Journal of proteome research

Technical Note

pubs.acs.org/jpr

Wavelet-Based Peak Detection and a New Charge Inference Procedure for MS/MS Implemented in ProteoWizard's msConvert

William R. French,[†] Lisa J. Zimmerman,[‡] Birgit Schilling,[§] Bradford W. Gibson,[§] Christine A. Miller,[∥] R. Reid Townsend,[⊥] Stacy D. Sherrod,[#] Cody R. Goodwin,[○] John A. McLean,[○] and David L. Tabb*,[†,‡]

[†]Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, Tennessee 37232-8340, United States

[‡]Jim Ayers Institute for Precancer Detection and Diagnosis, Vanderbilt-Ingram Cancer Center, Nashville, Tennessee 37232-6350, United States

[§]Buck Institute for Research on Aging, Novato, California 94945, United States

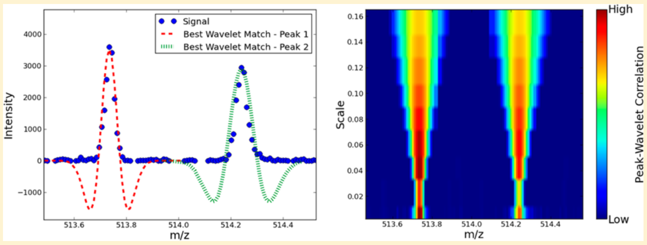
[∥]Agilent Technologies, Santa Clara, California 95051, United States

[⊥]Department of Cell Biology and Physiology, Washington University School of Medicine, St. Louis, Missouri 63110, United States,

[#]Department of Physics and Astronomy, [○]Department of Chemistry, Vanderbilt University, Nashville, Tennessee 37235, United States

S *Supporting Information*

ABSTRACT: We report the implementation of high-quality signal processing algorithms into ProteoWizard, an efficient, open-source software package designed for analyzing proteomics tandem mass spectrometry data. Specifically, a new wavelet-based peak-picker (CantWaiT) and a precursor charge determination algorithm (Turbocharger) have been implemented. These additions into ProteoWizard provide universal tools that are independent of vendor platform for tandem mass spectrometry analyses and have particular utility for intralaboratory studies requiring the advantages of different platforms convergent on a particular workflow or for interlaboratory investigations spanning multiple platforms. We compared results from these tools to those obtained using vendor and commercial software, finding that in all cases our algorithms resulted in a comparable number of identified peptides for simple and complex samples measured on Waters, Agilent, and AB SCIEX quadrupole time-of-flight and Thermo Q-Exactive mass spectrometers. The mass accuracy of matched precursor ions also compared favorably with vendor and commercial tools. Additionally, typical analysis runtimes (~1−100 ms per MS/MS spectrum) were short enough to enable the practical use of these high-quality signal processing tools for large clinical and research data sets.



KEYWORDS: *Continuous wavelet transformation, peak-picking, signal deconvolution, precursor charge determination, deisotoping, open-source software, mass spectrometry*

■ INTRODUCTION

Processing raw spectra from a tandem mass spectrometry (MS/MS) experiment is an essential analysis step that often goes overlooked. While advances in instrument technology have enabled unprecedented levels of protein detection and sensitivity,[1−3] it is vital to continue developing robust signal processing algorithms that take full advantage of improved instrument performance. Common signal processing steps include peak-picking, ion charge determination, and deisotoping. Each of these steps must be automated in order to process the huge number of spectra and peaks produced by modern instruments. Ideally, these processing steps are performed in a high-quality manner as efficiently as possible so that valuable information contained in raw data sets can be obtained in a reasonable amount of time. Achieving consistently high-quality analysis is especially challenging across data acquired on different instrument platforms, where disparities in signal quality, shape, and resolution require sufficiently universal processing approaches that can account for these differences. For this reason, signal processing software is often written, tested, and optimized to operate on data of a specific resolution or quality, and vendor software generally processes only raw data from its own instruments. This data-specific, vendor-specific approach can constrain users to a particular quality of analysis and can preclude unbiased comparisons of data sets spanning multiple platforms. Moreover, vendor and commercial software can often be expensive or difficult to access. To address these challenges, a set of open-source tools capable of performing reliable and efficient signal processing across a range of different instrument and data types is needed.

Received: August 24, 2014
Published: November 20, 2014

Peak-picking is a key signal processing step that, depending on the quality of an algorithm, can drastically impact downstream peptide identifications. Many approaches to peak-picking have been developed over the years. High-intensity peaks are generally straightforward to identify using some form of intensity or signal-to-noise ratio (SNR) threshold combined with signal-smoothing or noise-filtering;[4] however, this approach often breaks down for ions with peak intensities that are close to the noise level. When peaks are close to the noise threshold, shape-matching and curve-fitting methods can aid in discriminating true peaks from noise by evaluating the shape of the peak and/or the distribution of nearby peaks (in both the m/z and retention time domains). For example, MaxQuant[5] fits a Gaussian curve to the raw m/z points in each peak. MSInspect,[6] OpenMS,[7] and DyWave[8] all employ wavelet transformations to assess the correlation between a model wavelet and spectral data. Du and co-workers[9] developed a similar algorithm called MassSpecWavelet using continuous wavelet transformations (CWTs) and showed that this approach outperformed methods based on peak intensity and discrete wavelet transformations. Another feature that can be used to discern a true peak from noise is the presence of an isotopic envelope. For example, VIPER[10] and Hardklör[11] use the THRASH[12] algorithm, which performs isotope pattern matching for peak identification. Similarly, the commercial software package Mascot Distiller fits models of isotopic envelopes to profile data in order to discriminate between true peaks and noise.[13,14] A clear advantage of the isotope-based approach is the ability to simultaneously perform peak-picking, deisotoping, and ion charge determination.

