research articles Journal of proteome research

pNovo: *De novo* Peptide Sequencing and Identification Using HCD Spectra

Hao Chi,[†,‡] Rui-Xiang Sun,[†] Bing Yang,[§] Chun-Qing Song,[§] Le-Heng Wang,[†] Chao Liu,[†,‡] Yan Fu,[†] Zuo-Fei Yuan,[†,‡] Hai-Peng Wang,[†,‡] Si-Min He,[*,†] and Meng-Qiu Dong[*,§]

Key Lab of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing 100190, People's Republic of China, Graduate University of Chinese Academy of Sciences, Beijing 100049, People's Republic of China, and National Institute of Biological Sciences, Beijing, Beijing 102206, People's Republic of China

Received March 1, 2010

De novo peptide sequencing has improved remarkably in the past decade as a result of better instruments and computational algorithms. However, *de novo* sequencing can correctly interpret only ~30% of high- and medium-quality spectra generated by collision-induced dissociation (CID), which is much less than database search. This is mainly due to incomplete fragmentation and overlap of different ion series in CID spectra. In this study, we show that higher-energy collisional dissociation (HCD) is of great help to *de novo* sequencing because it produces high mass accuracy tandem mass spectrometry (MS/MS) spectra without the low-mass cutoff associated with CID in ion trap instruments. Besides, abundant internal and immonium ions in the HCD spectra can help differentiate similar peptide sequences. Taking advantage of these characteristics, we developed an algorithm called pNovo for efficient *de novo* sequencing of peptides from HCD spectra. pNovo gave correct identifications to 80% or more of the HCD spectra identified by database search. The number of correct full-length peptides sequenced by pNovo is comparable with that obtained by database search. A distinct advantage of *de novo* sequencing is that deamidated peptides and peptides with amino acid mutations can be identified efficiently without extra cost in computation. In summary, implementation of the HCD characteristics makes pNovo an excellent tool for *de novo* peptide sequencing from HCD spectra.

Keywords: tandem mass spectrometry • HCD • *de novo* sequencing • pNovo

1. Introduction

Both database search and *de novo* peptide sequencing can be used for protein identification. Thanks to the fast development of protein databases, such as IPI, Swiss-Prot, and RefSeq (reviewed in ref 1), database search has long been the dominant approach. A large number of database search algorithms and software tools are used in routine experiments, for example, Mascot,[2] SEQUEST,[3] X! Tandem,[4,5] pFind,[6–8] Phenyx,[9,10] and OMSSA.[11] Generally speaking, the essence of these methods is retrieving all candidate peptides from a specified database for each spectrum, followed by scoring of each peptide-spectrum match (PSM).[12] Only the precursor ion mass of each experimental spectrum is used to prune invalid peptides from the database, although detailed information in each spectrum such as the charge and m/z of fragment ion peaks could be used, too, to further filter invalid peptides from the database. As such, a spectrum may be matched with a huge number of peptides and it may be difficult to retrieve the correct interpretation efficiently and accurately. This situation is exacerbated if multiple post-translational modifications (PTMs) are searched for simultaneously, because the number of candidate peptides will increase exponentially and it will take a long time to score each spectrum against all candidates.[13] On the other hand, if the corresponding sequence of a given spectrum is not in the database, then the spectrum cannot be identified regardless of its quality.

An alternative approach is *de novo* peptide sequencing, which extracts a peptide sequence directly from a spectrum and hence does not require any protein database. *De novo* peptide sequencing is essential if there is no protein database available for a sample of interest.[12] Another potential advantage of *de novo* sequencing is to discover mutations and modifications, including unexpected or unknown ones. Multiple *de novo* peptide sequencing algorithms have been reported in recent years, such as PepNovo,[14] PEAKS,[15] SHERENGA,[13] Lutefisk,[16] AuDeNs,[17] MSNovo,[18] SeqMS,[19,20] PFIA,[21] and NovoHMM.[22] Most of them use spectrum graph or a similar approach, in which each original spectrum is transformed into a directed acyclic graph and the optimal paths are found via dynamic programming algorithms.[13,23–25]

Thanks to the advancement of the mass spectrometry technology, especially the emergence of new fragmentation techniques, for example, higher-energy collisional dissociation

* To whom correspondence should be addressed. Email: smhe@ict.ac.cn, dongmengqiu@nibs.ac.cn.
† Institute of Computing Technology, Chinese Academy of Sciences.
‡ Graduate University of Chinese Academy of Sciences.
§ National Institute of Biological Sciences, Beijing.

10.1021/pr100182k

Published on Web 03/23/2010

(HCD, also called higher-energy C-trap dissociation in earlier orbitrap instruments), electron capture dissociation (ECD), or electron transfer dissociation (ETD), and reduced cost and maintenance burden of high-precision mass spectrometers, novel computational methods are investigated to improve *de novo* peptide sequencing. Frank et al. proposed that precision mass spectrometry, available with Q-TOF, FT-ICR, and Orbitrap, can remarkably increase the ratio of identified amino acids and correct peptides.[26] Savitski et al. also presented strong evidence that mass accuracy plays an extremely important role in peptide sequencing.[27] Spengler proposed a strategy based on analysis of amino acid composition and high mass accuracy to reduce the possible combinations of amino acids.[28] Generally speaking, precision mass spectrometry decreases the complexity of common *de novo* sequencing algorithms by restricting the occurrence of random matches. Novel fragmentation methods complementary to the traditional collision-activated dissociation or collision-induced dissociation (CAD or CID) are also helpful to *de novo* sequencing. For instance, CID and ETD (or ECD) spectra belonging to the same precursor can be paired up to obtain more fragmentation information.[29] Horn et al. described an algorithm to distinguish N- and C-terminal fragments using CID and ECD spectra.[30] Savitski et al. developed a similar but more hierarchically structured method and used it in a proteomics-scale data analysis.[27] Datta and Bern proposed an algorithm to transform the information in each CID-ETD spectral pair into a higher-quality integrated spectrum using a Bayesian network.[31]

