analytical chemistry

Article

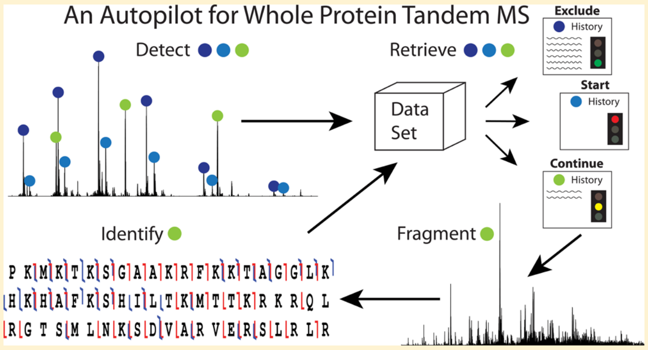
pubs.acs.org/ac

Autopilot: An Online Data Acquisition Control System for the Enhanced High-Throughput Characterization of Intact Proteins

Kenneth R. Durbin, Ryan T. Fellers, Ioanna Ntai, Neil L. Kelleher,* and Philip D. Compton*

Departments of Chemistry and Molecular Biosciences Northwestern University, 2145 North Sheridan Road, Evanston, Illinois 60208, United States

S *Supporting Information*


An Autopilot for Whole Protein Tandem MS
Detect
Retrieve
Data Set
Exclude
History
Start
History
Continue
History
Identify
Fragment
P K M K T K S G A A K R F K K T A G G L K
H K H A F K S H I L T K M T T K R K R Q L
R G T S M L N K S D V A R V E R S L R L R

ABSTRACT: The ability to study organisms by direct analysis of their proteomes without digestion via mass spectrometry has benefited greatly from recent advances in separation techniques, instrumentation, and bioinformatics. However, improvements to data acquisition logic have lagged in comparison. Past workflows for Top Down Proteomics (TDPs) have focused on high throughput at the expense of maximal protein coverage and characterization. This mode of data acquisition has led to enormous overlap in the identification of highly abundant proteins in subsequent LC-MS injections. Furthermore, a wealth of data is left underutilized by analyzing each newly targeted species as unique, rather than as part of a collection of fragmentation events on a distinct proteoform. Here, we present a major advance in software for acquisition of TDP data that incorporates a fully automated workflow able to detect intact masses, guide fragmentation to achieve maximal identification and characterization of intact protein species, and perform database search online to yield real-time protein identifications. On *Pseudomonas aeruginosa*, the software combines fragmentation events of the same precursor with previously obtained fragments to achieve improved characterization of the target form by an average of 42 orders of magnitude in confidence. When HCD fragmentation optimization was applied to intact proteins ions, there was an 18.5 order of magnitude gain in confidence. These improved metrics set the stage for increased proteome coverage and characterization of higher order organisms in the future for sharply improved control over MS instruments in a project- and lab-wide context.

The high throughput analysis of intact proteins by mass spectrometry has become increasingly relevant due to the recent and rapid acceleration in the development of enabling technologies. Front-end separations[1] have undergone a transformation as methods for solution-based intact protein separation have provided a viable alternative to the bottom up approach of in-gel digestion.[2] These new separation techniques are amenable to lower initial sample amounts and can be coupled to nano-LC prior to mass spectrometric analysis. Additionally, order of magnitude increases in sensitivity and speed enable modern instruments to more adequately handle the rigors of complex mixtures replete with high mass intact proteins.[3] These improvements in separations and instrumentation have transformed top down mass spectrometric analyses from single protein direct infusion experiments to high-throughput proteome-wide analyses up to 100 kDa.[4] Recent analyses have shown confident identification of thousands of proteoforms and nearly 2000 unique accession numbers.[4,5]

The focused efforts in separations and instrumentation, along with new and existing bioinformatics tools (e.g., ProSight PC 3.0[6]), have largely bypassed the realm of high-throughput mass spectrometric data acquisition. However, intelligent acquisition of mass spectrometric data holds great promise for increased experimental efficiency. Real-time adjustments to fragmentation parameters driven by decision logic have been successfully incorporated into bottom up proteomics workflows.[7,8] The incorporation of "intelligence" yielded sizable increases in the number of peptides identified in subsequent experiments. For intact protein analysis, intelligent acquisition strategies have been used to increase the number and quality of protein identifications resulting from an LC-MS analysis but relied on additional offline analysis.[9] Extending these efforts to an online data acquisition and analysis platform for TDP offers an exciting way to increase the value of these proteomics data sets.

Further development of data driven acquisition will not only increase the efficiency of data collection but also result in more effective use of the data that is generated during an analysis. High-throughput proteomics can generate immense quantities of data; however, each new fragmentation event is typically treated as a single, isolated event instead of part of a data set for a unique mass species. This leads to redundancy in data collection, which hinders deep exploration of the proteome.

Received: September 11, 2013
Accepted: January 8, 2014
Published: January 8, 2014

Additionally, project wide dynamic exclusion instead of an LC-MS run based dynamic exclusion would prevent much unneeded analysis of proteins that are already well characterized. Because of the mismatch of dynamic range between protein expression and mass spectrometers,[10] the very most abundant proteins (e.g., histone and heat shock proteins) can be fragmented thousands of times over the course of a proteomic investigation.[2] Unfortunately, these proteins are frequently fully characterized upon the first fragmentation so additional fragmentation of the protein produces no additional information.

Here, we demonstrate an upgraded TDP workflow driven by intelligent data collection through a software termed Autopilot. Online mass detection, supplemented with prior knowledge of past fragmentation events, guides precursor selection and fragmentation. Proteins not sufficiently characterized are subjected to repeat fragmentation with optimized parameters based on a fragmentation logic scheme. Spectral data are searched immediately and the information is added to a proteomic repository for reference in current and future runs. Each distinct mass builds a fragmentation history with new fragmentation data joined with previous data to produce the best possible sequence coverage of proteins. Through accretion of data from different fragmentation types, energies, and charge states, more complete protein characterization can be achieved. Furthermore, proteins which have exceeded a threshold of characterization are excluded from all forthcoming analysis to permit characterization of the maximal number of proteins in the given proteome analysis.

