Journal of proteome research

Article

pubs.acs.org/jpr

pNovo+: De Novo Peptide Sequencing Using Complementary HCD and ETD Tandem Mass Spectra

Hao Chi,[†,‡,⊥] Haifeng Chen,[†,‡,⊥,#] Kun He,[†,‡] Long Wu,[†,‡] Bing Yang,[§] Rui-Xiang Sun,[†] Jianyun Liu,[∥] Wen-Feng Zeng,[†,‡] Chun-Qing Song,[§] Si-Min He,[*,†] and Meng-Qiu Dong[*,§]

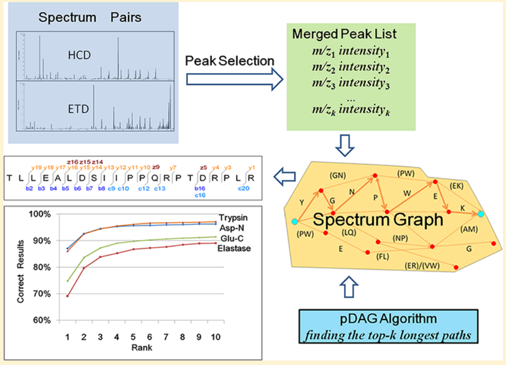
[†]Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, China

[‡]Graduate University of Chinese Academy of Sciences, Beijing 100049, China

[§]National Institute of Biological Sciences, Beijing, Beijing 102206, China

[∥]Laboratory of Intelligent Recognition and Image Processing, Beijing Key Laboratory of Digital Media, Beihang University, Beijing, 100191, China

Supporting Information



ABSTRACT: De novo peptide sequencing is the only tool for extracting peptide sequences directly from tandem mass spectrometry (MS) data without any protein database. However, neither the accuracy nor the efficiency of de novo sequencing has been satisfactory, mainly due to incomplete fragmentation information in experimental spectra. Recent advancement in MS technology has enabled acquisition of higher energy collisional dissociation (HCD) and electron transfer dissociation (ETD) spectra of the same precursor. These spectra contain complementary fragmentation information and can be collected with high resolution and high mass accuracy. Taking these advantages, we have developed a new algorithm called pNovo+, which greatly improves the accuracy and speed of de novo sequencing. On tryptic peptides, 86% of the topmost candidate sequences deduced by pNovo+ from HCD + ETD spectral pairs matched the database search results, and the success rate reached 95% if the top three candidates were included, which was much higher than using only HCD (87%) or only ETD spectra (57%). On Asp-N, Glu-C, or Elastase digested peptides, 69−87% of the HCD + ETD spectral pairs were correctly identified by pNovo+ among the topmost candidates, or 84−95% among the top three. On average, it takes pNovo+ only 0.018 s to extract the sequence from a spectrum or spectral pair on a common personal computer. This is more than three times as fast as other de novo sequencing programs. The increase of speed is mainly due to pDAG, a component algorithm of pNovo+. pDAG finds the k longest paths in a directed acyclic graph without the antisymmetry restriction. We have verified that the antisymmetry restriction is unnecessary for high resolution, high mass accuracy data. The extensive use of HCD and ETD spectral information and the pDAG algorithm make pNovo+ an excellent de novo sequencing tool.

KEYWORDS: *tandem mass spectrometry, de novo peptide sequencing, HCD, ETD, antisymmetry restriction, k longest paths*

1. INTRODUCTION

Database search and de novo sequencing are parallel methods used for peptide identification from tandem mass spectra. In database search, which has improved considerably in the past decades,[1] all candidate peptides from a specified database are retrieved for each spectrum, and each peptide-spectrum match is scored via a scoring function. There are many database search engines, such as Mascot,[2] SEQUEST,[3] X! Tandem,[4] OMSSA,[5] ByOnic,[6] pFind,[7] InsPecT,[8] and Phenyx.[9] They are widely used for routine analysis of tandem mass spectrometry (MS/MS) data. However, this approach fails when correct peptides are not present in the database. In situations like this, de novo sequencing is an indispensable and most valuable tool that can lead to correct peptide identification.

De novo sequencing extracts peptide sequences directly from tandem mass spectra without any protein sequence databases.[10,11] In 1999, Dancik et al. developed a de novo sequencing algorithm called Sherenga,[12] which utilized an offset frequency function (OFF) to determine which ion types ought to be considered for peaks in MS/MS spectra. Sherenga constructs a spectrum graph after converting peaks from an experimental spectrum to vertices and then derives peptide sequences from the spectrum graph.[13] Because it is impossible to know the ion type of every peak beforehand, different ion types are assumed. Thus, each peak generates several vertices. If the mass difference between two vertices equals the mass of

Received: July 24, 2012
Published: December 28, 2012

one or more amino acids, these two vertices are connected by a directed edge. It is stipulated that every peak in the spectrum may be interpreted as either an N-terminal ion or a C-terminal ion but not both. This is the basis of the antisymmetry restriction. The Sherenga algorithm was developed to find the antisymmetric longest path in the spectrum graph.

Many de novo peptide sequencing algorithms and tools have been published, including PepNovo,[14,15] PEAKS,[16] Lutefisk,[17] AuDeNs,[18] MSNovo,[19] SeqMS,[20,21] PFIA,[22] NovoHMM,[23] EigenMS,[24] PILOT,[25] pNovo,[26] Antilope,[27] and Vonode.[28] The majority of these algorithms are based on spectrum graphs, but alternative approaches are also used. Chen et al. proposed a dynamic programming algorithm to find optimal paths.[29−31] Frank et al. developed PepNovo using a probabilistic network to model the peptide fragmentation events in a mass spectrometer.[15] Zhang et al. designed a simple but effective divide-and-conquer algorithm to generate a list of sequence candidates.[32] PEAKS, developed by Ma et al., provides each sequenced result with confidence scores for the entire sequence and at individual residues.[16] For high-resolution MS/MS data, Spengler proposed an algorithm based on the analysis of amino acid composition and high mass accuracy to limit the amino acid combinations to be considered for a spectrum.[33] Another de novo sequencing approach introduced by Boersema et al. takes advantage of a special protease Lys-N to get nearly complete sequence ladders of *b*-ions in CID MS/MS data.[34]

However, de novo peptide sequencing has not yet become a mature method. A previous comparative study tested several de novo sequencing algorithms and found that no more than 50% of the peptides identified by database search can be correctly sequenced de novo.[35] Generally speaking, many spectra cannot be de novo sequenced due to incomplete fragmentation; that is, the fragment ion series contains too many gaps, or some gaps are too large.

As a result, many de novo sequencing algorithms tried to use complementary spectra belonging to the same precursor to obtain more fragmentation information. Horn et al. used complementary collisionally activated dissociation (CAD) and electron capture dissociation (ECD) spectra to distinguish N- and C-terminal fragments.[36] Savitski et al. developed an algorithm that integrates CAD and ECD information for peptide identification on a proteomics level.[37] Datta and Bern used a Bayesian network to combine information from several mass spectra of the same peptide.[38] Bertsch et al. developed a tool called CompNovo that used collisionally induced dissociation (CID) and electron transfer dissociation (ETD) to improve sequencing accuracy.[39] He et al. also developed a tool based on paired CID and ETD spectra, called ADEPTS.[40] A similar strategy is used to improve database search.[41]

Although the development of de novo peptide sequencing has improved peptide and protein identification and spectral interpretation, there are still many problems. Despite the fact that a pair of spectra can provide more information, de novo sequencing still yields fewer and less accurate peptides as compared to database search.[37] Besides, most de novo peptide sequencing algorithms are based on spectrum graph or other similar approaches to find the longest antisymmetric path,[12] which is an NP-hard problem.[42] Not surprisingly, most de novo peptide sequencing algorithms are time-consuming. Andreotti et al. recently compared four de novo sequencing tools and reported that for all, the speed is between 0.5 and 1.5 s per spectrum.[27]

In our previous work, we developed an algorithm pNovo[26] for de novo peptide sequencing on HCD spectra. Here, we have taken a step further by designing a new algorithm, called pNovo+, to make the best out of complementary high-resolution HCD and ETD spectral pairs. Immonium and internal ions in HCD spectra and hydrogen-rearranged fragment ions in ETD spectra are all taken into account. In addition, we verified that the antisymmetry restriction is not necessary for high-resolution, high-mass accuracy MS/MS data. Free from this constraint, we developed an efficient algorithm to find the *k* longest paths in a directed acyclic graph (DAG), which substantially accelerated the speed. pNovo+ correctly sequenced up to 95% of the spectra identified by database search at an average speed of less than 0.02 s per spectrum on a common personal computer.