Although development of mass spectrometry instruments and computation has improved spectral interpretation, *de novo* peptide sequencing is still far from being a mature method. Compared with database search, *de novo* peptide sequencing usually yields less accurate identifications. A comparative study showed that while more that 60% of the amino acid residues can be predicted by the most powerful software tools, only less than 30% of peptides can be correctly identified from the test data.[32] Another study examining several *de novo* sequencing algorithms found that no more than 50% of the peptide identifications were exactly right, no matter which algorithm was used to generate them.[33] Generally speaking, the performance of *de novo* sequencing algorithms deteriorates rapidly when longer sequences are required.[34] As such, *de novo* peptide sequencing is scarcely used in routine experiments. In most cases, *de novo* peptide sequencing is integrated with database search. With this hybrid approach, short and relatively reliable sequence tags or full-length *de novo* reconstructions are generated first, and then these sequences are used to filter candidate peptides in the database.[35–40] *De novo* sequencing of full-length peptides remains an immense challenge.

To conquer the difficulties described above and obtain more reliable results by *de novo* sequencing, a feasible approach is to utilize spectra containing peptide fragmentation information as complete as possible. Olsen et al. suggested that HCD spectra could facilitate *de novo* sequencing.[41] Here we find that HCD is indeed an excellent choice for *de novo* sequencing. In our data, around 48.6% of the HCD spectra that are reliably identified by database search contain full cleavage information of peptides, that is, all peptide-bond cleavages along a peptide backbone are represented by observed fragment ions. The spectra with only one missing cleavage account for another 31.2%. So together, 79.8% of the HCD spectra contain full or almost full cleavage information, substantially higher than that of CID (62.8%) or ETD (65.03%) spectra. Besides high mass accuracy of fragment ions and nearly complete ion series, the presence of many immonium ions and internal fragment ions in HCD spectra also improves *de novo* sequencing because it can be used to distinguish between similar candidate peptides.[42] We find that over 50% of dipeptide ions, as well as about 40% of tri- and tetra- peptide ions, are present in the HCD spectra. For some amino acids that are prone to produce immonium ions, such as Cys, Tyr, Trp, His, and Phe, their immonium ion peaks can be observed with a probability of over 95%. Other amino acids, such as Glu, Val, and Ile/Leu, also give a probability of 50−80% for the detection of their immonium ions. The internal and immonium ions can help distinguish between sequences with slight differences, so a more effective *de novo* peptide sequencing algorithm can be hoped for.

In this paper, we present an automated *de novo* algorithm, pNovo, which takes the characteristics of HCD spectra into full consideration. The average accuracy of pNovo results is ~96.2% for amino acid residues. From the test HCD spectra with reliable sequence identifications assigned by database search, pNovo obtained correct full-length sequences for at least 80% of them. The basic approach is similar to a spectrum graph but differs from it in some important details of realization, such as the application of low mass ions (below 500 Da) and careful consideration of mass accuracy. A novel scoring scheme has been developed to distinguish between similar peptides with minute differences. Spectra of highly charged peptides can also be handled efficiently by pNovo.

2. Algorithms

Data Preprocessing. The preprocessing consists of four steps. In Step 1, the charge state of each peak is determined by its isotopic peak cluster. HCD spectra are of high resolution, so doubly and triply charged ion peaks can be identified correctly (Figure 1a). If a peak cannot be assigned to an isotopic cluster, it is then treated as a singly charged ion. In Step 2, all absolute peak intensities are transformed into relative ranks. The reason is that some extremely strong peaks could be interpreted incorrectly as fragment ions that differ from other fragment ions by a certain amino acid(s), bringing in inaccurate results.[26] Therefore, the ranks are computed to smooth the intensity variation among different peaks. In Step 3, immonium ions are removed because some amino acids such as Cys, Phe, Ile/Leu, Tyr, and Trp tend to produce abundant immonium ions via HCD (Figure 1b). If these immonium ions are kept, the N- and C-terminal regions of the spectrum graph may become too complex. In the final step, K most intense peaks are picked out in each spectrum for the construction of a spectrum graph.

Constructing a Spectrum Graph. To select appropriate ion types used in the algorithm and learn the relationship between them, we used the *offset frequency function* (OFF), reported by Dančík et al.[13] Suppose a spectrum S consists of m observed peaks from s_1 to s_m, and the prefix residue masses of the ground-truth peptide is represented by p_1, p_2,.. ., p_n. Then the *OFF* is computed as follows. For every s_i and p_j, we calculate their distance δ with the accuracy of two decimal places, and plot the occurrence of different δ values. The suffix *OFF* is computed in a similar way. Finally, we let pNovo consider the six most abundant types of fragment ions, y^+, b^+, y^+-NH_3, y^+-H_2O, a^+ and y^{2+} (Figure S1 in the Supporting Information shows the prefix and suffix *OFF*s).