■ METHODS

Cell Culture. *Pseudomonas aeruginosa* PAO1 (ATCC 15692) was plated on Mueller Hinton II agar plates and incubated at 37 °C. A single colony was inoculated in 50 mL of Mueller Hinton broth (MHB) and incubated overnight at 37 °C. This starter culture was used to inoculate a larger culture at an inoculum to culture ratio of 1:500. Cells were harvested at midlogarithmic growth phase ($OD_{600} \approx 0.4$) by centrifugation at 5000 × *g*.

Sample Preparation. PAO1 cells were lysed in 4% SDS and 25 mM Tris, pH 7.4 with Halt protease inhibitors (Thermo Pierce, Rockford, IL) and 1 mM DTT. The cell lysate was acetone precipitated with three volumes of cold acetone and centrifuged at max speed for 10 min. The precipitated pellets were dried and 400−500 μg of protein was fractionated on a GELFREE 8100 Fractionation system (Expedeon, San Diego, CA) with either 10% or 12% cartridges. The collected fractions were precipitated with $MeOH/CH_3Cl/H_2O$ as previously described to remove SDS.[11] Prior to MS analysis, the samples were resuspended in 30 μL of buffer A (95% H_2O, 5% acetonitrile, 0.2% formic acid).

LC-MS/MS. Briefly, each GELFrEE fraction was injected onto a 2 cm, 150 μm i.d. PLRP-S trap column and a longer 10 cm, 75 μm i.d. column was used for the online separation as described previously.[12] Both the trap and analytical columns were packed in-house with 5 μm diameter, 1000 Å pore size PLRP-S. An Ultimate 3000 RPLCnano (Thermo Scientific Dionex) with 300 nL/min flow rate setup was used for online chromatography. An LTQ Velos Orbitrap Elite (Thermo Scientific) was the mass spectrometer used for all data acquisition. The resolution settings were 60 000 and 30 000 for precursor and fragmentation scans, respectively, at 400 *m/z*. An isolation window of 15 m/z was used for mass selection. For the Xcalibur data acquisition, the method implemented precursor and fragmentation scans with four μscans and dynamic exclusion with a repeat count of 1, a repeat duration of 240 s, and an exclusion duration of 5000 s. The maximum inject time for both precursor and fragmentation scans was 1000 and 400 ms for SIM scans. Autopilot utilized 4 μscans for precursor scans and SIM scans, 6 μscans for HCD fragmentation, and 12 μscans for ETD fragmentation. For Autopilot operation, the Thermo LTQ component object model (COM) is used in conjunction with an advanced user's license for the Orbitrap Elite. Several ion trap control language (ITCL) changes were applied to enable ETD and HCD fragmentation by Autopilot. The source code for the Autopilot logic interface is included in the Supporting Information. Instructions on the implementation of the fragmentation logic, an analysis to infer mass, online search are included.

Raw mass spectrometry data has been deposited in the IU Scholar Works Repository. The data is available at http://hdl.handle.net/2022/17234.

Online Data Analysis. Autopilot applies Xtract with a 3 signal/noise cutoff to full precursor scans, SIM scans, and fragmentation scans. For fragmentation data, the system performs online absolute mass searches against the PAO1 database with the precursor mass from Xtract. The search window is a 2000 Da tolerance around the precursor mass while the fragment tolerance is 10 ppm. The search is accomplished with the ProSight search engine and requires a minimum number of 4 matched fragments, with only the top hit (the most confident P-Score) returned from the search.

Offline Data Processing. The LC-MS/MS data files acquired by Xcalibur files were processed with an in-house software, cRAWler. The cRAWler produces ProSight upload format (.puf) files, which contain Xtract detected precursor masses linked with fragment ions from the corresponding data-dependent scan. The completed .puf files are then searched by ProSight PC 3.0 (Thermo Scientific, San Jose, CA) against the PAO1 database. A multitier search was implemented in ProSight PC with an initial small window absolute mass precursor search of 2.2 Da, followed by a large window absolute mass search of 100 000 Da if no hits were found in the first search. For data-dependent Xcalibur fragmentation scans with no associated precursor masses, a 100 000 Da window centered on 55 555.55 Da is searched.

■ RESULTS AND DISCUSSION

Characterization. Increased protein characterization in an automated fashion was the primary goal of this work. However, TDPs currently lacks a "characterization" score. Therefore, the P-score, which is a measure of the confidence of protein identification, was used as an indirect measure of characterization. As described previously, this score generally improves with an increasing number of matched fragment ions, which indicates better characterization. The score also decreases with an increasing number of unmatched fragment ions, breaking the direct link between characterization and this score. While it is recognized that this does not directly equate to characterization, the development of a characterization score is beyond the scope of this work.

Workflow. For mass spectrometric workflows to operate at optimal efficiency, the acquisition platform must utilize real-time and stored information during an analysis to enable the instrument to react to current acquisition needs in a productive manner. To properly guide acquisition, an accurate survey of

