Published in final edited form as:

Proteomics. 2013 August ; 13(16): 2386–2397. doi:10.1002/pmic.201300022.

A tool to evaluate correspondence between extraction ion chromatographic peaks and peptide-spectrum matches in shotgun proteomics experiments

Cristian I. Ruse¹, Samantha Peacock¹, Cornel Ghiban¹, Keith Rivera¹, Darryl J. Pappin¹, and Peter Leopold²

¹ Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, USA

² BioAnalyte, Inc., Portland, ME

Abstract

Chromatographed peptide signals form the basis of further data processing that eventually results in functional information derived from data-dependent bottom-up proteomics assays. We seek to rank LC/MS parent ions by the quality of their extracted ion chromatograms (ExICs.) Ranked extracted ion chromatograms (RExICs) act as an intuitive physical/chemical preselection filter to improve the quality of MS/MS fragment scans submitted for database search. We identify more than 4,900 proteins when considering detector shifts of less than 7ppm. High quality parent ions for which the database search yields no hits become candidates for subsequent unrestricted analysis for post-translational modifications. Following this rational approach, we prioritize identification of more than 5,000 spectrum matches from modified peptides and confirmed the presence of acetylaldehyde modified His/Lys. We present a logical workflow that scores data-dependent selected ion chromatograms and leverage information about semi-analytical LC/LC dimension prior to mass spectrometry. Our method can be successfully used to identify unexpected modifications in peptides with excellent chromatography characteristics, independent of fragmentation pattern and activation methods. We illustrate analysis of ion chromatograms detected in two different modes by RF linear ion trap and electrostatic field orbitrap.

Keywords

Multidimensional chromatography; statistical ranking; accurate mass measurement; fast chromatogram inversion

INTRODUCTION

The goal of any proteomics assay is to shed light on functional information associated with sample under analysis. Depth of coverage for stable isotope labeled-proteome[1] or overlapping sequence coverage of modified peptides[2] aided by fragmentation decision-

Corresponding author: Dr. Cristian I. Ruse, Proteomics Shared Resource, 1 Bungtown Rd, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724, Phone: +1 516-367-6806, FAX: +1 516-367-8873, ruse@cshl.edu.

Conflict of interest disclosure. Peter Leopold is employed by and owns more than 5% of BioAnalyte Inc., a company that was contracted by CSHL to develop the software described in this paper.

tree[3] are some of the recent examples of bottom-up proteomics assays that can derive function related information. When combined, quantitative proteomics screening[4] and large scale post-translational modification mapping[5, 6] provide biochemical description of biological systems at molecular level[7]. In bottom-up proteomics, peptide signals concentrated through nanoliter-range liquid chromatography form the basis of further data processing and eventually results in functional information.

MudPIT analysis is one of the standard protocols for protein identification using mass spectrometry[8]. The MudPIT protocol is an elegant multiplex chromatographic assay that supports a wide variety of sample preparation and separation techniques[9]. Sequential LC with orthogonal properties, RP and SCX, ensure comprehensive separations of complex proteomics mixtures. The final steps of the MudPIT protocol are always the same: submission of MS/MS fragment data to database search engines for identification of peptides[10].

MudPIT depends on data-dependent MS/MS features of modern mass spectrometers. The instrument operating software performs a rapid analysis of MS spectra and identifies (selects) potential parent ions for fragmentation based on preset parameters. In a typical study, 20-30% of the ions selected have no apparent redeeming features[11]. Traditionally, filtering strategies[12] have been developed to establish quality of fragmentation spectra either prior[13] or post[14] database searches. Tandem mass spectra were investigated for features of fragment ions either by sequence tagging[15], noise models[16] and statistical mixture models of discriminant scores[17]. Probability-based support vector machines[18] have been used to select phosphopeptide spectra from complex mixtures prior to database searches. However, all these approaches filter datasets based on features inferred from specific types of fragmentation spectra. Our goal is to introduce a rational approach for analysis of parent ions with excellent chromatographic characteristics as a first step of interrogating data.

We seek to establish a selection tool based on the application of physical and chemical characteristics of digested peptides. We characterize the chromatographic signals of data-dependent selected peptides across entire acquisition time for each particular mass channel rather than evaluating local-in-time chromatographic signals. The execution time must be relatively fast in order not to decelerate the overall work flow. Therefore, we introduce a novel method for LC/MS data inversion that meets this strict criterion. Ions with poor chromatographic characteristics are eliminated from further consideration. Ions with excellent chromatographic characteristics are searched and identified or, failing identification, are highlighted for further study.

Our goal is to interrogate ion species based on chromatographic properties detected during multidimensional separation of peptides, independent of activation/ionization methods and consequently of fragmentation pattern.

We present a data reduction method, describe the algorithm RExIC and its software implementation (www.bioanalyte.com/software/msredux/munk.shtml). The software reads the raw LC/MS-LC/MS/MS chromatograms and preselects ions using ranking by

chromatographic quality. This method allows identification of unexpected modifications in peptides with excellent chromatographic signals and improves the search time by 20% without compromising the overall quality of the subsequent identifications. More importantly, our approach is independent of any information regarding fragmentation of peptides modified either *in vivo* or through a chemical process.

MATERIALS AND METHODS

Cell Culture and Protein Harvesting

The HEK293 cells were maintained in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 5% dialyzed fetal bovine serum, 2 mM glutamax, 1% penicillin/streptomycin in a 5% CO₂, 37 °C incubator. Cells were cultured in 100 mm² dishes in DMEM and experiments were performed when the plates were 80% confluent. The cells were then quickly washed with ice-cold PBS with phosphatase and protease inhibitor cocktails (Sigma-Aldrich, MO). Trypsin solution containing 0.05% EDTA was briefly added to the cells, and then, the cells were removed from the plate with 10 mL of PBS with phosphatase inhibitors. The cells were centrifuged for 5 min at 500g. The cell pellet was resuspended in lysis buffer: 50mM TRIS, 10mM EDTA, 0.5% NP-40. Samples were reduced with 5mM Tris(2-carboxyethyl)phosphine and 10mM Methylmethanethiosulfonate sequentially. Lysates were then precipitated by the addition of methanol/chloroform (2:1 v/v) and the pellets reconstituted in 50mM TEAB, 0.1% ProteaseMax. After precipitation, samples were digested with trypsin digestion (1:50) at 37C overnight.

