only one peptide, about 10% of proteins show two peptides, and only one to a few percent of proteins have several peptides or more [11,12,78]. There was little probability that the authentic spectra were the same as random, noise or false positive data ($p<0.001$) in agreement with previous results [10–12,23,24,79]. It is important to note that the E3 cut off of parent ion intensity employed still captured 99.9% of the normal distribution of identified parent ions and so permitted complete sampling of the intensity distribution. However there were no proteins identified from blank runs with signals greater than E3 indicating that there were no noise spectra in the present data sets.

4.3. *Comparison of measured ion intensity values from chromatography fractions*

The efficacy of chromatographic separation was quantified with a simple ion trap by the measured ion current from tryptic digests of the column fractions. The discovery of secreted proteins by fractionation over partition chromatography [21,32] was quantified by ANOVA of the measured fragment ion intensity values of peptides from proteins. The ion currents between columns fractions showed similar trends to proteins assays of quaternary amine chromatography fractions with the lowest values in the 0 mM NaCl void volume, the highest protein concentrations in the between 200 and 600 mM NaCl and declining thereafter with little further elution beyond 1 M NaCl.

The chromatography fractions, protein accession numbers and peptide sequences provided an organization and framework to quantify the measured ion intensity values over the many LC-ESI-MS/MS experiments by statistical analysis. In