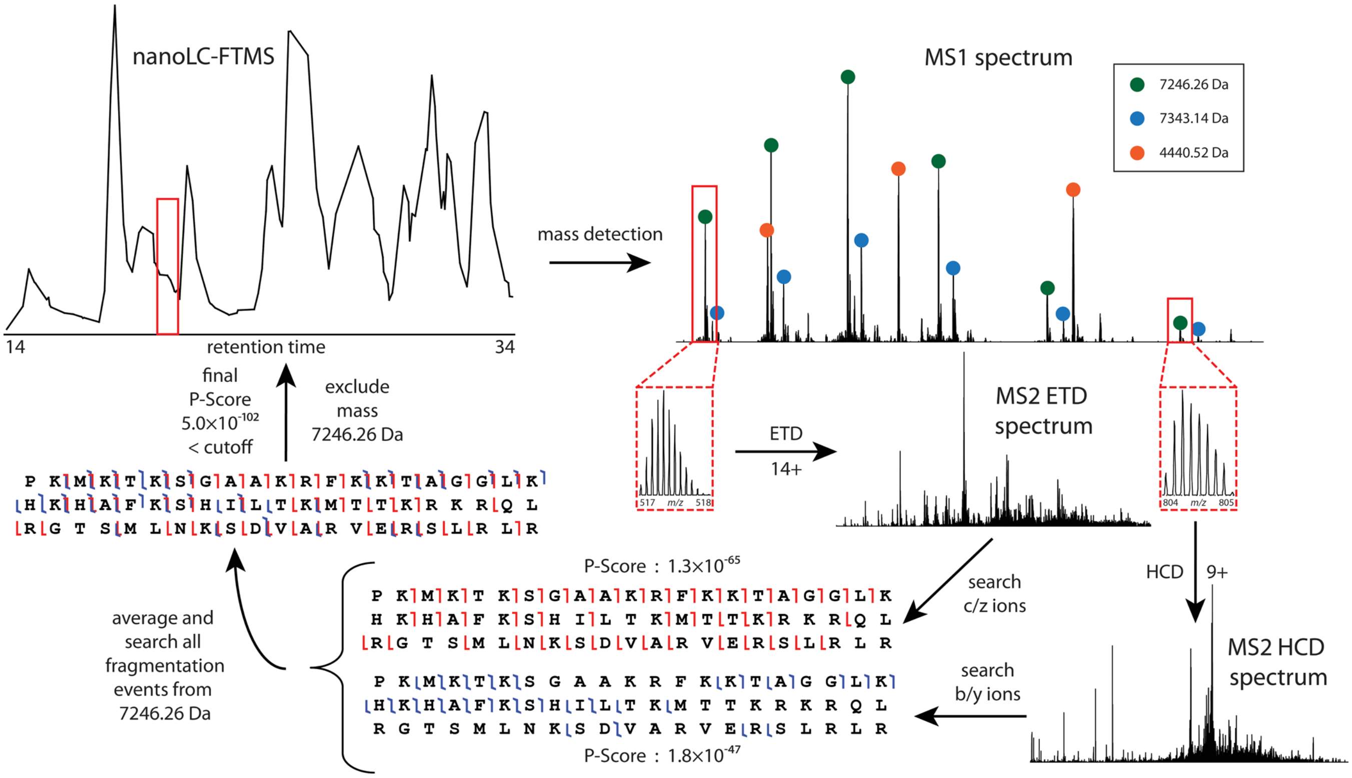

nanoLC-FTMS
14
retention time
34
mass detection
MS1 spectrum
● 7246.26 Da
● 7343.14 Da
● 4440.52 Da
517 m/z 518
ETD
14+
MS2 ETD spectrum
804 m/z 805
HCD
9+
MS2 HCD spectrum
final P-Score 5.0×10-102 < cutoff
exclude mass 7246.26 Da
P K M K T K S G A A K R F K K T A G G L K
H K H A F K S H I L T K M T T K R K R Q L
R G T S M L N K S D V A R V E R S L R L R
average and search all fragmentation events from 7246.26 Da
P-Score : 1.3×10-65
P K M K T K S G A A K R F K K T A G G L K
H K H A F K S H I L T K M T T K R K R Q L
R G T S M L N K S D V A R V E R S L R L R
P K M K T K S G A A K R F K K T A G G L K
H K H A F K S H I L T K M T T K R K R Q L
R G T S M L N K S D V A R V E R S L R L R
P-Score : 1.8×10-47
search c/z ions
search b/y ions

Figure 1. Data acquisition workflow on an example protein. A full precursor scan is taken, followed by HCD fragmentation of the 9+ charge state on the detected mass 7246.26 Da. After an online search, the software determines more analysis should be performed as the P-Score (1.8×10^{-47}) is not below the cutoff. An ETD scan of the highest charge state is taken and searched. The fragment ions are combined and the final P-Score of 5.0×10^{-102} is below the cutoff. All charge states of the 7246.26 Da species are permanently excluded from further fragmentation and the system goes in search of the next target mass.

the current protein elution landscape is of primary importance (Figure 1, top left). From a full MS spectrum, the data points can be directly piped from the instrument into an isotope determination routine, Xtract, which outputs a list of the average and monoisotopic masses present in the spectrum (Figure 1, top right and Figure S1A of the Supporting Information).[13] Along with mass information, Xtract includes the charge state distribution of the species, which enables charge states of a particular species to be grouped as a single “target”. The strategy of isotope analysis for mass detection is suitable for the mass range of proteins the Orbitrap Elite is able to isotopically resolve on an LC-MS time scale. For proteins no longer isotopically resolved (i.e., >30–35 kDa), a short transient FT scan or low resolution ion-trap scan can be combined with charge state deconvolution for intact mass determination.[14] However, this study focuses on lower molecular weight proteins that can be isotopically resolved by the Orbitrap mass spectrometer.

The protein species detected by Xtract represent the potential targets for focused fragmentation events. The mass list is parsed to group similar masses and account for simple hardware and software artifacts such as oxidation and off-by-one errors that result from incorrect isotope matches. The mass list is further culled to remove previously fragmented species identified at a high confidence level. Here, a very high degree of confidence cutoff (P-score $<1 \times 10^{-50}$) was selected to showcase the characterization abilities of the software. The remaining targets, sorted by decreasing abundance, are placed in a queue. This first-in, first-out queue supplies a new fragmentation scan definition to the instrument as soon as the data from the previous scan is returned.

Upon the initial fragmentation of a species, the lowest charge state is selected for HCD fragmentation with a normalized collision energy (NCE) of 25 (Figure 1, bottom right). The lowest charge is fragmented first as it typically will yield the largest number of unique fragments.[15] The underlying instrument control language translates the NCE into a HCD energy in eV to be applied based on a combination of the NCE value and the *m/z* of the fragmentation target. We apply Xtract to the fragmentation spectrum for fragment ion detection. After Xtract analysis is complete, a wide-window (2000 Da) absolute mass search of the fragment set is immediately started. The search is against the 5763 gene, 23 176 form pseudomonas PAO1 database with a 10 ppm fragment tolerance. The top hit is added to the fragmentation history of the mass species.

