

Proteomics-Grade de Novo Sequencing Approach

Mikhail M. Savitski,^{*,†} Michael L. Nielsen,[†] Frank Kjeldsen, and Roman A. Zubarev

Laboratory for Biological and Medical Mass Spectrometry, Uppsala University, Uppsala, Sweden

Received August 30, 2005

The conventional approach in modern proteomics to identify proteins from limited information provided by molecular and fragment masses of their enzymatic degradation products carries an inherent risk of both false positive and false negative identifications. For reliable identification of even known proteins, complete de novo sequencing of their peptides is desired. The main problems of conventional sequencing based on tandem mass spectrometry are incomplete backbone fragmentation and the frequent overlap of fragment masses. In this work, the first proteomics-grade de novo approach is presented, where the above problems are alleviated by the use of complementary fragmentation techniques CAD and ECD. Implementation of a high-current, large-area dispenser cathode as a source of low-energy electrons provided efficient ECD of doubly charged peptides, the most abundant species (65–80%), in a typical trypsin-based proteomics experiment. A new linear de novo algorithm is developed combining efficiency and speed, processing on a conventional 3 GHz PC, 1000 MS/MS data sets in 60 s. More than 6% of all MS/MS data for doubly charged peptides yielded complete sequences, and another 13% gave nearly complete sequences with a maximum gap of two amino acid residues. These figures are comparable with the typical success rates (5–15%) of database identification. For peptides reliably found in the database (Mowse score ≥ 34), the agreement with de novo-derived full sequences was $>95\%$. Full sequences were derived in 67% of the cases when full sequence information was present in MS/MS spectra. Thus the new de novo sequencing approach reached the same level of efficiency and reliability as conventional database-identification strategies.

Keywords: de novo sequencing • bioinformatics • mass spectrometry • ECD • FTMS

Introduction

The conventional approach in modern proteomics to identify proteins from limited sequence information provided by molecular and fragment masses of their enzymatic degradation products carries an inherent risk of both false positive and false negative identifications.¹ The first artifact may arise due to the similarity (homology) of many protein sequences. Of the estimated 40 000 protein families in nature, only approximately 20% are known and listed in databases; thus, many protein sequences currently considered unique have unknown closely related counterparts. False negative identifications may arise due to mutations, post-translational modifications, and unknown proteins. Of the estimated 10^{11} different protein amino acid sequences existing in nature, less than 10^6 nonrepetitive sequences ($<0.001\%$) are found in reliable protein databases. Libraries of coding DNA regions cover an order of magnitude more sequences, but they are not error-free. Thus, even for reliable identification of known proteins, complete de novo sequencing of their peptides is desired. Mass spectrometry has been known for decades to be capable of sensitive sequencing

of peptides and small proteins.^{2,3} However, the literature contains mostly favorable examples, while recent testing by several dozen MS labs resulted in a low success rate, revealing the lack of a proteomics-grade de novo sequencing approach.⁴ Proteomics-grade de novo sequencing should satisfy the criteria of efficiency, reliability, and speed. As a rule of thumb, the whole procedure should not take much longer than the typical LC/MS experiment (30–120 min), which limits the analysis time to less than 1 s per peptide sequence. As de novo data are to be compared to the sequences found in the database, the overall efficiency of the de novo procedure should not be less than the average success rate of a database search (5–15%).⁵ The validity requirement means that a proteomics-grade de novo sequencing algorithm should provide the same validity level as is generally required for false positive identification in the database search, which is $>95\%$.

In this work, such an approach is presented for the first time. The approach combines hardware and software improvements. The main problems of conventional de novo sequencing are incomplete backbone fragmentation and the frequent overlap of fragment masses.⁶ These problems are alleviated by the use of the complementary fragmentation techniques collision-activated dissociation (CAD) and electron-capture dissociation (ECD),^{7–9} as well as the high resolving power and mass accuracy provided by Fourier transform mass spectrometry (FTMS).^{10,11}

* Corresponding author: Laboratory for Biological and Medical Mass Spectrometry, Uppsala University, Box 583, S-75123 Uppsala, Sweden. Telephone, +46 (0) 18 471 5729; fax, +46 (0) 18 471 5729; e-mail, Mikhail.Savitski@bmms.uu.se.

[†] Both authors contributed equally to this work.

Implementation of a high-current, large-area dispenser cathode as a source of low-energy electrons¹² provided efficient ECD of doubly charged peptides, the most abundant species (65–80%) in a typical trypsin-based proteomics experiment. A new linear de novo algorithm is developed combining efficiency and speed. The necessity for the new algorithm was rationalized as follows.

Recent comparison of three de novo sequencing algorithms on 29 test peptides gave the average success rate of 30–35% in correctly identifying amino acid residues, with very few full sequences revealed.¹³ Although part of the problem was associated with the low mass accuracy and the use of just one fragmentation technique, it was not clear whether the design of the algorithms was optimal for a high-throughput task. After analysis of existing algorithms, including the one based on amino acid composition calculation as an intermediate step,¹⁴ we chose a linear design which starts with the most reliable fragment data and propagates by building a bridge of amino acid masses, filling eventual gaps with amino acid combinations. By constructing full sequences, the algorithm increases the probability for them to be correct, as will be shown later. Such an approach is optimal if the fragmentation is abundant, overlapping, and complementary, as is the case with CAD and ECD, and if the mass accuracy is high, as is the case with FTMS. The algorithm avoids total permutation of all amino acid combinations consistent with the measured molecular mass and thus is very fast. The latter parameter is important in proteomics, where data analysis remains the throughput-limiting bottleneck. The goal was to create an algorithm that would construct de novo sequences faster than the database search, which typically takes 10 min for a restricted search on 1000 MS/MS data files, on a devoted one processor 3 GHz personal computer, and 100 or more minutes if the search is unrestricted (enzyme not specified, multiple modifications allowed).