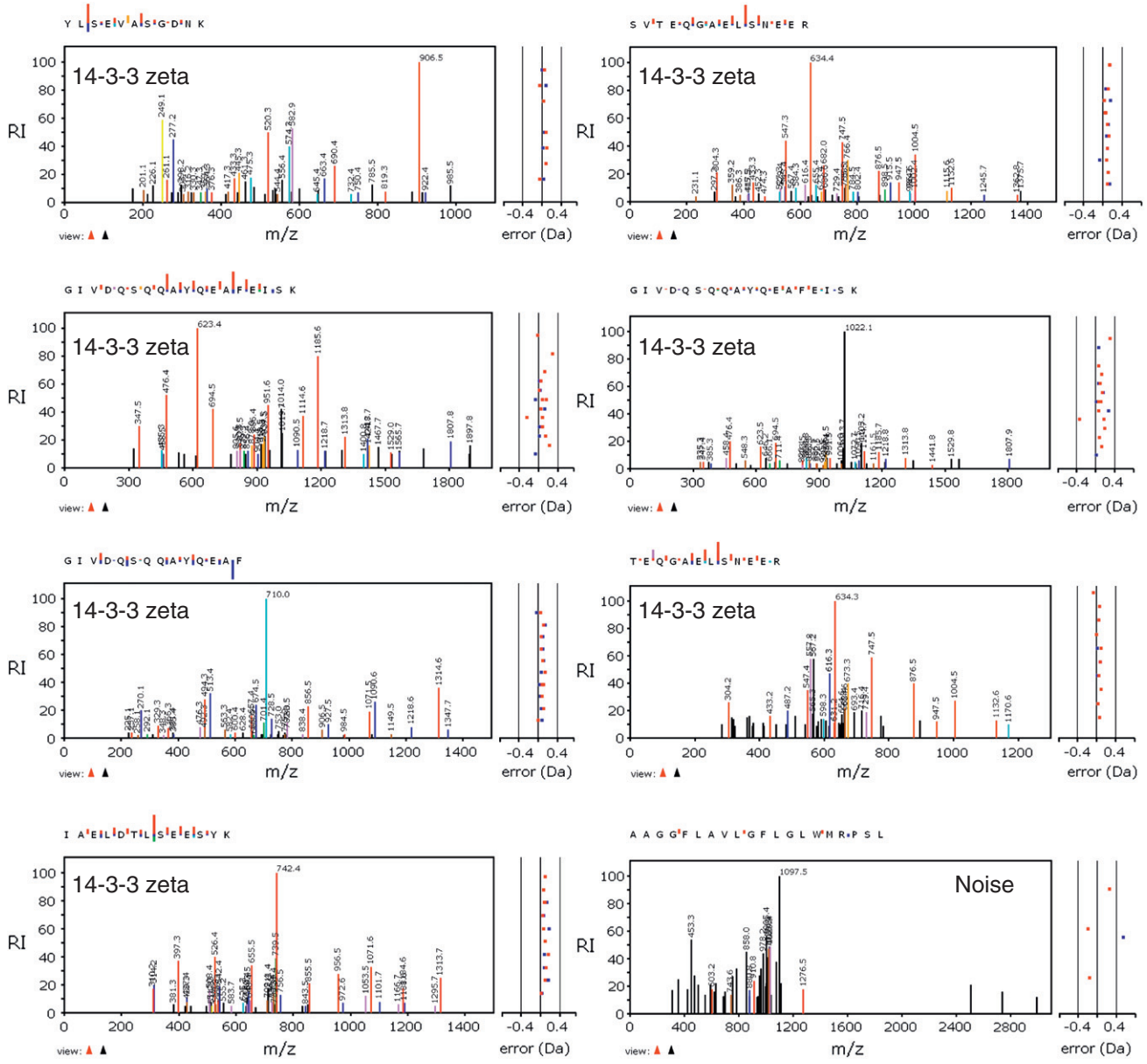

Fig. 5 – Tandem mass spectra from 2^+ peptides of 14-3-3 zeta. The peptide sequence and fitted b and y ions are indicated alongside each MS/MS spectra as provided by X!TANDEM. An example of the fit of a noise spectra from a parent ion of ≤E3 is shown for the sake of comparison.

this paper, ANOVA results of both the parent and fragment ion results together were analyzed first, but similar results can be obtained from the fragment ions alone. Matching the log fragment intensity values to the treatment, protein and peptides in SQL permitted the rapid assembly of ANOVA models of LC-ESI-MS/MS in SAS. There was no requirement for isotopic labelling, to compare or manipulate the source chromatograms, or to keep track of the chromatographic retention times [80]. The statistical model exploits the inherent structure of proteomic data where for every protein sequence there is a family of peptide(s) that might be redundantly detected: For each parent ion detected a sub family of many fragments and intensity values were observed. Statistical power for proteins and peptides builds rapidly within one experiment and true statistical power for treatments should build with replication of the whole experiment [11,12]. The approach is appropriate for the instant technical appraisal of chromatography fractions. The control and treatments might be arranged in replicate blocks for LC-ESI-MS/MS sampling to apply the method to biological study [22]. The ANOVA analysis indicated that there was significant variation in the ion currents between different peptides, proteins, or fractions. The method as shown is entirely satisfactory to determine which chromatography fraction(s) contain a

detectable amount of a particular protein without the use of isotopic labels, a result that is of great practical importance to proteomics [20].

4.4. *Analysis of complex experiments*

Complex experiments involving the analysis of multiple treatment fractions prior to digestion and separation of peptides by subsequent reversed-phase LC-ESI-MS/MS will require models that take into consideration the effect of protein fraction, the source protein, and the chemical differences in peptide sequence on measured ion intensity in calculated statistical significance. The capacity to completely examine the quality and quantity LC-ESI-MS/MS data statistically permit the interpretation of complex sets of LC-ESI-MS/MS experiments. Summarizing all of the peptide and protein data from a large set of treatments by automatically-generated statistical summary tables and graphical comparison plots should permit the relative quantification of proteins between samples even in the presence of minor variations in chromatographic separations. The application of statistical analysis to all the ion intensity of values of proteins and peptides from entire sets of samples should provide the capacity to completely analyze complex sets of LC-ESI-MS/MS experiments. The SQL and SAS data systems work well together and previously compared the blood peptides and proteins from all laboratories internationally [10,23]. Thus, there is every good reason to believe that the widely available SQL and SAS data systems that are already owned by most university, government and private research institutions, will be sufficient to analyze large proteomic experiments. Moreover these two data systems are already widely used and proven in clinical, laboratory, agricultural and engineering research. The SQL and SAS data system are so popular because they permit the researcher to rapidly make a custom-fit database and statistical solution for each experiment using point-and-click menus or common words and operators arranged in simple phrases. Thus the complete statistical analysis of many LC-ESI-MS/MS experiments was accomplished with the generic and automatic features of the SQL and SAS data systems with no special modification.