2. METHODS

2.1. Ion Types in HCD and ETD Spectra

To select the appropriate ion types to include in the algorithm, we used the OFF proposed by Dancik et al.[12] As described in pNovo,[26] a spectrum *S* consists of *m* peaks from s_1 to s_m, and prefix residue masses of the correct peptide are represented by p_1, p_2, ..., p_n. The prefix OFF is computed as follows. For every s_i and p_j, we calculated their distance δ with the accuracy of two decimal places and plotted the occurrence of different δ values. The suffix OFF is computed in a similar way. Figures S1 and S2 in the Supporting Information show prefix and suffix OFFs of HCD and ETD mass spectra, respectively. In Tables 1 and 2, the frequency of each ion type is calculated as the number of the observed ions divided by the number of all theoretical ions in the scanned mass range (100−2000 Da). As shown in Tables 1 and 2, y^+, b^+, a^+, $y^+{-}NH_3$, y^{+2}, $y^+{-}H_2O$, $b^+{-}H_2O$, and $b^+{-}NH_3$ are the most predominant ions in the HCD data, while in ETD data, c^+ and z^+ ions, as well as $c^+{-}H$ and $z^+{+}H$ ions, can be found in abundance, in agreement with previous findings.[43]

Table 1. Different Ion Types Learned from the OFF in the HCD Data

offset δ	prefix/suffix	ion type	frequency (obsd/theor) (%)
19.02	suffix	y^+	62.57
1.01	prefix	b^+	48.99
−26.98	prefix	a^+	31.35
1.99	suffix	$y^+{-}NH_3$	31.48
10.01	suffix	y^{+2}	23.22
1.01	suffix	$y^+{-}H_2O$	21.25
−16.99	prefix	$b^+{-}H_2O$	21.09
−16.01	prefix	$b^+{-}NH_3$	16.04

Table 2. Different Ion Types Learned from the OFF in the ETD Data

offset δ	prefix/suffix	ion type	frequency (obsd/theor) (%)
4.00	suffix	$z^+{+}H$	31.52
3.00	suffix	z^+	31.49
18.03	prefix	c^+	28.84
17.02	prefix	$c^+{-}H$	13.10

2.2. Selecting Peaks

Peak selection is one of the most important steps in de novo peptide sequencing, and many MS/MS preprocessing methods are devoted to it.[44,45] In our algorithm, we address several critical problems that can affect correct detection of signal peaks as opposed to noise peaks. First, the weight of each peak is set as the natural logarithm of its intensity. Second, monoisotopic peaks are selected and assigned charge states. If c is the charge state of a precursor ion, and a fragment ion appears as an isotopic cluster, the algorithm assigns a charge state to the fragment by finding the best-fitting one from c, $c-1$, $c-2$, ... till reaching +1. From an isotopic cluster, we select the peak p with the lowest mass-to-charge ratio (m/z). In addition, the peaks with intensities greater than p in ETD are also selected so that c, $c-$H, z and z+H ions are all included.[26] For peaks not associated with any isotopic clusters, we assume that they are singly charged in ETD spectra, but in HCD spectra, we treat them as both singly and doubly charged since y^{+2} is also an abundant ion type in HCD spectra (Table 1).

Then, all peaks are transformed to +1 charge according to their charge states. If two or more peaks are of equal mass within a given tolerance range (±20 ppm for HCD and ETD MS/MS data), they are merged together as a new peak, taking the average mass and weight of these peaks. Lastly, precursor ion peaks are detected and deleted, because they are often the most abundant peaks in MS/MS spectra and yet are useless and even misleading for de novo peptide sequencing. Precursor ion peaks with neutral losses such as the loss of water or ammonia are also removed.

2.3. Constructing a DAG for Each Spectrum or Spectral Pair

2.3.1. Generating Graph Vertices. In general, the ion types of the peaks in tandem mass spectra are unknown, so all valid fragment ion types have to be considered. For HCD, y, b, a, $y-NH_3$, $y-H_2O$, $b-H_2O$, and $b-NH_3$ ions are considered, while for ETD, z+H, z, c, and $c-$H ions are taken into account. For instance, if there is a singly charged peak located at m/z 796.54 in a spectrum whose precursor MH$^+$ is 1387.76 Da, the following b ions are generated: m/z 592.22 (assuming the original peak is a y ion), m/z 796.54 (assuming the original peak is a b ion), m/z 824.53 (assuming a), m/z 609.25 (assuming $y-NH_3$), m/z 610.23 (assuming $y-H_2O$), m/z 814.55 (assuming $b-H_2O$), and m/z 813.57 (assuming $b-NH_3$). In this way, each HCD peak is transformed to seven singly charged b ions, and each ETD peak is transformed to four singly charged b ions. Then, all b ions thus generated from an HCD + ETD spectral pair are integrated, so that in the new spectrum all peaks are singly charged b ions. If two or more peaks are of equal mass within a given tolerance range, they are merged as a new peak.

For each vertex in the spectrum graph, its nominal mass and weight are the same as its corresponding peak. In addition, we add a source vertex and a destination vertex to each spectrum graph. The mass of the source vertex is zero, and the destination vertex is the precursor mass subtracted by the mass of a water molecule. Because all proper paths should contain the source and destination vertices, their weights do not influence path ranking and thus are set to zero in our algorithm.

2.3.2. Generating Graph Edges. For a pair of vertices, if the mass difference is equal to the mass of one or two amino acid residues, they are connected by a directed edge. Every edge $e(u, v)$ is assigned a weight equal to the weight of vertex v. The weight of a path is defined as the sum of the weights of the traversed edges; that is, the weight of a path is the sum of the weights of the traversed vertices because the weight of the source vertex is 0. Figure 1 is an example of a spectrum graph translated from a pair of HCD and ETD spectra. In addition, leucine and isoleucine are not distinguished, for they have exactly the same mass.

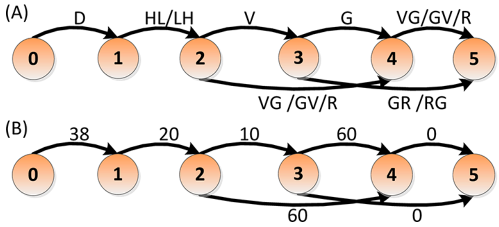


Figure 1. (A) Example of an initial spectrum graph constructed by pNovo+. (B) DAG translated from the spectrum graph above. The number written beside each edge is the weight of the edge. The masses of the six vertices are 0.000, 115.022, 365.168, 464.237, 521.259, and 677.359 Da. The longest, best-scoring paths DHLVGR and DLHVGR from D(HL/LH)VGR have the same score, so they are randomly assigned to be the top1 and top2 ranked sequences from this graph.

2.4. Removing the Antisymmetry Restriction

As described above, each peak, because of uncertainty in its ion type (e.g., b or y), is converted to more than one vertex. In most cases, however, a peak, if matched, is just one type of ion (e.g., only y). Thus, each correct vertex has at least one "fake" vertex associated with it, which is why an antisymmetry path-finding problem is proposed.[12] Unfortunately, the antisymmetry longest path-finding problem is NP-hard,[42] so algorithms based on finding antisymmetric paths are time-consuming.

On the other hand, a significant number of spectra contain peaks that can be matched, with confidence from high mass accuracy, to two types of ions or more. For example, in Figure 2, b_6^+ and y_{12}^{++} have the same m/z and are matched to the same peak. This phenomenon is common in HCD and ETD MS/MS data. We have counted from the Worm data set (described in section 3) the number of spectra containing at least one peak to which two or more fragment ions match within the tolerance of ±20 ppm (Table 3). We find that they account for at least 11.8% of the HCD spectra and 9.2% of the ETD spectra and can be as high as 15.0%. Therefore, it is sometimes incorrect to limit the interpretation of a peak to just one type of ion.

We also set the tolerance to ±0.5 Da to simulate the condition found in low-resolution CID/ETD data. As shown in Table 3, the number of spectra containing at least one peak matching two or more fragment ions is ~1.2 times more than that when ±20 ppm is used. This increase comes from fragment ions that are absent from the spectra but whose absence cannot be made sure in low-resolution data. Allowing two types of ions to match these peaks are obviously inappropriate and would lead to erroneous results. Therefore, considering the antisymmetric restriction is sometimes essential for processing low-resolution MS/MS data to reduce incorrect matches as previously reported.[12,15] In pNovo+, because both HCD and ETD data are of high resolution, the antisymmetry restriction is lifted. Benefiting from such freedom, we designed an efficient algorithm pDAG to find the k longest paths in a DAG, as shown in the next section.