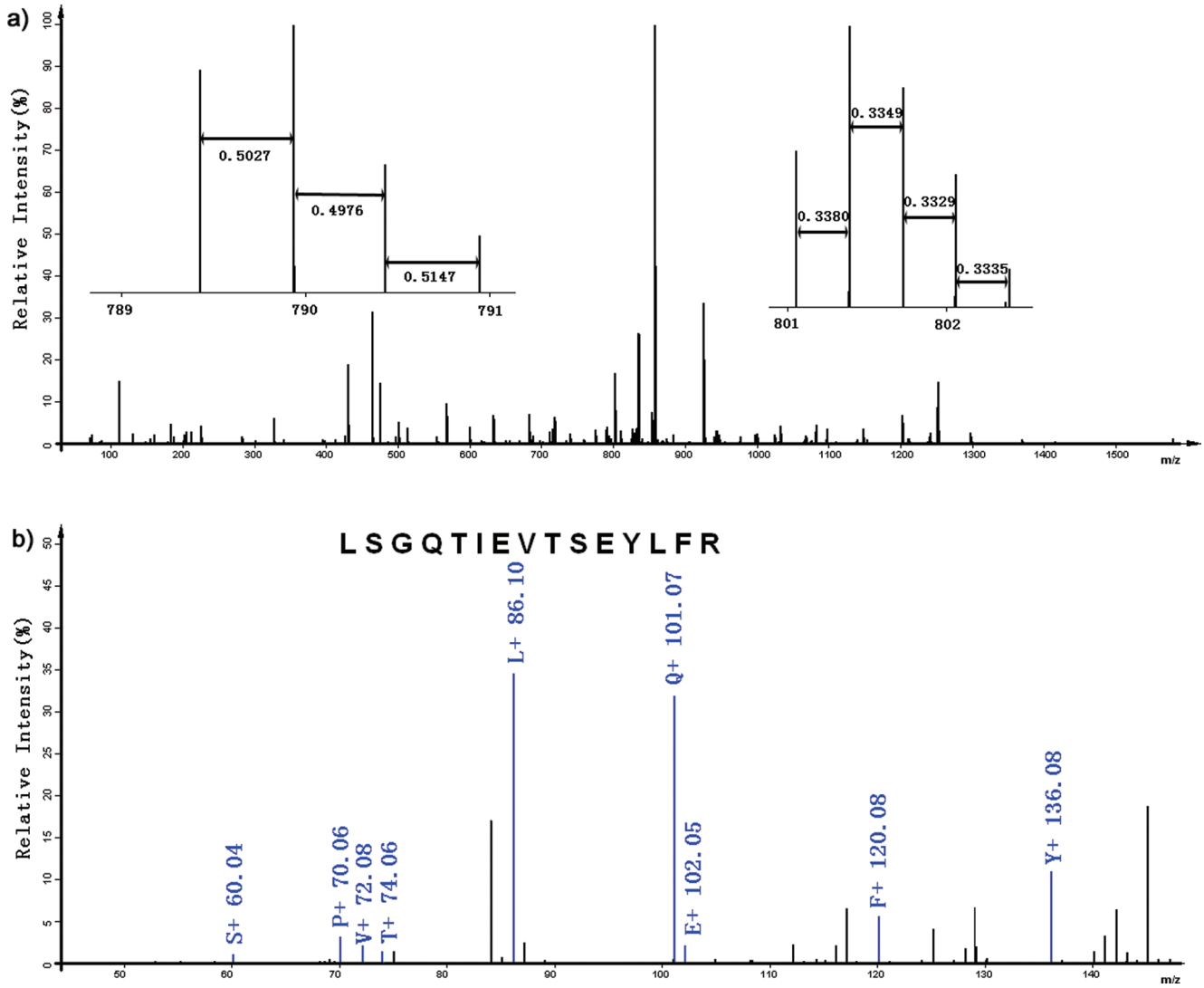

Figure 1. (a) High mass accuracy is of great help to determine the charge states of peaks in HCD MS/MS spectra. The monoisotopic peak p_1 located at *m*/*z* of 789.4269 is doubly charged, and the average distance between two adjacent peaks in the isotopic cluster is 0.505 *m*/*z*. The monoisotopic peak p_2 at *m*/*z* of 801.0531 is triply charged, for its isotopic peaks are spaced 0.3353 *m*/*z* apart. The average deviation associated with the isotopic peak spacing is 0.00665 Da for the former and 0.00255 Da for the latter. (b) HCD MS/MS spectra are rich in information of immonium ions. The peptide LSGQTIEVTSEYLFR is assigned to this spectrum by pFind with an e-value of 1.13×10^{-5}. Present in the spectrum are all the immonium ions of amino acids found in this peptide except for Gly (the immonium ion of Gly is below the scan range) and Arg (the immonium ion of Arg is always very weak or absent[42]).

The construction procedure of a spectrum graph is as follows: in the first step, each peak is split into *k* vertexes in general cases, where *k* is the number of the selected ion types ($k = 6$ in this paper, for y^+, b^+, y^+-NH_3, y^+-H_2O, a^+, and y^{2+} ions). For instance, if there is a peak located at *m*/*z* 796.54 in a spectrum whose MH^+ is 1387.76 Da, and the possibilities of both *b*- and *y*- ions are taken into consideration, then two vertexes, located at *m*/*z* 795.54 and *m*/*z* 591.22, are generated respectively. For convenience, we also call these *m*/*z* values the "masses" of the vertexes. The weight of each vertex is the intensity of its corresponding peak. For each peak, one or more of the *k* vertexes may not be generated in the algorithm. For example, the appearance of the *y*-H_2O-ions depends on the appearance of their cognate *y*-ions, and *a*-ions are often gathered in the low and medium mass region of a spectrum. As a result, only if a peak is associated with a probability greater than 0.1 as an assumed ion type, the vertex can be generated accordingly (the fragment ion frequencies are shown in Table 1). A conditional probability greater than 0.8 for each derived ion type is also necessary, as described in Table S1 (Supporting Information).

Table 1. Information of Different Ion Types Learned from the Offset Frequency Function (OFF)

ion	offset	prefix/suffix	mass deviation	frequency[a]
y	19.0158	suffix	−0.0021	0.668
b	1.0065	prefix	−0.0009	0.286
y-NH_3	1.9903	suffix	−0.0009	0.177
y-H_2O	1.0065	suffix	−0.0008	0.155
a	−26.9885	prefix	0.0009	0.200
b-H_2O	−17.0044	prefix	0.0011	0.121
y^{2+}	10.0102	suffix	−0.0024	0.100[b]

[a] Frequency of each ion type is calculated as no. observed ions/no. total ions in the scanned mass range. [b] Although y^{2+} ions appear as a lower frequency, we also choose to consider it for constructing spectrum graph because in triply charged spectra, 38.8% of y^{2+} ions can be observed.