Determining ion charge is another important signal processing step, as it enables a precursor m/z value to be converted to the mass of an unknown peptide for database searching. The general procedure for determining precursor charge involves analyzing a small isolation window within the survey scan from which the precursor ion is derived. Charge is determined by computing the reciprocal of m/z spacing between peaks and, in most cases, evaluating the relative intensities of these peaks. Averagine[15] is perhaps the most widely used model (e.g., see OpenMS[7] or Hardklör[11]) for predicting the relative intensities of the peaks within an isotopic envelope, although alternate methods exist. For instance, MSInspect[6] models the intensities with a Poisson distribution and then uses the Kullback−Leibler divergence to evaluate the similarity between the modeled and observed intensities. In addition to the previously mentioned isotopic modeling techniques, other precursor charge determination algorithms include a customized isotopic wavelet method,[16] a Fourier transformation technique,[17] and scoring approaches.[18]

In this work, we present results from implementing and applying high-quality signal processing algorithms in ProteoWizard,[19] an open-source, proteomics MS/MS analysis package. Specifically, a peak-picker entitled CantWaiT builds on the CWT method (MassSpecWavelet) developed by Du et al.[9] by tuning it for computational efficiency. For charge determination, we introduce a new algorithm called Turbocharger that scores groups of candidate isotopic peaks based on their m/z spacing, relative intensity distribution, and total intensity. We compared the performance of CantWaiT and Turbocharger with that of vendor and commercial signal processing software; the results demonstrated that our processing software produces identifications and mass accuracy on par with vendor and commercial software across numerous types of samples and instruments.

■ METHODS

The algorithms described below can be run using the msConvert tool as part of the freely distributed ProteoWizard package (binaries and source code are available at www.proteowizard.sourceforge.net). ProteoWizard supports commonly used vendor raw files and many open formats (e.g., mzML, mzXML, mz5) and is executed as native C++ code in order to maximize runtime performance. As such, this software package provides an ideal programming environment for running efficient cross-platform analyses. ProteoWizard also provides seamless access to many vendor peak-picking libraries, enabling users to compare results across different algorithms.

CantWaiT: A Wavelet Peak-Picker

The key signal processing algorithm that we incorporated into ProteoWizard was a spectral peak-picker. In order to maximize processing quality and cross-platform performance, we selected the continuous wavelet transformation (CWT) method for peak selection. The CWT method is noted for its performance in noisy, low-resolution spectra while also performing well on smooth, high-resolution data.[9,20,21] For the parent wavelet, we chose the Ricker ("Mexican Hat") wavelet for its unimodal and symmetric properties, providing a generic means of detecting peptides regardless of the strength of their isotopic distributions.

CantWaiT is based on the MassSpecWavelet algorithm introduced by Du et al.,[9] but we have built upon their implementation to reduce analysis runtime and to better handle spectra of varying resolution. Previously, the method's biggest disadvantage was slow runtime, with the original developers reporting a processing time of ∼30 s for a single spectrum, which renders it impractical for use with large clinical and research data sets. While this runtime was reported in 2006 when processor speeds were slower than they are today, the authors themselves acknowledged that code enhancements for improving the algorithm's efficiency were needed. We reduced runtime by limiting the CWT computation to 10 carefully chosen wavelet scales. A challenge for choosing the appropriate scales is the tendency for peaks and m/z spacing to widen with increasing m/z; consequently, a good set of scales at low m/z is seldom optimal at high m/z values. CantWaiT addresses this problem by applying a novel, adaptive method in which the wavelet scales that are sampled depend on m/z. Each spectrum is quickly scanned to determine the appropriate scales to sample as a function of m/z, by computing the average m/z spacing in a 10-point sliding window for each m/z point. The wavelet transformation is then performed using wavelet scales that vary linearly between one and seven times the m/z spacing computed during preprocessing. CWT computations are performed with wavelets centered on each m/z point and midpoint in the spectrum. In addition to being computationally efficient, CantWaiT also handles irregularly spaced data or spectra with zero-intensity points removed. Thus, no additional signal preprocessing such as smoothing or resampling is needed prior to peak-picking.

Unlike MassSpecWavelet, CantWaiT does not perform ridge line identification in wavelet space for locating peaks, which further improves CantWaiT's computational efficiency. Ridge line identification is one of the more computationally intense steps of MassSpecWavelet in which local maxima are connected along both the m/z and wavelet scale axes. CantWaiT avoids this step by simply scanning the correlation matrix a single time, locating local maxima that are separated by at least 0.1 m/z. While this is a less rigorous procedure for filtering true peaks from noise, we found that it provided consistent results with the ridge line identification procedure at a fraction of the computational cost. The local maxima identified by CantWaiT are next filtered via SNR thresholding, where the signal is defined as the maximum wavelet correlation value in the ridge line. To calculate the noise, CantWaiT separates a spectrum into bins composed of 300 m/z points each. The noise level for each point in a bin is then set to the 95-percentage quantile of correlation at the smallest scale within that bin. In the final peak list, CantWaiT records the m/z and intensity values corresponding to the maximum intensity value within each peak. Pseudocode for CantWaiT is including in the Supporting Information, Algorithm S1.

Turbocharger: A Precursor Charge Determination Algorithm

In MS/MS experiments, mass analyzers catalog intense peaks in a survey scan and then isolate these ions for fragmentation in tandem MS. Ions are generally isolated for fragmentation within a small window (e.g., ±1.25 m/z) centered at a target m/z value. Analyzing the isotopic content of this region is the most conventional means of determining a precursor ion charged state. This process is not always straightforward. For example, multiple ions with overlapping peaks may be present (i.e., chimeric spectra[22]), or strong evidence for isotopic features may be lacking. A robust algorithm must cope with these situations while remaining applicable among many different instruments. Turbocharger achieves this by using three criteria (m/z spacing, relative intensity distribution, and total intensity rank) to score how well groups of peaks conform to expected isotopic behavior. When the software is first initiated, it performs simulations to determine how to convert scores on the three criteria described below to p-values for evaluating isotopic packets.

