analytical chemistry

Article

pubs.acs.org/ac

UVnovo: A *de Novo* Sequencing Algorithm Using Single Series of Fragment Ions via Chromophore Tagging and 351 nm Ultraviolet Photodissociation Mass Spectrometry

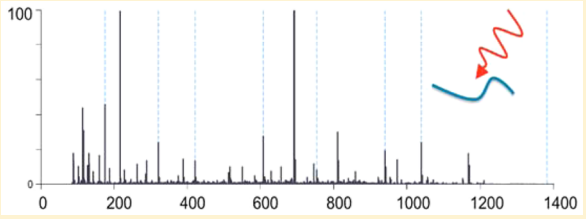
Scott A. Robotham,[†,§] Andrew P. Horton,[‡,§] Joe R. Cannon,[†] Victoria C. Cotham,[†] Edward M. Marcotte,[*,‡] and Jennifer S. Brodbelt[*,†]

[†]Department of Chemistry, University of Texas, Austin, Texas 78712, United States

[‡]Center for Systems and Synthetic Biology, Department of Molecular Biosciences, University of Texas, Austin, Texas 78712, United States

Supporting Information

ABSTRACT: *De novo* peptide sequencing by mass spectrometry represents an important strategy for characterizing novel peptides and proteins, in which a peptide's amino acid sequence is inferred directly from the precursor peptide mass and tandem mass spectrum (MS/MS or MS^3) fragment ions, without comparison to a reference proteome. This method is ideal for organisms or samples lacking a complete or well-annotated reference sequence set. One of the major barriers to *de novo* spectral interpretation arises from confusion of N- and C-terminal ion series due to the symmetry between *b* and *y* ion pairs created by collisional activation methods (or *c*, *z* ions for electron-based activation methods). This is known as the "antisymmetric path problem" and leads to inverted amino acid subsequences within a *de novo* reconstruction. Here, we combine several key strategies for *de novo* peptide sequencing into a single high-throughput pipeline: high-efficiency carbamylation blocks lysine side chains, and subsequent tryptic digestion and N-terminal peptide derivatization with the ultraviolet chromophore AMCA yield peptides susceptible to 351 nm ultraviolet photodissociation (UVPD). UVPD-MS/MS of the AMCA-modified peptides then predominantly produces *y* ions in the MS/MS spectra, specifically addressing the antisymmetric path problem. Finally, the program UVnovo applies a random forest algorithm to automatically learn from and then interpret UVPD mass spectra, passing results to a hidden Markov model for *de novo* sequence prediction and scoring. We show this combined strategy provides high-performance *de novo* peptide sequencing, enabling the *de novo* sequencing of thousands of peptides from an *Escherichia coli* lysate at high confidence.



The breadth of proteomic studies has never been greater, as a growing trend in proteomics research pushes mass spectrometry experiments beyond the study of model organisms, proteotypic peptides, and common posttranslational modifications. This strains the limits of traditional spectral interpretation using sequence databases, and it has driven development of more flexible search methods and proteogenomic pipelines. *De novo* peptide and protein sequencing is one potential strategy for characterizing novel peptides.[1] Rather than comparing a peptide spectrum to theoretical candidate spectra from a reference protein sequence database, *de novo* analysis directly infers a peptide sequence from the precursor peptide mass and tandem mass spectrum (MS/MS or MS^3) fragment ions.[2] This method is ideal for organisms or samples lacking a complete or well-annotated reference sequence set. In the event that gene sequences are available, *de novo* approaches are well-suited for interpreting unidentified spectra and discovering unknown splice variants, intergenic peptides, sequence polymorphisms, and other novel peptides.

Given an ideal MS/MS spectrum, *de novo* peptide sequence assignment is a trivial exercise. Such a spectrum would exhibit a complete series of ions, all of a single-fragment type (N-terminal *a*/*b*/*c* or C-terminal *x*/*y*/*z* ions) and known charge state, that span an entire precursor peptide. The sequence could then be read directly from the spectrum by matching the mass difference between each consecutive ion pair to its corresponding amino acid. Technological developments, notably high-resolution MS/MS acquisition and concurrent collection of complementary fragmentation spectra (e.g., paired collision-induced dissociation (CID)/electron transfer dissociation (ETD) mass spectra), have greatly improved the potential of *de novo* peptide sequencing, but spectra still suffer from incomplete peptide fragmentation, complex fragmentation patterns and neutral losses, and uninterpretable noise. CID,[3,4] HCD,[5] ETD,[6,7] and dual fragmentation (EThcD, ETciD),[8] have all been applied for *de novo* sequencing. Infrared multiphoton dissociation (IRMPD) and ultraviolet photo-

Received: January 20, 2016
Accepted: March 2, 2016
Published: March 3, 2016

DOI: 10.1021/acs.analchem.6b00261
Anal. Chem. 2016, 88, 3990−3997

dissociation (UVPD)[9−12] are also emerging as viable alternatives for tandem mass spectrometry of peptides.

One of the major barriers to *de novo* spectral interpretation arises from confusion of N- and C-terminal ion series due to the symmetry between *b* and *y* ion pairs created by collisional activation methods (or *c*, *z* ions for electron-based activation methods). This is known as the "antisymmetric path problem" and leads to inverted amino acid subsequences within a *de novo* reconstruction.[13] A related difficulty arises when fragment ions with similar m/z values cannot be independently resolved.[14] Biased peptide backbone fragmentation, the most serious problem, leads to spectral regions without fragment ion evidence and precludes definition of a complete amino acid sequence. These issues have made it unrealistic in practice to assign full and accurate peptide sequences in an automated *de novo* fashion. Therefore, database searches still greatly outperform *de novo* in any complex bottom-up shotgun proteomics experiment for which representative sequence data are available. Many modern *de novo* algorithms compensate by reporting tens or hundreds of putative sequences for a single peptide spectrum or only partial peptide sequences containing gaps where amino acids cannot be derived.[15,16] The results are most useful after manual curation or homology-based database comparisons, where such hybrid sequence tag-based homology searching combines the flexibility of *de novo* sequencing with the identification power provided through database comparison.