MudPIT

Linear ion trap-orbitrap analysis—We used a 2D Vented Column Setup with a Proxeon nano-flow HPLC pump[19]. The buffer solutions were 5% ACN/0.1% formic acid (buffer A) and 90% ACN/0.1% formic acid (buffer B). Salt buffer solution was 500mM ammonium acetate/5% ACN/0.1% formic acid. Samples were loaded onto a 180-µm i.d. fused silica capillary column containing 3 cm of 5-µm Aqua C 18 material (Phenomenex, Ventura, CA) followed by 3 cm of 5-µm Luna strong cation exchanger (Phenomenex) and 2 cm of 5-µm Aqua C 18 material. The triphasic column was then connected to an analytical column of a 100-µm i.d. capillary with a 5-µm pulled tip and packed with 15-17cm cm of 3-µm Aqua C 18 material (Phenomenex, Ventura, CA). After loading step, 16-salt steps were typically applied using the autosampler: 0%, 5%, 10%, 15%, 20%, 25%, 30%, 40%, 50%, 60%, 80%, 100%, 95% salt-5%B, 90% salt-10%B, 85% salt-15%B, followed by a last step of 80% salt/20% B for complete elution of strong cation exchange resin[20]. For each salt step, peptides were then separated using the following gradient: 5 min at 8% B, 90 min to 35%, 5 min to 70%, 5 min at 70%, 2 min to 5% followed by 5 min of reequilibration at 5% B, total 112 min. The flow rate was at 300nL/min (splitless).

Peptides eluted from the microcapillary fritless column were directly electrosprayed into a linear ion trap-orbitrap mass spectrometer (LTQ Orbitrap XL) with the application of a distal 2.3 kV spray voltage and a capillary temperature of 150C. A cycle of one full-scan mass spectrum (380-1700 m/z) followed by 6 data-dependent MS/MS spectra at a 35% normalized collision energy was repeated continuously throughout each step of the

multidimensional separation. Full scan MS was acquired in orbitrap analyzer at resolution 60,000 with Full FTMS maximum injection time (IT) of 1000ms. Tandem MS2 were acquired in the linear ion trap analyzer for a minimum signal threshold of 5000[21]. We used monoisotopic precursor selection for charge states 2+, 3+ and 4+ with preview mode enabled. Unassigned charged states and charge state 1+ were excluded from MS2 analysis. Dynamic exclusion parameters were set at repeat count of 1 in 30s, exclusion list size 500 and exclusion duration 90s. All data were acquired in centroid mode.

Database searches for linear ion trap-orbitrap spectra—Acquired files (.RAW) were extracted using ProteomeDiscoverer v. 1.3.0.339. SpectrumSelector extracted all CID MS2 spectra with a minimum peak count of 1 and S/N> 1.5 in full scan MS (FT mode). We used a workflow to search against IPI human database v3.75 using SEQUEST, with a fixed MMTS-derived modification of cysteine (methylthio, C=45.988Da). Trypsin specificity was allowed with up to 2 missed cleavages and a precursor mass tolerance of 200ppm. Peptide FDR was set to 1% using Percolator with validation based on q-value[22]. Results were filtered using a precursor mass tolerance of 5ppm. Total search time was recorded in .msf ProteomeDiscoverer file.

Unrestricted searches for post-translational modifications—A subset database was created from proteins identified by ProteomeDiscoverer-SEQUEST workflow. RExIC-selected spectra with no associated IPI database peptide were searched with MSAlignment[23] under ProteoSAFe v 1.2.3[24]. We used a subset database with the following parameters: protease trypsin, parent mass tolerance 0.5Da, ion tolerance 0.5Da, instrument high accuracy LTQ, modifications of up to 50Da and spectrum level FDR 0.05. Results were presented in default format with no further filtering, either for mass accuracy or localization of the modification site.

Directed searches for analysis of acetaldehyde presence were performed using ProteomeDiscoverer v. 1.3.0.339. We searched against IPI human database v3.75 using SEQUEST, with a fixed MMTS-derived modification of cysteine (methylthio, C=45.988Da) and dynamic modification +26.016 Da (H, K). Trypsin specificity was allowed with up to 2 missed cleavages and a precursor mass tolerance of 5ppm and peptide FDR was set to 1% using Percolator.

Linear ion trap analysis—An identical sequence of 16 salt steps was used for linear ion trap analyses. Reversed-phase gradient conditions were slightly modified compared with linear ion trap-orbitrap analyses by adding 8 min of equilibration time at the end of the gradient (Supplementary material).

Peptides eluted from the microcapillary fritless column were directly electrosprayed into a linear ion trap mass spectrometer (LTQ) with the application of a distal 2.4 kV spray voltage and a capillary temperature of 200C. A cycle of one full-scan mass spectrum (400-1400 m/z) followed by 5 data-dependent MS/MS spectra at a 35% normalized collision energy was repeated continuously throughout each step of the multidimensional separation. Full MS IT was set at 50ms. As above, dynamic exclusion parameters were set at repeat count of 1 in 30

s, exclusion list size 500 and exclusion duration 90s. All data were acquired in centroid mode.

Acquired files (.RAW) were extracted, searched and FDR filtered as above. Precursor mass tolerance was set at 3Da. Results were further filtered for high confidence peptides between 6 and 30 amino acids.

RExIC algorithm: The creation of one XIC for each parent ion is roughly equivalent to the inversion of the LC/MS from a series of m/z scans indexed by retention time to a series of RT scans indexed by m/z . The process is computationally intensive and is limited by available RAM and RAM addressing capabilities of the libraries and operating system used.