Turbocharger first uses CantWaiT to perform peak-picking within a window surrounding the instrument-reported target isolation m/z. Next, Turbocharger builds chains of peaks whose interpeak spacing is equal to $m_{\mathrm{neutron}}/q \pm \varepsilon$, where m_{neutron} is the mass of a neutron (here computed as the difference between the ^{13}C and ^{12}C masses: 1.00335 Da), q is the charge state associated with the chain (generally any positive integer corresponding to a peptide's charged state), and ε is a fixed m/z tolerance of 0.06. This tolerance is wide enough to catch isotopic peaks whose peak spacing is offset compared to the theoretical spacing, but it is not so wide that the number of chains becomes unnecessarily large (as estimated empirically using the data sets presented in this paper). All possible subchains within a chain are also scored; for instance, if peaks A, B, and C make up a chain, then chains A−B and B−C also receive their own scores. The maximum chain length is capped at five, while chains of charge five or greater must contain at least three peaks.

The first scoring metric used by Turbocharger is m/z spacing. While previous algorithms have not scored isotope distributions on the basis of m/z spacing, we have found this to be an important metric for evaluating whether a set of peaks belong to an isotope envelope or appeared at this spacing by random chance. Turbocharger employs a sum of squared errors (SSE) method for scoring the fidelity of m/z spacing within a chain, similar to a published strategy from sequence tagging.[23] Specifically, each peak's m/z value is used to estimate the leading (i.e., the monoisotopic peptide peak for conventional peptides) peak m/z; this is done by subtracting km_{neutron}/q from each peak, where k is the peak number with respect to the monoisotopic peak. These estimates are applied to compute the SSE of the monoisotopic peak, which is then converted to a p-value by comparing to a simulated distribution of SSEs. In these simulations, the SSEs are computed for chains consisting of randomly positioned (i.e., spaced by $m_{\mathrm{neutron}}/q \pm \varepsilon\gamma$, where γ is a uniformly distributed value between zero and one) peaks; each possible chain length is simulated 10 000 times. This simulation technique is used to estimate the null distribution of SSE scores. The p-value is the fraction of simulated SSE scores that produce lower SSE values than computed for a set of putative isotopes.

The relative intensity distribution is the second scoring metric evaluated by Turbocharger. We chose to implement the Poisson model introduced by Breen et al.[24] for predicting the relative intensities of peptide isotopes as a function of m/z and charge. Specifically, Turbocharger follows the implementation in MSInspect[6] where the Kullback−Leibler (KL) divergence is used to evaluate the similarity between the modeled and observed intensities. The KL score for each chain is converted to a p-value by comparing to a simulated distribution of KL scores. In these simulations, CantWaiT first identifies the 200 most intense peaks in each of nine survey scans (one at every 10th percentile in retention time). Turbocharger then builds hypothetical isolation windows around each of the intense peaks, and the KL score is computed for random sets of peaks within these windows, with 10 000 samples used for each chain length.

The final scoring metric used by Turbocharger is based on the intensity rank sum of the chain. This criterion is meant to capture the tendency for an abundant precursor to produce more fragments than a less abundant precursor. Turbocharger ranks each peak in the precursor isolation window and then sums the ranks of the peaks included in a putative isotope set to evaluate the total intensity score of a chain. The rank-sum score is converted to a p-value by computing the fraction of possible rank-sums less than the calculated value.

Once scoring is complete, the chains are ranked by each metric, and these ranks are summed together to produce the final chain score. Turbocharger does not always select the lowest scoring chain for assigning charge and the monoisotopic m/z. For instance, Turbocharger requires a relative intensity p-value of less than 0.30. Once Turbocharger has found the lowest scoring chain that satisfies this criterion, Turbocharger checks if another chain in the top eight consists of the same peaks plus an additional leading peak. If such a chain is found, and the relative intensity p-value is less than 0.30, then this chain is selected in place of the original. This is intended to correct the monoisotope m/z value in the event that a subchain scores better than a longer version of this chain that also scores well. If no chain among the top eight has a relative intensity p-value of less than 0.30, then Turbocharger returns the default charge(s) and target isolation m/z. The default charges may be specified by the user. By default, Turbocharger assigns no charge to indeterminate spectra. Pseudocode for Turbocharger is included in the Supporting Information, Algorithm S2.

Table 1. Summary of Samples Used for Evaluating ProteoWizard Signal Processing Algorithms[a]

instrument	sample	replicates	min. charge	max. charge
Agilent 6530	Bovine 6	8	+1	+5
Agilent 6550	Serum (Human)	4	+2	+14
AB SCIEX Triple TOF 5600	UPS1 (Human)	5	+2	+8
AB SCIEX Triple TOF 5600	*E. coli*	3	+2	+8
AB SCIEX Triple TOF 5600	Rat Liver	3	+2	+8
Thermo Q-Exactive	Bovine 6	5	+2	+6
Thermo Q-Exactive	Jurkat (Human)	1	+2	+6
Waters Synapt G2	Bovine 6	3	+1	+8
Waters Synapt G2	Yeast	5	+1	+8
Waters Synapt G2	*E. coli*	1	+1	+8
Waters Synapt GS-S	K562 (Human)	3	+1	+8

[a] For each vendor, sample complexity increases moving from top to bottom. The Turbocharger precursor charge search range is also listed and was taken from the vendor-reported ranges for each dataset.

Deisotoping

Peaks from MS/MS spectra are deisotoped using the Poisson model approach combined with the KL divergence score, as described in the previous section. Peaks are chained together in the same way described above except that here the interpeak spacing must be within 100 ppm of m_{neutron}/q. Note that a fixed interspeak spacing was applied by Turbocharger to facilitate the simulation of SSE scores. By default, the Poisson deisotoper checks charge states between one and three. If a chain yields SSE and KL scores below the assigned cutoff values, then the nonmonoisotopic peaks within the chain are removed from the peak list. The SSE and KL cutoffs depend on the length of the chain; the values that we applied are listed in the Supporting Information (Table S1).