This requirement arose from the desire to obtain high efficiency for the overall peptide analysis routine, of which de novo sequencing is one of the first steps. This step follows LC–MS/MS experiment and compiling a component list that includes neutral monoisotopic mass, abundance, and retention time of each peptide in the mixture. Then the quality factor (“S-score”)¹⁵ is determined for every MS/MS data (“dta” format) file, and low-quality data ($S \leq 1$) are filtered out.¹⁵ The subsequent steps include database search and comparison with the de novo-derived sequences. Conflicting cases are resolved based on the database-fit quality score (“M-score”)¹⁶ as well as on the yet-to-be-developed quality score for de novo sequencing. In cases when complete or nearly complete de novo sequences were derived but the database search was unsuccessful, a homology search was performed.

The above-envisioned procedure clarifies at least two aspects important for the de novo sequencing algorithm. The first aspect is the scope: the algorithm must be applicable to the most common species in a typical proteomics experiment, which are doubly charged peptides. These species compose up to 80% of all multiply charged ions from tryptic peptide mixtures (ECD requires multiply protonated species). The implication is that ECD efficiency has to be high enough for this species, which is a challenge since the electron capture cross section scales as charge squared.⁸

The second aspect is that the algorithm should preferentially yield a unique sequence. If such a sequence is provided by the algorithm, the probability of being correct should be >95%.

Typically, de novo algorithms provide a multitude of sequences to which a large range of probabilities are assigned.^{13,17} This aspect is important for the above analysis procedure to be efficient and useful, as handling multiple choices is both time-consuming and troublesome.

As shown below, incomplete determined sequences are rarely reliable. Thus, it was important to know in advance whether an efficient algorithm aimed at full or nearly full sequencing can be designed given the typical MS/MS data. Preliminary experiments revealed that 23.6% of all MS/MS CAD data for doubly charged peptides, reliably identified by Mascot (Mowse score $M \geq 34$), contain full sequence information (cleavage between all interresidues). The average sequence coverage was 84%, the corresponding figure for ECD mass spectra was 5%, and the average sequence coverage was 58%. The combination of CAD and ECD data gave full sequences in 32.4% of the cases and average sequence coverage of 91%, raising a hope for a proteomics-grade de novo sequencing algorithm.

Below, we present the details of the realization of such an algorithm and its validation. Ideologically, the algorithm is similar to the one suggested by Horn et al.¹⁸ but differs from it in the application area and some important details of realization. The current algorithm is more hierarchically structured (four steps instead of two), treats ECD data with greater caution (allows for hydrogen loss as well as gain), and also utilizes peaks without isotope distributions. While Horn’s algorithm has been trained on a few highly charged proteins and took minutes to provide a single sequence,¹⁸ the current algorithm is optimized for speed, efficiency, and validity on thousands of MS/MS spectra of multiply charged peptides. This optimization led to some unique filtering routines that significantly improve the sequence validity.

Methods

Sample Preparation. A-431 human epidermoid carcinoma whole cell lysate (200 μg) (Santa Cruz Biotechnology, CA) was loaded onto a one-dimensional SDS-PAGE gel (~30–200 kDa). The protein bands were visualized with colloidal Coomassie blue, and 30 equally sized fractions were excised from the gel. Subsequently, the proteins were reduced and alkylated, as well as in-gel digested with modified sequence-grade trypsin (Promega, Madison, WI), as previously described in the literature.¹⁹ Finally, the samples were vacuum-centrifuged to remove all organic solvents and reconstituted prior to analysis in 20 μL of HPLC water containing 0.1% trifluoroacetic acid (TFA, Sigma-Aldrich).

Mass Spectrometry. The mass spectrometric experiment, including the preliminary data processing, is described in detail elsewhere.¹¹ Briefly, an LTQ FT (Thermo, Bremen, Germany) was used equipped with a nano-flow liquid chromatograph HP1100 (Agilent) and electrospray interface (Proxeon, Odense, Denmark). Commercial electron source was used based on a large-area (>20 mm²) dispenser cathode that provides a large current of low energy electrons (<1 eV) and confinement of the fragments and unreacted ions inside the electron beam.²⁰ Such a design provides high efficiency (>20%) ECD for doubly charged ions.