Parse .raw file for LC/MS and LC/MS/MS data: We read the raw chromatogram and parse it for index information relating to addresses of MS and MS/MS scans, values for parent ions, calibration parameters, etc.

Fast chromatogram inversion: The calculation of a single XIC trace requires the querying of each MS scan for m/z values in a given range. Ordinarily, this requires searching the entire chromatogram scan-by-scan. Depending on complexity, an XIC calculation takes 1s to 15s to probe the entire LC/MS chromatogram. Since roughly 80% of all scans are MS/MS, and since each MS/MS scan triggers a separate XIC calculation, the pedestrian approach to a global XIC calculation is prohibitively slow, e.g., $10s \times 2000 \text{ scans} = 20,000 \text{ s} = 5.5 \text{ hr}$. To avoid this process of re-reading each of the 2000 MS scans 4000 times, we use the following approach. We sequentially select each MS centroid scan (skipping profile and MS/MS scans) and create a master list of data consisting of ordered triplets (m/z , *intensity*, *rt*). The master list requires substantial RAM memory, e.g., 1-2 GB. We sort this list by increasing m/z . For each parent ion μ at retention time τ , we select a subset of this list that satisfies the requirement $\mu - \delta\mu \leq m/z \leq \mu + \delta\mu$, where $\delta\mu = ppm \times 10^{-6} \times \mu$ and *ppm* is a user-supplied m/z tolerance. Finally, we sort the subset list by retention time. When two elements have the same retention time, we simply add their intensities together. The sorted subset becomes the $xic(t; \mu, \delta\mu)$ for the parent ion μ with m/z tolerance $\delta\mu$. This process is ~2000 faster than performing individual costly XIC calculations for each parent ion.

Scoring: For each $xic(t; \mu, \delta\mu)$, we calculate a distribution of integrated chromatographic intensities corresponding to a nominal chromatographic peak width a specified by the user. A window of width a is moved along the retention time axis, and all intensities in the window are summed to create a data point of the distribution. All data points sorted by intensity, and the one corresponding the parent ion of interest μ at the retention time of interest τ is assigned a percentile score (Figure 1).

Formally, we perform a boxcar transform on a given $xic(t; \mu, \delta\mu)$ in order to evaluate the significance of a particular time interval specified by the MS/MS selection event. The kernel of the boxcar transform is given by the step-function $f(t-t', a) = 0$ for $|t-t'| > a$ and $f(t-t', a) = 1$ for $|t-t'| \leq a$.

The boxcar transform of $xic(t; \mu, \delta\mu)$ is $b(t; a, \mu, \delta\mu) = \int dt' f(t-t', a) xic(t'; \mu, \delta\mu)$. We treat $xic(t'; \mu, \delta\mu)$ as a discrete set of measurements, so $b(t; a, \mu, \delta\mu)$ is also a discrete set with the same discretization. We then sort the intensities $b(t; a, \mu, \delta\mu)$ in descending order and assign a percentile score to each value based on the fraction of all other transform intensities that are below any value of $b(t; a, \mu, \delta\mu)$. The percentile score for the measurement of interest (μ, τ) is based on the transform value $b_\tau = b(\tau; a, \mu, \delta\mu)$. It is subject to the user-parameters a and $\delta\mu$, where $\delta\mu$ is calculated for each μ using the *ppm* mass tolerance parameter. The boxcar transform has a number of advantages. The characteristic peak width a is the only parameter, and its meaning is intuitive. It makes no assumptions about chromatographic peak shape or background. The parametric dependence on a can even be removed by calculating b for several values of a and selecting the value for which b is optimized. However, we have not applied this calculation approach here.

We introduce here a robust scoring method to estimate the significance of a measurement, within a single XIC. Our approach emphasizes the rank of an interval next to all other peak sizes within the same XIC. We use the boxcar transform that effectively exaggerates local maxima and it emphasizes broad peaks over narrow ones or noise spikes[25]. Notably, it does not require an arbitrary cutoff parameter based on some fraction of a base-peak. The percentile score of boxcar-sliced areas of XIC is invulnerable to random noise since a few noise spikes do not change the percentile ranking significantly. Likewise, the presence of many XIC features of the same intensity and duration as the putative signal *do* reduce the ranking of the signal, as one would expect. Ranking is self-normalizing and is consistent with the interpretation of a mass spectrometry experiment as a highly-multiplexed, many-analyte detection system.

RExIC algorithm has been implemented as the software package Munk available from www.bioanalyte.com/software/msredux/munk.html. It currently reads XCalibur data format 63 from LTQ Orbitrap XL and LTQ instruments. Documentation on how to use Munk is included in Supplementary Information. Munk output generates lists of MS2 scans from ions with set RExIC score either 70 > or 80 > percentile. Included in the supplementary material is a perl script that use Munk-generated lists to filter .mgf files.

RESULTS

1. Analysis of FT-MS chromatograms before database searches

A. Chromatographic profiles—We calculate an extracted ion chromatogram (XIC) for each ion species triggering an MS2 spectrum. For this calculation we use a mass tolerance of 10ppm for orbitrap detection and a chromatographic peak width of 0.2 min. In this case, a mass tolerance of 10ppm refers to the difference between m/z value determined by a short prescan required to decide isolation for further fragmentation by MS2 and m/z determined in full scan FT spectrum over the entire acquisition time[26].

Next, we assign a quantitative value to the reference point (MS2 associated-RT) and then we build a population of quantitated areas for $m/z \pm dm/z$ within the same XIC. We rank quantified areas using statistics applied in an individual XIC. Each RT being an individual

assay for a given $m/z \pm dm/z$ and therefore an XIC is built by a succession of retention times.

Each chromatographic signal derived from an MS2-triggering ion is evaluated against all separated ion species with the same mass within the experimental tolerance. The chromatographic peaks extracted within a mass width are ranked to percentage of the highest signal within 10ppm tolerance across the entire acquisition time (112 min per salt step) by using calculated area under the peak. Figure 2 A-B illustrates a chromatography profile ranked at 75 percentile and 59 percentile respectively. We refer to these two quantities as RExIC 75% and RExIC 59% scores.