DATASETS

In order to assess our software's cross-platform performance, we tested numerous simple and complex sample types (Table 1) from the following instruments: AB SCIEX Triple TOF 5600 quadupole time-of-flight (QqTOF), Agilent 6530/6550 QqTOF, Thermo Scientific Q-Exactive, and Waters Synapt G2/G2-S QqTOF. All raw data sets are publicly available through the mass spectrometry interactive virtual environment (MassIVE) at http://massive.ucsd.edu/ProteoSAFe. Please note that in many cases these data sets were collected solely for software evaluation purposes and were not necessarily acquired under conditions representative of optimal instrument performance. For instance, the Waters Synapt G2/G2-S is typically run under ion-mobility enhanced data-independent acquisition to maximize proteome coverage and quantitative reproducibility. Moreover, common samples were not necessarily selected to be run across all of the platforms, as interplatform comparisons are not and should not be inferred from these data. Rather, the capabilities of ProteoWizard signal processing is compared with the specific software typically used with each specific platform on which the data were generated to assess the universality of our software's performance versus what would ordinarily be achieved with each specific platform. Details related to sample preparation and data acquisition are provided briefly in the Supporting Information.

PEPTIDE DATABASE SEARCHING AND PROTEIN ASSEMBLY

Results from signal processing in msConvert were searched using two peptide database engines: MyriMatch[25] and MS-GF+.[26] These two tools employ different matching approaches, with MyriMatch modeling the number of matched vs unmatched peaks using the multivariate hypergeometric distribution, and MS-GF+ scoring MS/MS spectra using a prefix residue mass graph. Testing these two search engines allowed us to evaluate the robustness of results from ProteoWizard signal processing.

Database searches were configured similarly between MyriMatch and MS-GF+. Files were searched against full proteome-level FASTA databases, which also included the reverse sequences for each protein in order to compute the false discovery rate (FDR); the peptide search space was generated by allowing a maximum of two missed cleavages and one missed termini cleavage (semitryptic digest), with a minimum peptide length of 5 amino acids and a maximum length of 75 amino acids. In addition, the following static modifications were applied: carboxymethylation of cysteine at any position for the bovine 6 (simple mixture of 6 bovine proteins) data and carbamidomethylation of cysteine at any position for all other files. Dynamic modifications for all files included oxidation of methionine at any position and ammonia loss from glutamine at the peptide N-terminus; deamindation of glutamine or asparagine at the protein N-terminus was also applied for bovine 6 data. Searching tolerances were set to 50 and 100 ppm (in m/z) for precursors and fragments, respectively, with the exception of the Thermo Q-Exactive files, which were searched using 10 and 20 ppm. MyriMatch and MS-GF+ can both be configured to search precursor m/z values that are shifted to different isotopic peaks, which is meant to account for incorrectly assigned monoisotopic peaks. While checking for misidentified monoisotopic peaks generally improves overall identifications, we chose to disable this option in order to compare how well the various software platforms are able to locate the monoisotopic peak.

The results from peptide database searches were imported into IDPicker[27] (version 3.0), which assembles all matched peptides into their parsimonious proteins. The default import settings were applied in all cases. Specifically, a maximum Q-value of 2% (for peptide spectrum matches) and a minimum of two distinct peptides per protein group were required.

ANALYSIS METHODS

To evaluate the performance of the PreoteoWizard signal processing software, we compared results to vendor and commercial software. We applied the same four processing steps employed by ProteoWizard signal processing software:

peak-picking of MS/MS spectra, precursor *m/z* assignment, precursor monoisotope assignment, and deisotoping. Thermo and Agilent data were processed using the vendor peak-picking and precursor charge determination libraries available through ProteoWizard's msConvert tool; neither of these vendor libraries include deisotoping, so we also applied deisotoping on the Thermo and Agilent peak lists using ProteoWizard's Poisson deisotoper. AB SCIEX raw files were analyzed with the AB SCIEX MS Data Converter using the "proteinpilot" peak list option. Waters data were processed using the commercial software package Mascot Distiller; default processing parameters were applied, with the following exceptions: maximum precursor charge of +8, precursor search tolerance of ±1.5 *m/z*, precursor grouping tolerance of 0.05 *m/z*, minimum precursor mass of 1.0 Da, maximum subscan summation time of 30 s, and minimum peak list size of one. The results from vendor (or commercial in the case of Waters data) signal processing were next searched using MyriMatch and MS-GF+ and then assembled using IDPicker 3.0.

For analysis of raw data with our new signal processing software, we applied CantWaiT, Turbocharger, and the Poisson deisotoper as filters accessed through msConvert; see Table S2 in the Supporting Information for examples of how to apply these filters within msConvert and Tables S3−S5 for lists of parameters and their allowed values. Note that CantWaiT peak-picking was performed only on MS/MS spectra, although it may also be applied to MS1 spectra, if desired. Waters data also required subscan summation prior to these three processing steps; we implemented a simple algorithm (which is also available within msConvert) that summed the intensity at identical *m/z* values across MS/MS spectra with the same precursor value (within 0.05 *m/z* and 30 s in retention time, consistent with the Mascot Distiller settings). Turbocharger was applied by searching the charge ranges listed in Table 1. Two survey spectra were analyzed for determining precursor charge for each MS/MS spectrum; two spectra immediately preceding (in retention time) the MS/MS spectrum were used for Thermo and Waters data, whereas one spectrum preceding and one spectrum proceeding the MS/MS spectrum were used for AB SCIEX and Agilent data. Isolation widths of ±1.00 and ±1.25 *m/z* were applied for AB SCIEX and Thermo files, respectively, whereas ±1.50 *m/z* was applied for all other data sets. These settings for survey spectra and isolation width were applied because they led to optimal peptide identification rates and are summarized in Table S6. For cases where Turbocharger could not identify a reliable isotopic distribution in the isolation widow, default charges were set to values consistent with those used from vendor and commercial software: +2 for Thermo, +2 and +3 for Agilent and Waters, and no charge for AB SCIEX.