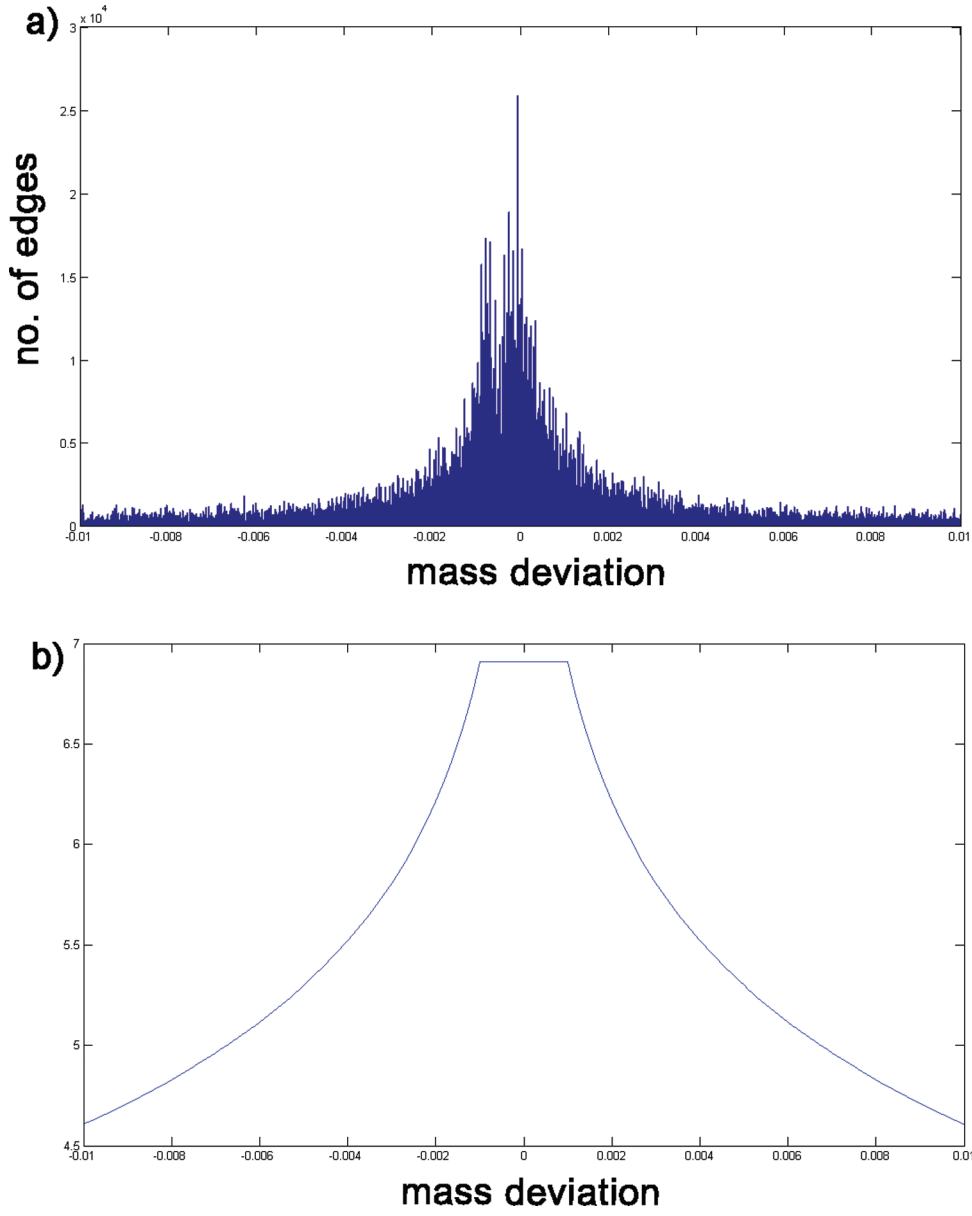

Figure 2. (a) Distribution of mass deviations for the number of edges in all the optimal paths in the spectrum graphs. The range is from −0.001 to 0.01 Da, in which 2 056 391 different mass deviations of edges are counted in the histogram. Of these, 1 455 036 deviations (70.9% of the total), are in the range from −0.001 to 0.001 Da. (b) Curve of the penalty function used in the scoring scheme of pNovo.

In the second step, if two or more vertexes are of equal mass within a tolerance range, then a merging algorithm is used to integrate them together. Adapted from a reported algorithm,[13] the weight of each merged vertex is assigned the sum of the weight of each component vertex.

In the third step, we add special vertexes, including the source and target vertexes with the mass of 0 and $M - 18$, respectively, where M denotes the peptide mass, and some characteristic vertexes as appropriate for enzyme specificity. For example, if trypsin is used in the experiment to digest proteins, two vertexes, with the mass of $M - 128.09$ and $M - 156.10$, respectively, should be added. After that, pNovo connects two vertexes if and only if their distance in mass is equal to the sum of one or more residue masses within a tolerance range. Note that there may be some edges with more than one amino acid residue combination. For instance, the mass of the sum of Ser and Leu is equal to that of Thr and Val. Then all combinations are recorded for further generation of candidate peptides. The weight of each edge is computed by adding the weights of the two corresponding vertexes.

In the last and most important step, we reassign the weights of the edges with considerations of mass accuracy and observed internal and immonium ions. As shown in Figure 2a, although the maximum mass deviation of HCD spectra is usually ±0.01 Da, over 70% of the edges in the spectrum graph are within a much narrower tolerance window, from −0.001 to 0.001 Da. Therefore, the weight of each edge should be multiplied by a penalty factor correlated with mass deviation. Suppose e stands for an edge, v and v' for the corresponding vertexes of e, tol for the narrower tolerance and δ for the mass deviation of e, the following functions are used to recalculate the weight of e:

$$\text{Penalty}(\delta, tol) = \begin{cases} -\log(abs(\delta)) \text{ if } \delta \geq tol \\ -\log(abs(tol)) \text{ otherwise} \end{cases}$$
$$\text{Weight}(\delta e, tol) = \text{Penalty}(\delta e, tol) \times (\text{weight}(v) + \text{weight}(v'))$$

In the functions above, weight(v) denotes the weight of the vertex v, that is, the accumulative intensity of all of its corresponding edges.