Among the many *de novo* programs available today, a few of the more popular established and emerging options include PEAKS, PepNovo, NovoHMM, pNovo, DirecTag, and Novor for bottom-up proteomics and Twister for top-down analysis.[14,17−22] Most such tools use statistical models of peptide fragmentation for spectral interpretation prior to sequence generation or for scoring candidate *de novo* sequences constructed from simple initial assumptions and rules. These fragmentation models are rooted in the idea of the offset frequency function (OFF), introduced by Dančík et al. in 1999.[23] Fundamentally, OFF treats fragmentation as a stochastic process whereby specific ions (example, b^+, y^+-NH_3) have a certain chance for being observed from each peptide residue position. These models are highly dependent on the type of spectra used during construction, limiting the application of existing software for new spectral paradigms.

In parallel to the continued development of *de novo* interpretation software, considerable effort has focused on creating "ideal" spectra for *de novo* sequencing through novel sample preparation and instrumentation methods.[2] Most of these methods have been implemented to overcome the antisymmetric path problem or, more generally, the issue of discerning product ion type. Differential labeling between two samples, via isotopic or chemical modification of peptide N- or C-termini, is applied to evoke a mass difference between product ion pairs and allows MS^2 ion type annotation.[24,25] Alternatively, spectral simplification through chemical derivatization and charge sequestration can either enhance or eliminate a particular fragment ion series. In particular, peptide termini may be modified to increase the relative abundance of either the N- or C-terminal ion series.[26,27] Changing the basicity or charge of a peptide terminus influences the charge localization of and charge mobility within a peptide and, consequently, produces a more prominent series of fragment ions from the end where charge is concentrated.

We recently demonstrated marked spectral simplification through a combination of chromophore derivatization and UVPD-MS.[28,29] The simplification mechanism, fundamentally different from those described above, destroys rather than neutralizes redundant fragment ions. By attaching the chromophore 7-amino-4-methylcoumarin 3-acetic acid (AMCA) to a peptide N-terminus, the peptide becomes susceptible to 351 nm photoactivation. The selectivity of 351 nm UVPD ensures that only AMCA-derivatized peptides undergo photodissociation, and successive laser pulses effectively eliminate N-terminal chromophore-containing ions. C-terminal product ions (without a chromophore) remain unaffected by the UVPD, and the process yields a clean series of *y* ions uniformly distributed along the entire peptide length.

In this work, we combine three key strategies for *de novo* peptide sequencing into a single high-throughput pipeline: (i) covalent modification of peptides and (ii) 351 nm UVPD fragmentation to favor elimination of N-terminal products and survival of C-terminal fragment ions with (iii) a dedicated software platform, UVnovo, to interpret these data. We introduce an improved strategy for selective peptide N-terminal AMCA derivatization. This is accomplished through highly efficient carbamylation of lysine side-chain amines[30] prior to tryptic digestion and AMCA labeling. LC-UVPD-MS/MS of the AMCA-modified peptides then predominantly produces *y* ions in the MS/MS spectra, specifically addressing the antisymmetric path problem. Finally, the program UVnovo applies a random forest (RF) algorithm[31] to automatically learn from and then interpret UVPD spectra, passing results to a hidden Markov model (HMM) for *de novo* sequence prediction and scoring. We show this combined strategy provides high-performance *de novo* peptide sequencing.

MATERIALS AND METHODS

Materials. Trypsin Gold, mass spectrometry grade, was purchased from Promega (Madison, WI, USA). LC-MS grade acetonitrile and water were purchased from EMD Millipore (Darmstadt, Germany). Phosphate-buffered saline (PBS) and dimethyl sulfoxide (DMSO) were purchased from Thermo Fisher Scientific Inc. (San Jose, CA, USA). Sulfosuccinimydyl-7-amino-4-methylcoumarin-3-acetic acid (Sulfo-NHS-AMCA) was purchased from Pierce Biotechnology (Rockford, IL, USA). *Escherichia coli* (*E. coli*) lysate was graciously donated by Dr. M. Stephen Trent's research group at the University of Texas at Austin.

Modification of *E. coli* Lysate. Figure 1 shows the process for N-terminal AMCA peptide derivatization. A 50 μg amount of *E. coli* lysate in 100 μL of 50 mM sodium carbonate and 8 M urea was heated at 80 °C for 4 h to carbamylate lysine side chains (ε-amines) and the N-termini primary amine of each protein, blocking subsequent reaction with AMCA. (The carbamylation reaction of all primary amines means that the N-terminus and N-terminal peptide of each protein will not be characterized by UVPD-MS). Urea was removed through PBS buffer exchange, and proteins were then digested using trypsin at 37 °C overnight. After digestion, 25 μL of 20 mM sulfo-NHS-AMCA in DMSO was added to approximately 270 μL of the digest to label the primary N-terminal amine of each peptide (except the N-terminus of each protein which was already blocked by carbamylation in the first step), and the solution was kept in the dark overnight at room temperature. Samples were cleaned using a C18 SPE cartridge to facilitate removal of unreacted AMCA, evaporated to dryness, and resuspended for LC-MS/MS (98% water/2% acetonitrile with 0.1% formic acid). We anticipate that carbamylation could also

DOI: 10.1021/acs.analchem.6b00261
Anal. Chem. 2016, 88, 3990−3997

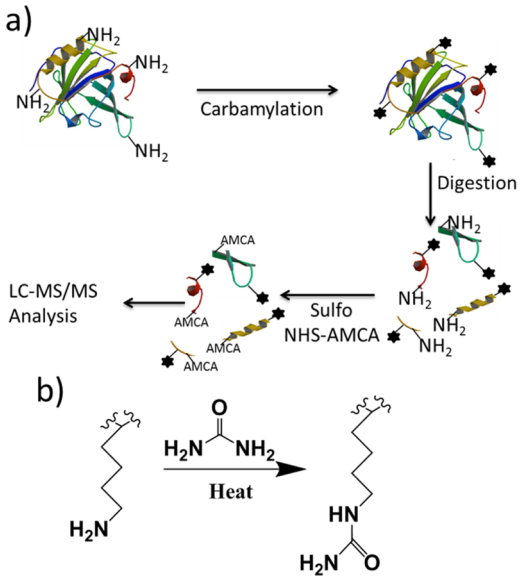


Figure 1. (a) Workflow for carbamylation/AMCA modification and (b) carbamylation reaction.

react with monomethyl-lysine, following slower kinetics, and if the presence of substantial monomethyl-lysines is expected, then ArgC could be used as an alternative protease.