2.5. Finding the k Longest Paths

Finding the k longest paths in a DAG has many practical applications. For example, it can be used in timing analysis tools that examine paths in an integrated circuit to determine the worst case delay. Several algorithms have been proposed in the past decades. Yen et al. presented an algorithm to find the k

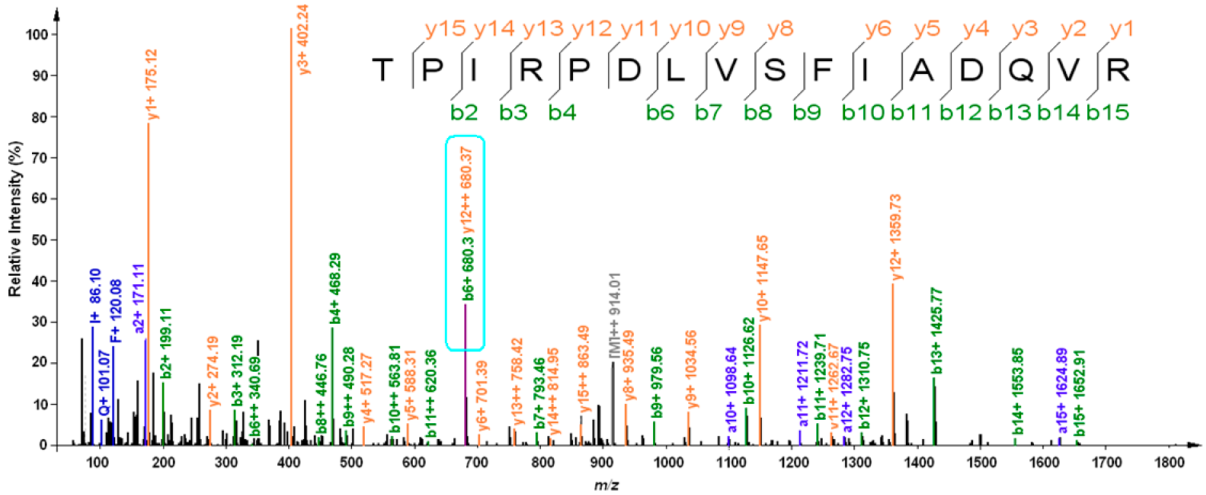


Figure 2. Peptide-spectrum match in which two fragment ions have the same m/z. In this example, b_6^+ and y_{12}^{++} are both found at m/z 680.3713.

Table 3. Number of Spectra Containing at Least One Peak to Which Two or More Fragment Ions Match on the Worm Data Set[a]

	±20 ppm		±0.5 Da	
	HCD	ETD	HCD	ETD
Asp-N (2367)[b]	334 (14.1%)	317 (13.4%)	782 (33.0%)	900 (38.0%)
elastase (1161)	159 (13.7%)	133 (11.5%)	365 (31.4%)	282 (24.3%)
Glu-C (1523)	179 (11.8%)	140 (09.2%)	588 (38.6%)	505 (33.2%)
trypsin (3626)	545 (15.0%)	494 (13.6%)	1249 (34.5%)	1209 (33.3%)

[a] In the peptide-spectrum match, y, b, a, $y-NH_3$, y^{+2}, $y-H_2O$, $b-H_2O$, and $b-NH_3$ ion types were considered for HCD, while z+H, z, c, and $c-H$ ion types were considered for ETD. [b] Total number of spectra analyzed from a particular enzyme digestion.

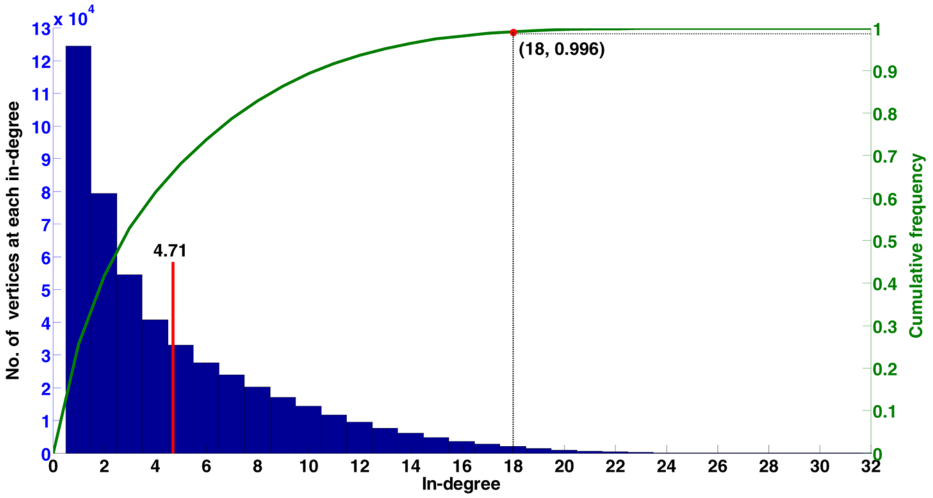


Figure 3. Number of vertices at each in-degree. The green curve is the cumulative frequency of the histogram. The figure shows that 99.6% of vertices have an in-degree of 18 or less. The average in-degree of these vertices is 4.71, while most of the spectrum graphs in our data set have more than 300 vertices; therefore, the spectrum graph is a sparse graph. The time complexity of our algorithm is $O(|E| \log d + k|V| \log d')$. Thus, the algorithm requires about linear time with respect to the number of vertices.

longest paths in DAG in 1989.[46] Ju and Saleh reported an incremental algorithm for enumeration of paths in 1991.[47] Kundu also proposed a sophisticated algorithm with well-contained memory growth in 1994.[48]

We have developed an efficient algorithm called pDAG to find the k longest paths in DAG, which is similar to Kundu's. The difference is that we use a maximum priority queue to significantly reduce the time required. The pseudo code of the pDAG algorithm is shown in the Supporting Information. Here, we describe how it works by taking Figure 1B as an example. Suppose that $L(u, j)$ is the weight of the j-th longest path to vertex u from source vertex 0 and $weight(u, v)$ is the weight of directed edge from u to v. First, to calculate $L(5, 1)$, we should first calculate $L(3, 1)$ and $L(4, 1)$, because vertices 3 and 4 are predecessor vertices of vertex 5. Then, we compare $L(3, 1) + weight(3, 5)$ and $L(4, 1) + weight(4, 5)$, and the greater of the two is $L(5, 1)$. Second, as shown in Figure 1B, $L(5, 1)$ comes from vertex 4. Therefore, to calculate $L(5, 2)$, we should first calculate $L(4, 2)$, then compare $L(4, 2) + weight(4, 5)$ and $L(3, 1) + weight(3, 5)$ and choose the greater one. Similarly, when calculating the k-th longest path to vertex v, we should first calculate $L(u, j)$ if $L(v, k-1)$ comes from $L(u, j-1)$.

The time complexity analysis of pDAG is shown below. We denote by E the edge set of the graph to find the k longest paths and denote by V the vertex set of the graph. Define d as the maximum in-degree of the graph and define d' as the average in-degree. It can be proved that the time complexity of finding the longest path and initializing the maximum priority queue is $O(|E| \log d)$, and the time complexity of finding one of any other paths is $O(|V| \log d')$. Then, the time complexity of our algorithm is $O(|E| \log d + k|V| \log d')$, which is given by Theorem 1: The time complexity of algorithm pDAG is $O(|E| \log d + k|V| \log d')$. The proof of this theorem is shown in the Supporting Information.

For a DAG transformed from a spectrum, the maximum in-degree and the average in-degree are usually very small. As shown in Figure 3, while most of the spectrum graphs in our data set have more than 300 vertices, the maximum in-degree for a vertex is 32. Moreover, 99.6% of the vertices have an in-degree of 18 or less, and the average in-degree of these vertices is 4.71. Thus, the spectrum graph is a sparse one. So, our algorithm has a linear time complexity with both k and the number of vertices. In addition, our algorithm uses a maximum priority queue, which saves about half the time needed by Kundu's algorithm.

2.6. Ranking Candidate Peptides

We use the breadth first search method to generate all candidate peptides from the k longest paths. In this step, peptides whose precursor masses are outside the given tolerance range are eliminated.