Database Search. The data were searched using the Mascot search engine (Matrix Science, U.K.)¹⁶ against the whole NCBI non redundant database with a mass accuracy of ± 5 ppm for molecular mass determination and ± 0.02 Da for fragment masses. Carbamidomethyl on cysteines was set as a fixed

modification, and oxidized methionine was allowed as a variable modification. Tryptic constraints were applied allowing up to two missed cleavages.²¹

Preliminary Data Processing. The preliminary data processing included extraction of the so-called dta-files that contain the mass and the charge state of the precursor, as well as the m/z values and intensities of all the fragment ion peaks in the spectrum above a certain cutoff intensity value. The extraction was performed using TurboSequest Dta (BioWorks software, Thermo, Bremen). For every precursor, two dta-files were present, one representing CAD and the other ECD fragmentation. The de novo sequence was derived from the list of possible fragments which was prepared as follows. The isotopic clusters found in the CAD and ECD dta-files were deisotoped and charge-deconvoluted to the neutral state.^{11,15} However, fragments below ca. 800 Da often appear without their heavier isotopes due to the low abundance of the latter and the noise cutoff. When heavier isotopes were missing, the neutral masses were derived assuming that the charge state of the peak cannot exceed that of the precursor ion if the peak originates from a CAD dta-file and will be less than the charge state of the precursor if the peak originates from an ECD dta-file (due to the charge reduction in ECD). Thus, a peak without isotopes located at $m/z = 500.2$ in a CAD spectrum of a 2+ precursor could have the neutral mass of 500.2 and 1000.4 Da. Both these values will appear in the list of possible fragments with equal abundances. The incorrect value will likely be filtered out on subsequent steps of data processing.

De Novo Sequencing Algorithm. Because of the use of two complementary fragmentation techniques, not all fragment masses have an equal value in de novo sequencing. The masses that are doubly confirmed (i.e., information on them appear in both ECD and CAD spectra) are naturally much more reliable than the rest of the data. The algorithm uses these very reliable fragment masses to create a “backbone” for the sequence. Known amino acid masses are then fitted to this backbone to construct a “reliable sequence tag” RST as has been described earlier.¹⁵ If the RST does not cover the whole sequence (which is usually the case), less reliable fragment masses are added to the fragment mass list, and the amino acid masses are fitted between the old and the new entries and so on until the full sequence is obtained. This algorithm is linear, as it progresses in only one direction and employs only local loops and permutations, providing high speed. The hierarchic structure of the data included in the fragment mass list and the use of doubly confirmed data for reliable backbone construction, as well as multiple data filtering, ensure high reliability.

Briefly, the RST is constructed from neutral fragment masses (M_c) that simultaneously fulfill two requirements. The first is to contribute to a golden complementary pair rule,¹⁸ fulfilled if one of the following four equations is satisfied:

$$M_c - M_e = -17.0265 \quad (1)$$

$$M_c - M_e = 16.0187 \quad (2)$$

$$M_c + M_e - 17.0265 = M \quad (3)$$

$$M_c + M_e + 16.0187 = M \quad (4)$$

Here, the neutral mass M_c originates from a CAD dta-file, and the neutral mass M_e originates from an ECD dta-file. The second requirement dictates that M_c must be part of a complementary pair, thus satisfying the following equation:

$$M_c + M_e = M \quad (5)$$

The masses which comply with these two requirements simultaneously are stored in the de novo vector (DNV) as neutral b-fragment masses, even if the original mass was due to a y-ion.¹⁵ The conversion of all types of fragments to one type of fragment is done to simplify the algorithm, decreasing the amount of information contained in the peak list and thus amounts to data filtering. Along with the fragment masses, two other masses are added to the list, 0 Da and $(M - 18.01055)$ Da. The latter number is the neutral mass of a water molecule, which is subtracted from the molecular mass as b-ions are dehydrated peptides.

In the second step, neutral masses satisfying *any* of the conditions 1–4 are selected. Subsequently, these masses (which are less reliable than RST-contributing masses) are fitted into the DNV. This is done as follows:

Each mass m_i is treated as a “candidate mass”, to which a score F_i is assigned which is initially zero. For each mass, its distance L to the closest lower mass in DNV and distance H to the closest higher mass are calculated. Then the score F_i accepts the value $F_i = H^{-1} + L^{-1}$, with L^{-1} or H^{-1} equal to zero if L or H is greater than some threshold value. The threshold value of 554 Da was chosen empirically; a higher number did not provide substantial gain and slowed the algorithm, while a lower number caused an inferior sequencing success rate. If either L or H are <554 Da and are not equal to a combination of amino acid masses, the mass m_i is erased from the candidate list. After the calculation of F_i is completed for all candidate masses, the mass with the greatest score is selected and added to the DNV. This operation is repeated until all candidate masses have been either included into the DNV or rejected and erased. In rare cases, the score F_i never gets a greater than zero value, in which cases the algorithm simply proceeds to the next step.

In the next (third) step, masses fulfilling condition 5 are considered. These are complementary fragments whose origin (i.e., b- or y-ion) is unknown. The procedure for them is similar to the one above, except that, after finding the mass m_i with the greatest F_i among those that fit as b-ions, a check is made whether the same mass m_i could be fitted as a y-ion (peptides with at least one b-ion having close mass with a y-ion appear rather frequently⁶). The mass m_i that can potentially be due to both b- and y-ion is not accepted in the DNV, but not erased from the candidate list either unless it coincides with a mass already present in the DNV. Instead, the mass with the next highest F_i is considered under the same conditions. In the following round, the previously rejected mass m_i is considered again since the incorporation of some new mass in the DNV can remove the ambiguity of the origin of m_i . In the rare cases when a peptide that has a y-ion mass which coincides with a b-ion mass is sequenced, a blind spot will arise unless the y- and b-ions are confirmed as golden complementary pairs.

The next and final (fourth) step fits masses which have not been confirmed by other fragments to fill the remaining gaps in the peptide sequence. Here, the chance for an error is the greatest, and thus, one should proceed with maximum caution. Only DNVs where the differences between the adjacent, sorted in increasing order, masses do not exceed 600 Da qualify for that last step, where the following additional conditions are imposed.