The curve of the penalty function is shown in Figure 2b.

Immonium and internal ions are also used to adjust the weights of the edges. As a preliminary task, all theoretical internal ions below 500 Da, including both *ay*- and *by*- ion types,[42] are enumerated first. Because the masses of all internal ions could be precalculated and indexed, the internal ions that appear in the spectra could be retrieved with a linear time complexity. Then all edges are tested whether to reassign their weights. If there is an edge e' (to be more accurate, it comes from only the N-terminal fragment ions) starting from the source vertex whose mass distance equals to the mass of an immonium or internal ion present in the spectrum, the weight of e increases with the intensity of the corresponding immonium or internal ion peak $p_{e'}$, multiplied by the penalty function.

Lastly, the weight of an edge is calculated using the following formula, where $Int(p_{e'})$ denotes the intensity of $p_{e'}$:

$$\text{Weight}(\delta_e, \delta_{e'}, tol) = \text{Penalty}(\delta_e, tol) \times (\text{weight}(v) + \text{weight}(v')) + \text{Penalty}(\delta_{e'}, tol) \times Int(p_{e'}))$$

Generating Candidate Peptides. After a spectrum graph is constructed, the state-of-the-art algorithms in the graph theory can be used to generate optimal paths. The score of each path is defined as the sum of the weight of each edge in the path. Like other *de novo* sequencing algorithms, only antisymmetric paths are generated in our algorithm;[13] however, unlike the traditional dynamic programming approach, a depth-first search (DFS) with an efficient pruning strategy, is used. The pruning strategy is described as follows. First, we define *Best_Score* of a vertex as the expected highest score from this vertex to the target vertex. The *Best_Score* of each vertex can be computed using the backward dynamic programming approach:

$$\text{Best Score}(v) = \max\{\text{Best Score}(v') + \text{weight}(e_{v,v'}), \text{if there is an edge from } v \text{ to } v'\}$$

After the computation of the Best Score of each vertex, all optimal paths can be retrieved from the spectrum graph using a DFS algorithm. For instance, if only one path is to be retrieved from a spectrum graph, we assume that only the topmost path with the highest score is to be found. In the DFS algorithm, we define Pre Score of a vertex as the score of the path from the source to itself, that is, the sum of the weight of each edge from the source to this vertex. If a path is found from the source vertex to the target vertex and its score is the highest of all the paths found up to this point, then this path as well as its score is recorded. For an arbitrary vertex v which follows vertex w on the spectrum graph, we require that the sum of Pre Score(v) and Best Score(v) must be greater than the recorded highest score, so that a path with a potentially higher score could be found; otherwise the algorithm will trace back to w, and another vertex which follows w on the spectrum graph and to which there is an edge extending from w will be considered. Similarly, the second, third, fourth. .. and nth best path can be retrieved from each spectrum graph. In our study the DFS algorithm together with the pruning strategy is more time-efficient than other approaches we investigated. This approach also fits other additive scoring schemes and could be easily extended to finding top-k paths.

Finally, peptide candidates are generated using all of the retrieved optimal paths. As mentioned above, there may be some edges that are marked by different combinations of amino acids with the same mass. In this step, all possible peptides in the optimal paths are enumerated and then matched with the spectrum.

Scoring Candidate Peptides. Designing a good scoring scheme is of prime importance in both database search and *de novo* peptide sequencing. In conventional approaches, dot product and probability-based approaches are most widely used. However, candidate peptide sequences interpreted from the same spectrum always bear a high degree of similarity with each other; hence, the aforementioned approaches may be unable to distinguish them if only a few conditions are considered. This situation is especially troublesome in *de novo* peptide sequencing, for the candidate peptides are from the "theoretical database" that contains all possible sequences. Additional information is needed to effectively discriminate slight differences between peptide sequences. Fortunately, high mass accuracy and ample information of immonium and internal ions in HCD spectra provide such help. In this section, we choose several key features to construct a PSM scoring scheme. First, it is known that the percentage of the matched high-intensity peaks properly reflects the quality of the PSM. This feature is also used in machine learning,[43,44] but in our algorithm we consider more ion types including internal ions and some backbone-derived ions with neutral losses. Assuming that all peaks p_1, p_2, p_3.. ., p_m in a spectrum S are sorted by their intensities from the strongest to the weakest and the weakest peak matching a fragment ion of peptide P is p_k, $1 \leq k \leq m$, we calculate S_H as below:

$$\text{match}(p) = \begin{cases} 1 \text{ if } p \text{ matches with a fragment ion} \\ 0 \text{ otherwise} \end{cases}$$
$$S_H(S, P) = \frac{1}{k}\sum_{i=1}^{k}\frac{1}{i}\sum_{j=1}^{i}\text{match}(pj)$$

Second, the cleavage information is also utilized in the scoring step. In general, a peptide tends to be a reliable candidate if it has many fragmentation sites supported by the spectrum and if it has a long consecutive sequence tag. Let c_f denote the total count of observed cleavage signals of P in the spectrum S and t_f denote the length of the longest sequence tags, then we calculate S_F to evaluate the fragmentation of the peptide P in the spectrum S:

$$S_F(S, P) = \frac{\sqrt{cf \cdot tf}}{\text{length}(P) - 1}$$

Third, mass deviation is also useful to differentiate two peptide sequences that resemble each other. Suppose that T is the specified maximum mass deviation and md is the function for computing the mass deviation between an ob-

served peak p and its corresponding ion, we compute the value of S_{MD} as follows:

$$S_{MD}(S,P) = \left(T - \sqrt{\frac{\sum_{K\text{most intense peaks } p} md^2(p)}{K}}\right) / T$$