The main problem in this step is how to correctly rank the candidate peptides. For a candidate peptide, if the number of amino acids equals the number of edges in the corresponding path, then all of the fragmentation sites in the peptide have at least one peak as evidence. Therefore, such peptides are usually more reliable. For example, DHLGVR is less reliable than DHLRR in path 0−1−2−4−5 in Figure 1B. We define GAP_{pep} as follows:

$$\text{GAP}_{\text{pep}} = L_{\text{pep}} - L_{\text{path}} \quad (1)$$

In eq 1, L_{pep} is the number of amino acids in the peptide, and L_{path} is the number of edges in the path. If a peptide has a lower GAP, it is considered to be more reliable. So, we first select top m (m was equal to 200 in pNovo+) candidate peptides with the lowest GAP. Then, we used a scoring function similar to pNovo[26] to rank the peptides that have the same GAP, in which immonium and internal ions in HCD spectra and hydrogen-rearranged fragment ions in ETD spectra are taken into account.

3. MATERIALS AND RESULTS

3.1. MS/MS Data

The performance of pNovo+ is tested on HCD and ETD spectral pairs from two data sets. One is called the Worm data set, which is from a whole cell lysate of *C. elegans* analyzed on a LTQ-Orbitrap XL mass spectrometer equipped with ETD using a six-step Multidimensional Protein Identification Technology (MudPIT).[49] The other one is called 8-protein STD, part of which was described by Chi et al.[26] Four enzymes, Asp-N, elastase, Glu-C, and trypsin, are used separately during sample preparation, resulting in four MS/MS data sets for each sample. For all experiments, the MS and MS/MS resolutions were set to 60000 and 7500, respectively.

3.2. Benchmark Data Sets

pFind 2.6[50] was used to search all of the MS/MS data. The database search parameters are listed in Table 4. A software package called pBuild was used to filter the results, with FDR controlled at 1% at the spectrum level. Then, we selected spectral pairs with matched HCD and ETD identifications and required that the identified sequences each contained 6−19 amino acids. In the end, 8677 spectral pairs from the Worm data set and 913 pairs from 8-protein STD and were selected. The numbers of spectral pairs from different types of enzymatic digestion are shown in Table 5.

Table 4. Parameters for Database Search by pFind 2.6

item	setting
enzyme	Asp-N, elastase, Glu-C, or trypsin, separately
maximum missed cleavage sites	2
precursor tolerance	±20 ppm
fragment tolerance	±20 ppm
variable modifications	carbamidomethyl (C), oxidation (M)

Table 5. Number of HCD and ETD Spectral Pairs Selected from Each Enzymatic Digestion

data set	Asp-N	elastase	Glu-C	trypsin	sum
Worm	2367	1161	1523	3626	8677
8-protein STD	170	388	149	206	913

We compared the performance of pNovo+ with PEAKS from PeaksStudio 5.3 on the data sets described in Table 5. For pNovo+, an error tolerance of ±20 ppm was used for both precursor and fragment ions. Carbamidomethylation on cysteines and oxidation on methionines were set as variable modifications. The PEAKS parameters are the same as those of pNovo+ except that the fragment mass tolerance is ±0.02 Da (ppm is not supported). To be noted, pNovo+ is an upgrade of pNovo, which was designed for HCD data. Now, pNovo+ supports de novo sequencing from paired HCD + ETD spectra or only HCD or ETD spectra. The spectrum graph construction and the scoring function are the same for all three scenarios; the only difference is that ion types used in HCD and ETD spectra are set separately if no spectral pairs are to be expected. In this section, the performance of pNovo+ using spectral pairs is compared with that of pNovo+ using HCD or ETD spectra alone.

3.3. Comparison between HCD + ETD and HCD or ETD Alone

Table 6 summarizes the performance of pNovo+ and PEAKS when tested on the Worm data set. For each spectrum or spectral pair, we ask whether the correct peptide sequence (obtained from database search) can be found among the top *three* candidate sequences obtained by de novo sequencing. De novo sequencing on HCD and ETD spectral pairs is indicated by (HCD + ETD), which is a unique function of pNovo+. De novo sequencing on HCD or ETD spectra alone, or a simple union of their top three results, are denoted by (HCD), (ETD), or (HCD ∪ ETD). In the case of (HCD ∪ ETD), there may be up to six different candidate sequences. From Table 6, it is

Table 6. Comparing Successful de Novo Peptide Sequencing Results between pNovo+ and PEAKS on the Worm Data Set[a]

	pNovo+				PEAKS		
	(HCD + ETD)[b]	(HCD ∪ ETD)[c]	(HCD)	(ETD)	(HCD ∪ ETD)	(HCD)	(ETD)
Asp-N (2367)[d]	2238 (94.6)[e]	2034	1888	1504	1856	1765	848
elastase (1161)	974 (83.9)	704	513	494	485	444	200
Glu-C (1523)	1329 (87.3)	1058	848	760	793	714	330
trypsin (3626)	3425 (94.5)	3252	3159	2073	3098	3013	1275
sum (8677)	7966 (91.8)	7048 (81.2)	6408 (73.9)	4831 (55.7)	6232 (71.8)	5936 (68.4)	2653 (30.6)

[a] For a given spectrum or spectral pair, de novo sequencing is considered successful if the correct peptide sequence is among the top three candidates. [b] (HCD + ETD) indicates that paired HCD and ETD spectra are used for obtaining complementary fragmentation information in de novo sequencing, which is a feature of pNovo+. [c] (HCD ∪ ETD) represents the union of the (HCD) top three and the (ETD) top three candidates, and de novo sequencing is considered successful if one of the six candidates is correct. [d] The total number of spectra in each test set is indicated in the leftmost column. [e] In parentheses is the percentage of spectra that are correctly sequenced de novo.

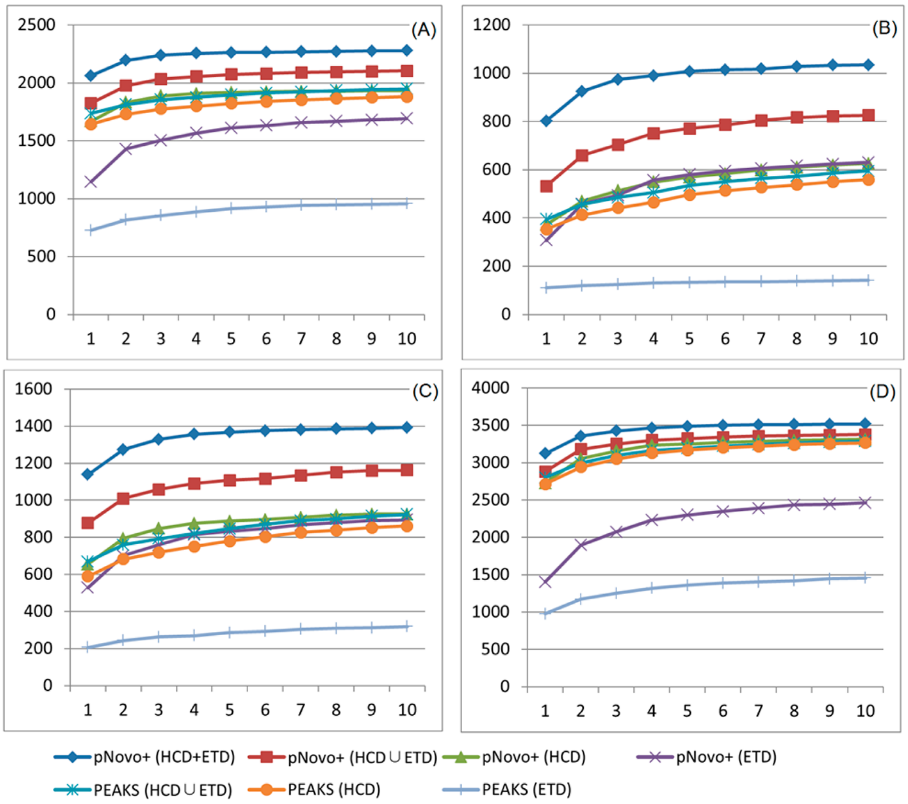


Figure 4. Cumulative curves of the number of correct sequences found among the top one to top 10 candidates obtained using pNovo+ (HCD + ETD), pNovo+ (HCD ∪ ETD), pNovo+ (HCD), pNovo+ (ETD), PEAKS (HCD ∪ ETD), PEAKS (HCD), and PEAKS (ETD) on the Worm data set. (A) Asp-N, (B) elastase, (C) Glu-C, and (D) trypsin.

obvious that although combining results obtained separately from HCD and ETD spectra can increase the number of correct de novo identifications, the best performance comes from pNovo+ using (HCD + ETD) spectral pairs, which correctly sequenced 7966 out of 8677 pairs, achieving an overall success rate of 91.8% and reaching as high as ~95% for trypsin and Asp-N peptides. This is 10–20% more than simply combining separate HCD and ETD results. Such improvement is seen across all of the data sets including the Glu-C or elastase digested samples, with the success rates of de novo sequencing ranging from ~84 to ~95%. On (HCD + ETD) spectral pairs, pNovo+ generated 20.0% more correct sequences than PEAKS HCD and ETD results combined or 10.6% more correct sequences than pNovo+ HCD and ETD results combined. Moreover, pNovo+ obtained 25.1% more correct sequences than PEAKS did using only ETD spectra or 5.5% more using only HCD spectra.