When a precursor mass of a previously fragmented species is redetected by Xtract, whether in the same run or a subsequent injection, the history of the species is queried from the proteomic data set. The best previous fragmentation event for the species is compared to an identification confidence cutoff. If the confidence of the identification exceeds the cutoff, the target is placed on an exclusion list. If it is below the set score cutoff, a fragmentation decision tree (Figure S1B of the Supporting Information) charts the course of action to optimize the energy, charge state, and fragmentation technique chosen for the reanalysis of the protein. The resultant spectrum from refragmentation is then processed in a similar fashion to the first fragmentation and the database record for the target is updated accordingly.

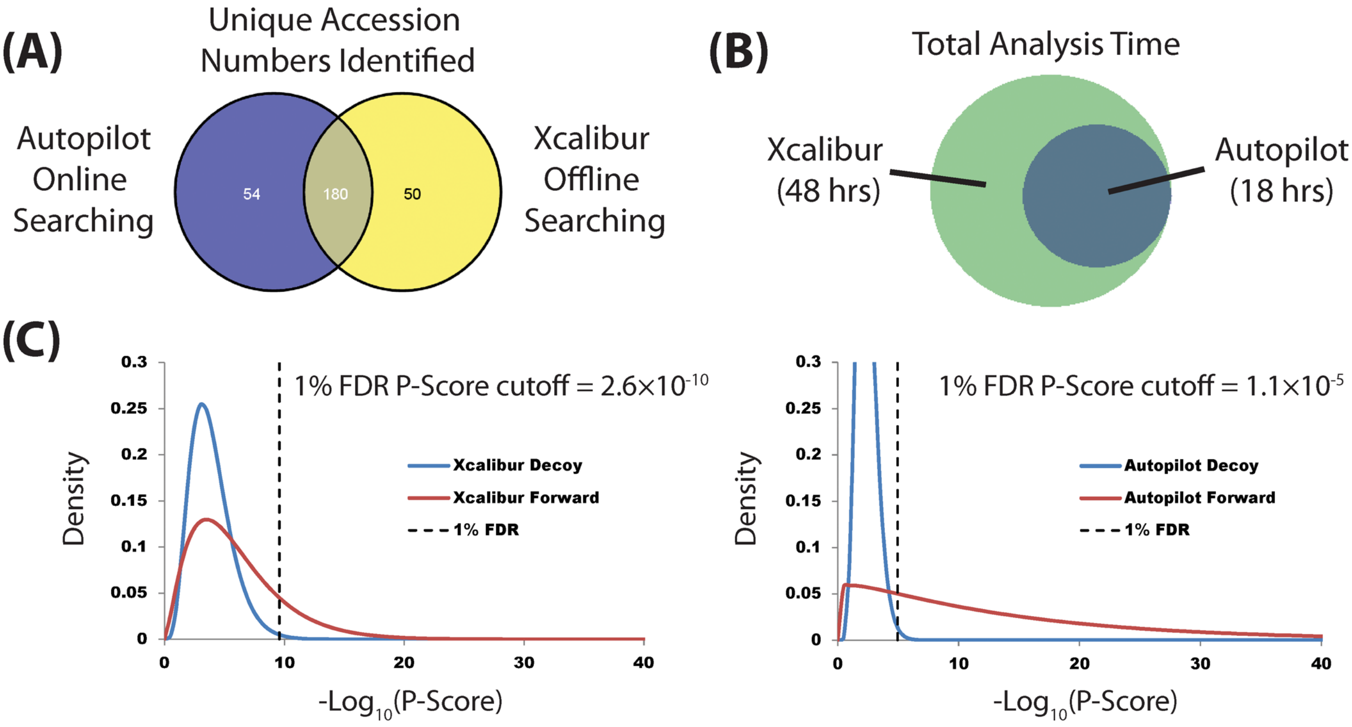


Figure 2. Comparison between Autopilot and the standard data acquisition software for Thermo mass spectrometers, Xcalibur. (A) All Autopilot identifications were generated online during data acquisition with a 2000 Da absolute mass search mode. Meanwhile, the Xcalibur identifications were generated with offline searches with a search tree capable of up to 100 000 Da absolute mass searches. At a 1% FDR cutoff, Autopilot was able to identify 234 unique accession numbers compared to 230 by Xcalibur. (B) The total analysis time for the online data acquisition and offline data analysis with the Xcalibur acquisition mode was 48 h, with 18 h of instrument time. The time for each run is the LC-MS acquisition time of 90 min plus 30 min for spectral summing and ion detection and another 1−2 h for database searches and reports. Because Autopilot processes the data concurrently with acquisition, the time from first injection to end of analysis is 18 h, which is a 270% increase in overall experimental throughput. (C) The decoy and forward gamma distributions from the calculated instantaneous FDRs for every identification event from both data acquisition methods are shown.

If no new targets are found in the precursor scan, the platform queues a scan event for the acquisition of a 15 *m/z* SIM scan. The region selected for the SIM scan is from a part of the spectrum not occupied by an Xtract detected mass. The SIM scan is for discovery purposes, as SIM scans can provide the increased ion statistics required for Xtract to discover new targets. With the noise reduction of the LTQ software, there are frequently areas of the spectrum with no signal and a dedicated SIM scan of these areas can often reveal protein species in this space. In Figure S2 of the Supporting Information, a SIM scan of the highlighted region was taken. While that region had very little signal in the full precursor scan, Xtract detection on the SIM scan produced nine precursor masses, including two methylations. This discovery mode effectively increases the dynamic range of the experiment by preventing the monopolization of the ion current in a given scan by highly abundant protein species.

Acquisition Mode Comparison Study. A comparison of the data acquisition control system described above, named Autopilot, to the standard acquisition software of Thermo mass spectrometers, Xcalibur, was conducted on the low mass GELFrEE fractions from a 10% T GELFrEE fractionation of PAO1. Each fraction was analyzed in technical triplicate with both systems. At a 1% FDR cutoff, Autopilot enabled the confident identification of 234 unique accession numbers while the standard data dependent software identified 230 unique accession numbers, with 180 accession numbers shared between the two data sets (Figure 2A and Table S1 of the Supporting Information).