For CAD fragments, if the charge state is ambiguous (no isotopic peaks), it is only assumed to be 1+ and the corre-

sponding neutral mass is derived under that assumption. The reason for that is the predominance of 1+ fragments in CAD of a doubly charged precursor which prompts us to disregard the less probable possibility of a 2+ fragment. ECD fragments are considered only if an isotopic distribution is present. The risk for a mass error is higher for ECD than for CAD due to, for example, frequent hydrogen loss from radical fragments and hydrogen rearrangement between radical and even-electron fragments.

The selection of masses in the last step is different from the previous steps. A mass m_i from the CAD dta-file qualifies only if it can be unequivocally fitted as a y- or b-ion. The qualified mass is not added immediately to the DNV, but instead is added to a temporary vector (TV). The subsequent mass m_{i+1} is fitted in the same fashion and is “unaware” of the previous mass m_i , since it is only interacting with DNV. A mass m_j from the ECD dta-file is treated in a similar fashion; only here, it is fitted as a c- or z-ion, and it is verified that the mass cannot be of a complementary type. The masses that pass the test are added to the TV vector. After all masses have been considered, the two vectors DNV and TV are merged, and possible amino acid sequences are fitted to the merged vector anew. If several alternative amino acid sequences are possible, they are qualitatively ranked according to the following rules:

(1) A sequence with a greater number of suggested cleavages outranks a sequence with fewer cleavages.

(2) If the number of cleavages is the same, the sequence that is fully compliant with the enzyme cleavage rules is preferred.

(3) If the number of cleavages is the same and rule 2 does not resolve the conflict, the abundances A_i of the fragments corresponding to masses m_i are added together and divided by the number of cleavages (average fragment abundance, AFA). The sequence with the highest AFA is deemed as the more reliable of the two.

The sequence deemed most reliable is reported as the final answer.

Additional Filters. The algorithm described above relies heavily on the correct identification of the type of ions (b- and y-ion for direction-confirmed ions and by-ions for direction-unknown fragments). Masses suggested in step one are extremely reliable¹⁵ and require no filtering. Masses in step two and three pass an additional criterion, as they have to be in agreement with previously fitted masses. But a special filter is required for singly confirmed masses entering at the stage two to remove possible incorrect assignments occurring from the following scenario. It is common to have weak CAD-type fragmentation in conjunction with ECD; thus, y-ions (or b-ions) can be present in both the ECD and CAD dta-files. If these ions exhibit the common NH_3 loss in CAD, then the fragment corresponding to this loss together with the fragment in ECD will satisfy eq 1 and the mass derived from the NH_3 loss fragment will be erroneously interpreted as a b-ion. To avoid this kind of error, the CAD dta-files are checked for masses which are greater by the mass of NH_3 than singly confirmed b-ions and are more abundant. If this is the case, then the singly confirmed b-ion is removed from the list. Doubly confirmed masses are practically immune to this mishap because of the fulfillment of eq 5.

In step four, where the lowest-quality data are considered, even stricter filters are used. One filter removes masses that may correspond to NH_3 and H_2O losses from CAD fragments by finding mass pairs with corresponding mass differences and removing the lower mass if the higher one is more abundant.

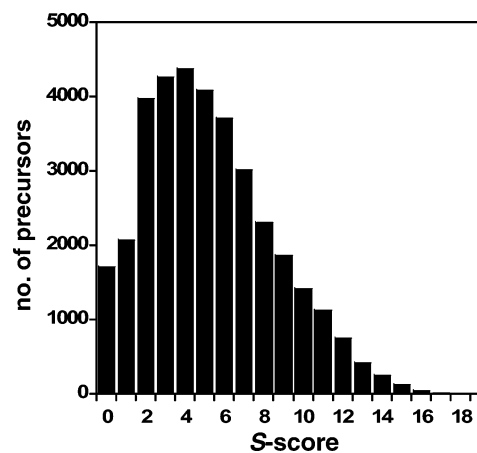


Figure 1. Distributions of S -scores for all 35 549 ECD/CAD spectra of doubly charged peptides.

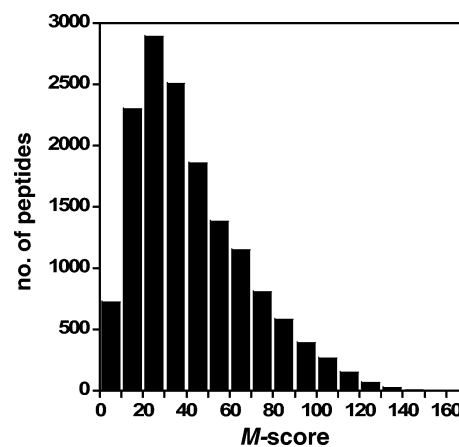


Figure 2. Distribution of Mowse scores for dta-files for which Mascot found a peptide sequence.

Finally, a filter removes ECD masses if they coincide with any CAD mass to eliminate the possibility of incorrect assignment of their type.

Results

The performance of the above algorithm was tested on 35 549 ECD/CAD dta-file pairs of doubly charged peptides from an A-431 human carcinoma cell lysate sample collected during 21 separate LC/MS experiments. The S -score distribution¹⁵ of all data is shown in Figure 1. The distribution is bimodal; although, this may not be apparent from Figure 1. The threshold value of $S = 1$ removes 10.6% of low-quality spectra with $S \leq 1$ from further processing.¹⁵ The remaining 31 763 MS/MS data sets were preprocessed and merged as described in¹¹ and then submitted to the Mascot search engine.