A typical total ion chromatogram is shown in Supplementary Figure 1A. We apply our strategy to determine RExIC percentile score for all MS2 triggering ions in a single salt step from MudPIT analysis. We observe a distribution of RExIC scores above 70% from 20 min to 110 min retention times (Supplementary Figure 1B). The 90 min interval corresponds with higher intensity region of TIC trace, where most of peptides are eluted (Supplementary Figure 1A).

B. Evaluation of MudPIT steps using RExIC percentile scoring—Using the RExIC percentile score, we investigated a complete MudPIT profiling experiment. In our experimental workflow, we collected a total of 15 salt steps following gradient loading of peptides on the cation exchange material. On average, approximately 11,000 MS2 events were collected per fraction to a total of 179,079 MS2 spectra for entire MudPIT profiling (Table 1). Next, we investigated the characteristics of fragmented ion species based on RExIC percentile scores. At 4ppm mass equivalence between 61% and 73% peaks showed chromatograms characteristics depending on filtering on RExIC scores. For comparison, we present the results for filtering with RExIC scoring at 7ppm and 10ppm mass equivalence in Table 1.

C. Database searches against IPI human database v 3.75 of MudPIT fractions

—We searched both unfiltered spectra and spectra selected with RExIC using SEQUEST followed by Percolator analysis to establish FDR. Results are summarized in Table 2 with 23,623 peptides identified from 5,054 proteins. We then compared results from different filtering conditions. Using RExIC 70% we identified 94% of proteins (89% of peptides) at 4 ppm mass equivalence compared to unfiltered spectra. Total search time was only 72% of search time of unfiltered spectra. At RExIC 80 percentile scoring and a mass equivalence of 4 ppm, we identified 90% of proteins (82% of peptides) compared to unfiltered spectra. Total search time decreased to 65% of unfiltered tandem mass spectra. Notably, every identified protein had a set of chromatographic peaks attached to it.

Filtering using ranked chromatograms showed the same effect on peptide identifications across MudPIT-salt steps (Figure 3 and Supplementary Figure 2). The number of spectra submitted to search for each chromatographic step is detailed Table 1. Next, we compared results from unfiltered and filtered searches at protein level as well as physico-chemical properties of peptides. We compared identified proteins with 1-4 peptides or more than 4 peptides. Typically, we observed a variation of 1% for all these classes among different

mass equivalence and RExIC filter conditions (Supplementary Figure 3). When using more stringent filtering (4ppm mass equivalence and 80% RExIC), we observed a variation of close to 2% for proteins described either by 1 peptide or by more than 4 peptides.

Filtering of data by RExIC at all three mass equivalence values resulted in less than 1% variation of tryptic peptides as well as peptides with 1 or 2 missed tryptic sites (Supplementary Figure 4). The ratio of peptides to cysteine-containing peptides was 5.8 for unfiltered data and varied between 5.6 and 5.7 for all filtering conditions (Supplementary Figure 5), a variation of up to 3%. Identified peptides showed highly similar isoelectric point profiles (Supplementary Figure 6). Overall, application of RExIC filtering at different mass equivalences showed no bias towards the major physico-chemical properties (number of missed tryptic sites, cysteine content relative to all identified peptides and peptide isoelectric points) of identified proteome. Variations of less than 3% of major peptide properties showed that application of RExIC scoring had minimal effect as related to the properties of identified proteome. Thus, selection of RExIC scored chromatograms resulted in minimal impact for the results of database searches against IPI human database v3.75 as compared with unfiltered data.

D. Unrestricted search for post-translational modifications—To illustrate the use of RExIC, we performed an unrestricted search using MSAlignment under ProteoSAFE v1.2.3 package. The process of selecting spectra for unrestricted MSAlignment search is illustrated in Figure 4. We created a subset database of 4558 proteins (Table 2 - *RExIC* 80%, 4ppm column) identified from 19328 peptides (36,383 PSM) mapped to IPI human database. After subtracting matched PSMs from total spectra 108,553 (Table 1 - *RExIC* 80%, 4ppm column), we analyzed 72,200 MS2 spectra associated with high scoring chromatograms. Thus, RExIC filtering combined with database searches resulted in selection of 40% of total acquired spectra 179079 (Table 1 - *unfiltered* column) for further unrestricted searches.

Modified peptides identified by MSAlignment (modifications 50Da) were organized in tables and presented in *Supplementary dataset 1*. A summary of identified peptides is presented in Table 3. We highlighted several modifications usually associated with sample preparation part of the experiment: +16 (oxidation), +42 (acetylation/trymethylation), +46 (methylthio), -17 (pyro-Glu). Interestingly, we detected 582 spectra with peptides modified by a positive mass shift +26. There are several possible amino acids substitutions with +26 shift as well as acetaldehyde modification of Lys and His (www.unimod.org, #254).

Next, we followed up on the possibility of acetaldehyde modification of Lys and His. We identified 234 spectra of 146 acetaldehyde-modified His/Lys-peptides with a mass accuracy of 5ppm or less while searching against entire IPI human database (*Supplementary dataset 2*).

2. Analysis of low resolution-linear ion trap chromatograms before database searches

We applied the algorithm established above to analyze MudPIT datasets acquired on 2D linear ion trap[27]. For simplicity, we performed an identical sequence of MudPIT salt steps as before. We calculated extracted ion chromatograms with a mass tolerance of 50ppm a chromatographic peak width of 0.2 min. Initial experiments on prototype low resolution 3D

ion trap characterized this detector with a unit resolving power and 20-40ppm accuracy for the quadrupole ion trap[28]. In our case, a mass tolerance of 50ppm was measured in reference to the precursor picked for fragmentation within a 0.4 min time window. We associated an absolute area derived from boxcar chromatograms to each ion triggering an MS2 event. Examples of RExIC scores and calculated areas derived from linear ion trap detector are presented in Figure 5 A-D.