CantWaiT accepts two command-line arguments for adjusting peak-filtering: a minimum SNR (in wavelet space) and minimum *m/z* peak spacing. By default, CantWaiT applies a minimum SNR of 1.0 and peak spacing of 0.1 *m/z*. We have found that for complex samples a minimum SNR of 1.0 generally results in identifications within 5% of the maximum number of identifications across a range of minimum SNR values (see Figure S1 in the Supporting Information); however, we recommend testing different values for analysis of new data sets. We have observed optimal values ranging from 0.0 (no filtering of peaks based on the SNR) to above 2.0. Adjusting the peak spacing parameter can be useful for high-resolution data, where peaks may be resolved within narrow *m/z* ranges; for instance, we applied a value of 0.01 *m/z* for analysis of Thermo Q-Exactive data.

■ RESULTS

Peak List Size Distributions

Because the number of peptide identifications may be sensitive to the minimum SNR, we compared identifications on the basis of equal peak list size. This was accomplished by adjusting the minimum SNR until the total number of peaks was roughly equal to that of vendor or commercial software. Note that this approach still resulted in a different number of peaks in individual MS/MS spectra when comparing CantWaiT and vendor/commercial software. This is apparent in Figure 1, which shows the peak list size distributions for complex samples measured on four instruments. The discrepancies in Figure 1 do not necessarily reflect gaps in algorithm quality, but rather highlight differences in peak resolution and in peak-picking/filtering methods. For instance, the close agreement in the Thermo panel is a result of highly resolved (17 500 for MS/MS spectra) peaks, which makes peak identification more straightforward and thus more consistent between CantWaiT and Thermo software. For less ideal data, CantWaiT is versatile enough to produce reasonable agreement in peak list distributions across different instrument and software platforms. The main feature of CantWaiT that makes it well suited for achieving consistent performance across different platforms is the calculation of correlation values between the Ricker wavelet and raw profile data. This procedure accounts for peak shape and makes CantWaiT less sensitive to an instrument's sampling rate than methods based on peak intensity alone.

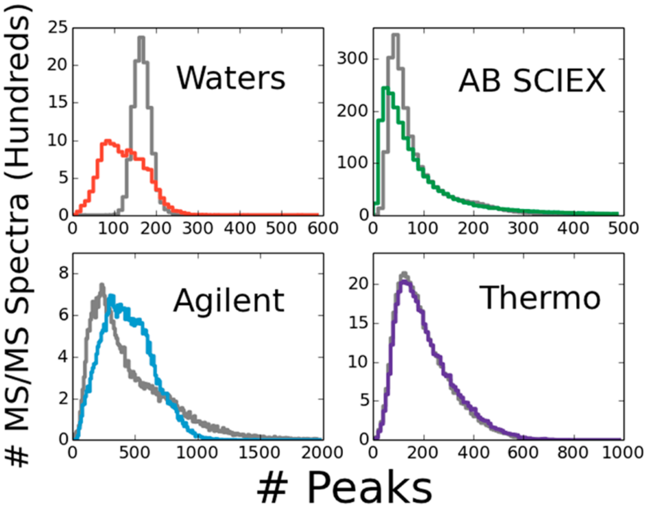


Figure 1. Comparison of peak list size distributions for the most complex samples analyzed on four instruments. The *x* axis represents the number of peaks in an individual MS/MS spectrum. Gray curves correspond to results from vendor or commercial software, and colored curves correspond to results produced by CantWaiT peak-picking. Note that the samples used for each platform are different and therefore comparisons should be made only within each individual panel.

Precursor Mass Accuracy

An important metric for evaluating signal processing performance is mass accuracy, which measures how close the observed masses of matched ions are to the predicted locations. Figure 2

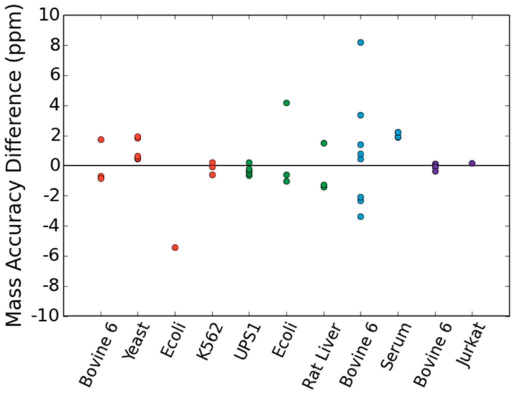


Figure 2. Median precursor mass accuracy difference between precursors matched from ProteoWizard vs vendor/commercial software. Positive values indicate that ProteoWizard's median mass accuracy is better (i.e., smaller on an absolute scale) than the median mass accuracy of vendor/commercial software (value on y axis = $|MA_{vendor}| - |MA_{Pwiz}|$, where MA is the median mass accuracy). Red data correspond to Waters Synapt G2/G2-S; green, AB SCIEX Triple TOF 5600; blue, Agilent 6530/6550 QqTOF; and purple, Thermo Q-Exactive. Sample complexity increases for each vendor moving from left to right.

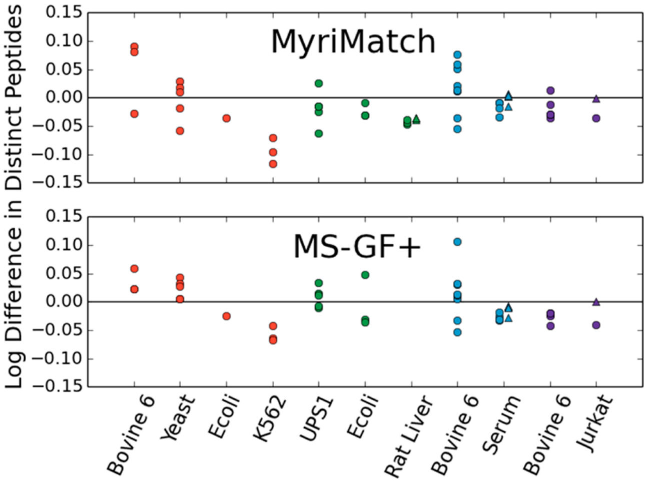


Figure 3. Logarithm of the ratio of the number of distinct peptides identified by ProteoWizard to the number of peptides identified by vendor/commercial software and searched using (Top) MyriMatch and (Bottom) MS-GF+. Positive values indicate that identifications were higher with ProteoWizard signal processing. Circles indicate that all signal processing was performed within ProteoWizard, and triangles correspond to cases where ProteoWizard peak lists were combined with vendor-reported precursor charges and monoisotopic m/z values. The symbol colors indicate the vendor; see the caption to Figure 2 for details. Note that rat liver data analyzed by MS-GF+ were removed due to suspected software errors.