Figure 4 shows the cumulative curves of the number of correct sequences found among the top one through top 10 candidate sequences on the Worm data set. The exact numbers are shown in Tables S1–S4 in the Supporting Information. pNovo+ (HCD + ETD) obtained the highest number of correct sequences in all four subdata sets (Asp-N, elastase, Glu-C, and tryptic peptides). In particular, pNovo+ (HCD + ETD) successfully sequenced 38.4% more spectra than pNovo+ (HCD ∪ ETD)—the second best—on elastase-digested peptides.

Longer peptides are more challenging for de novo sequencing than shorter peptides. Shown in Figure 5 and Table S13 in the Supporting Information are the distributions of peptide lengths of correct de novo sequencing results

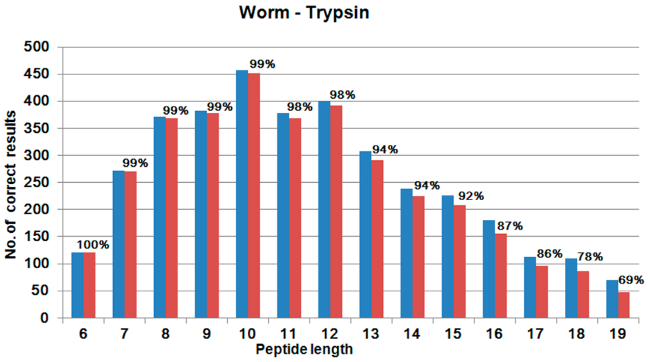


Figure 5. Peptide length affects the performance of pNovo+. Blue bars represent database search results. Red bars and the percentage labels indicate the numbers and percentages of correct sequences obtained by pNovo+. Spectra of tryptic peptides in the Worm data set were used in this analysis.

Table 8. Distribution of the Levenshtein Distances between the Top de Novo Candidates and the Correct Peptide Sequences

	pNovo+					
	(HCD + ETD)		(HCD)		(ETD)	
LD	no. of spectra	cumulative percentage (%)	no. of spectra	cumulative percentage (%)	no. of spectra	cumulative percentage (%)
0	7125	82.1	5423	62.7	3385	39.9
2	1035	94.0	1338	78.1	1912	62.4
3	99	95.2	197	80.4	301	66.0
4	188	97.4	507	86.3	650	73.6
5	73	98.2	230	88.9	362	77.9
6	54	98.8	243	91.7	323	81.7
7	29	99.1	197	94.0	262	84.8
8	29	99.5	137	95.6	258	87.8
9	15	99.6	97	96.7	230	90.5
10	8	99.7	87	97.7	183	92.7

obtained by pNovo+ on the Worm data set. For tryptic peptides (Figure 5) of no more than 12 amino acids, pNovo+ achieved a success rate of 98% or higher. However, the success rate decreased on longer peptides, especially those containing more than 17 amino acids. For example, only 69% of the 19-aa peptides in this data set were correctly sequenced.

We also compared the top three results of pNovo+ and PEAKS on the 8-protein STD data set (Table 7). As shown, de novo sequencing using pNovo+ on (HCD + ETD) spectral pairs is remarkably better than using HCD or ETD spectra alone or a simple combination of the two (87.7 vs 73.5% or lower). This is similar to what is observed on the Worm data set. Comparison between the top 10 results of pNovo+ and PEAKS on this data set is shown in Tables S5−S8 in the Supporting Information.

Table 8 shows from another angle that spectral pairing effectively improves the accuracy of de novo sequencing. Here, we measure accuracy using the Levenshtein distance (LD), which is a string metric for measuring the amount of difference between two sequences.[51] For each spectrum (or HCD + ETD spectral pair for pNovo+), we calculated the LD between the top candidate and the correct sequence. LD cannot be one because adding, deleting, or replacing one amino acid will change the peptide mass (no distinction made between I and L). As shown in Table 8, de novo sequencing results using spectral pairs (HCD + ETD) are much closer to the correct sequences, as compared with those obtained using only HCD or ETD spectra. About 94% of the top candidates generated by pNovo+ (HCD + ETD) have a LD of 2 or less, and the average LD is only 0.55, which means that these sequences are much more accurate than those from HCD (average LD = 1.69) or ETD data (average LD = 3.16).

Underlying the high success rates of pNovo+ is the use of complementary information from HCD and ETD spectra.[52] Many spectra cannot be de novo sequenced because the fragment ion series are incomplete. Luckily, incomplete fragmentation in one spectrum is often complemented by its cognate spectrum under a different fragmentation mechanism. As shown by the example in Figure 6, the subsequences “PPQR” and “PTD” have no fragmentation information in the HCD spectrum (A) and the subsequences “LLE”, “ALD”, “LLPP”, and “RPT” have no fragmentation information in the ETD spectrum (B). In the regular setting of pNovo+, edges with more than two amino acids are not considered in the spectrum graph construction. Therefore, the correct peptide cannot be identified by pNovo+ using HCD or ETD spectrum alone. However, if two spectra are used together, the fragmentation information is present for nearly all of the amino acid residues, so that the correct peptide can be identified by pNovo+.

To assess the contribution of the complementarity of HCD and ETD in a quantitative manner, we examined spectra containing gaps that probably compromised the accuracy of de novo sequencing. Table 9 shows that HCD and ETD pairs contain fewer gaps, as compared with HCD or ETD spectra separately. For example, 99.0% of the HCD and ETD pairs from the Worm data set have a maximum gap length of no more than one amino acid, and the percentage decreases to 92.2% for HCD spectra or 81.1% for ETD spectra. Because one or two amino acids are considered when vertices are connected in the spectrum graph construction, nearly all of the HCD and ETD pairs could be de novo sequenced, theoretically (provided that from separate HCD and ETD spectra the same peptide

Table 7. Comparing Successful de Novo Peptide Sequencing Results between pNovo+ and PEAKS on the 8-Protein STD Data Set

	pNovo+				PEAKS		
	(HCD + ETD)	(HCD ∪ ETD)	(HCD)	(ETD)	(HCD ∪ ETD)	(HCD)	(ETD)
Asp-N (170)	151 (88.8)	130	102	86	129	102	68
elastase (388)	349 (90.0)	288	226	207	281	196	177
Glu-C (149)	106 (71.1)	69	59	40	71	58	30
trypsin (206)	195 (94.7)	184	177	129	188	179	71
sum (913)	801 (87.7)	671 (73.5)	564 (61.8)	462 (50.6)	669 (73.3)	535 (58.6)	346 (37.9)