Results of database searches from 2D linear ion trap experiments are summarized in Table 4. RExIC filtering reduced the search input by 28% for MudPIT with linear ion trap detection. Less than 7% of proteins were lost upon RExIC filtering of MudPIT dataset detected with a lower resolution ion trap. Total search time decreased by at least 25% (Supplementary Table 1). Therefore, in addition to FT-MS detection, we illustrate the applicability of RExIC algorithm to centroid acquired data on a linear ion trap.

3. Comparison RExIC-scored ion chromatograms with S/N ratio threshold for selection of peptide spectrum matches

We further investigated the results of database searches for different values for RExIC scores. For comparison, we include thresholds determined from S/N ratio of identical extracted ion chromatograms as calculated by MASIC (<http://omics.pnl.gov/software/MASIC.php>). Therefore, we were able to establish a correspondence between RExIC score and a common figure of merit, S/N.

Results from linear ion trap-orbitrap analysis are summarized in Figure 6. Relative to unfiltered dataset, RExIC 90% score selected approximately the same number of identified spectra as S/N 20. RExIC 85% score and S/N 10 virtually selected the same number of peptide spectrum matches. RExIC 80% score and S/N ratio 5 led to selection of similar numbers of identified spectra ~67% relative to unfiltered dataset number of peptides (23623 column 2, Table 2). RExIC 70% was roughly equivalent with S/N ratio 2 with 77% of peptide spectrum matches detected relative to unfiltered dataset.

RExIC scoring captures the complexity of sample per each mass-chromatographed channel selected for data dependent fragmentation, rather than deriving local-in-time S/N chromatographic parameter. Within percentile RExIC score we leverage information on ions assayed only by mass measurement in relation to data-dependent ions selected for sequencing fragmentation.

DISCUSSION

Interrogation of MudPIT steps from splitless-flow chromatograms with RExIC provides a logical step in selecting MS2 spectra for database search. Our approach prioritizes spectra for multi-pass searches[29] based on chromatographic characteristics. Our MudPIT FT-MS setup acquired a total of 179,079 MS2 spectra (Table 1). A decade earlier[8], the original MudPIT work showed a peak capacity of 23,000 MS2 spectra[8, 30]. When filtering with a mass equivalence of 4ppm and at least 80% RExIC score, we obtained 61% of spectra compared to unfiltered data. From this total of 108,583 (Table 1, 4ppm, 80 RExIC column), after database searches we identified 36,383 peptide-spectrum matches and 4558 proteins.

Approximately ~66% of chromatographic peaks with good characteristics (4ppm, 80 RExIC) were not associated with any identified peptides. We can arrive to a set of modifications using scoring from the first analytical dimension of the experiment: multidimensional chromatography (Figure 4). In this work, we use an innovative rational approach to detect the presence of acetaldehyde modified His/Lys (Table 3). Reduction and alkylation reagents are a possible source for this modification[31]. It is worth mentioning that acetaldehyde is the first metabolite of ethanol thereby detection of acetaldehyde peptides might be of interest for certain biological studies[32].

We rank MS1 chromatogrammed signals from the analytical dimension of MudPIT assay prior to tandem mass spectrometry using signals from detectors based on two different modes: a) image current from orbitrap detector[28] and b) mass selective radial ejection[27]. In principle, axial mass selective ejection[33] is also possible for orbitrap analyzer[34]. Our score is independent of fragmentation pattern and activation method being utilized. We statistically ranked multiplexed chromatographic FT-MS experiments (XICs) before database searches. Duration of our MudPIT experiments was > 36hrs and we investigated several different values of mass equivalence. Orbitrap mass errors below 5ppm have been shown in stability tests for longer than 20hrs[35]. We identified between 4558 proteins to 4962 proteins from soluble HEK proteome depending on the mass equivalence values and RExIC scores (Table 2). Increasing the mass equivalence of extracted chromatograms from 4ppm to 7ppm resulted in more identified proteins for either RExIC 70% or RExIC 80%. Only higher quality ion chromatograms (RExIC 80%) resulted in more protein identifications when increasing the mass equivalence from 7ppm to 10ppm. These results suggest stability of mass accuracy influences the optimal value of $\delta\mu$ for extraction of ion chromatograms. Optimal values of $\delta\mu$ might also be dependent on the acquisition mode, centroid or profile[36].

Lower resolution detection with first generation ion trap led to identification of more 4000 protein species after ranking extracted chromatograms. These results compare favorably to MudPIT separations detected by linear ion trap-orbitrap. Optimization of experimental parameters[37] can improve number of ions selected for fragmentation. In our study, dynamic exclusion parameters had very similar values since they belonged to the same analytical stage of the experiment. While dual linear ion trap-orbitrap increases substantially the speed of spectral acquisition[38], it is intriguing to speculate that further optimization of first stage FT mass measurement mode might have similar impact.

RExIC uses a global, full RT range estimator of chromatographic peak significance. Our estimator omits detailed peak characterization. It is robust because it relies solely on the ranking of the selected peak area to every other interval sampled by the MS detector. Approaches that use parent ions to reconstruct only partial SICs of the MS data calculate common estimators and estimate noise[39, 40]. MASIC is one such approach that calculate local-in-time signal-to-noise to estimate chromatographic peak significance[41]. Since RExIC uses IO libraries optimized for speed and a rapid LCMS inversion algorithm, it runs ~40x faster than MASIC[41]. We compared RExIC scoring with a convenient figure of merit derived by MASIC, S/N ratio. Our approach is focused on chromatographed signals and avoids any subjective choice in measuring noise. Sources and distribution of noise in

multidimensional LC coupled with MS detectors preclude direct implementation of routines for peak characterization developed for detectors such as UV[42]. Evaluation and description of sources of noise have been presented for RF ion traps[43] and electrostatic field orbitrap[28]. Filtering using median statistical values improved peak detection in 3D low resolution ion trap LC-MS data[25]. More recently, for FT orbitrap mass spectrometer, removal of noise in LC-MS/MS can be accomplished through a planar orthogonal range query[44, 45]. In our approach, we include all signals arrived at detectors(s). We partition entire chromatographic time in boxcar-slices of 0.4 min. Therefore, we obtain a population of measured areas anchored at an MS2-triggered event, for each mass channel. Then, we use a statistical ranking score to describe the distribution of these areas, RExIC score.