Importantly, the identifications from Autopilot were found with searches performed during data acquisition by an absolute mass search with a 2000 Da precursor tolerance. This search strategy was chosen to sufficiently limit the search space to enable searches to be performed on the same desktop computer that controls acquisition and finish within the cycle time of the instrument.

The identification mode for Xcalibur utilized an in-house software for precursor and fragment detection followed by a ProSight PC 3.0 search. The ProSight search incorporated a search tree with a first-pass 2.2 Da precursor tolerance absolute mass search and a larger window absolute mass search with a 100 000 Da precursor tolerance if needed. Additionally, for fragmentation scans without an associated precursor, the large window 100 000 Da precursor search was performed.

Figure 2A and B demonstrate the advantages of an online search. Even with the much larger search space used in the offline searches, Autopilot produced more protein identifications with the simple online search. Further, Figure 2B compares the total amount of processing time required to obtain a meaningful list of identifications. With Autopilot searches performed simultaneously with acquisition instead of the typical online acquisition, offline analysis approach, the number of identifications/time increases from 4.8 identifications/hour with the conventional workflow to 13 identifications/hour with Autopilot, a 270% increase in throughput.

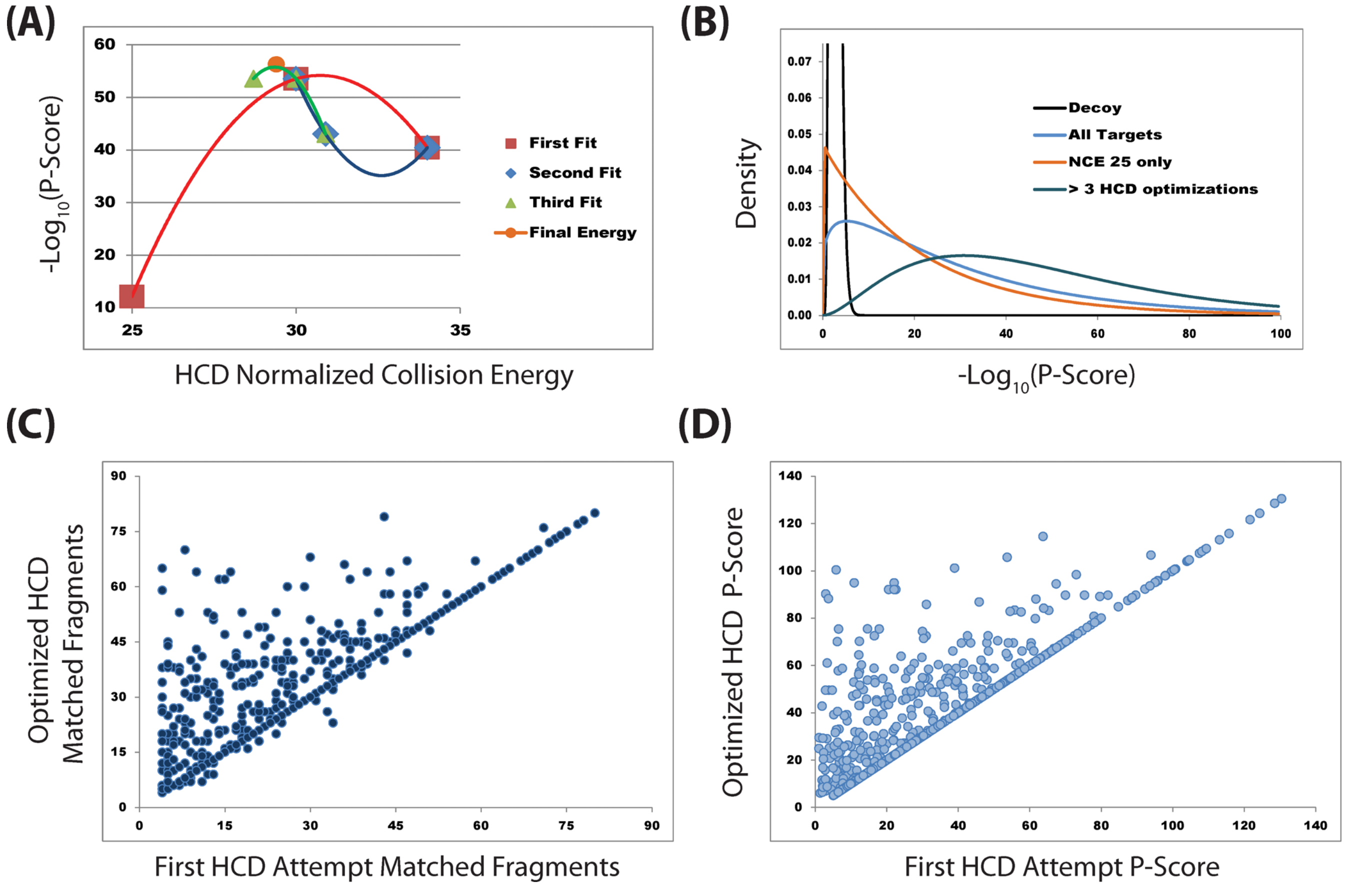


Figure 3. (A) Example of the HCD optimization process is shown for the 10+ charge state at 1032 *m*/*z* of the 50S ribosomal protein L7 (RL7_PSEAE). Each colored line is a different quadratic fit to find the energy that will produce the best P-Score. The maximum for each fit is the energy applied in the new fragmentation scan. After three quadratic optimizations, the best energy for this protein was determined. Table S2 of Supporting Information lists the energy, matched fragments, and P-Score for each fragmentation event. (B) The gamma distributions for the decoy and forward search spaces. The distribution for those species that were optimized more than three times undergoes a sizable shift to the right, which results in a greater percentage of the masses moving past the 1% cutoff. (C) The number of matched fragments from the most optimized scan are compared to those from the first pass, an HCD fragmentation scan with 25 NCE. (D) A similar graph to (C) but with $-\text{Log}_{10}$(P-Score) shown instead.