LC-MS/MS Analysis of *E. coli* Lysate. All mass spectra were acquired using a Thermo Scientific Velos Pro dual linear ion trap mass spectrometer (Thermo Scientific, San Jose) modified for UVPD by addition of a Coherent 351 nm excimer laser (Coherent, Santa Clara, CA, USA) to allow 351 nm UV excitation of ions present in the ion trap.[32] The laser was set to 3 mJ/pulse at 500 Hz, with 15 pulses/scan. Peptides were separated by reverse phase chromatography and eluted into the mass spectrometer using a Dionex NSLC 3000 nanoLC system (Thermo Scientific, Waltham, MA, USA). We used a 15 cm capillary column (75 μm i.d.) packed with 3.5 μm particles (C18 stationary phase) with a pore size of 140 Å, loading 5 μg of peptide mixture (via 1 μL injection). Sample elution followed a 360 min gradient starting at 3% B and increasing to 50% B with a flow rate of 300 nL/min; solvent A was water with 0.1% formic acid (v/v), and solvent B was acetonitrile with 0.1% formic acid (v/v).

SEQUEST. In order to obtain a list of high-confidence peptide spectral matches, raw spectra were analyzed using the SEQUEST and Percolator nodes of Proteome Discoverer v. 1.4 (Thermo Fisher, San Jose). AMCA was required as a fixed N-terminal modification, and optional oxidized methionine in any position was allowed. The precursor mass tolerance was set at ±1.6 Da due to the low resolution of ion trap spectra. Because trypsin does not cleave at carbamylated lysines, SEQUEST protease specificity was set to trypsin (R) and included the proline rule. We considered only *y* ion fragments for the UVPD data sets, searching spectra against the UniProt *E. coli* strain K12 reference proteome.

***De Novo* Analysis Using UVnovo.** We implemented UVnovo, a *de novo* sequencing program for analysis of UVPD spectra, in the MATLAB programming language. All top-ranked high-confidence SEQUEST peptide spectrum matches (PSMs) from doubly charged precursor ions (2+) were used to train and validate UVnovo using 3-fold cross-validation as follows:

Spectral Partitioning and Preprocessing. Spectra were randomly partitioned into three sets. All spectra from a given peptide, collapsing PTM variants, were allocated to the same set, preventing their use for both training and validation. During each of the three cross-validation rounds, a different partition was treated as an "unknown" test set, and the "known" spectra in the remaining two partitions were used for model training. We repeated this three times, withholding a different test partition each time, to evaluate the performance of UVnovo against the high-confidence SEQUEST PSMs.

Thermo *.raw files were converted to the mzXML format using MSConvert with peak picking, and peaks with an intensity < 5 were removed. Through an unexplained artifact of UVPD spectral generation, all fragment ions in the MS^2 spectra

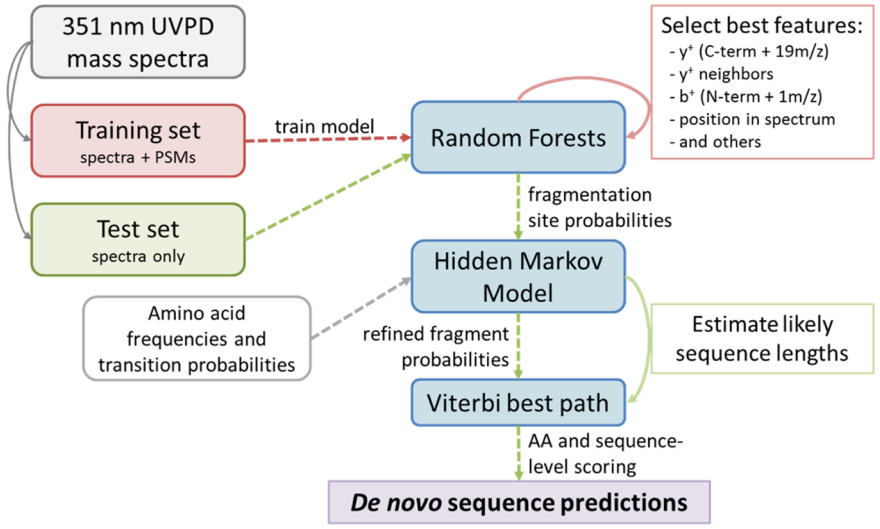


Figure 2. UVnovo workflow for *de novo* sequencing. Spectra are divided into training and test sets. A random forest, trained on known spectra, transforms an unknown spectrum into a simplified representation of peptide cleavage site probabilities. At each position in this "simplified spectrum", a hidden Markov model (HMM) refines the probability, also incorporating amino acid frequencies and requiring valid mass transitions. The best valid path through the HMM yields the *de novo* sequence prediction, and the individual fragmentation site probabilities provide a means to score each sequence.

DOI: 10.1021/acs.analchem.6b00261
Anal. Chem. 2016, 88, 3990−3997

from all precursors less than m/z 817.2 were systematically shifted up 0.16 m/z by the instrument, whereas peaks of the remaining spectra displayed no such systematic mass error. This was corrected in preprocessing by subtracting 0.16 m/z from all peaks of the affected spectra. Additionally, because our goal was to evaluate the potential of UVPD-MS/MS for automated *de novo* peptide sequencing, we chose to marginalize the effect of incorrect precursor mass on *de novo* sequence assignment. We set the precursor mass for each spectrum to the integer mass nearest its respective SEQUEST PSM. Thus, our results should be understood as being contingent on an accurate definition of precursor mass, to the nearest 1 Da, well within the capacity of modern high-resolution instruments.