Finally, the *C-Score* of a peptide-spectrum match is defined as the geometric mean of the S_H, S_F and S_{MD}, multiplied by S_O, which is the normalized score of the path from which the peptide is generated:

$$C\text{−Score}(S,P) = \sqrt[3]{S_H(S,P) \times S_F(S,P) \times S_{MD}(S,P)} \times S_O$$

3. Experiments and Results

Materials and MS/MS Data. Two kinds of biological samples were used, one simple and the other complex. The simple one was a mixture of Bio-Rad unstained low- and high-range protein standards (called 8-protein STD) consisting of Myosin, Glycogen phosphorylase, Serum albumin, Beta-galactosidase, Carbonic anhydrase, Trypsin inhibitor, Ovalbumin, and Lysozyme. This 8-protein mixture was digested with trypsin and analyzed by LC−MS/MS on a LTQ-Orbitrap mass spectrometer equipped with ETD (Thermo-Fisher Scientific). A C18 reverse-phase column (100 μm ID and 8 cm in length) connected to an Agilent 1200 quaternary HPLC was used to separate peptides. MS/MS spectra were acquired in a data-dependent acquisition mode. Full scans were acquired in the Orbitrap and the two most intense precursor ions from each full scan were isolated to generate five MS/MS spectra for each. The five MS/MS events are low-mass HCD (mass range 50−2000), HCD (mass range 100−2000), CID detected in LTQ, ETD detected in orbitrap, and ETD detected in LTQ. Only the HCD data were used in *de novo* analysis. Two HCD MS/MS spectra are necessary to cover the mass range from 50 to 2000 because low-mass HCD spectra (50−2000 m/z) are almost devoid of fragment ions above 1000 m/z. All tandem mass spectra were extracted by Xcalibur 2.0.7 as RAW files. The .ms2 file containing MS/MS spectra was generated by RawXtract 1.9.3. Then different types of MS/MS spectra were separated by an in-house software tool MS2Extractor. Each pair of HCD spectra were integrated into a single spectrum by gathering all the peaks in the two spectra and merging peaks with identical m/z values within a tolerance window of ±0.01 Da. The intensity of the peaks that are merged together are summed up and given to the resultant peak.

The other sample was a tryptic digest of a whole-cell lysate of *C. elegans*. This extremely complex mixture (40 μg) was analyzed on a LTQ-Orbitrap mass spectrometer using a 12-step MudPIT method similar to what had been described before.[45] Briefly, a 250 μm (ID) × 2 cm (length) desalting column was packed with 5 μm, 125 anstrong Aqua C18 resin (Phenomenex). The analytical reverse phase column was 100 μm (ID) × 9 cm (length) with a pulled tip, packed with 3 μm, 125 anstrong Aqua C18 resin (Phenomenex). Between the desalting column and the analytical column is a strong cation exchange column (SCX), 250 μm (ID) by 2 cm (length), containing 5 μm, 120 anstrong Partisphere SCX material (Whatman). The salt pulses of these 9-step MudPIT experiments were set at 0, 5, 10, 15, 20, 30, 40, 50, 60, 70, 80, and 100, expressed as the percentage of buffer C. In the MudPIT experiment, the five most intense precursor ions from each full scan were isolated to generate three MS/MS spectra for each: low-mass HCD (mass range 50−2000), HCD (mass range 100−2000), and CID (detected in LTQ). Only the HCD data were used in *de novo* analysis. Three of the 12 RAW files were used for the performance test of pNovo and the remaining 9 files served as a training set to determine the ion types present in HCD spectra.

Database Search and Data Sets. Two database search software tools, Mascot v2.1.03 and pFind v2.1, were used in this paper to generate the test data sets and compared with pNovo. The protein sequence database and parameters used in database search are listed in the Supporting Information.

Three test data sets and one training set were used in this work. The first two, STD-951 and STD-208, were from the analysis of the 8-protein sample and the rest were from the MudPIT analysis of the *C. elegans* lysate.

STD-951: A total of 951 spectra were extracted directly from the original RAW file without any special filtering. Two consecutive HCD spectra from the same precursor were merged together.

STD-208: This data set contains 208 HCD spectra for which pFind and Mascot agree completely on their sequence identities under the 1% FDR control at the spectrum level, of which 197 are doubly charged peptides, and the rest triply charged. In short, STD-951 is the original data generated from the "8-protein STD" sample, and STD-208 is a subset of STD-951 with reliable database identification results.

WORM-767: This data set contains 767 HCD spectra. A total of 1214 HCD spectra were identified from the three RAW files by Mascot and pFind with identical results under 1% FDR. After removing duplicate peptides, we retained 767 HCD spectra. Out of the 767 peptides, 58 were triply charged, and the rest doubly charged.

Training set: As mentioned before, the remaining nine RAW files out of a total of 12 from the worm sample were used for training. This data set contained 4718 spectra that were identified by both Mascot and pFind with identical results under 1% FDR.

***De novo* Peptide Sequencing and Protein Identification.** The pNovo algorithm was tested on three data sets described above. Of the 20 standard amino acids and their combinations, only Leu and Ile are considered as the same. In preprocessing, 150 most intense peaks in each spectrum are kept for later steps. To compute *C-Score*, the original spectrum with all the peaks, rather than the top 150, is used because some low intensity internal ions could provide extra information to distinguish between similar sequences. When the algorithm constructs the spectrum graph, it uses a tolerance window of ±0.01 Da to determine whether or not to connect an edge between two vertexes. However, a narrower tolerance window of ±0.001 Da is used in the penalty function. No more than 100 paths are generated from the spectrum graph to balance the speed and the accuracy of the algorithm.