To validate our search results, the instantaneous FDR, a Bayesian posterior *p* value to measure the FDR for each identification event, was calculated for each Poisson-based P-Score.[4] The instantaneous FDR plots in Figure 2C indicate our directed fragmentation strategy significantly improves the characterization of the identified proteins as the calculated FDRs for each data set are substantially different. The 1% FDR cutoff for Autopilot was 1.1×10^{-5} while the cutoff for Xcalibur was 2.6×10^{-10}. This difference can be attributed to the shift of the decoy distribution toward higher P-Scores with a subsequent shift of the forward distribution toward lower, more confident P-Scores. In other words, there are more matched fragments with a corresponding lower number of unmatched fragments, which reinforces the use of P-score as a reasonable substitute in the absence of a true characterization score.

Fragmentation and Reanalysis of Protein Forms. The increase in characterization seen in the comparison experiment is a result of directed reanalysis of protein targets. Because not every protein behaves in a predictable manner, each fragmented species must be examined to determine if additional fragmentation is needed. After the online protein identification discussed above, the top P-Score proteoform from the search is considered the correct hit for the mass. If the top hit is above the confidence cutoff, the mass is excluded from all future analysis; meaning, when detected again, no fragmentation event will be triggered on any charge state of the species. Alternatively, if the fragmentation led to mediocre results that missed the cutoff for a characterized identification, another fragmentation event can be queued. The subsequent fragmentation of a mass target follows a decision tree composed of ETD fragmentation with a variety of different reaction times and HCD fragmentation with energy optimization (Figure S1 of the Supporting Information).

The fast speed of HCD coupled with generally robust fragmentation coverage makes HCD an ideal starting point for characterization. The first fragmentation event of any detected mass is an HCD scan on the lowest charge state with an NCE of 25. If more fragmentation is required, a different fragmentation technique is used for the first reanalysis. As ETD and HCD have been shown to be complementary fragmentation techniques,[16] the combination of the two can increase overall sequence coverage. An ETD scan of the highest charge state is therefore applied for the reanalysis. The scan has twice as many μscans as the HCD scans to maximize ion statistics of the many, low abundant fragment ions produced from ETD. The highest charge state is selected because the ETD reaction becomes faster and more efficient as the charge

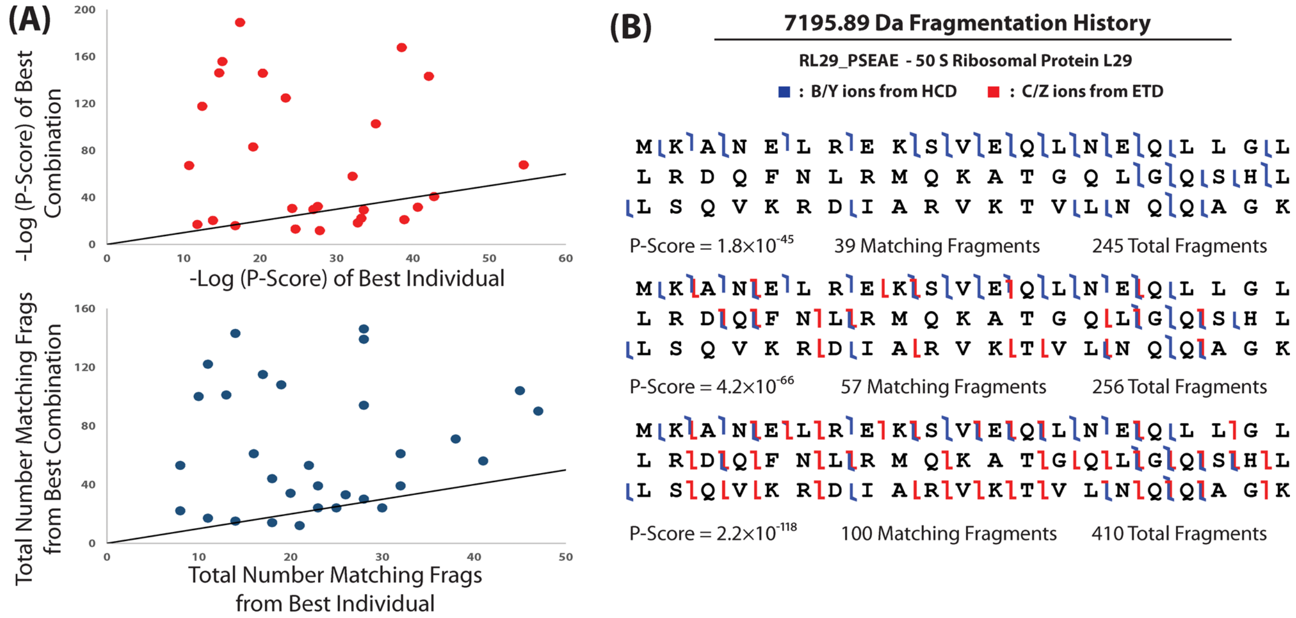


Figure 4. Comparison of targeted masses with greater than one fragmentation event. (A) The utilization of a fragmentation history results in greater sequence coverage (i.e., more matching fragments) and a more confident P-Score than the top hit from the best fragmentation event. (B) The fragmentation maps of the 50 S ribosomal protein L29 depicts the fragmentation history of the 7195.89 Da species. The top map shows the results from the initial HCD fragmentation. The map in the middle is the HCD fragmentation together with the fragments from the first ETD pass. The bottom map is the culmination of all the fragmentation data from the species, with the total combination yielding a confidence score of 2.2 × 10^{-118}.

increases on the protein. McLuckey and co-workers have demonstrated that the rate of an ion−ion reaction is dominated by Coulombic interaction.[17] Therefore, the reaction time is calculated by a simple expression: $t_{\text{precursor}} = z_{\text{precursor}}^2 \times (t_{\text{opt}}/z_{\text{opt}}^2)$, where opt indicates a measured optimal reaction time for a species with a known charge state. In this case, 60 ms for a +3 ion was utilized. If the first ETD attempt does not yield the desired coverage, the software will try various reaction times in an effort to better maximize the ETD reaction and fragmentation. When the ETD optimization is complete, the program will shift efforts to HCD fragment energy optimization if the confidence cutoff has still not been reached.