UVnovo Overview. Figure 2 presents the overall software workflow. Following data import and preprocessing, spectral interpretation follows four main steps: (1) transform each MS/MS spectrum into a spectral representation of peptide cleavage site probability at each possible mass position (this applies a random forest model for peptide fragmentation pattern deconvolution); (2) refine the backbone cleavage site predictions using a hidden Markov model; (3) identify amino acid sequences that best fit the predictions; (4) score and rank the *de novo* sequence reconstructions.

Details of the fragment pattern deconvolution, fragment site scoring, HMM construction, and assignment and scoring of *de novo* sequences are described in full in the Supporting Information.

RESULTS AND DISCUSSION

We based our strategy to enhance *de novo* peptide spectrum interpretation on the ability of UVPD to efficiently generate C-terminal fragment ion (y ion) series while eliminating N-terminal ions (a, b ions). This strategy required efficient attachment of a UV chromophore to the N-terminus of each peptide in order to target them by 351 nm UVPD, while avoiding labeling of lysine side chains that would result in indiscriminant chromophore attachment. We describe a sample processing scheme that accomplishes these goals and enables UVPD-based *de novo* peptide sequencing. We also introduce the *de novo* sequencing program UVnovo, as to date there is no *de novo* sequencing program suitable for analysis of 351 nm UVPD mass spectra.

Lysine Capping with Carbamylation. In order to confine AMCA modification to the N-termini of peptides, the ϵ-amino group on lysine side chains must first be blocked. We have previously employed lysine guanidination for this purpose, converting lysines into homoarginines via reaction with *O*-methylisourea in the presence of 7 M ammonium hydroxide.[8,29] Here, we improve on this strategy and instead convert lysine to homocitrulline via carbamylation. This provides a quick and efficient alternative to guanidination for blocking the reactive ϵ-amino group on lysine side chains. Heating samples at 80 °C for 4 h in an 8 M urea solution resulted in complete reaction of reactive primary amines on model proteins, including the N-termini and lysine side chains.[30] As a proof of concept we evaluated carbamylation efficiency on intact myoglobin molecules before and after the carbamylation reaction, using direct injections of the intact protein into a high-resolution Thermo Orbitrap Elite. With 19 lysine residues and a free N-terminus, myoglobin has 20 amine reactive sites (Supporting Information Figure S1a). We observed a mass shift of 860.09 Da between the modified and unmodified forms, very close to the 860.116 Da expected from complete carbamylation (20 × 43.0058 Da). We estimated it to be nearly complete based on the ESI mass spectrum shown in Figure S1b,c. A similar analysis of intact ubiquitin (data not shown) also revealed complete lysine carbamylation.

351 nm UVPD Spectra. Figure 3 presents a representative UVPD mass spectrum for a peptide from *E. coli* elongation

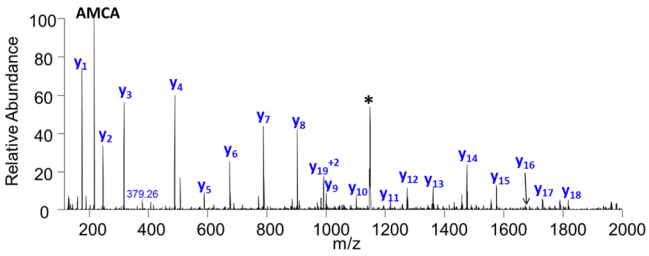


Figure 3. UVPD (3 mJ/pulse, 15 pulses) mass spectrum of elongation factor G peptide V[AMCA]YSGVVNSGDTVLNSVK[carbamyl]AAR (2+) from a trypsin-digested *E. coli* lysate. The precursor is labeled with an asterisk.

factor G protein. The clean series of y ions is consistent with 351 nm UVPD and demonstrates the effective annihilation of b ions during the activation period (i.e., 15 laser pulses). The b ions (which contain the N-terminus) retain the AMCA chromophore and are susceptible to photoabsorption and dissociation during successive laser pulses. Very few internal fragment ions are observed. While fragment ions are often diminished C-terminal to proline (akin to conventional collisional activation), peptide cleavage otherwise produces a comprehensive series of observable y ions. In general, photoactivation using 351 nm photons results in cleavage of the C−N backbone bonds analogous to that observed upon collisional activation. There is no evidence for production of a/x and c/z ions more commonly observed upon 157 nm UVPD or 193 nm UVPD.[10,12]

In one regard, spectral symmetry is beneficial for low-resolution data because b and y ion pairs provide the most effective means for correct *de novo* precursor mass assignment.[19,23,33,34] The lack of complementary ion pairs and other telltale MS/MS signatures of precursor mass in our data precluded effective mass error correction. After a baseline correction of systematic error, only 63% of the *E. coli* lysate spectra we used for benchmarking (described below) had a mass within ±0.5 Da of the SEQUEST PSM. In all results below, the precursor mass was therefore assigned as the integer nearest the PSM mass.

However, the benefits of the UVPD method for *de novo* sequencing are 2-fold, and they cannot be overstated. First, with a complete y ion ladder, full-length, gapless *de novo* reconstructions are frequently attainable for nontrivial peptides. Second, the spectra display an ion ladder from only the C-terminus, eliminating the computationally intractable antisymmetric path problem (where mirror-image sequences propagate along both N-terminal and C-terminal ion ladders). *De novo* algorithms commonly address this problem by making imprecise assumptions, such as requiring that b and y ions not share the same mass node. Such assumptions are unnecessary with 351 nm UVPD mass spectra.