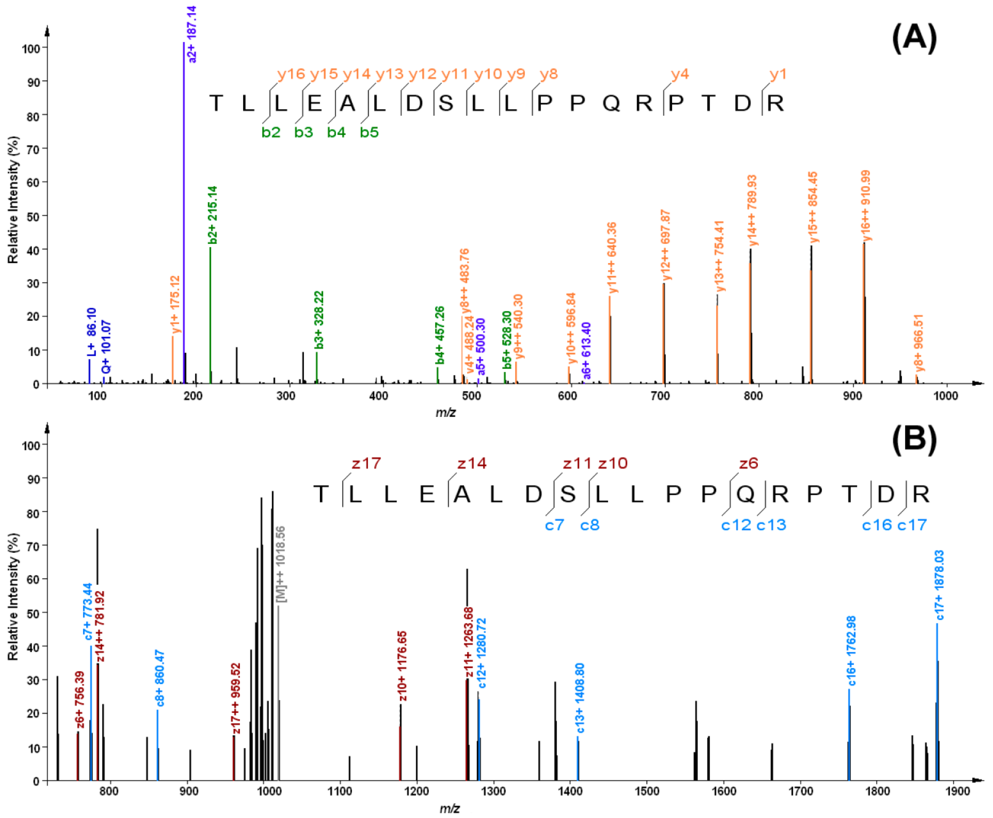


Figure 6. HCD and ETD spectral pair belonging to the same precursor. De novo sequencing failed on either HCD (A) or ETD (B) spectrum but succeeded on the pair of them, as the fragmentation information is present for nearly all of the amino acids in one spectrum or the other.

Table 9. Number of Spectra with a Maximum Gap Length of No More than Two Amino Acids in the Worm Data Set[a]

	HCD + ETD	HCD	ETD
Asp-N (2367)[b]	2346 (99.1%)	2217 (94.5%)	1999 (85.2%)
elastase (1161)	1148 (98.9%)	906 (78.9%)	895 (78.0%)
Glu-C (1523)	1488 (97.7%)	1239 (83.3%)	1143 (76.8%)
trypsin (3626)	3612 (99.6%)	3561 (98.6%)	2936 (81.3%)
sum (8677)	8594 (99.0%)	7923 (92.2%)	6973 (81.1%)

[a] Maximum gap is defined as the maximum length of consecutive fragmentation sites not supported by any of the peaks in the corresponding spectrum. [b] The number in parentheses represents the total number of spectra from the corresponding enzymatic digestion.

sequences could be identified through database search at 1% FDR). The HCD and ETD spectra of trypsin or Asp-N peptides contain fewer gaps than those of Glu-C or elastase peptides, consistent with the result that more trypsin and Asp-N peptides are sequenced successfully using pNovo+.

Mass accuracy is another important factor. We have compared the results obtained by pNovo+ using different mass tolerance windows for fragment ions. As shown in Figure 7, the best result was achieved at ±20 ppm. A sharp decrease of correct de novo sequencing results occurred at ±300 ppm or larger, which simulates low mass accuracy MS/MS data. Thus, high mass accuracy is critical for de novo sequencing, consistent with previous reports.[26,37]

3.4. Run Time Comparison

The run time of pNovo+ on each data set was compared with that of PEAKS. As shown in Table 10, the average run time of pNovo+ is 0.018 s per spectrum, much faster than PEAKS (0.191 for HCD and 0.186 for ETD). With the pDAG algorithm, pNovo+ is very efficient at finding the *k* longest paths. As compared with the run time of other de novo sequencing algorithms, pNovo+ is 3−100 times faster.[27,37] It can be reasonably deduced that 50 or more MS/MS spectra can be sequenced per second, as fast as, if not faster than, the data acquisition speed of any high-resolution mass spectrometers available today. Theoretically, it could be employed for real-time spectral data analysis in shot gun proteomics.

To better evaluate how the release of the antisymmetry restriction might affect de novo sequencing, we used depth-first search (DFS) with an efficient pruning strategy, which we had used in pNovo,[26] to replace the pDAG algorithm to find the *k* longest paths. Antisymmetry restriction is considered in the DFS algorithm. Table 11 shows the run time comparison between pDAG and DFS in finding the *k* longest paths, and pDAG is ~20 times faster than DFS. With respect to accuracy, there is hardly any difference between pDAG and DFS on (HCD + ETD) data and no more than 1% difference on HCD or ETD alone, or (HCD ∪ ETD) (Table S9−S12 in the Supporting Information). Therefore, we conclude that releasing the antisymmetry restriction markedly increases the speed and hardly affects the accuracy of de novo sequencing.

4. DISCUSSION

In this paper, we present a de novo sequencing algorithm called pNovo+ based on complementary HCD and ETD spectra. With high mass accuracy HCD and ETD spectral pairs, missing fragmentation information in one spectrum may be found in the other, and the antisymmetry restriction becomes unnecessary. Thus, we removed the antisymmetry restriction and

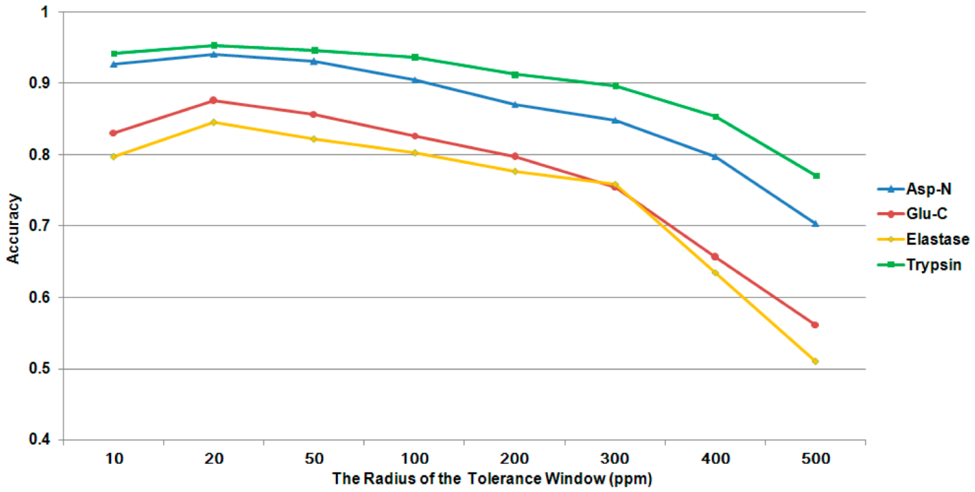


Figure 7. A 20 ppm mass accuracy window is optimal for the performance of pNovo+. The top three candidates of every spectrum in the Worm data set were included in the statistical analysis.

Table 10. Run Time of pNovo+ and PEAKS[a]

time (s)	pNovo+ (HCD + ETD)	pNovo+ (HCD)	pNovo+ (ETD)	PEAKS[b] (HCD)	PEAKS (ETD)
Asp-N (2367)[c]	45.879	40.747	36.738	439	387
elastase (1161)	20.779	20.187	18.595	274	253
Glu-C (1523)	30.42	26.972	26.427	373	319
trypsin (3626)	62.571	59.998	48.906	575	656
average (8677)	0.018[d]	0.017	0.015	0.191	0.186

[a] All of the run time tests were performed on the same PC (Dell Precision T1500, Inter(R) Core(TM) i7 CPU 860 at 2.80 GHz). [b] A module for run time calculation is embedded into pNovo+. However, we are unable to obtain the exact run time of PEAKS; therefore, it was calculated manually by recording the start and end times for each run, and the accuracy is up to 1 s. [c] The number in parentheses indicates the total number of spectra generated in each test set. [d] This value is the average run time per merged spectrum from an HCD and ETD spectral pair.