HCD Optimization. A past attempt was made to generalize the HCD fragmentation energy selection for proteins with a method to estimate optimal energy as a function of the protein's mass and m/z.[18] The equation to govern the choice of energy outlined in that article was developed on a small data set from standard proteins. We applied this equation, along with several other energies (NCE of 20, 25, and 30), across the set of detected proteins from the PAO1 proteome during high-throughput Top Down proteome analysis. We then visualized the P-Scores obtained from the varied HCD energies across the data set with a random set of proteins from the run (Figure S3 of the Supporting Information). The general shape of the energy distributions in relation to the corresponding P-Scores is a defined maximum with steep falloffs on each side.

We adapted this new knowledge into our HCD optimization. To accomplish our optimization, different fragmentation energies are surveyed until a local maximum is found. A quadratic is then fit to the maximum and the closest points on each side. The software will continue to fit the points until the best fit is found. In Figure 3A, the software takes HCD scans of 25, 30, and 34 NCE. The scan of 30 NCE produces a P-Score better than the 25 NCE scan. A higher energy scan (34 NCE) is taken in an effort to produce even better coverage. When the 34 NCE scan returns a P-Score worse than the 30 NCE scan, the quadratic is fit to the three points. This fit is repeated two more times until the optimal value is discovered (orange dot, Figure 3A).

Figure 3B−D showcases the result of this optimization throughout the entire data set. From 1342 unique mass species, 408 were targeted for refragmentation at least once and 204 were retargeted more than three times. Figure 3B illustrates the shift in the forward P-Score gamma distributions upon optimization. After full HCD optimization, the forward distribution has shifted very significantly to the right of the previous distributions. If applied to a completely optimized, full proteomic investigation, many species would be shifted from the unacceptable FDR range to a confident value. Additionally, the $-\log_{10}$ of the P-Scores were continually increased on average as the proteins were reanalyzed with different HCD energies (Figure 3D). The mean change to the $-\log_{10}$ value of the P-Score was 4.0, 6.0, and 18.5 after one, two, and three or more reanalyses, respectively. As evidenced by panel C, there is an almost universal improvement in the characterization of proteins with repeated, directed fragmentation. Furthermore, panel D indicates that, generally, the increase in number of matched fragment ions does not come at the expense of more unmatched fragment ions.

While the end result of Figure 3A is a modestly improved P-Score, the difference between a protein falling outside of the FDR cutoff can often be only a couple of fragment ions. A fragmentation optimization of proteins on the boundary of the confidence cutoff can easily lead to gains that can bring the protein identification into acceptable confidence levels.

Fragmentation History. By analyzing data and intelligently directing acquisition online, metadata generated at runtime may be used to improve the quality of information extracted from a data set. If fragmentation scans were previously performed (during any analysis in a project) on the mass of interest, the former fragment spectra corresponding to that mass can be combined with the new fragmentation spectrum through spectral averaging of the centroid data. The summation of scans increases signal-to-noise of real ions and can lead to detection of new fragment ions. From there, the averaged scan is Xtracted and searched. An improved version of the ProSight search engine, capable of database retrieval on the different sets of ions generated by the different fragmentation techniques in the same search, completes the search. The combination of CID or HCD with ETD ions from the same protein form bolsters identification confidence and improves overall protein characterization. The confidence score for the combined ion types is unchanged if we make a simplification to the score assignment. The score used is similar to the Poisson distribution calculation outlined in Meng et al.:[19] $P_{f,n} = ((xf)^n \times e^{-xf}/n!)$ where $x = (1/111.1) \times 2 \times (M_a \times 2)$, f is the number of detected fragments, and n is the number of random fragment ion matches. The first factor of 2 in x is the number of different ion types in the search type. With separate ETD and HCD searches, the final results can be pooled together and scored with this P-Score calculation if Y ions generated during ETD are ignored. That is, the ETD search will have two ion types, C and Z, and the HCD search will have two ion types, B and Y. Typically, only a small number of Y ions are produced by ETD so any detrimental effects of extra unmatched fragment ions will be minor to the overall score.

On one PAO1 GELFrEE run, significant improvements were produced in the coverage of the masses that were targeted again (Figure 4). The median improvement for the overall combined history in comparison to the individual target was 13 orders of magnitude. The mean improvement was even larger, with a gain of >42 orders of magnitude in confidence. In terms of matched fragments, the median coverage had 29.0 more matched fragments while the mean had 41.1 more matched fragments. A few species had lower P-Scores or less matched fragments because of a poorly fragmented protein spectrum being averaged with the best previous individual fragmentation spectrum. However, this combined result would be ignored in favor of the best individual spectrum upon selection for further fragmentation, but is included for illustration purposes here. Also, an increase in number of matched fragments is not always accompanied by a commensurate increase in P-Score. This can be attributed to a higher number of unmatched fragments detected in the averaged scan, thereby lowering the P-Score.

■ CONCLUSION

This work demonstrates that the implementation of several important data acquisition features can produce large enhancements in the characterization of proteins observed during TDPs investigation. When past data from the same species is coupled with current fragmentation data, a vast improvement to the overall fragmentation coverage of many proteins in our samples was achieved. Further, performing data analysis in parallel with data acquisition enabled increased characterization of proteins in significantly less overall time.

While this work focused on increased characterization of proteins, the benefits of this software in terms of increased unique protein accession numbers will be more visible and directly tractable on newer mass spectrometers that feature tremendous improvements in speed paired with quadrupolar selection of mass species. The speed is desirable for different MS^1 modes, such as SIM scans. As shown in Figure S2 of the Supporting Information, the SIM scan discovery mode can uncover many species previously undetectable because of dynamic range issues. With the quadrupole on these instruments, such as the Thermo Q Exactive, precise selection of these new SIM mass targets should enable large increases in the number of identified proteins. The tight isolation windows will prevent the current problem on hybrid instruments of nearby, abundant proteins being isolated and fragmented with the target species of interest. The abundant fragment ions drown the signal of the target species, making identification of the lower abundance species difficult. With quadrupole isolation, this limitation is all but eliminated.