UVnovo. We developed UVnovo to *de novo* interpret AMCA-treated UVPD spectra. As illustrated in Figure 2, the UVnovo spectral processing pipeline progresses through four main steps for each MS/MS spectrum, described with further

DOI: 10.1021/acs.analchem.6b00261
Anal. Chem. 2016, 88, 3990−3997

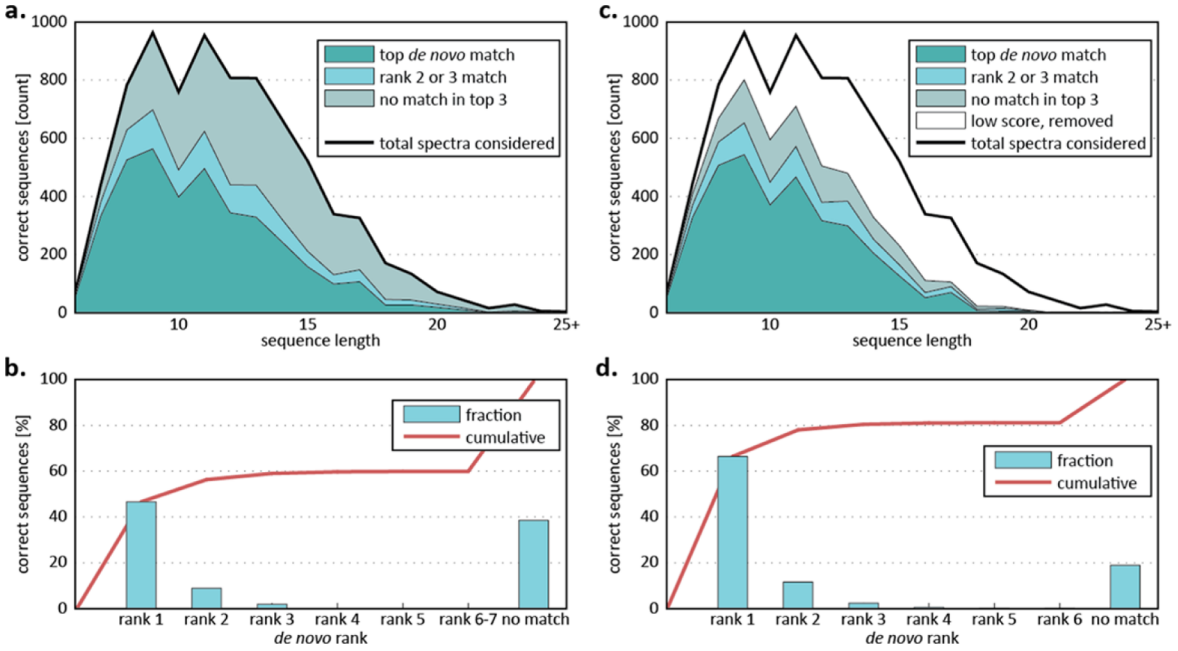


Figure 4. UVnovo *de novo* results for the *E. coli* lysate test set. A correct sequence matches the SEQUEST PSM exactly with no gaps. UVnovo scores each sequence reconstruction and ranks it relative to others from the same spectrum. (a) Number of correct sequences versus peptide length for the top-ranked *de novo* result and for the top three *de novo* results. (b) Fraction of correct sequences versus *de novo* rank. (c, d) Filtering of low-scoring *de novo* predictions improves sequence-level precision. 5062 of the original 7911 spectra remain, and over 75% of those removed had no correct match.

detail in the Supporting Information. Briefly, the spectrum is simplified using a random forest (RF) classifier.[31] At each integer mass position along the spectrum, the RF merges evidence from 30 spectral features to predict whether that position falls at a peptide bond of the precursor peptide backbone. Next, a hidden Markov model is used to estimate the probability that each site corresponds to a true fragment ion.[35] Each spectrum is then assigned one or more potential sequences using the Viterbi algorithm, with a single best sequence generated for each likely spectrum peptide length.[36] The candidate *de novo* sequences are scored and ranked using the HMM fragment node probabilities.

Validation of UVnovo Using *E. coli* Lysate. In order to measure performance on a complex protein sample, we applied the AMCA-UVPD strategy on a full *E. coli* lysate. The lysate was carbamylated, digested, and derivatized with AMCA and analyzed via LC-MS/MS with 351 nm UVPD. Spectra from triplicate *E. coli* runs were processed with Proteome Discoverer SEQUEST using the Percolator node and allowing a ±1.6 Da precursor mass tolerance. Limiting the results to doubly charged precursors and top-ranked matches, 7911 high-confidence identifications matching 2983 unique peptides were obtained from the 106,870 spectra collected across all three samples. We benchmark UVnovo against these high-confidence PSMs using 3-fold cross-validation (CV) to maintain independence between training and test sets.

Each CV repetition was trained independently for UVPD spectral interpretation. During random forest generation, 30 predictor variables were automatically identified as the most important out of a total space of 407 potential features. Feature scoring and selection was largely consistent between the three CV repetitions, and 27 of the 30 selected features were the same between each of the CV repetitions (Supporting Information Table S1). These primarily represented spectral peaks at specific mass offsets relative to the base position, and as expected, the most important feature corresponded with y^+ ions. Many of the features have not been used in prior *de novo* software, although the fact that they independently emerged among the most important from each of the CV rounds shows their value and the power of an open machine learning approach to spectral analysis.

UVnovo generated a list of sequences for each spectrum, typically with no more than seven candidates, and sequences were ranked based on descending confidence score. We required a "correct" sequence reconstruction to exactly match its corresponding SEQUEST PSM, after allowing for indistinguishable residue pairs I/L and $F/M^{oxidation}$. No sequence gaps or truncations were permitted.

UVnovo produced correct top-ranked sequences for 47.4% of the *E. coli* mass spectra, and when considering the top three *de novo* sequences for each spectrum, 59.8% had a match to the corresponding SEQUEST PSM (Figure 4a). The number of correct reconstructions drops substantially with decreasing *de novo* sequence rank (Figure 4b). Peptides with correct sequences ranged in size between 6 and 24 amino acids and had an average length of 11.0 residues. This compares to an average peptide length of 11.8 from the total set of SEQUEST PSMs, only two of which were longer than 24 residues. Exclusion of spectra without high-scoring *de novo* reconstructions dramatically improved sequencing precision. This filtered out two-thirds of the false positive predictions while retaining 85.5% of true predictions, boosting the precision to 66.4% and 80.4% for the top one and top three *de novo* sequence sets, respectively (Figure 4c,d). Our ability to identify correct full-length sequences from the majority of the test set demonstrates the benefits of AMCA-UVPD for comprehensive and interpretable peptide fragmentation.