The M -score distribution of identified peptides is shown in Figure 2. The bimodal character is not obvious from the plot. The Mascot-suggested peptide threshold of $M = 34$ yielded 7852 identifications, of which at least 95% should be the correct result. The actual figure found in our previous work¹¹ was close to 99%. We used these positive Mascot identifications to test the reliability of our de novo sequencing program, which yielded 1704 full sequences of peptides (22% of all identified by Mascot). The time required by the algorithm to sequence 1000 spectra de novo was less than 60 s on a standard one

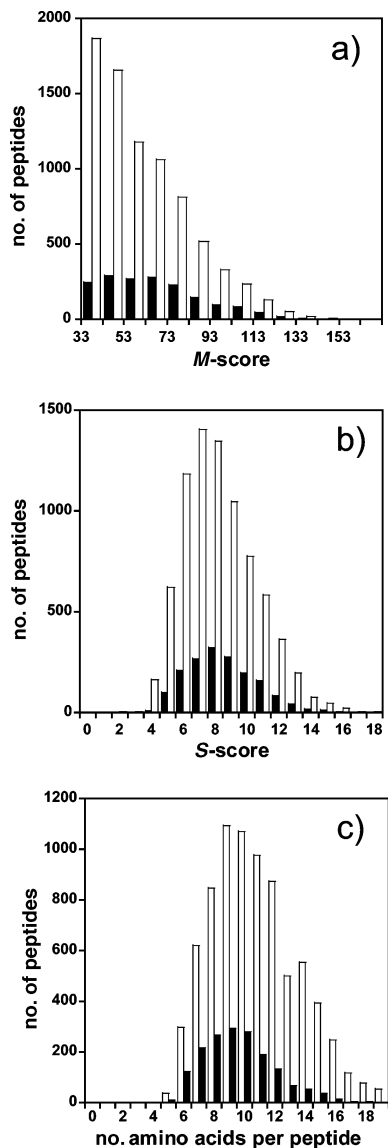


Figure 3. Distributions for all MS/MS datasets (open columns) and for completely de novo sequenced peptides (filled columns): (a) of Mowse scores; (b) of *S*-scores; (c) of the number of amino acid residues in the Mascot-suggested sequences. Only MS/MS datasets identified by Mascot with the Mowse score above the threshold value of 33 are shown in all three plots.

processor 3 GHz PC, which was much shorter than a restricted Mascot search took on the same computer (10 min).

Comparison of the *M*-score, *S*-score, and size distributions of Mascot-identified and completely de novo-sequenced peptides from these MS/MS scans (Figure 3a–c) showed that the means of the fully de novo sequenced peptides had slightly higher values for Mowse and *S*-scores, 67.0 versus 60.7 for the Mowse scores, 8.5 versus 8.2 for the *S*-scores, while the number of amino acid residues was slightly lower, 10.9 versus 9.5 amino acid residues for the size. These results were expected, since de novo sequencing requires more information than a Mascot search. The difference was not very large, supporting the efficiency of the de novo algorithm.

Besides the peptides with Mascot-derived sequences, the de novo algorithm provided 583 full sequences of peptides either not identified by Mascot or those with below-threshold scores. Thus, the total number of fully sequenced peptides was 2278,

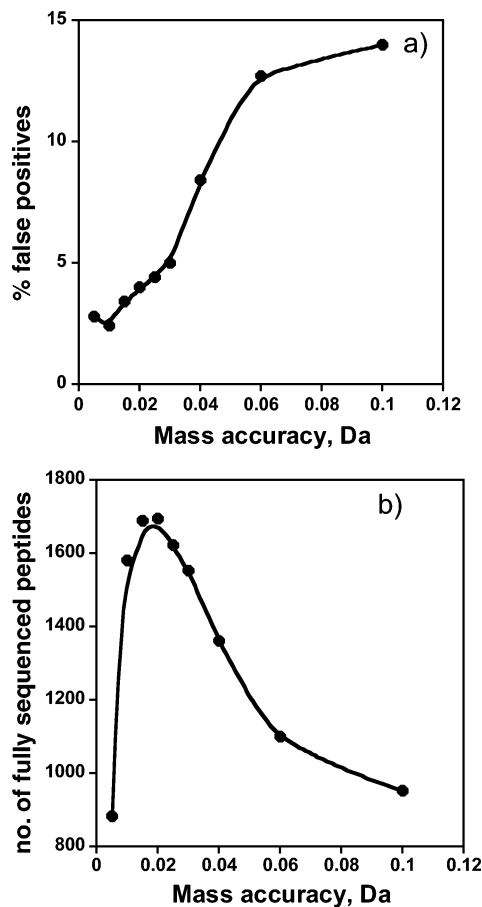


Figure 4. Relation between the mass accuracy and (a) the percentage of false positives and (b) the number of fully sequenced peptides.