Here it is shown that the log transformed peptide and fragment ions had a normal distribution, have been thoroughly sampled and quantitatively analyzed at the level of thousands of peptides nested within hundreds of proteins simultaneously using a simple ion trap under conditions that showed low false positive rates. Hence the false positive rate of both identification and quantification of complex sets of LC-ESI-MS/MS experiments can be determined using only the goodness of fit tests and ANOVA with Tukey–Kramer HSD. The SQL and SAS analysis shown here provides qualitative and quantitative information about complex LC-ESI-MS/MS experiments without the requirement for accurate mass values, keeping track of retention times, or the use of heuristic or multivariate statistics. The approach utilizes only standardized software packages common to all fields of science [46]. All of the observations were consistent with the conclusion that tandem mass spectra collected with a high signal to noise ratio under controlled conditions area reliable means to identify and quantify peptides and proteins across treatments.

Supplementary materials related to this article can be found online at doi:10.1016/j.jprot.2011.11.002.

Acknowledgement

This work was supported with a discovery grant to JGM from The Natural Science and Engineering Research Council of Canada.

REFERENCES

[1] Griffin NM, Yu J, Long F, Oh P, Shore S, Li Y, et al. Label-free, normalized quantification of complex mass spectrometry data for proteomic analysis. Nat Biotechnol 2011;28:83–9.
[2] Simpson KL, Whetton AD, Dive C. Quantitative mass spectrometry-based techniques for clinical use: biomarker identification and quantification. J Chromatogr B Analyt Technol Biomed Life Sci 2009;877:1240–9.
[3] Schenk S, Schoenhals GJ, de Souza G, Mann M. A high confidence, manually validated human blood plasma protein reference set. BMC Med Genomics 2008;1:41.
[4] Mueller LN, Brusniak MY, Mani DR, Aebersold R. An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data. J Proteome Res 2008;7:51–61.
[5] States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, et al. Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study. Nat Biotechnol 2006;24:333–8.
[6] Mortensen P, Gouw JW, Olsen JV, Ong SE, Rigbolt KT, Bunkenborg J, et al. MSQuant, an open source platform for mass spectrometry-based quantitative proteomics. J Proteome Res 2010;9:393–403.
[7] Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383–92.
[8] Craig R, Beavis RC. A method for reducing the time required to match protein sequences with tandem mass spectra. Rapid Commun Mass Spectrom 2003;17:2310–6.
[9] Craig R, Beavis RC. TANDEM: matching proteins with tandem mass spectra. Bioinformatics 2004;20:1466–7.
[10] Bowden P, Beavis R, Marshall J. Tandem mass spectrometry of human tryptic blood peptides calculated by a statistical algorithm and captured by a relational database with exploration by a general statistical analysis system. J Proteomics 2009;73:103–11.
[11] Zhu P, Bowden P, Tucholska M, Marshall JG. Chi-square comparison of tryptic peptide-to-protein distributions of tandem mass spectrometry from blood with those of random expectation. Anal Biochem 2011;409:189–94.
[12] Zhu P, Bowden P, Tucholska M, Zhang D, Marshall JG. Peptide-to-protein distribution versus a competition for significance to estimate error rate in blood protein identification. Anal Biochem 2011;411:241–53.
[13] Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. Statistical analysis of membrane proteome expression changes in *Saccharomyces cerevisiae*. J Proteome Res 2006;5:2339–47.
[14] Griffin TJ, Gygi SP, Rist B, Aebersold R, Loboda A, Jilkine A, et al. Quantitative proteomic analysis using a MALDI quadrupole time-of-flight mass spectrometer. Anal Chem 2001;73:978–86.