For those spectra without a correct full-length identification, the highest scoring prediction often differed from its matching PSM at only a single fragmentation site, corresponding to a

DOI: 10.1021/acs.analchem.6b00261
Anal. Chem. 2016, 88, 3990−3997

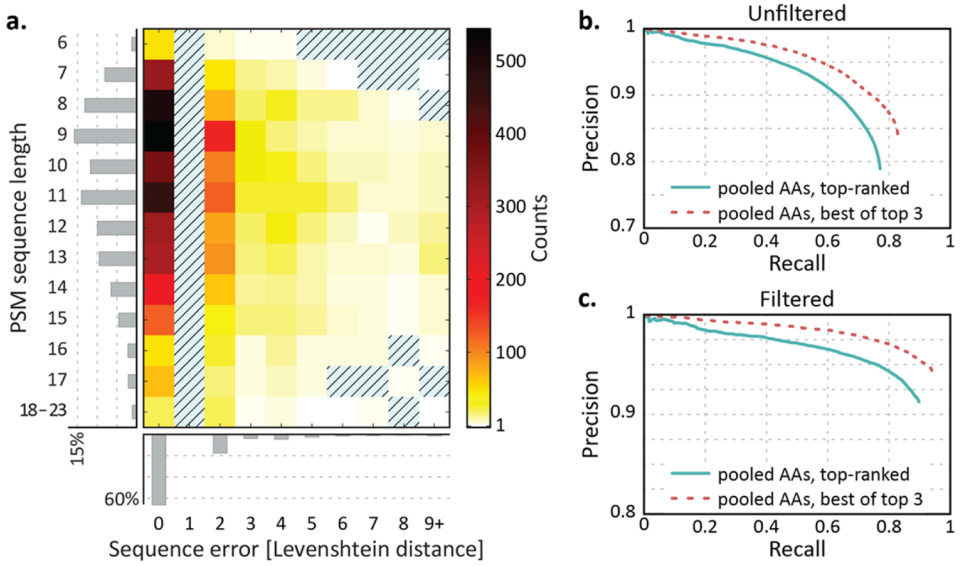


Figure 5. UVnovo performance for the *E. coli* lysate *de novo* reconstructions. (a) Amino acid error versus peptide length for top-ranked *de novo* sequences from the filtered set of higher confidence predictions. Most sequences are correct with no insertions or deletions. Incorrect sequences tend to diverge from SEQUEST PSMs by only two residues (a single fragmentation site misprediction). Histograms show fractional counts in each dimension. (b, c) Amino acid precision recall for the complete and filtered *de novo* results. AAs are pooled and sorted by residue-level score from (blue) the top-ranked *de novo* predictions for each spectrum or (dashed red) the best match among the top three predictions for each spectrum.

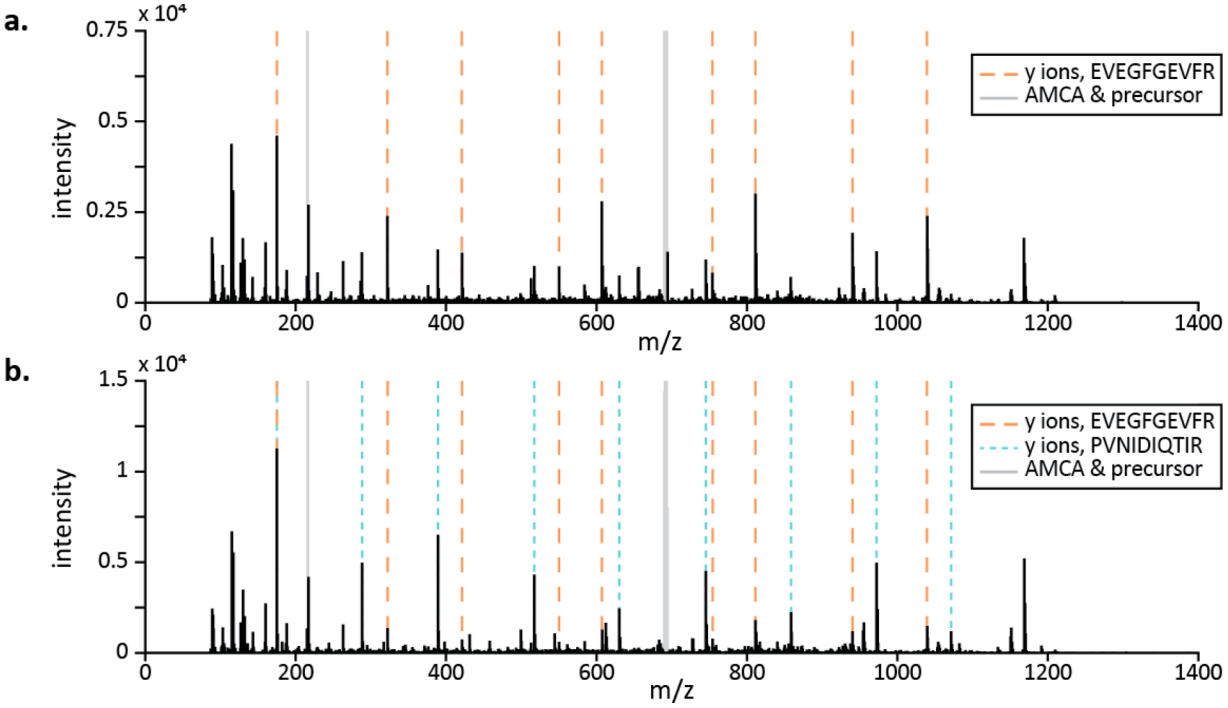


Figure 6. Co-eluting *E. coli* peptides independently identified between UVnovo and SEQUEST. (a) UVnovo and SEQUEST both assign the sequence EVEGFGEVFR. (b) Spectrum is acquired 49 s after that of panel a. Here, UVnovo assigns PVNIDIQTIR, conflicting with the SEQUEST identification, EVEGFGEVFR. Both sequences are presented within the *E. coli* reference database.