compares the median mass accuracy of matched precursors in each data set resulting from MyriMatch searches. The agreement between CantWaiT peak positions and vendor software is high, with the majority of the results falling within 3 ppm. This result establishes that CantWaiT selects peak positions accurately and, in many cases, outperforms software tools that were optimized for a particular quality of spectra.

Distinct Peptide Identifications

We also compared the total number of distinct peptides resulting from application of ProteoWizard vs vendor/commercial signal processing. Figure 3 plots the $\log_{10}$ difference between distinct peptides identified from ProteoWizard and vendor/commercial processing. For reference, log difference values of 0.05, 0.10, and 0.15 indicate that ProteoWizard identified 12, 26, and 41% more peptides than that of vendor/commercial software, respectively.

Overall, ProteoWizard signal processing software produced identifications that were comparable to those of vendor and commercial software, with the number of identified peptides generally within ∼10%. This is true for both search engines, across different sample complexities measured on numerous instruments. There are exceptions (discussed below) where ProteoWizard lags behind vendor/commercial software more significantly (e.g., Waters G2-K562 data); however, there are also cases where the opposite is true. While there appear to be some minor differences between using MyriMatch vs MS-GF+ in Figure 3, overall these differences are insignificant.

A perhaps less obvious trend in Figure 3 is the slight drop in relative identifications for complex samples. To investigate this behavior further, we isolated different areas of the signal processing workflow by combining CantWaiT peak lists with vendor precursor charges and monoisotopic m/z values. We performed this analysis for all vendors except Waters, where Mascot Distiller's summation of MS/MS subscans resulted in a different number and order of spectra compared to that from ProteoWizard.

The general result of exchanging precursor information (i.e., charge and monoisotopic m/z value) from Turbocharger with those from vendor software was an increase in the number of identified peptides. In the Thermo Jurkat sample, exchanging precursor information produced consistent identifications with peak lists from Thermo software, suggesting that CantWaiT peak-picking is on par with Thermo software for the Jurkat sample. Exchanging precursor information produced less drastic improvements for the AB SCIEX and Agilent samples. Overall, CantWaiT appears to produce identifications on par with Agilent while slightly lagging behind AB SCIEX peak-picking.

The Effect of Precursor Charge Assignment

To examine the performance of Turbocharger, we compared the precursor charges reported by vendor software with those predicted by Turbocharger. Figure 4 plots the percent agreement between vendor-reported and Turbocharger-reported charge states on complex samples from all vendors except Waters (Waters data were again not included due to differences in the number and order of spectra between Mascot Distiller and ProteoWizard). We first analyzed results from the Thermo Jurkat data, which is a valuable benchmark, as acquisition-time processing required that the precursor isotope cluster in the MS1 spectrum match an averagine-based theoretical cluster. Hence, it is very unlikely for the Thermo software's precursor charge or monoisotopic peak assignment to be wrong. The agreement with Thermo software is high (>95%) across all of the reported charge states, indicating that Turbocharger assigned reasonable charges for a vast majority of MS/MS spectra. Interestingly, this small discrepancy in charge assignment still resulted in a noticeable disagreement in the number of identified peptides (Figure 3).

For Agilent and AB SCIEX data, the level of agreement between Turbocharger and vendor software was lower, in particular for high charge states. This does not necessarily indicate poor algorithm performance on the part of Turbocharger, but rather differences in the way these

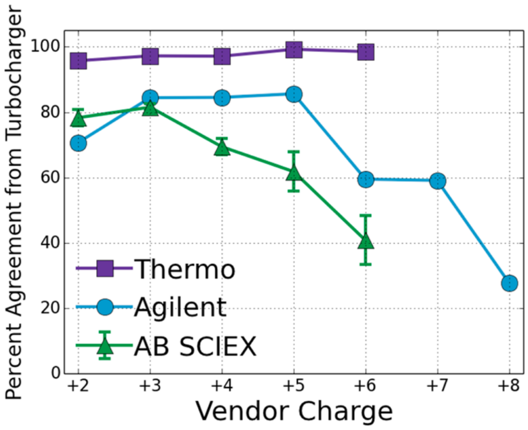


Figure 4. Percent of vendor-assigned precursor charges that were assigned the same charge from Turbocharger for the most complex sample analyzed from each vendor (except for Waters). Error bars span ±1 SD and are visible when greater than the symbol size. For each charge state, a sample size of at least 200 was required for inclusion. Note that comparisons should not be inferred between vendors, as each data set is different in sample and size, but should be made only to the agreement of Turbocharger with the vendor charge state assignments.

algorithms handle chimeric spectra (i.e., isolation windows containing multiple precursor ions). In spite of the disagreement in charge states, the results in Figure 3 suggest that Turbocharger is still able to identify a comparable number of peptides for these samples. In the case of the Agilent serum data, ProteoWizard signal processing actually produced more spectrum matches than vendor software for all four replicates. However, ProteoWizard slightly underperformed vendor software on the basis of distinct peptides, suggesting that Agilent software was better able to identify new peptides that are not repeatedly identified across multiple MS/MS spectra. An analysis of the overlap in distinct matches between ProteoWizard and vendor software corroborated this point (see Figure S2 in the Supporting Information).