[15] Dicker L, Lin X, Ivanov AR. Increased power for the analysis of label-free LC–MS/MS proteomics data by combining spectral counts and peptide peak attributes. Mol Cell Proteomics 2011;9:2704–18.
[16] Freue GV, Sasaki M, Meredith A, Gunther OP, Bergman A, Takhar M, et al. Proteomic signatures in plasma during early acute renal allograft rejection. Mol Cell Proteomics 2010;9: 1954–67.
[17] Ardekani AM, Liotta LA, Petricoin III EF. Clinical potential of proteomics in the diagnosis of ovarian cancer. Expert Rev Mol Diagn 2002;2:312–20.
[18] Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, et al. Use of proteomic patterns in serum to identify ovarian cancer. Lancet 2002;359:572–7.
[19] Petricoin III EF, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, et al. Serum proteomic patterns for detection of prostate cancer. J Natl Cancer Inst 2002;94: 1576–8.
[20] Eckel-Passow JE, Mahoney DW, Oberg AL, Zenka RM, Johnson KL, Nair KS, et al. Bi-Linear Regression for O Quantification: Modeling across the elution profile. J Proteomics Bioinform 2010;3:314–20.
[21] Marshall J, Kupchak P, Zhu W, Yantha J, Vrees T, Furesz S, et al. Processing of serum proteins underlies the mass spectral fingerprinting of myocardial infarction. J Proteome Res 2003;2:361–72.
[22] Florentinus AK, Jankowski A, Petrenko V, Bowden P, Marshall JG. The Fc receptor-cytoskeleton complex from human neutrophils. J Proteomics 2011;75:450–68.
[23] Bowden P, Pendrak V, Zhu P, Marshall JG. Meta sequence analysis of human blood peptides and their parent proteins. J Proteomics 2010;73:1163–75.
[24] Tucholska M, Bowden P, Jacks K, Zhu P, Furesz S, Dumbrovsky M, et al. Human serum proteins fractionated by preparative partition chromatography prior to LC-ESI-MS/MS. J Proteome Res 2009;8:1143–55.
[25] Williams D, Ackloo S, Zhu P, Bowden P, Evans KR, Addison CL, et al. Precipitation and selective extraction of human serum endogenous peptides with analysis by quadrupole time-of-flight mass spectrometry reveals posttranslational modifications and low-abundance peptides. Anal Bioanal Chem 2010;396:1223–47.
[26] Benjamini Y, Hochberg Y. Controlling false discovery rate: a practical approach to multiple testing. J R Stat Soc 1995;57: 289–300.
[27] Anderle M, Roy S, Lin H, Becker C, Joho K. Quantifying reproducibility for differential proteomics: noise analysis for protein liquid chromatography–mass spectrometry of human serum. Bioinformatics 2004;20:3575–82.
[28] Baggerly KA, Morris JS, Wang J, Gold D, Xiao LC, Coombes KR. A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples. Proteomics 2003;3:1667–72.
[29] Sorace JM, Zhan M. A data review and re-assessment of ovarian cancer serum proteomic profiling. BMC Bioinformatics 2003;4:24.
[30] Baggerly KA, Morris JS, Coombes KR. Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. Bioinformatics 2004;20:777–85.
[31] Eckel-Passow JE, Oberg AL, Therneau TM, Bergen III HR. An insight into high-resolution mass-spectrometry data. Biostatistics 2009;10:481–500.
[32] Marshall J, Jankowski A, Furesz S, Kireeva I, Barker L, Dombrovsky M, et al. Human serum proteins preseparated by electrophoresis or chromatography followed by tandem mass spectrometry. J Proteome Res 2004;3:364–82.
[33] Tucholska M, Florentinus A, Williams D, Marshall JG. The endogenous peptides of normal human serum extracted from the acetonitrile-insoluble precipitate using modified aqueous buffer with analysis by LC-ESI-Paul ion trap and Qq-TOF. J Proteomics 2010;73:1254–69.