Last but not least, by characterizing identified fragmentation spectra associated with lower RExIC scores one can start to describe ion species with borderline signal thresholds independent of the activation method. We expect other proteomic fractionation techniques could benefit from application of RExIC as a first step in interrogating ubiquitous RPLC-MS/MS datasets[46]. One can envision coupling our approach with further modeling of high mass accuracy MS data[47].

It is important to emphasize that we apply ordered statistics to amounts of chromatographed signals captured by boxcar transform. We analyze directly the signal without any further processing of individual peaks.

CONCLUSIONS

We present a logical workflow that scores data-dependent selected ion chromatograms and provides similar depth of coverage of proteome to unfiltered searches. We score chromatographic quantities of ions selected for fragmentation against all ions within a detector mass tolerance across chromatographic time. We leverage information about analytical dimension prior to mass spectrometry. Therefore, we establish a relationship between ions selected for MS/MS and identified as belonging to protein species and unselected and/or unidentified ions characterized only by mass measurement.

Scored chromatograms of data-dependent selected ions introduce a deductive step that prioritizes identification of peptides. Following this rational approach, we identified more than 5000 spectrum matches from modified peptides with excellent chromatographic characteristics (*Table 3*). From these, 5-10% spectral matches belonged to acetyldehydro-Lys/His, a modification associated with a metabolic process. Our method could be successfully used to identify unexpected modifications in peptides with excellent chromatography characteristics, independent of fragmentation pattern and activation methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We acknowledge Benjamin Weindorf for technical help and Monica Ruse for help in preparing the manuscript.

REFERENCES

1. de Godoy LM, Olsen JV, Cox J, Nielsen ML, et al. Comprehensive mass-spectrometry-based proteome quantification of haploid versus diploid yeast. *Nature*. 2008; 455:1251–1254. [PubMed: 18820680]
2. Molina H, Horn DM, Tang N, Mathivanan S, Pandey A. Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry. *Proc Natl Acad Sci U S A*. 2007; 104:2199–2204. [PubMed: 17287340]
3. Swaney DL, Wenger CD, Coon JJ. Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *J Proteome Res*. 2010; 9:1323–1329. [PubMed: 20113005]
4. Ross PL, Huang YN, Marchese JN, Williamson B, et al. Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents. *Mol Cell Proteomics*. 2004; 3:1154–1169. [PubMed: 15385600]
5. Lemeer S, Heck AJ. The phosphoproteomics data explosion. *Curr Opin Chem Biol*. 2009; 13:414–420. [PubMed: 19620020]
6. Mischerikow N, Heck AJ. Targeted large-scale analysis of protein acetylation. *Proteomics*. 2011; 11:571–589. [PubMed: 21246731]
7. Cox J, Mann M. Quantitative, high-resolution proteomics for data-driven systems biology. *Annu Rev Biochem*. 2011; 80:273–299. [PubMed: 21548781]
8. Washburn MP, Wolters D, Yates JR 3rd. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat Biotechnol*. 2001; 19:242–247. [PubMed: 11231557]
9. Yates JR, Ruse CI, Nakorchevsky A. Proteomics by mass spectrometry: approaches, advances, and applications. *Annu Rev Biomed Eng*. 2009; 11:49–79. [PubMed: 19400705]
10. Swanson SK, Washburn MP. The continuing evolution of shotgun proteomics. *Drug Discov Today*. 2005; 10:719–725. [PubMed: 15896685]
11. Nesvizhskii AI. A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics. *J Proteomics*. 2010; 73:2092–2123. [PubMed: 20816881]
12. Salmi J, Nyman TA, Nevalainen OS, Aittokallio T. Filtering strategies for improving protein identification in high-throughput MS/MS studies. *Proteomics*. 2009; 9:848–860. [PubMed: 19160393]
13. Bern M, Goldberg D, McDonald WH, Yates JR 3rd. Automatic quality assessment of peptide tandem mass spectra. *Bioinformatics*. 2004; 20(Suppl 1):i49–54. [PubMed: 15262780]
14. Nesvizhskii AI, Roos FF, Grossmann J, Vogelzang M, et al. Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics*. 2006; 5:652–670. [PubMed: 16352522]
15. Ma ZQ, Chambers MC, Ham AJ, Cheek KL, et al. ScanRanker: Quality assessment of tandem mass spectra via sequence tagging. *J Proteome Res*. 2011; 10:2896–2904. [PubMed: 21520941]
16. Xu H, Freitas MA. A dynamic noise level algorithm for spectral screening of peptide MS/MS spectra. *BMC Bioinformatics*. 2010; 11:436. [PubMed: 20731867]
17. Wong JW, Sullivan MJ, Cartwright HM, Cagney G. msmsEval: tandem mass spectral quality assignment for high-throughput proteomics. *BMC Bioinformatics*. 2007; 8:51. [PubMed: 17291342]
18. Lu B, Ruse CI, Yates JR 3rd. Colander: a probability-based support vector machine algorithm for automatic screening for CID spectra of phosphopeptides prior to database search. *J Proteome Res*. 2008; 7:3628–3634. [PubMed: 18563924]
19. Taylor P, Nielsen PA, Trelle MB, Horning OB, et al. Automated 2D peptide separation on a 1D nano-LC-MS system. *J Proteome Res*. 2009; 8:1610–1616. [PubMed: 19178303]