In order to better understand the impact of chimeric spectra on ProteoWizard signal processing, we computed the frequency of the presence of multiple precursors in all data sets. MS/MS spectra were classified as chimeric if at least two of the top-eight-ranked isotope chains possessed different charge states, with relative intensity scores of less than 0.30 and total intensity scores of less than 0.10. For all tested samples, we found a maximum frequency of 51% in the Waters K562 data; recall that ProteoWizard performance (relative to Mascot Distiller) was lowest in this data set among all that we tested. Other complex samples for Thermo, AB SCIEX, and Agilent were found to contain multiple charges in the MS1 isolation window in 13, 27, and 35% of MS/MS spectra, respectively, further suggesting that the discrepancies in Figure 4 are due to the presence of multiple precursor ions in the MS1 isolation window. An example of overlapping isotopic distributions from the Agilent serum data is provided in the Supporting Information (see Figure S3 and Table S7), along with a more detailed discussion of how Turbocharger handles situations where multiple precursor ions are present.

The Effect of Deisotoping

The final signal processing algorithm we tested was the Poisson deisotoper. Deisotoping can improve identifications by reducing the total number of peaks in a MS/MS spectrum while retaining the number of matched peaks (assuming none of the matched peaks are incorrectly removed); in other words, deisotoping can improve a peptide search engine's confidence that matched peaks are not false positives. Indeed, we found that deisotoping increased the number of identified peptides for all replicates in the most complex samples tested. The increases were marginal in some cases (e.g., 0.1%), but ranged as high as 5.6%. See Figure S4 in the Supporting Information for a plot of the percent increase in distinct peptide identifications across all complex data sets. In addition to improving CantWaiT peak lists, the Poisson deisotoper can also be applied to improve nondeisotoped vendor peak lists. For instance, peptide identifications were improved by 8.2% by applying deisotoping to vendor peak lists of one of the Agilent serum samples.

Analysis Runtime

Computational efficiency was one of the primary design goals of the ProteoWizard signal processing software. The runtime for analysis depends on a number of factors, primarily the number of points in a typical peak and in the entire spectrum. As a test case, in the Supporting Information (Figure S5) we have included a plot showing the CantWaiT analysis time as a function of the number of m/z points in a spectrum for the Thermo Jurkat sample. The runtime ranged from ~1−5 ms for spectra containing less than ~2000 points to ~80 ms for a spectrum containing ~20 000 points. Meanwhile, Turbocharger executed in ~5−15 ms per spectrum, depending of the density of m/z points within the isolation window and number of isotope chains that are built and scored.

Note that runtime also depends on the hardware employed for analysis. The timings reported in the previous paragraph are from benchmarks run on a 3.5 GHz Intel Core i7-4770K processor. We additionally incorporated support for multi-threading in msConvert, allowing multiple spectra to be analyzed simultaneously if multiple CPU cores are available on a processor.

■ CONCLUSIONS

We have incorporated high-quality signal processing tools into ProteoWizard that can be easily accessed and applied through msConvert for analyzing files in a variety of raw and open file formats (please refer to Table S2 in the Supporting Information for examples of how to apply the tools through msConvert). Comparisons to vendor and commercial signal processing algorithms showed that these new tools perform well across many different types and qualities of spectral data, both in terms of precursor mass accuracy and identified peptides. This, combined with the high computational efficiency of CantWaiT and Turbocharger, makes the tools an attractive option for analysis of proteomic MS/MS data.

In the future, we plan to improve these signal processing tools further. Current plans include exploring strategies for improving how chimeric spectra are handled by Turbocharger and incorporating peak information across adjacent scans in retention time to improve CantWaiT performance.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

Pseudocode for CantWaiT and Turbocharger; KL and SSE score cutoffs used for the Poisson deisotoper; a table of example commands in msConvert; a table of parameters for CantWaiT, Turbocharger, and Poisson deisotoper; parameters applied for Turbocharger analysis; a plot of peptide identification as a function of the minimum SNR for complex samples; peptide overlap between search results on ProteoWizard peak lists vs vendor/commercial software peak lists; an example of a real chimeric MS/MS spectrum, with a detailed discussion of how Turbocharger treats cases where multiple precursor ions are present; a plot showing the percent increase in peptide identifications with deisotoping applied for complex samples; a plot showing the CantWaiT runtime as a function of the number of m/z points in a spectrum; and data set details. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Phone: 615-936-0380; Fax: 615-343-8372; E-mail: david.l.tabb@vanderbilt.edu.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We thank the Proteomics Platform group at the Broad Institute of MIT and Harvard for providing the Q-Exactive Jurkat lysate tryptic digest dataset, Arthur Moseley at Duke University for sharing the Waters yeast data, Roy Martin for providing the K562 cell lysate data from the Waters Synapt G2-S, John Cottrell for assisting us with evaluation licenses of Mascot Distiller, and John Wikswo for helpful discussion. We gratefully acknowledge support from the National Institutes of Health through grant no. 5UH2TR000491-2 (W.R.F., S.D.S., C.R.G., J.A.M.) and the Department of Defense through DARPA-11-73-MPSys-FP-011 (W.R.F., S.D.S., C.R.G., J.A.M., D.L.T.) and CBMXCEL-XL1-2-0001 (W.R.F., S.D.S., C.R.G., J.A.M.). W.R.F., L.J.Z., and D.L.T. received support from the National Cancer Institute through grant no. U24 CA159988. B.W.G. acknowledges support by an National Center for Research Resources shared instrumentation grant for the Triple TOF 5600 (1S10 OD016281). R.R.T. acknowledges the support of the National Institute of General Medical Sciences (8 P41 GM103422-35) from the National Institutes of Health.

■ ABBREVIATIONS

AGC, automatic gain control; CWT, continuous wavelet transformation; DDA, data-dependent acquisition; DTT, dithiothreitol; IAA, iodoacetic acid; KL, Kullback−Leibler; LC, liquid chromatography; MS, mass spectrometry; MS1, mass spectrum of precursor ions; MS/MS, tandem mass spectrometry or tandem mass spectrum; m/z, mass over charge; QqTOF, quadrupole time-of-flight; SNR, signal-to-noise ratio; SSE, sum of squared errors; TCEP, tris(2-carboxyethyl)phosphine; TOF, time-of-flight; UHPLC, ultra-high-performance liquid chromatography

■ REFERENCES

(1) Angel, T. E.; Aryal, U. K.; Hengel, S. M.; et al. Mass spectrometry-based proteomics: existing capabilities and future directions. *Chem. Soc. Rev.* **2012**, *41*, 3912−28.