We chose the Levenshtein distance[46] with two extensions[33] to measure the similarity between the answer (from the database search) and the pNovo sequencing result. Similarity ratio is defined by the following formula (the function *LD* denotes the extended Levenshtein distance between two peptide sequences):

Table 2. Comparison of *de novo* peptide sequencing algorithms on STD-208

				percentage of identifications with a subsequence of at least x amino acids long							
algorithms	correct peptides	correct aa	predicted aa	$x = 3$	$x = 4$	$x = 5$	$x = 6$	$x = 7$	$x = 8$	$x = 9$	$x = 10$
pNovo	**181**	**1837** (96.2%)	1910	93.8	**90.9**	**90.4**	**88.0**	**63.9**	**54.8**	**51.9**	**45.2**
PepNovo	117	1454 **(98.4%)**	1478	75.0	59.1	45.2	32.7	16.8	13.0	7.2	3.9
PEAKS	147	1784 (92.5%)	**1928**	97.5	80.6	50.3	36.3	21.4	14.9	7.5	4.5

Table 3. Comparison of *de novo* Peptide Sequencing Algorithms on WORM-767

				percentage of identifications with a subsequence of at least x amino acids long							
algorithms	correct peptides	correct aa	predicted aa	$x = 3$	$x = 4$	$x = 5$	$x = 6$	$x = 7$	$x = 8$	$x = 9$	$x = 10$
pNovo	**612**	8190 (94.9%)	8633	99.1	98.4	97.4	**95.8**	92.7	86.0	74.2	60.2
PepNovo	320	6695 **(95.7%)**	6993	98.7	97.1	94.8	91.5	88.3	68.2	46.0	30.1
PEAKS	538	8136 (93.5%)	**8699**	99.6	98.8	97.9	95.4	91.1	82.3	69.8	55.8

$$\text{Similarity ratio}(a, b) = \begin{cases} \dfrac{\text{length}(a) - LD(a, b)}{\text{length}(a)} & \text{if length}(a) > LD(a, b) \\ 0 \text{ otherwise} \end{cases}$$

For instance, if two sequences are identical, the Similarity ratio is 1.

Following *de novo* sequencing of peptides, a mapping algorithm is used to map the peptides to proteins. For example, if the top 10 candidate peptides are kept for each spectrum, then a set of k-length subsequences are generated based on each peptide sequence, where k is an empirical parameter usually set to 3 or 4. The whole set of these k-length strings form a dictionary. Then Aho-Corasick algorithm[47] is used to find out whether the proteins in the database can each find a match of one or more sequences in the dictionary in a linear time complexity. If a match is retrieved, the short sequence in the original match is extended to verify whether the original peptide sequence can be mapped to a protein with a considerable Similarity ratio (for the experiments in this paper, the threshold of Similarity ratio is set as 0.85), which is measured by the extended Levenshtein distance mentioned above.

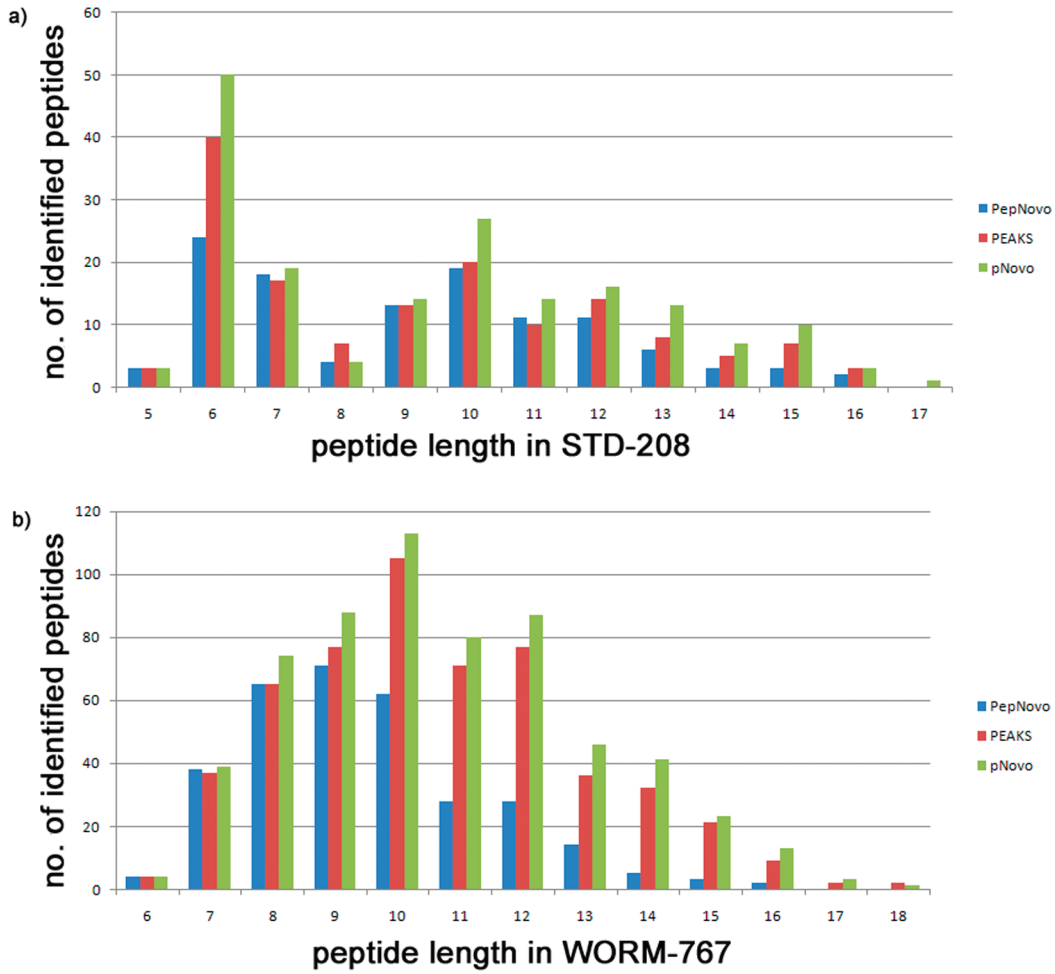

Figure 3. Peptide length distribution of the correct full-length sequences generated by different *de novo* sequencing algorithms on (a) STD-208 and (b) WORM-767.