that is, ca. 6.4% of the total number of MS/MS spectra and 7.2% of those with $S > 1$. In 80 additional cases, the algorithm suggested two sequences instead of a unique one. In those rare cases, the qualitative ranking was applied to select one of the two sequences. This is comparable to the typical success rate of database searches (5–15%) reported in the literature.⁵ The average *S*-score of the de novo sequences not reliably identified or found by Mascot was $S = 6.2$ with an average length $L = 7.4$, suggesting that the unidentified data were of lesser quality and belonged to shorter peptides than the identified data. It is worth mentioning here that peptides containing seven or more amino acid residues are usually unique in the human genome and thus are often sufficient for unique protein identification.²²

Mass Accuracy. The effect of mass accuracy on the false positive rate (de novo sequence conflicted with Mascot) and the number of de novo sequenced peptides was studied (Figure 4). The false positive rate increases dramatically, Figure 4a, from 4% to 14%, when the mass accuracy requirement is relaxed from ± 0.02 Da for fragment ions to ± 0.1 Da. At the same time, the number of fully sequenced peptides decreases, Figure 4b, since errors occur more frequently during the selection of the reliable building blocks (golden and complementary pairs), and the linear sequence building procedure more often chooses a wrong path and halts. When the mass accuracy window was narrowed to below ± 0.02 Da, the false positive rate dropped below 4%, but the number of sequences dropped as well because many fragment masses fell outside the mass accuracy window. This result means that mass accuracy is extremely

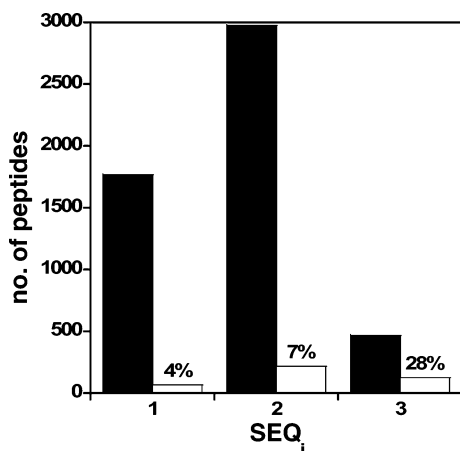


Figure 5. Number of de novo sequenced peptides in the three different classes SEQ₁, SEQ₂, SEQ₃ (filled columns) and the number of incorrect (false positives) de novo-derived sequences in each class (open columns). The false positives rate is indicated on top of the open columns for each class. Only MS/MS datasets identified by Mascot with $M > 33$ are shown in both plots. De novo sequence was considered correct if it agreed with the Mascot-suggested sequence.

important for proteomics-grade de novo sequencing and that mass accuracy better than ± 0.04 Da is required. This, in turn, means that only high-resolution mass spectrometers are suitable for this task, at least at the same sensitivity level.

Incomplete Sequences. Fully and partially sequenced peptides were classified as follows. Complete sequences (all gaps between fragment masses filled by one amino acid) were attributed to class SEQ₁, while sequences with at least one gap that could be filled by two amino acid residues belong to class SEQ₂ and so on. For Mascot-identified ($M > 33$) mass spectra, Figure 5 shows the number of sequences in each class (filled columns) which complied with Mascot-suggested sequences and the number of conflicting sequences (open columns). The maximum allowed gap size was 398 Da (if the gap was larger, the MS/MS data were not considered to be amenable for sequencing). The value 398 Da was chosen empirically. The nominal mass of the glycine residue (57 Da) times 7 equals 399 Da, so 398 guarantees that a maximum of six residues can be fitted in the gap. The false positive rate is indicated for each class. The probability of false positives increases intact with the class index as one would expect. The elements of class SEQ₂ have a false positive rate of 7.2%, which is above the 5% threshold but can be tolerable in many cases. These sequences can be useful, since often one only has to decide between two alternative orders of neighboring amino acid residues. This is a problem which can be solved by conducting MS BLAST^{23–25} searches against relevant databases. Class SEQ₂ has 2972 elements among MS/MS data with a suggested Mascot sequence with $M > 33$ and 1643 elements among residual data, which amounts to a total of 13% of the entire MS/MS data. If classes SEQ₁ and SEQ₂ are combined, 19.4% of all MS/MS spectra produced complete or nearly complete sequences.

Completeness Issue. Even if each gap between fragment masses can be explained by one amino acid, an element of uncertainty still remains for some sequences regardless of the mass accuracy. This is because amino acid asparagine (N) has the same elemental composition as two glycines (G + G); glutamine Q can be represented as A + G and tryptophan (W) as (A + D) or (E + G). The situation is aggravated by the

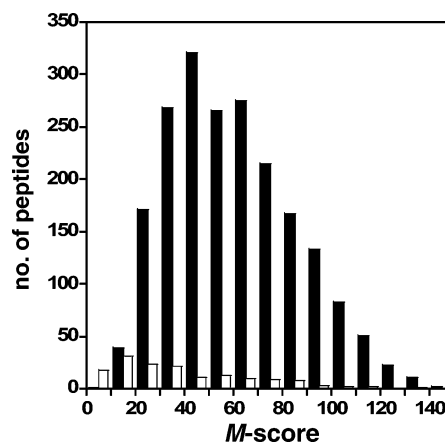


Figure 6. The distribution of Mowse scores for completely de novo-sequenced peptides which coincided with the Mascot-suggested sequence (filled columns) and conflicted sequences (open columns).

frequent losses of 57 Da from ECD fragments, which, for example, in the case of N, can be interpreted as a cleavage between two glycines. The 57 Da loss can also be an indication for the presence of isoaspartic acid.²⁶ Furthermore, arginine has a very similar mass as the (G + V) combination. Without hot ECD (HECD^{27,28}), it is also impossible to distinguish isoleucine (I) from leucine (L).