[34] Oberg AL, Mahoney DW, Eckel-Passow JE, Malone CJ, Wolfinger RD, Hill EG, et al. Statistical analysis of relative labeled mass spectrometry data from complex samples using ANOVA. J Proteome Res 2008;7:225–33.
[35] Xia JQ, Sedransk N, Feng X. Variance component analysis of a multi-site study for the reproducibility of multiple reaction monitoring measurements of peptides in human plasma. PLoS One 2011;6:e14590.
[36] Aguilera R, Hatton CK, Catlin DH. Detection of epitestosterone doping by isotope ratio mass spectrometry. Clin Chem 2002;48:629–36.
[37] Arneberg R, Rajalahti T, Flikka K, Berven FS, Kroksveen AC, Berle M, et al. Pretreatment of mass spectral profiles: application to proteomic data. Anal Chem 2007;79: 7014–26.
[38] Hastings CA, Norton SM, Roy S. New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data. Rapid Commun Mass Spectrom 2002;16:462–7.
[39] van Hemert MJ, Steensma HY, van Heusden GP. 14-3-3 proteins: key regulators of cell division, signalling and apoptosis. Bioessays 2001;23:936–46.
[40] Powell DW, Rane MJ, Joughin BA, Kalmukova R, Hong JH, Tidor B, et al. Proteomic identification of 14-3-3zeta as a mitogen-activated protein kinase-activated protein kinase 2 substrate: role in dimer formation and ligand binding. Mol Cell Biol 2003;23:5376–87.
[41] Pozuelo Rubio M, Geraghty KM, Wong BH, Wood NT, Campbell DG, Morrice N, et al. 14-3-3-affinity purification of over 200 human phosphoproteins reveals new links to regulation of cellular metabolism, proliferation and trafficking. Biochem J 2004;379:395–408.
[42] Gao J, Garulacan LA, Storm SM, Opiteck GJ, Dubaquie Y, Hefta SA, et al. Biomarker discovery in biological fluids. Methods 2005;35:291–302.
[43] Sardana G, Marshall J, Diamandis EP. Discovery of candidate tumor markers for prostate cancer via proteomic analysis of cell culture-conditioned medium. Clin Chem 2007;53: 429–37.
[44] Declan Williams PZ, Bowden Peter, Stacey Catherine, McDonell Mike, Kowalski Paul, Kowalski Jane Marie, et al. Comparison of methods to examine the endogenous peptides of fetal calf serum clinical proteomics. Clin Proteomics 2007;2:67–89.
[45] Zhu P, Bowden P, Pendrak V, Thiele H, Zhang D, Siu M, et al. Comparison of protein expression lists from mass spectrometry of human blood fluids using exact peptide sequences versus BLAST. Clin Proteomics 2007;2:185–203.
[46] Gupta N, Bandeira N, Keich U, Pevzner PA. Target-decoy approach and false discovery rate: when things may go wrong. J Am Soc Mass Spectrom 2011;22:1111–20.
[47] Zhang R, Barker L, Pinchev D, Marshall J, Rasamoelisolo M, Smith C, et al. Mining biomarkers in human sera using proteomic tools. Proteomics 2004;4:244–56.
[48] Jankowski A, Zhu P, Marshall JG. Capture of an activated receptor complex from the surface of live cells by affinity receptor chromatography. Anal Biochem 2008;380:235–48.
[49] Kolialexi A, Tsangaris GT, Papantoniou N, Anagnostopoulos AK, Vougas K, Bagiokos V, et al. Application of proteomics for the identification of differentially expressed protein markers for Down syndrome in maternal plasma. Prenat Diagn 2008;28:691–8.
[50] Xing G, Zhang J, Chen Y, Zhao Y. Identification of four novel types of in vitro protein modifications. J Proteome Res 2008;7: 4603–8.
[51] Wang KY, Chuang SA, Lin PC, Huang LS, Chen SH, Ouarda S, et al. Multiplexed immunoassay: quantitation and profiling