Table 11. Run Time Comparison between pDAG and DFS Algorithms in Finding the *k* Longest Paths

	pDAG algorithm			DFS algorithm		
time (s)	pNovo+ (HCD + ETD)	pNovo + (HCD)	pNovo + (ETD)	pNovo+ (HCD + ETD)	pNovo+ (HCD)	pNovo+ (ETD)
Asp-N (2367)	1.192	1.384	0.876	32.675	25.177	17.867
elastase (1161)	0.511	0.535	0.422	15.409	10.253	6.859
Glu-C (1523)	0.852	0.775	0.578	30.238	18.730	15.025
trypsin (3626)	2.016	1.946	3.180	47.444	37.984	25.372
sum (8677)	4.571	4.64	5.056	125.766	92.144	65.123
average (8677)	0.001	0.001	0.001	0.014	0.011	0.008

developed an efficient algorithm pDAG to find the *k* longest paths, and this significantly improved the speed of de novo peptide sequencing. Immonium and internal ions in HCD spectra and hydrogen rearranged fragment ions in ETD spectra are also considered in the spectrum graph construction and the design of the scoring function. When tested on two different data sets, each with four types of enzymatic digestions, up to 95% HCD + ETD spectral pairs were correctly sequenced by pNovo+ at a rapid speed.

In the course of this study, we found that the antisymmetry restriction affects de novo sequencing differently depending on the type of MS/MS data. For low-resolution MS/MS data, imposing the antisymmetry restriction is critical; otherwise, many incorrect paths will be retrieved. In contrast, the gain in accuracy from such restriction is diminished for high-resolution MS/MS data and disappeared entirely for high-resolution HCD + ETD spectral pairs. An incorrect peak-ion match, if encountered, can often be distinguished by its mass deviation, which is typically larger than that of a correct match; a peak-ion match with a larger mass deviation is given a smaller weight. This is used in pNovo, pNovo+, and a number of other algorithms such as Vonode.[28]

Identifying novel proteins (that is, to infer the function of a novel protein of interest by identifying its true homologous proteins) from disconnected short peptides remains a challenge. To obtain longer sequences, several studies relied on de novo sequencing of peptides from a variety of enzymatic digestions and assembling them into contigs.[53,54] However, local sequencing errors, frequent on the N termini of peptides, make it difficult to obtain long and accurate sequences. In a previous work, we have shown that de novo sequencing results on separate ETD and HCD data helped identify novel proteins by providing assurance in sequence accuracy and better overlap.[55] In this study, we report further improvement with ETD and HCD spectral pairs, which we recommend for all de novo sequencing applications.

In several studies, the "golden complementary pairs" rule has been used for ion type determination in complementary CAD and ECD spectra.[36,37] In this work, we also attempted to determine the ion types of peaks recorded in paired HCD and ETD spectra. While this is successful for a subset of peaks, for the vast majority, the ion types cannot be determined. Furthermore, knowing the ions types of a small fraction of the peaks did not help de novo sequencing based on spectrum graphs (data not shown), in which the different types of ions are treated as different vertices. Nevertheless, further refinement of ion type determination is under investigation. A method that can accurately determine the charge states of peaks not associated with any isotopic peak series could be helpful, too. pNovo+ can be downloaded for free from http://pfind.ict.ac.cn/software/pNovo/index.html.

■ ASSOCIATED CONTENT

S Supporting Information

Figures of HCD and ETD prefix and suffix offset frequency function, algorithm pDAG, and the proof of theorem 1. This material is available free of charge via the Internet at http://pubs.acs.org.

■ AUTHOR INFORMATION

Corresponding Author

*Tel: +86-10-62601042. Fax: +86-10-62601356. E-mail: smhe@ict.ac.cn (S.-M.H.). Tel: +86-10-80726688-8515. Fax: +86-10-80706053. E-mail: dongmengqiu@nibs.ac.cn (M.-Q.D.).

Present Address

#H.C. is now a Ph.D. student at the Program in Computational Biology and Bioinformatics, University of Southern California, Los Angeles, California 90089, United States.

Author Contributions

⊥These authors contributed equally to this work.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Key Basic Research and Development Program of China (973) under Grant Nos. 2010CB912701, 2010CB835203, and 2012CB910602 and the CAS Knowledge Innovation Program under Grant Nos. KGCX1-YW-13 and 20120633. We thank Zuo-Fei Yuan, Chao Liu, Sheng-Bo Fan, and Kun Zhang for valuable discussions.

■ REFERENCES

(1) Aebersold, R.; Mann, M. Mass spectrometry-based proteomics. *Nature* **2003**, *422* (6928), 198−207.

(2) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551−3567.

(3) Eng, J. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994**, *5* (11), 976−989.

(4) Craig, R.; Beavis, R. C. TANDEM: Matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466−1467.

(5) Geer, L. Y.; Markey, S. P.; Kowalak, J. A.; Wagner, L.; Xu, M.; Maynard, D. M.; Yang, X.; Shi, W.; Bryant, S. H. Open mass spectrometry search algorithm. *J. Proteome Res.* **2004**, *3* (5), 958−964.

(6) Bern, M.; Cai, Y.; Goldberg, D. Lookup peaks: A hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.* **2007**, *79* (4), 1393−1400.

(7) Fu, Y.; Yang, Q.; Sun, R.; Li, D.; Zeng, R.; Ling, C. X.; Gao, W. Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. *Bioinformatics* **2004**, *20* (12), 1948−1954.

(8) Tanner, S.; Shu, H.; Frank, A.; Wang, L. C.; Zandi, E.; Mumby, M.; Pevzner, P. A.; Bafna, V. InsPecT: Identification of posttranslationally modified peptides from tandem mass spectra. *Anal. Chem.* **2005**, *77* (14), 4626−4639.

(9) Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. OLAV: Towards high-throughput tandem mass spectrometry data identification. *Proteomics* **2003**, *3* (8), 1454−1463.

(10) Allmer, J. Algorithms for the de novo sequencing of peptides from tandem mass spectra. *Expert Rev. Proteomics* **2011**, *8* (5), 645−657.

(11) Ma, B.; Johnson, R. De novo sequencing and homology searching. *Mol. Cell Proteomics* **2012**, *11* (2), O111.014902.

(12) Dancik, V.; Addona, T. A.; Clauser, K. R.; Vath, J. E.; Pevzner, P. A. De novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **1999**, *6* (3−4), 327−342.

(13) Bartels, C. Fast Algorithm for Peptide Sequencing by Mass-Spectroscopy. *Biomed. Environ. Mass Spectrom.* **1990**, *19* (6), 363−368.

(14) Frank, A. M.; Savitski, M. M.; Nielsen, M. L.; Zubarev, R. A.; Pevzner, P. A. De novo peptide sequencing and identification with precision mass spectrometry. *J. Proteome Res.* **2007**, *6* (1), 114−123.

(15) Frank, A.; Pevzner, P. PepNovo: De novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **2005**, *77*, 964−973.

(16) Ma, B.; Zhang, K. Z.; Hendrie, C.; Liang, C. Z.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: powerful software for peptide de novo sequencing by MS/MS. *Rapid Commun. Mass Spectrom.* **2003**, *17*, 2337−2342.

(17) Taylor, J. A.; Johnson, R. S. Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **1997**, *11* (9), 1067−1075.

(18) Grossmann, J.; Roos, F. F.; Cieliebak, M.; Liptak, Z.; Mathis, L. K.; Muller, M.; Gruissem, W.; Baginsky, S. AUDENS: A tool for automated peptide de novo sequencing. *J. Proteome Res.* **2005**, *4* (5), 1768−1774.

(19) Mo, L.; Dutta, D.; Wan, Y.; Chen, T. MSNovo: A dynamic programming algorithm for de novo peptide sequencing via tandem mass spectrometry. *Anal. Chem.* **2007**, *79* (13), 4870−4878.

(20) Fernandez-de-Cossio, J.; Gonzalez, J.; Betancourt, L.; Besada, V.; Padron, G.; Shimonishi, Y.; Takao, T. Automated interpretation of high-energy collision-induced dissociation spectra of singly protonated peptides by "SeqMS", a software aid for de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **1998**, *12* (23), 1867−1878.

(21) Fernandez-de-Cossio, J.; Gonzalez, J.; Satomi, Y.; Shima, T.; Okumura, N.; Besada, V.; Betancourt, L.; Padron, G.; Shimonishi, Y.; Takao, T. Automated interpretation of low-energy collision-induced dissociation spectra by SeqMS, a software aid for de novo sequencing by tandem mass spectrometry. *Electrophoresis* **2000**, *21* (9), 1694−1699.

(22) Jagannath, S.; Sabareesh, V. Peptide Fragment Ion Analyser (PFIA): a simple and versatile tool for the interpretation of tandem mass spectrometric data and de novo sequencing of peptides. *Rapid Commun. Mass Spectrom.* **2007**, *21* (18), 3033−3038.