In this work, we liberally assumed the above uncertainties inherent to MS/MS (although future research will undoubtedly solve some, if not all, of them).

Efficiency. To evaluate the relative efficiency of the linear algorithm, the 7851 CAD/ECD MS/MS spectra with $M \geq 34$ were inspected in terms of the *theoretically* present sequence information (i.e., if all peaks with masses corresponding to cleavage sites of Mascot-identified peptides are considered). Of these, 2543 data sets (32.4%) contained full sequence information. The relative efficiency of the described algorithm for *full* sequencing was thus 67% (72% if only MS/MS data with at least one doubly confirmed fragment were considered). These figures must be viewed keeping in mind the reliability issue. While it is relatively easy to increase the algorithm efficiency by removing some or all of the filters, this would reflect negatively in the data validity. The desire to have both proteomics-grade reliability *and* efficiency dictated the need to compromise.

De Novo and below-Threshold Mascot Scores. Mascot sequences with a below-threshold score can be “rescued” if they coincide with the full sequence suggested de novo. As illustrated in Figure 6, a significant part of below-threshold peptides have Mascot-suggested sequence that coincide with de novo sequences. The “rescue operation” is of particular utility in the field of peptidomics, where peptides usually are not produced by strict enzyme rules, and thus, the Mascot-suggested threshold is typically much higher than in proteomics.

Modifications. The developed software allows the user to utilize any conceivable set of building blocks for de novo sequencing. It also allows the user to focus on interesting modifications and ignore other data. For instance, if methylation of glutamic acid is of interest, the mass of the modified amino acid is added to the building block list and the software is asked to output only de novo sequences containing this modification. Application of this procedure successfully identified seven peptides containing methylated glutamic acid

Table 1. The List of Completely Sequenced Peptides Containing a Methylation on the Glutamic Acid Identified by the de Novo Sequencing Software^a

peptides containing methylated glutamic acid, E _m
FE _m NICK
SPE _m DIER
AVE _m HINK
FHVE _m EEGK
IQSIGTENTE _m ENR
IAE _m QAER
VGGTSDVE _m VNEK

^a Here, asparagine (N) could not be fully distinguished from G + G, glutamine (Q) from A + G and arginine (R) from the G + V combination.

(Table 1). Their authenticity was verified by the presence in the mixture and reliable Mascot identification of the same peptides in unmodified state, as well as significant identification of the modified peptides by conducting an additional Mascot search with the selected modification.

Conclusions

A new sequencing algorithm is designed combining high speed with proteomics-grade efficiency and reliability. The algorithm derives a substantial part (>67%) of theoretically possible sequence information, sacrificing some of the efficiency to obtain >95% reliability. A significant part of the remaining 33% of information is utilized to produce nearly-complete sequences with a maximum gap of two amino acid residues. The algorithm was designed as a part of the ongoing project to fully automate the MS-based proteomics experiment and may become an irreplaceable part of the standard proteomics analysis. The first results are very encouraging; however, more needs to be done. In particular, we plan to implement and validate quantitative scoring of de novo-derived sequences, where not only the full sequence but various parts of it would be assigned a measure of validity.

This study highlighted previously unnoticed (or unreported) difficulties with traditional de novo sequencing. It is now clear that, using CAD alone, it is extremely difficult to obtain proteomics-grade performance. Surprisingly even two complementary fragmentation techniques do not necessarily guarantee obtaining full theoretically possible sequence information, which is a prerequisite for de novo sequencing of proteins and novel peptides. This problem can be addressed in two ways. The first one is to improve the performance of both CAD and ECD (for which there are instrumental as well as theoretical limitations), and the second one is to try to employ a third technique complementary to the first two. As possible candidates, one could name hot ECD and EDD.^{29,30} We are currently investigating these possibilities.

Finally, mass accuracy is found to play an extremely important role. Since the 95% reliability threshold is only reached at better than ±0.04 Da mass accuracy, high resolution is an absolute requirement for proteomics-grade sequencing. To be sure, full permutation of amino acid residues, a much slower procedure than the linear algorithm, will undoubtedly discover the correct sequence even at much lower level of mass accuracy. But it will hardly be able to distinguish the correct answer from the multitude of alternative sequences despite clever scoring techniques and probability assessments. Thus, we predict that high-resolution instrumentation will remain indispensable in proteomics and peptidomics even after commercialization of ECD/ETD technology on radio frequency mass spectrometers.

Acknowledgment. This work was supported by the Knut and Alice Wallenberg Foundation and Wallenberg Consortium North (Grant WCN2003-UU/SLU-009 to R.A.Z.) as well as by Swedish research council (Grants 621-2002-5025 and 621-2003-4877). Thomas Köcher, Oleg Silivra, Alexander Misharin, and Chris Adams are acknowledged for fruitful discussions.