difference of two amino acids. Figure 5a displays the frequency and extent of amino acid sequencing errors versus peptide length in the filtered set. Over half (52.0%) of the misidentified sequences differ from the SEQUEST PSM by only two amino acids, meaning that only one fragmentation site per peptide is recognized incorrectly. Furthermore, each amino acid in a sequence reconstruction has an associated score. Pooling all residue predictions and sorting by descending score allowed us to plot the amino acid-level precision and recall of residue assignments, a common metric for *de novo* algorithm performance.[19,21,37] In brief, precision is measured as the fraction of correct predictions out of all amino acid predictions, and recall is the fraction of all amino acids in the test set that are correctly identified. Correct counts must match both the residue assignment and mass position along the spectrum. Shown for the total set in Figure 5b and the filtered set in Figure 5c, the UVnovo precision−recall curves confirm high sequence coverage and low error at the amino acid level.

DOI: 10.1021/acs.analchem.6b00261
Anal. Chem. 2016, 88, 3990−3997

Co-eluting peptides in our data sometimes manifest as differences between the *de novo* sequence and SEQUEST PSM for a spectrum. In some cases, this resulted in a hybrid *de novo* sequence blended from the two precursor peptides. Ideally, however, the differing *de novo* results complement the SEQUEST identification, and both are correct. As an example, Figure 6 presents a pair of co-eluting peptides observed across two spectra. Both SEQUEST and UVnovo identified the first spectrum as EVEGFGEVFR (1383.62 Da). The second spectrum, acquired 49 s later in the same injection, took the same SEQUEST PSM, while UVnovo assigned the sequence PVNLDLQTIR (1383.73 Da). Both are present within the *E. coli* sequence database, though the latter was not included in the SEQUEST search due to the presence of proline C-terminal to the tryptic arginine residue. We also observed other "incorrect" *de novo* identifications with exact matches to semitryptic *E. coli* peptides. Such examples indicate inflated error rate estimates in our results and point to the power of *de novo* methods in general for identifying unanticipated peptide variants.

Finally, we note that our results compare favorably to the performance of leading *de novo* sequencing algorithms on high-resolution data sets in general, although specific comparisons on this data set were not feasible due to the nature of the modifications and ion series employed here. For example, while UniNovo was designed to interpret novel fragmentation spectra, it does not permit user-defined peptide modifications or custom protease specificities.[14] More generally, most available *de novo* software is designed to recognize peptide fragmentation patterns generated through HCD, CID, or ExD, very different from the single *y* ion series we observe, and many of these programs address the antisymmetric path problem with assumptions that would negatively affect results for spectra with unambiguous directionality. Nonetheless, by employing stringent benchmarking criteria (e.g., requiring complete peptide sequence predictions that exactly match corresponding database PSMs), our data show that UVPD/UVnovo accurately identifies peptide sequences in complex samples and cell lysate contexts through a fully *de novo* sequencing approach.

CONCLUSIONS

We describe new experimental methods and the UVnovo software package for *de novo* peptide sequencing by UVPD. High-efficiency carbamylation blocks lysine side chains, and subsequent tryptic digestion and N-terminal peptide derivatization with the UV chromophore AMCA yield peptides susceptible to 351 nm ultraviolet activation. The UVPD mass spectra, primarily composed of *y* ions, are particularly well suited for *de novo* sequencing. As illustrated in the present study, 351 nm UVPD alleviates two of the fundamental limitations for *de novo* sequencing of standard spectra: incomplete or biased peptide sequence coverage and spectral symmetry due to observation of both N- and C-terminal ions. Because of the proclivity to generate abundant *y* ions, the spectral peaks are easier to interpret, and the antisymmetric path problem is nonexistent. Additionally, the comprehensive peptide backbone cleavage of UVPD provides the means to reconstruct full or nearly full sequences for most high-quality peptide spectra.

Development of UVnovo was motivated by a lack of appropriate tools for analysis of 351 nm UVPD peptide mass spectra. UVnovo combines random forests and hidden Markov models to simplify and interpret UVPD fragmentation spectra, enabling the *de novo* sequencing of thousands of peptides from an *E. coli* lysate at high confidence. UVnovo performance, seen here for low-resolution ion trap spectra, broadly matches that of leading *de novo* programs on high-resolution MS/MS spectra. Due to the full sequence coverage provided through UVPD, our workflow offers unprecedented capability for full-length peptide *de novo* sequencing. Further refinement of the UVnovo algorithm is underway and will capitalize on integrating CID and UVPD paired spectra.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.analchem.6b00261.

Additional figures including the ESI mass spectra of myoglobin prior to and after carbamylation and heat maps showing the amino acid error versus peptide length compiled for thousands of UVPD spectra of peptides and text giving more detailed description of the UVnovo algorithm (PDF)

Table of predictor variables and feature scoring (XLSX)

AUTHOR INFORMATION

Corresponding Authors

*E-mail: jbrodbelt@cm.utexas.edu (J.S.B.).

*E-mail: marcotte@icmb.utexas.edu (E.M.M.).

Author Contributions

§S.A.R. and A.P.H. contributed equally to this work.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

Funding from the NSF (Grant CHE-1402753) and the Welch Foundation (Grant F-1155) to J.S.B. and from the NIH, NSF, ARO Award W911NF-12-1-0390, Welch Foundation (Grant F-1515), DARPA, and DTRA to E.M.M. is acknowledged. J.R.C. acknowledges support from NIH Grant 1K12GM102745. Our data is publicly available at ProteomeXchange with accession ID: PXD003767. UVnovo is available for download at https://github.com/marcottelab/UVnovo. We thank Dr. William Press for helping to guide our earliest efforts in computational *de novo* sequence analysis, including suggesting the method of spectral normalization and the use of the HMM.

REFERENCES

(1) Ma, B.; Johnson, R. *Mol. Cell. Proteomics* **2012**, *11*, 014902.