of serum biomarkers using magnetic nanoprobes and MALDI-TOF MS. Anal Chem 2008;80:6159–67.
[52] Nicol GR, Han M, Kim J, Birse CE, Brand E, Nguyen A, et al. Use of an immunoaffinity-mass spectrometry-based approach for the quantification of protein biomarkers from serum samples of lung cancer patients. Mol Cell Proteomics 2008;7: 1974–82.
[53] Marchi N, Mazzone P, Fazio V, Mekhail T, Masaryk T, Janigro D. ProApolipoprotein A1: a serum marker of brain metastases in lung cancer patients. Cancer 2008;112:1313–24.
[54] Luque-Garcia JL, Zhou G, Spellman DS, Sun TT, Neubert TA. Analysis of electroblotted proteins by mass spectrometry: protein identification after Western blotting. Mol Cell Proteomics 2008;7:308–14.
[55] Hao Y, Yu Y, Wang L, Yan M, Ji J, Qu Y, et al. IPO-38 is identified as a novel serum biomarker of gastric cancer based on clinical proteomics technology. J Proteome Res 2008;7:3668–77.
[56] Barba de la Rosa AP, Lugo-Melchor OY, Briones-Cerecero EP, Chagolla-Lopez A, De Leon-Rodriguez A, Santos L, et al. Analysis of human serum from women affected by cervical lesions. J Exp Ther Oncol 2008;7:65–72.
[57] Morgan PE, Sturgess AD, Hennessy A, Davies MJ. Serum protein oxidation and apolipoprotein CIII levels in people with systemic lupus erythematosus with and without nephritis. Free Radic Res 2007;41:1301–12.
[58] Haqqani AS, Hutchison JS, Ward R, Stanimirovic DB. Biomarkers and diagnosis; protein biomarkers in serum of pediatric patients with severe traumatic brain injury identified by ICAT-LC-MS/MS. J Neurotrauma 2007;24:54–74.
[59] Yokoi K, Shih LC, Kobayashi R, Koomen J, Hawke D, Li D, et al. Serum amyloid A as a tumor marker in sera of nude mice with orthotopic human pancreatic cancer and in plasma of patients with pancreatic cancer. Int J Oncol 2005;27:1361–9.
[60] Malik G, Ward MD, Gupta SK, Trosset MW, Grizzle WE, Adam BL, et al. Serum levels of an isoform of apolipoprotein A-II as a potential marker for prostate cancer. Clin Cancer Res 2005;11:1073–85.
[61] Zhao X, Okeke NL, Sharpe O, Batliwalla FM, Lee AT, Ho PP, et al. Circulating immune complexes contain citrullinated fibrinogen in rheumatoid arthritis. Arthritis Res Ther 2008;10:R94.
[62] Reichel C. Identification of zinc-alpha-2-glycoprotein binding to clone AE7A5 antihuman EPO antibody by means of nano-HPLC and high-resolution high-mass accuracy ESI-MS/MS. J Mass Spectrom 2008;43:916–23.
[63] Plavina T, Hincapie M, Wakshull E, Subramanyam M, Hancock WS. Increased plasma concentrations of cytoskeletal and Ca2+-binding proteins and their peptides in psoriasis patients. Clin Chem 2008;54:1805–14.
[64] Hammerer-Lercher A, Halfinger B, Sarg B, Mair J, Puschendorf B, Griesmacher A, et al. Analysis of circulating forms of proBNP and NT-proBNP in patients with severe heart failure. Clin Chem 2008;54:858–65.
[65] Gramolini AO, Kislinger T, Alikhani-Koopaei R, Fong V, Thompson NJ, Isserlin R, et al. Comparative proteomics profiling of a phospholamban mutant mouse model of dilated cardiomyopathy reveals progressive intracellular stress responses. Mol Cell Proteomics 2008;7:519–33.