References

- (1) Aebersold, R.; Mann, M. *Nature* **2003**, *422*, 198–207.
- (2) Lasonder, E.; Ishihama, Y.; Andersen, J. S.; Vermunt, A. M. W.; Pain, A.; Sauerwein, R. W.; Eling, W. M. C.; Hall, N.; Waters, A. P.; Stunnenberg, H. G.; Mann, M. *Nature* **2002**, *419*, 537–542.
- (3) Hunt, D. F.; Yates, J. R.; Shabanowitz, J.; Winston, S.; Hauer, C. R. *Proc. Natl. Acad. Sci. U.S.A.* **1986**, *83*, 6233–6237.
- (4) Turck, C. W.; Falick, A. M.; Kowalak, J. A.; Lane, W. S.; Neubert, T. A.; Phinney, B. S.; Weintraub, S. T.; West, K. A. Presented at the 53rd ASMS Conference on Mass Spectrometry, San Antonio, TX, 2005.
- (5) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. *Anal. Chem.* **2002**, *74*, 5383–5392.
- (6) Budnik, B. A.; Nielsen, M. L.; Olsen, J. V.; Haselmann, K. F.; Horth, P.; Haehnel, W.; Zubarev, R. A. *Int. J. Mass Spectrom.* **2002**, *219*, 283–294.
- (7) Zubarev, R. A.; Kelleher, N. L.; McLafferty, F. W. *J. Am. Chem. Soc.* **1998**, *120*, 3265–3266.
- (8) Zubarev, R. A.; Horn, D. M.; Fridriksson, E. K.; Kelleher, N. L.; Kruger, N. A.; Lewis, M. A.; Carpenter, B. K.; McLafferty, F. W. *Anal. Chem.* **2000**, *72*, 563–573.
- (9) McLafferty, F. W.; Fridriksson, E. K.; Horn, D. M.; Lewis, M. A.; Zubarev, R. A. *Science* **1999**, *284*, 1289–1290.
- (10) Marshall, A. G.; Hendrickson, C. L. *Int. J. Mass Spectrom.* **2002**, *215*, 59–75.
- (11) Nielsen, M. L.; Savitski, M. M.; Zubarev, R. A. *Mol. Cell. Proteomics* **2005**, *4*, 835–845.
- (12) Haselmann, K. F.; Budnik, B. A.; Olsen, J. V.; Nielsen, M. L.; Reis, C. A.; Clausen, H.; Johnsen, A. H.; Zubarev, R. A. *Anal. Chem.* **2001**, *73*, 2998–3005.
- (13) Grossmann, J.; Roos, F. F.; Cieliebak, M.; Liptak, Z.; Mathis, L. K.; Muller, M.; Widmayer, P.; Gruitsem, W.; Baginsky, S. J. *Proteome Res.* **2005**, *4*, 1768–1774.
- (14) Spengler, B. *J. Am. Soc. Mass Spectrom.* **2004**, *15*, 703–714.
- (15) Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A. *Mol. Cell. Proteomics* **2005**, *4*, 1180–1188.
- (16) Perkins, D. N.; Pappin, D. J. C.; Creasy, D. M.; Cottrell, J. S. *Electrophoresis* **1999**, *20*, 3551–3567.
- (17) Ma, B.; Zhang, K. Z.; Hendrie, C.; Liang, C. Z.; Li, M.; Doherty-Kirby, A.; Lajoie, G. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2337–2342.
- (18) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97*, 10313–10317.
- (19) Wilms, M.; Shevchenko, A.; Houthaave, T.; Breit, S.; Schweigerer, L.; Fotsis, T.; Mann, M. *Nature* **1996**, *379*, 466–469.
- (20) Tsybin, Y. O.; Hakansson, P.; Budnik, B. A.; Haselmann, K. F.; Kjeldsen, F.; Gorshkov, M.; Zubarev, R. A. *Rapid Commun. Mass Spectrom.* **2001**, *15*, 1849–1854.
- (21) Olsen, J. V.; Ong, S. E.; Mann, M. *Mol. Cell. Proteomics* **2004**, *3*, 608–614.
- (22) Rappsilber, J.; Mann, M. *Trends Biochem. Sci.* **2002**, *27*, 74–78.
- (23) Shevchenko, A.; de Sousa, M. M. L.; Waridel, P.; Bittencourt, S. T.; de Sousa, M. V. J. *Proteome Res.* **2005**, *4*, 862–869.
- (24) Habermann, B.; Oegema, J.; Sunyaev, S.; Shevchenko, A. *Mol. Cell. Proteomics* **2004**, *3*, 238–249.
- (25) Shevchenko, A.; Sunyaev, S.; Loboda, A.; Shevchenko, A.; Bork, P.; Ens, W.; Standing, K. G. *Anal. Chem.* **2001**, *73*, 1917–1926.
- (26) Cournoyer, J. J.; Pittman, J. L.; Ivleva, V. B.; Fallows, E.; Waskell, L.; Costello, C. E.; O'Connor, P. B. *Protein Sci.* **2005**, *14*, 452–463.
- (27) Kjeldsen, F.; Haselmann, K. F.; Budnik, B. A.; Jensen, F.; Zubarev, R. A. *Chem. Phys. Lett.* **2002**, *356*, 201–206.
- (28) Kjeldsen, F.; Haselmann, K. F.; Sorensen, E. S.; Zubarev, R. A. *Anal. Chem.* **2003**, *75*, 1267–1274.
- (29) Kjeldsen, F.; Silivra, O. A.; Ivonin, I. A.; Haselmann, K. F.; Gorshkov, M.; Zubarev, R. A. *Chem.—Eur. J.* **2005**, *11*, 1803–1812.
- (30) Budnik, B. A.; Haselmann, K. F.; Zubarev, R. A. *Chem. Phys. Lett.* **2001**, *342*, 299–302.

PR050288X