In addition to cross-platform capabilities, the data acquisition structure of Autopilot is extensible by nature and can eventually be operated in more complex ways. For instance, after initial determination of the protein form with a first pass fragmentation mode, the system could precisely fragment the protein based on primary sequence composition. Also, our flexible database allows for offline searches to be imported into the database which will enable time and computationally intensive searches, such as biomarker searches, to be done offline on a cluster or supercomputer. The results of those searches can then inform future data acquisition decisions in real-time. Lastly, the database is compatible with our previously collected data, which will allow past data sets to be seamlessly incorporated into future proteomic studies.

As TDPs continues to advance, a necessary part of its evolution is more efficient and intelligent data acquisition. The characterization improvements presented here, coupled with advanced instrumentation, will serve as the foundation for a more complete elucidation of the complex proteoforms that comprise higher organisms.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Additional materials as described in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: n-kelleher@northwestern.edu.
*E-mail: philip-compton@northwestern.edu.

Notes

Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation, the National Center for Genome Analysis Support, or Indiana University.

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Institute of General Medical Sciences of the National Institutes of Health under award number R01 GM067193 (N.L.K.). We would also like to acknowledge Bryan Early, Paul Thomas, and Yihan Li, as well as the rest of the members of the Kelleher Research Group, for their help with this work. This research is based upon work supported by the National Science Foundation under grant No. ABI-1062432 to Indiana University.

■ REFERENCES

(1) Tran, J. C.; Doucette, A. A. *Anal. Chem.* **2009**, *81*, 6201−6209.

(2) Lee, J. E.; Kellie, J. F.; Tran, J. C.; Tipton, J. D.; Catherman, A. D.; Thomas, H. M.; Ahlf, D. R.; Durbin, K. R.; Vellaichamy, A.; Ntai, I.; Marshall, A. G.; Kelleher, N. L. *J. Am. Soc. Mass Spectrom.* **2009**, *20*, 2183−2191.

(3) Ahlf, D. R.; Compton, P. D.; Tran, J. C.; Early, B. P.; Thomas, P. M.; Kelleher, N. L. *J. Proteome Res.* **2012**, *11*, 4308−4314.

(4) Tran, J. C.; Zamdborg, L.; Ahlf, D. R.; Lee, J. E.; Catherman, A. D.; Durbin, K. R.; Tipton, J. D.; Vellaichamy, A.; Kellie, J. F.; Li, M.; Wu, C.; Sweet, S. M. M.; Early, B. P.; Siuti, N.; LeDuc, R. D.; Compton, P. D.; Thomas, P. M.; Kelleher, N. L. *Nature* **2011**, *480*, 254−U141.

(5) Catherman, A. D.; Durbin, K. R.; Ahlf, D. R.; Early, B. P.; Fellers, R. T.; Tran, J. C.; Thomas, P. M.; Kelleher, N. L. *Mol. Cell. Proteomics* **2013**, *12*, 3465−3473.

(6) Zamdborg, L.; LeDuc, R. D.; Glowacz, K. J.; Kim, Y. B.; Viswanathan, V.; Spaulding, I. T.; Early, B. P.; Bluhm, E. J.; Babai, S.; Kelleher, N. L. *Nucleic Acids Res.* **2007**, *35*, W701−W706.

(7) Bailey, D. J.; Rose, C. M.; McAlister, G. C.; Brumbaugh, J.; Yu, P. Z.; Wenger, C. D.; Westphall, M. S.; Thomson, J. A.; Coon, J. J. *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109*, 8411−8416.

(8) Graumann, J.; Scheltema, R. A.; Zhang, Y.; Cox, J.; Mann, M. *Mol. Cell. Proteomics* **2012**, *11*, No. M111.013185.

(9) Wenger, C. D.; Boyne, M. T., II; Ferguson, J. T.; Robinson, D. E.; Kelleher, N. L. *Anal. Chem.* **2008**, *80*, 8055−8063.

(10) Zubarev, R. A. *Proteomics* **2013**, *13*, 723−726.

(11) Wessel, D.; Flugge, U. I. *Anal. Biochem.* **1984**, *138*, 141−143.

(12) Catherman, A. D.; Li, M. X.; Tran, J. C.; Durbin, K. R.; Compton, P. D.; Early, B. P.; Thomas, P. M.; Kelleher, N. L. *Anal. Chem.* **2013**, *85*, 1880−1888.

(13) Zabrouskov, V.; Senko, M. W.; Du, Y.; Leduc, R. D.; Kelleher, N. L. *J. Am. Soc. Mass Spectrom.* **2005**, *16*, 2027−2038.

(14) Durbin, K. R.; Tran, J. C.; Zamdborg, L.; Sweet, S. M. M.; Catherman, A. D.; Lee, J. E.; Li, M. X.; Kellie, J. F.; Kelleher, N. L. *Proteomics* **2010**, *10*, 3589−3597.

(15) Fabris, D.; Kelly, M.; Murphy, C.; Wu, Z. C.; Fenselau, C. *J. Am. Soc. Mass Spectrom.* **1993**, *4*, 652−661.

(16) Skinner, O. S.; Catherman, A. D.; Early, B. P.; Thomas, P. M.; Compton, P. D.; Kelleher, N. L. Unpublished work, 2013.

(17) McLuckey, S. A.; Stephenson, J. L.; Asano, K. G. *Anal. Chem.* **1998**, *70*, 1198−1202.

(18) Patrie, S. M.; Ferguson, J. T.; Robinson, D. E.; Whipple, D.; Rother, M.; Metcalf, W. W.; Kelleher, N. L. *Mol. Cell. Proteomics* **2006**, *5*, 14−25.

(19) Meng, F. Y.; Cargile, B. J.; Miller, L. M.; Forbes, A. J.; Johnson, J. R.; Kelleher, N. L. *Nat. Biotechnol.* **2001**, *19*, 952−957.