(23) Fischer, B.; Roth, V.; Roos, F.; Grossmann, J.; Baginsky, S.; Widmayer, P.; Gruissem, W.; Buhmann, J. M. NovoHMM: a hidden Markov model for de novo peptide sequencing. *Anal. Chem.* **2005**, *77* (22), 7265−7273.

(24) Bern, M.; Goldberg, D. De novo analysis of peptide tandem mass spectra by spectral graph partitioning. *J. Comput. Biol.* **2006**, *13* (2), 364−378.

(25) DiMaggio, P. A., Jr.; Floudas, C. A. De novo peptide identification via tandem mass spectrometry and integer linear optimization. *Anal. Chem.* **2007**, *79* (4), 1433−1446.

(26) Chi, H.; Sun, R. X.; Yang, B.; Song, C. Q.; Wang, L. H.; Liu, C.; Fu, Y.; Yuan, Z. F.; Wang, H. P.; He, S. M.; Dong, M. Q. pNovo: De novo peptide sequencing and identification using HCD spectra. *J. Proteome Res.* **2010**, *9* (5), 2713−2724.

(27) Andreotti, S. K.; Reinert, G. W.; Antilope, K.; Lagrangian, A. Relaxation Approach to the de novo Peptide Sequencing Problem. *IEEE/ACM Trans. Comput. Biol. Bioinf.* **2011**, *9* (2), 385−394.

(28) Pan, C.; Park, B. H.; McDonald, W. H.; Carey, P. A.; Banfield, J. F.; VerBerkmoes, N. C.; Hettich, R. L.; Samatova, N. F. A high-throughput de novo sequencing approach for shotgun proteomics

using high-resolution tandem mass spectrometry. *BMC Bioinf.* **2010**, *11*, 118.

(29) Chen, T.; Kao, M. Y.; Tepel, M.; Rush, J.; Church, G. M. A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **2001**, *8* (3), 325−337.

(30) Lu, B.; Chen, T. A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **2003**, *10* (1), 1−12.

(31) Lu, B.; Chen, T. Algorithms for de novo peptide sequencing via tandem mass spectrometry. *Biosilico* **2004**, *2* (2), 85−90.

(32) Zhang, Z. Prediction of low-energy collision-induced dissociation spectra of peptides. *Anal. Chem.* **2004**, *76* (14), 3908−3922.

(33) Spengler, B. De novo sequencing, peptide composition analysis, and composition-based sequencing: a new strategy employing accurate mass determination by fourier transform ion cyclotron resonance mass spectrometry. *J. Am. Soc. Mass Spectrom.* **2004**, *15* (5), 703−714.

(34) Boersema, P. J.; Taouatas, N.; Altelaar, A. F.; Gouw, J. W.; Ross, P. L.; Pappin, D. J.; Heck, A. J. Mohammed, S., Straightforward and de novo peptide sequencing by MALDI-MS/MS using a Lys-N metalloendopeptidase. *Mol. Cell Proteomics* **2009**, *8* (4), 650−660.

(35) Pevtsov, S.; Fedulova, I.; Mirzaei, H.; Buck, C.; Zhang, X. Performance evaluation of existing de novo sequencing algorithms. *J. Proteome Res.* **2006**, *5* (11), 3018−3028.

(36) Horn, D. M.; Zubarev, R. A.; McLafferty, F. W. Automated de novo sequencing of proteins by tandem high-resolution mass spectrometry. *Proc. Natl. Acad. Sci. U.S.A.* **2000**, *97* (19), 10313−10317.

(37) Savitski, M. M.; Nielsen, M. L.; Kjeldsen, F.; Zubarev, R. A. Proteomics-grade de novo sequencing approach. *J. Proteome Res.* **2005**, *4* (6), 2348−2354.

(38) Datta, R.; Bern, M. Spectrum fusion: Using multiple mass spectra for de novo Peptide sequencing. *J. Comput. Biol.* **2009**, *16* (8), 1169−1182.

(39) Bertsch, A.; Leinenbach, A.; Pervukhin, A.; Lubeck, M.; Hartmer, R.; Baessmann, C.; Elnakady, Y. A.; Muller, R.; Bocker, S.; Huber, C. G.; Kohlbacher, O. De novo peptide sequencing by tandem MS using complementary CID and electron transfer dissociation. *Electrophoresis* **2009**, *30* (21), 3736−3747.

(40) He, L.; Ma, B. ADEPTS: Advanced peptide de novo sequencing with a pair of tandem mass spectra. *J. Bioinform. Comput. Biol.* **2010**, *8* (6), 981−994.

(41) Kim, S.; Mischerikow, N.; Bandeira, N.; Navarro, J. D.; Wich, L.; Mohammed, S.; Heck, A. J.; Pevzner, P. A. The generating function of CID, ETD, and CID/ETD pairs of tandem mass spectra: applications to database search. *Mol. Cell Proteomics* **2010**, *9* (12), 2840−2852.

(42) Gabow, H. N.; Maheshwari, S. N.; Osterweil, L. J. On Two Problems in the Generation of Program Test Paths. *IEEE Trans. Softw. Eng.* **1976**, *2* (3), 227−231.

(43) Sun, R. X.; Dong, M. Q.; Song, C. Q.; Chi, H.; Yang, B.; Xiu, L. Y.; Tao, L.; Jing, Z. Y.; Liu, C.; Wang, L. H.; Fu, Y.; He, S. M. Improved peptide identification for proteomic analysis based on comprehensive characterization of electron transfer dissociation spectra. *J. Proteome Res.* **2010**, *9* (12), 6354−6367.

(44) Gentzel, M.; Kocher, T.; Ponnusamy, S.; Wilm, M. Preprocessing of tandem mass spectrometric data to support automatic protein identification. *Proteomics* **2003**, *3* (8), 1597−1610.

(45) Zhang, J.; He, S.; Ling, C. X.; Cao, X.; Zeng, R.; Gao, W. PeakSelect: Preprocessing tandem mass spectra for better peptide identification. *Rapid Commun. Mass Spectrom.* **2008**, *22* (8), 1203−1212.

(46) Yen, S. H.; Du, D. H.; Ghanta, S. Efficient algorithms for extracting the K most critical paths in timing analysis. *Proc. ACM/IEEE Des. Autom. Conf.* **1989**, 649−654.

(47) Ju, Y. C.; Saleh, R. A. Incremental Techniques for the Identification of Statically Sensitizable Critical Paths. *In Design Autom. Conf.* **1991**, 541−546.

(48) Kundu, S. An incremental algorithm for identification of longest (shortest) paths. *Integr. VLSI J.* **1994**, *17* (1), 25−31.

(49) McDonald, W. H.; Ohi, R.; Miyamoto, D. T.; Mitchison, T. J.; Yates, J. R. Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: Single-dimension LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *Int. J. Mass Spectrom.* **2002**, *219* (1), 245−251.

(50) pFind Studio: A computational solution for mass spectrometry-based proteomics (http://pfind.ict.ac.cn).

(51) Levenshtein, V. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* **1966**, 707−710.

(52) Zubarev, R. A.; Zubarev, A. R.; Savitski, M. M. Electron capture/transfer versus collisionally activated/induced dissociations: Solo or duet? *J. Am. Soc. Mass Spectrom.* **2008**, *19* (6), 753−761.

(53) Bandeira, N.; Clauser, K. R.; Pevzner, P. A. Shotgun protein sequencing: Assembly of peptide tandem mass spectra from mixtures of modified proteins. *Mol. Cell Proteomics* **2007**, *6* (7), 1123−1134.

(54) Liu, X.; Han, Y.; Yuen, D.; Ma, B. Automated protein (re)sequencing with MS/MS and a homologous database yields almost full coverage and accuracy. *Bioinformatics* **2009**, *25* (17), 2174−2180.

(55) Zhao, Y.; Sun, W.; Zhang, P.; Chi, H.; Zhang, M. J.; Song, C. Q.; Ma, X.; Shang, Y.; Wang, B.; Hu, Y.; Hao, Z.; Huhmer, A. F.; Meng, F.; L'Hernault, S, W.; He, S. M.; Dong, M. Q.; Miao, L. Nematode sperm maturation triggered by protease involves sperm-secreted serine protease inhibitor (Serpin). *Proc. Natl. Acad. Sci. U.S.A.* **2012**, *109* (5), 1542−1547.