20. Motoyama A, Venable JD, Ruse CI, Yates JR 3rd. Automated ultra-high-pressure multidimensional protein identification technology (UHP-MudPIT) for improved peptide identification of proteomic samples. *Anal Chem.* 2006; 78:5109–5118. [PubMed: 16841936]
21. Wong CC, Cociorva D, Venable JD, Xu T, Yates JR 3rd. Comparison of different signal thresholds on data dependent sampling in Orbitrap and LTQ mass spectrometry for the identification of peptides and proteins in complex mixtures. *J Am Soc Mass Spectrom.* 2009; 20:1405–1414. [PubMed: 19467883]
22. Kall L, Canterbury JD, Weston J, Noble WS, MacCoss MJ. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods.* 2007; 4:923–925. [PubMed: 17952086]
23. Tsur D, Tanner S, Zandi E, Bafna V, Pevzner PA. Identification of post-translational modifications by blind search of mass spectra. *Nat Biotechnol.* 2005; 23:1562–1567. [PubMed: 16311586]
24. Huang, T.; Farcas, C.; Carver, J.; Castellana, N., et al. Proceedings of the 59th ASMS Conference on Mass Spectrometry and Allied Topics; 2011;
25. Hastings CA, Norton SM, Roy S. New algorithms for processing and peak detection in liquid chromatography/mass spectrometry data. *Rapid Commun Mass Spectrom.* 2002; 16:462–467. [PubMed: 11857732]
26. Makarov A, Denisov E, Kholomeev A, Balschun W, et al. Performance evaluation of a hybrid linear ion trap/orbitrap mass spectrometer. *Anal Chem.* 2006; 78:2113–2120. [PubMed: 16579588]
27. Schwartz JC, Senko MW, Syka JE. A two-dimensional quadrupole ion trap mass spectrometer. *J Am Soc Mass Spectrom.* 2002; 13:659–669. [PubMed: 12056566]
28. Perry RH, Cooks RG, Noll RJ. Orbitrap mass spectrometry: instrumentation, ion motion and applications. *Mass Spectrom Rev.* 2008; 27:661–699. [PubMed: 18683895]
29. Ahrne E, Muller M, Lisacek F. Unrestricted identification of modified proteins using MS/MS. *Proteomics.* 2010; 10:671–686. [PubMed: 20029840]
30. Wolters DA, Washburn MP, Yates JR 3rd. An automated multidimensional protein identification technology for shotgun proteomics. *Anal Chem.* 2001; 73:5683–5690. [PubMed: 11774908]
31. Saati, A.; Friesse, O.; Patel, H.; Boisvert, D., et al. Proceedings of the 60th ASMS Conference on Mass Spectrometry and Allied Topics; 2012;
32. Niemela O. Aldehyde-protein adducts in the liver as a result of ethanol-induced oxidative stress. *Front Biosci.* 1999; 4:D506–513. [PubMed: 10352137]
33. Hager JW. A new linear ion trap mass spectrometer. *Rapid Commun Mass Sp.* 2002; 16:512–526.
34. Makarov A. Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Anal Chem.* 2000; 72:1156–1162. [PubMed: 10740853]
35. Makarov A, Denisov E, Lange O, Horning S. Dynamic range of mass accuracy in LTQ Orbitrap hybrid mass spectrometer. *J Am Soc Mass Spectrom.* 2006; 17:977–982. [PubMed: 16750636]
36. Xia YQ, Lau J, Olah T, Jemal M. Targeted quantitative bioanalysis in plasma using liquid chromatography/high-resolution accurate mass spectrometry: an evaluation of global selectivity as a function of mass resolving power and extraction window, with comparison of centroid and profile modes. *Rapid Commun Mass Spectrom.* 2011; 25:2863–2878. [PubMed: 21913265]
37. Andrews GL, Dean RA, Hawkridge AM, Muddiman DC. Improving proteome coverage on a LTQ-Orbitrap using design of experiments. *J Am Soc Mass Spectrom.* 2011; 22:773–783. [PubMed: 21472614]
38. Olsen JV, Schwartz JC, Griep-Raming J, Nielsen ML, et al. A dual pressure linear ion trap Orbitrap instrument with very high sequencing speed. *Mol Cell Proteomics.* 2009; 8:2759–2769. [PubMed: 19828875]
39. Wong JW, Sullivan MJ, Cagney G. Computational methods for the comparative quantification of proteins in label-free LCn-MS experiments. *Brief Bioinform.* 2008; 9:156–165. [PubMed: 17905794]
40. Smith CA, Want EJ, O'Maille G, Abagyan R, Siuzdak G. XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification. *Anal Chem.* 2006; 78:779–787. [PubMed: 16448051]

41. Monroe ME, Shaw JL, Daly DS, Adkins JN, Smith RD. MASIC: a software program for fast quantitation and flexible visualization of chromatographic profiles from detected LC-MS(/MS) features. *Comput Biol Chem.* 2008; 32:215–217. [PubMed: 18440872]
42. Shellie RA, Haddad PR. Comprehensive two-dimensional liquid chromatography. *Anal Bioanal Chem.* 2006; 386:405–415. [PubMed: 16927069]
43. Blackler AR, Klammer AA, MacCoss MJ, Wu CC. Quantitative comparison of proteomic data quality between a 2D and 3D quadrupole ion trap. *Anal Chem.* 2006; 78:1337–1344. [PubMed: 16478131]
44. Khan Z, Amini S, Bloom JS, Ruse C, et al. Accurate proteome-wide protein quantification from high-resolution 15N mass spectra. *Genome Biol.* 2011; 12:R122. [PubMed: 22182234]
45. Khan Z, Bloom JS, Garcia BA, Singh M, Kruglyak L. Protein quantification across hundreds of experimental conditions. *Proc Natl Acad Sci U S A.* 2009; 106:15544–15548. [PubMed: 19717460]
46. Di Palma S, Boersema PJ, Heck AJ, Mohammed S. Zwitterionic hydrophilic interaction liquid chromatography (ZIC-HILIC and ZIC-cHILIC) provide high resolution separation and increase sensitivity in proteome analysis. *Anal Chem.* 2011; 83:3440–3447. [PubMed: 21443167]
47. Jaitly N, Mayampurath A, Littlefield K, Adkins JN, et al. Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics.* 2009; 10:87. [PubMed: 19292916]

Figure 1.