(2) Seidler, J.; Zinn, N.; Boehm, M. E.; Lehmann, W. D. *Proteomics* **2010**, *10* (4), 634−649.

(3) Wells, J. M.; McLuckey, S. A. Collision-Induced Dissociation (CID) of Peptides and Proteins. In *Methods in Enzymology*; Burlingame, A. L., Ed.; Academic Press: San Diego, California, 2005; Vol. *402*, pp 148−185, DOI: 10.1016/S0076-6879(05)02005-7.

(4) Laskin, J.; Futrell, J. H. *Mass Spectrom. Rev.* **2003**, *22* (3), 158−181.

(5) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. *Nat. Methods* **2007**, *4* (9), 709−712.

(6) Mikesh, L. M.; Ueberheide, B.; Chi, A.; Coon, J. J.; Syka, J. E. P.; Shabanowitz, J.; Hunt, D. F. *Biochim. Biophys. Acta, Proteins Proteomics* **2006**, *1764* (12), 1811−1822.

(7) Wiesner, J.; Premsler, T.; Sickmann, A. *Proteomics* **2008**, *8* (21), 4466−4483.

DOI: 10.1021/acs.analchem.6b00261
Anal. Chem. 2016, 88, 3990−3997

(8) Madsen, J. A.; Brodbelt, J. S. *Anal. Chem.* **2009**, *81* (9), 3645−3653.
(9) Brodbelt, J. *J. Am. Soc. Mass Spectrom.* **2011**, *22* (2), 197−206.
(10) Reilly, J. P. *Mass Spectrom. Rev.* **2009**, *28* (3), 425−447.
(11) Ly, T.; Julian, R. R. *Angew. Chem., Int. Ed.* **2009**, *48* (39), 7130−7137.
(12) Brodbelt, J. S. *Chem. Soc. Rev.* **2014**, *43*, 2757−2783.
(13) Song, Y.; Yu, M. *Inf. Process. Lett.* **2015**, *115* (2), 377−381.
(14) Chi, H.; Chen, H.; He, K.; Wu, L.; Yang, B.; Sun, R.-X.; Liu, J.; Zeng, W.-F.; Song, C.-Q.; He, S.-M.; Dong, M.-Q. *J. Proteome Res.* **2013**, *12* (2), 615−625.
(15) Jeong, K.; Kim, S.; Pevzner, P. A. *Bioinformatics* **2013**, *29* (16), 1953−1962.
(16) Kim, S.; Bandeira, N.; Pevzner, P. A. *Mol. Cell. Proteomics* **2009**, *8* (6), 1391−1400.
(17) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2337−2342.
(18) Frank, A.; Pevzner, P. *Anal. Chem.* **2005**, *77* (4), 964−973.
(19) Fischer, B.; Roth, V.; Roos, F.; Grossmann, J.; Baginsky, S.; Widmayer, P.; Gruissem, W.; Buhmann, J. M. *Anal. Chem.* **2005**, *77* (22), 7265−7273.
(20) Tabb, D. L.; Ma, Z.-Q.; Martin, D. B.; Ham, A.-J. L.; Chambers, M. C. *J. Proteome Res.* **2008**, *7* (9), 3838−3846.
(21) Ma, B. *J. Am. Soc. Mass Spectrom.* **2015**, *26*, 1885−1894.
(22) Vyatkina, K.; Wu, S.; Dekker, L. J. M.; VanDuijn, M. M.; Liu, X.; Tolić, N.; Dvorkin, M.; Alexandrova, S.; Luider, T. M.; Paša-Tolić, L.; Pevzner, P. A. *J. Proteome Res.* **2015**, *14*, 4450−4462.
(23) Dančík, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. *J. Comput. Biol.* **1999**, *6* (3−4), 327−342.
(24) Richards, A. L.; Vincent, C. E.; Guthals, A.; Rose, C. M.; Westphall, M. S.; Bandeira, N.; Coon, J. J. *Mol. Cell. Proteomics* **2013**, *12* (12), 3812−3823.
(25) Devabhaktuni, A.; Elias, J. E. *J. Proteome Res.* **2016**, *15*, 732−742.
(26) Keough, T.; Youngquist, R. S.; Lacey, M. P. *Proc. Natl. Acad. Sci. U. S. A.* **1999**, *96* (13), 7131−7136.
(27) Robinson, M. R.; Madsen, J. A.; Brodbelt, J. S. *Anal. Chem.* **2012**, *84*, 2433−2439.
(28) Wilson, J. J.; Brodbelt, J. S. *Anal. Chem.* **2007**, *79* (20), 7883−7892.
(29) Robotham, S. A.; Kluwe, C.; Cannon, J. R.; Ellington, A.; Brodbelt, J. S. *Anal. Chem.* **2013**, *85* (20), 9832−9838.
(30) Angel, P. M.; Orlando, R. *Rapid Commun. Mass Spectrom.* **2007**, *21* (10), 1623−1634.
(31) Breiman, L. *Mach. Learn.* **2001**, *45* (1), 5−32.
(32) Gardner, M. W.; Smith, S. I.; Ledvina, A. R.; Madsen, J. A.; Coon, J. J.; Schwartz, J. C.; Stafford, G. C.; Brodbelt, J. S. *Anal. Chem.* **2009**, *81* (19), 8109−8118.
(33) Datta, R.; Bern, M. *J. Comput. Biol.* **2009**, *16* (8), 1169−1182.
(34) Grossmann, J.; Roos, F. F.; Cieliebak, M.; Lipták, Z.; Mathis, L. K.; Müller, M.; Gruissem, W.; Baginsky, S. *J. Proteome Res.* **2005**, *4* (5), 1768−1774.
(35) Rabiner, L. R. *Proc. IEEE* **1989**, *77* (2), 257−286.
(36) Forney, G. D. *Proc. IEEE* **1973**, *61* (3), 268−278.
(37) Mo, L.; Dutta, D.; Wan, Y.; Chen, T. *Anal. Chem.* **2007**, *79* (13), 4870−4878.

DOI: 10.1021/acs.analchem.6b00261
Anal. Chem. 2016, 88, 3990−3997