[66] Wilson AM, Kimura E, Harada RK, Nair N, Narasimhan B, Meng XY, et al. Beta2-microglobulin as a biomarker in peripheral arterial disease: proteomic profiling and clinical studies. Circulation 2007;116:1396–403.
[67] Kulasingam V, Smith CR, Batruch I, Buckler A, Jeffery DA, Diamandis EP. "Product ion monitoring" assay for prostate-specific antigen in serum using a linear ion-trap. J Proteome Res 2008;7:640–7.
[68] Yang Z, Harris LE, Palmer-Toy DE, Hancock WS. Multilectin affinity chromatography for characterization of multiple glycoprotein biomarker candidates in serum from breast cancer patients. Clin Chem 2006;52:1897–905.
[69] Reynolds MA, Kirchick HJ, Dahlen JR, Anderberg JM, McPherson PH, Nakamura KK, et al. Early biomarkers of stroke. Clin Chem 2003;49:1733–9.
[70] Fortin T, Salvador A, Charrier JP, Lenz C, Lacoux X, Morla A, et al. Clinical quantitation of prostate-specific antigen biomarker in the low nanogram/milliliter range by conventional bore liquid chromatography–tandem mass spectrometry (multiple reaction monitoring) coupling and correlation with ELISA tests. Mol Cell Proteomics 2009;8:1006–15.
[71] Qian M, Sleat DE, Zheng H, Moore D, Lobel P. Proteomics analysis of serum from mutant mice reveals lysosomal proteins selectively transported by each of the two mannose 6-phosphate receptors. Mol Cell Proteomics 2008;7:58–70.
[72] Plavina T, Wakshull E, Hancock WS, Hincapie M. Combination of abundant protein depletion and multi-lectin affinity chromatography (M-LAC) for plasma protein biomarker discovery. J Proteome Res 2007;6:662–71.
[73] Zenzmaier C, Marksteiner J, Kiefer A, Berger P, Humpel C. Dkk-3 is elevated in CSF and plasma of Alzheimer's disease patients. J Neurochem 2009;110:653–61.
[74] Ma Y, Peng J, Liu W, Zhang P, Huang L, Gao B, et al. Proteomics identification of desmin as a potential oncofetal diagnostic and prognostic biomarker in colorectal cancer. Mol Cell Proteomics 2009;8:1878–90.
[75] Chelius D, Huhmer AF, Shieh CH, Lehmberg E, Traina JA, Slattery TK, et al. Analysis of the adenovirus type 5 proteome by liquid chromatography and tandem mass spectrometry methods. J Proteome Res 2002;1:501–13.
[76] Moore RE, Young MK, Lee TD. Qscore: an algorithm for evaluating SEQUEST database search results. J Am Soc Mass Spectrom 2002;13:378–86.
[77] Park SK, M.K.. Random Number Generators: Good Ones Are Hard To Find. Commun ACM 1988;31:1191–201.
[78] Cargile BJ, Bundy JL, Stephenson Jr JL. Potential for false positive identifications from large databases through tandem mass spectrometry. J Proteome Res 2004;3:1082–5.
[79] Peihong Zhu, PB, Voitek Pendrak, Herbert Thiele, Du Zhang, Michael Siu, Eleftherios P, Diamandis, John Marshall. Comparison of protein expression lists from mass spectrometry of human blood fluids using exact peptide sequences versus BLAST. Clin Proteomics 2006;3–4:185–203.
[80] Mann B, Madera M, Sheng Q, Tang H, Mechref Y, Novotny MV. ProteinQuant Suite: a bundle of automated software tools for label-free quantitative proteomics. Rapid Commun Mass Spectrom 2008;22:3823–34.