A selected parent ion is ranked using the following algorithm. The parent ion's MS-only extracted ion chromatogram is constructed from the raw MS data. The retention time interval associated with the parent ion's MS/MS fragmentation spectrum is identified, and the measured XIC intensities are integrated in that interval. The integrated signal is then compared the list of integrated signals in all intervals of similar duration. If the interval of interest is the highest one on the list, it is assigned a percentile rank of 99. If it is the lowest, it is assigned a percentile rank of 0. If the interval ranks below a threshold, its associated MS/MS fragmentation spectrum is eliminated from database search.

Figure 2.

Illustration of XIC peaks for two different ions in FT-MS chromatograms. RExIC percentile score is calculated across 112min within m/z (629.337-629.349) window leading to a 75% score for precursor ion fragmented at 53.12 min. Similarly, we calculate a RExIC score of 59% within m/z (597.833-596.8455) for precursor ion selected for fragmentation at 25.96 min.

Figure 3.

Comparison of search results between unfiltered spectra and spectra selected from chromatograms within 70 or 80 percentiles at mass equivalence 4ppm.

Figure 4.

Data analysis workflow for unrestricted searches for peptide modifications. RExIC percentile scoring selects fragmentation spectra based on the quality of individual ion chromatograms. Selected spectra are searched only with fixed modifications introduced during sample processing against IPI human database v3.75. Upon elimination of peptide spectrum matches from initial database search, unassigned spectra with high scoring ion chromatograms were further analyzed by unrestricted MSAlign search against subset database. We establish the frequency of modifications and validate them by database search against IPI human database v3.75. We establish and identify a proteome dataset upon selection of quality tandem mass spectra following scoring of ion chromatograms. Our score is independent of fragmentation pattern of modified peptides and the choice of activation method.

Figure 5.

Illustration of XIC peaks for two different ions in IT-MS chromatograms. (A) RExIC percentile score of 99% for ion fragmented at 43.95min. (B) Selected signal for calculation of area corresponding to ion fragmented at 43.95min. (C) RExIC percentile score of 83% and (D) calculated area for ion fragmented at 48.01min.

Figure 6.

Peptide spectrum matches determined by selection with RExIC scoring (open squares) and S/N ratio of chromatographic trace (open circles). Percentage of matches is set relative to peptide spectrum matches from unfiltered linear ion trap-orbitrap dataset.

Table 1

Comparison of MS2 parent ions based on ranking FT-MS chromatograms at 70% and 80% RExIC filter for 4ppm, 7ppm and 10ppm mass equivalence.

Salt steps	Number of MS2 events						
	Unfiltered	70 RExIC			80 RExIC		
		4ppm	7ppm	10ppm	4ppm	7ppm	10ppm
load	6770	4378	5031	5269	3633	4194	4429
5C	13378	9027	10450	10998	7451	8712	9270
10C	14213	10428	11907	12462	8863	10266	10898
15C	13298	10156	11498	11937	8601	9948	10467
20C	13063	9751	10983	11482	8192	9478	10012
25C	13437	9552	10642	11124	8045	9154	9634
30C	12273	8828	9990	10496	7355	8491	9061
40C	11793	8733	9843	10289	7274	8388	8872
50C	11855	8961	10088	10502	7533	8677	9112
60C	11308	8375	9428	9788	6974	8004	8409
80C	10274	7722	8758	9104	6546	7514	7883
100C	9602	7016	8011	8344	5819	6807	7182
95-5	10360	7350	8216	8579	5969	6857	7184
90-10	10125	7285	8124	8462	5952	6809	7119
85-15	9684	7127	7943	8252	5791	6588	6972
80-20	7646	5591	6283	6474	4585	5257	5465
Total of MS2 spectra	179079	130280	147195	153562	108583	125144	131969

Table 2

Summary of protein and peptide identifications from orbitrap dataset. For a mass equivalence value (4ppm, 7ppm or 10ppm), RExIC scoring of a 16-step MudPIT dataset takes less than 30 min. The analysis calculates RExIC scoring of chromatograms for the entire percentile score interval, including 70% and 80% scores. Once RExIC scores are calculated, the threshold can be decided at no additional analysis time.

	Unfiltered	70 RExIC			80 RExIC		
		<i>4ppm</i>	<i>7ppm</i>	<i>10ppm</i>	<i>4ppm</i>	<i>7ppm</i>	<i>10ppm</i>
Total peptides	23623	21048	22752	23006	19328	21220	21875
Total protein groups	5054	4759	4962	4948	4558	4748	4822
Search Time	553m56s	402m11s	455m7s	485m8s	363m57s	388m19s	403m12s
RExIC Analysis time	-	<30m	<30m	<30m			

Table 3
Summary of identified peptides from unrestricted search using MS-Alignment and 80% RExIC score, 4ppm mass equivalence.

Salt steps	load	5C	10C	15C	20C	25C	30C	40C	50C	60C	80C	100C	95-5	90-10	85-15	80-20
Total Peptides	177	571	477	532	510	643	612	517	514	531	480	365	532	538	478	241
Modified peptides	141	340	295	342	346	426	402	362	366	381	326	237	364	343	337	179
Mod +16	16	64	67	77	87	134	106	94	95	102	89	71	117	115	103	62
Mod +26	4	35	44	49	48	39	46	47	51	43	32	36	40	29	30	9
Mod +42	27	23	10	13	13	7	9	14	8	6	4	2	3	3	10	1
Mod +46	9	17	19	22	20	34	34	23	22	29	21	10	28	28	23	10
Mod -17	22	17	19	18	19	23	19	17	16	18	13	9	18	13	14	16

Table 4

Summary of identifications from linear ion trap experiments at 70%RExIC score, 50 ppm mass equivalence.

	Unfiltered	RExIC filter
Search Input	511019	364534
Total peptides	22150	20546
Total protein groups	4890	4572