(2) Nesvizhskii, A. I.; Vitek, O.; Aebersold, R. Analysis and validation of proteomic data generated by tandem mass spectrometry. *Nat. Methods.* **2007**, *4*, 787−797.

(3) Bantscheff, M.; Lemeer, S.; Savitski, M. M.; Kuster, B. Quantitative mass spectrometry in proteomics: critical review update from 2007 to the present. *Anal. Bioanal. Chem.* **2012**, *404*, 939−65.

(4) Mantini, D.; Petrucci, F.; Pieragostino, D.; et al. LIMPIC: a computational method for the separation of protein MALDI-TOF-MS signals from noise. *BMC Bioinf.* **2007**, *8*, 101.

(5) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367−72.

(6) Bellew, M.; Coram, M.; Fitzgibbon, M.; et al. A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC−MS. *Bioinformatics* **2006**, *22*, 1902−9.

(7) Sturm, M.; Bertsch, A.; Gröpl, C.; et al. OpenMS—an open-source software framework for mass spectrometry. *BMC Bioinf.* **2008**, *9*, 163.

(8) Wang, P.; Yang, P.; Arthur, J.; Yang, J. Y. H. A dynamic wavelet-based algorithm for pre-processing tandem mass spectrometry data. *Bioinformatics* **2010**, *26*, 2242−9.

(9) Du, P.; Kibbe, W. A.; Lin, S. M. Improved peak detection in mass spectrum by incorporating continuous wavelet transform-based pattern matching. *Bioinformatics* **2006**, *22*, 2059−2065.

(10) Monroe, M. E.; Tolić, N.; Jaitly, N.; Shaw, J. L.; Adkins, J. N.; Smith, R. D. VIPER: an advanced software package to support high-throughput LC−MS peptide identification. *Bioinformatics* **2007**, *23*, 2021−3.

(11) Hoopmann, M. R.; Finney, G. L.; Maccoss, M. J. High speed data reduction, feature detection, and MS/MS spectrum quality assessment of shotgun proteomics datasets. *Anal. Chem.* **2007**, *79*, 5620−5632.

(12) Horn, D. M.; Zubarev, R. A.; Mclafferty, F. W. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J. Am. Soc. Mass Spectrom.* **2000**, *11*, 320−332.

(13) Berndt, P.; Hobohm, U.; Langen, H.; Technologies, G.; Roche, F. H. Reliable automatic protein identification from matrix-assisted laser desorption/ionization mass spectrometric peptide fingerprints Proteomics and 2-DE. *Electrophoresis* **1999**, *20*, 3521−3526.

(14) Gras, R.; Müller, M.; Gasteiger, E.; et al. Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis* **1999**, *20*, 3535−3550.

(15) Senko, M. W.; Beu, S. C.; Mclafferty, F. W. Determination of monoisotopic masses and ion populations for large biomolecules from resolved isotopic distributions. *J. Am. Soc. Mass Spectrom.* **1995**, *6*, 229−233.

(16) Hussong, R.; Tholey, A.; Hildebrandt, A.; et al. Efficient analysis of mass spectrometry data using the isotope wavelet. *AIP Conf Proc.* **2007**, *940*, 139−149.

(17) Tabb, D. L.; Shah, M. B.; Strader, M. B.; Connelly, H. M.; Hettich, R. L.; Hurst, G. B. Determination of peptide and protein ion charge states by Fourier transformation of isotope-resolved mass spectra. *J. Am. Soc. Mass Spectrom.* **2006**, *17*, 903−15.

(18) Zhang, Z.; Marshall, A. G. A universal algorithm for fast and automated charge state deconvolution of electrospray mass-to-charge ratio spectra. *J. Am. Soc. Mass Spectrom.* **1998**, *9*, 225−33.

(19) Chambers, M. C.; Maclean, B.; Burke, R.; et al. A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* **2012**, *30*, 918−20.

(20) Yang, C.; He, Z.; Yu, W. Comparison of public peak detection algorithms for MALDI mass spectrometry data analysis. *BMC Bioinf.* **2009**, *10*, 4.

(21) Emanuele, V. A.; Gurbaxani, B. M. Benchmarking currently available SELDI-TOF MS preprocessing techniques. *Proteomics* **2009**, *9*, 1754−62.

(22) Houel, S.; Abernathy, R.; Renganathan, K.; Meyer-arendt, K.; Ahn, N. G.; Old, W. M. Quantifying the impact of chimera MS/MS spectra on peptide identification in large-scale proteomics studies. *J. Proteome Res.* **2010**, *9*, 4152−4160.

(23) Tabb, D. L.; Ma, Z.; Martin, D. B.; Ham, A. L.; Chambers, M. C. DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring. *J. Proteome Res.* **2008**, *7*, 3838−3846.

(24) Breen, E. J.; Hopwood, F. G.; Williams, K. L.; Wilkins, M. R. Automatic Poisson peak harvesting for high throughput protein identification proteomics. *Electrophoresis* **2000**, *21*, 2243−2251.

(25) Tabb, D. L.; Fernando, C. G.; Chambers, M. C. MyriMatch: highly accurate tandem mass spectral peptide identification by multivariate hypergeometric analysis. *J. Proteome Res.* **2007**, *6*, 654−661.

(26) Kim, S.; Gupta, N.; Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **2008**, *7*, 3354−3363.

(27) Ma, Z.; Dasari, S.; Chambers, M. C.; et al. IDPicker 2.0: improved protein assembly with high discrimination peptide identification filtering. *J. Proteome Res.* **2009**, *8*, 3872−3881.