Performance on STD-208 and WORM-767. *De novo* peptide sequencing results on STD-208 and WORM-767 are shown in Tables 2 and 3, respectively. All three algorithms, pNovo, PEAKS (PeaksStudio5.1) and PepNovo (latest release on 2009.10.29), achieved high accuracy in predicting amino acid residues, which strongly suggests that HCD is well suited for *de novo* sequencing. Because HCD spectra are of high mass accuracy and high resolution, the complexity of the spectrum graph is sharply decreased, and optimal paths can be retrieved more precisely. In both data sets, PepNovo predicted the least number of amino acid residues, although its accuracy is the highest. Both PEAKS and pNovo made longer-peptide predictions, and pNovo results were more precise. Compared with PEAKS and PepNovo, pNovo achieved superior results with a larger number of correct full-length sequences (87.0% of the total spectra in STD-208 and 79.8% in WORM-767). The pNovo scoring scheme takes mass accuracy into consideration, so low intensity fragment ions with high mass accuracy also contribute to the score with a proper weight. Because of this, even the N- or C- terminus of a peptide sequence, which has been difficult to predict, can be determined efficiently. pNovo also takes advantage of internal fragment ions to enhance the reliability of predicted sequences. With respect to the average length of correct subsequences, pNovo performed the best, making accurate predictions for more than half of the sequence tags containing as many as eight amino acids (e.g., 54.8% on STD-208), which is much better than PEAKS (14.9%) or PepNovo (13.0%). In database search, sequence tags can be used to filter candidate sequences in the database, and longer tags do so much more efficiently than shorter ones (usually of length 3),[40] Thus, long and accurate subsequences generated by pNovo should be useful in tag-based database search as well.

As shown in Figure 3, pNovo has achieved a higher accuracy on longer peptides compared with PepNovo and PEAKS on both data sets. In general, both pNovo and PEAKS can efficiently sequence peptides of varying lengths, and of these two, pNovo generates a larger number of correct full-length sequences. The performance of PepNovo falls as the length of a peptide increases, especially for peptides longer than 9 amino acids on WORM-767.

Compared with other *de novo* algorithms, peptides of charge states higher than 2+ can also be sequenced efficiently by pNovo. In WORM-767, 58 different peptides of 3+ charge were retained under 1% FDR, and their spectra were sequenced by pNovo with an overall accuracy of 0.89. With respect to these peptides, the pNovo results are 100% correct for 34 of them (~59%). Only 11 full-length sequences (~19.0%) extracted by PepNovo are 100% correct, although the accuracy of its shorter-sequence predictions is still as high as 0.92. PEAKS generated 27 correct sequences (~46.6%) from these +3 spectra with an overall accuracy of 0.87.

According to the performance of pNovo, PEAKS, and PepNovo on both doubly and triply charged spectra, it is evident that all three algorithms can achieve substantially high prediction accuracy. PepNovo tends to produce shorter sequence tags, and pNovo generates more full-length sequences than either PepNovo or PEAKS.

Mass Accuracy. As depicted in Figure 2a, although ±0.01 Da is used as the normal mass tolerance width, ~80% of the mass deviations are within a much narrower tolerance window of ±0.001 Da. Figure 4a shows the relationship between mass tolerance and identification results. As the tolerance window increases from ±0.01 Da to ±0.5 Da, the number of identifications decreases by 78.5%. On the other hand, if the tolerance window is narrowed to ±0.001, only 63 correct sequences are generated, and the vast majority of the spectra do not give any answer due to a lack of backbone cleavage information. Therefore, we conclude that fragment ion mass tolerance greatly affects the performance of the *de novo* sequencing algorithm.

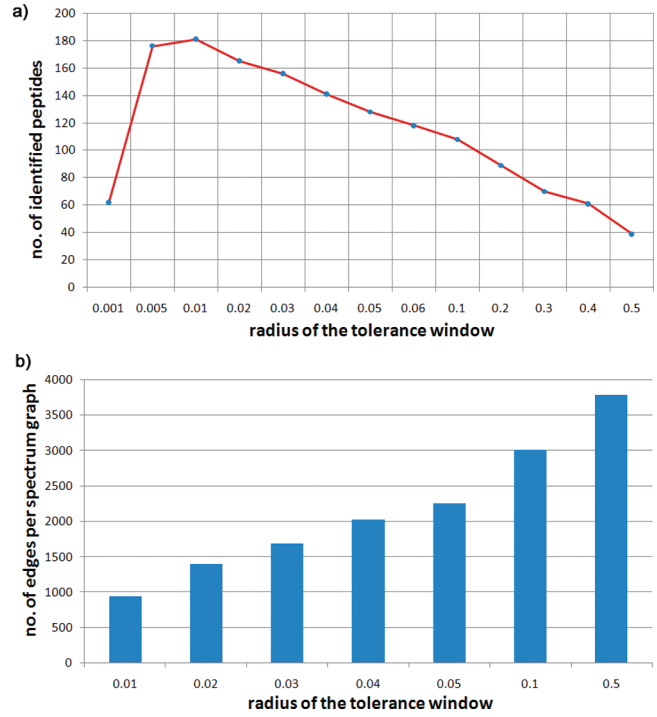

Figure 4. (a) Above the threshold of ±0.01 Da, the number of correct sequences arrived at by *de novo* sequencing decreases as fragment ion mass tolerance increases. The tolerance window for penalty is set as ±0.001 Da. (b) Number of edges in an average spectrum graph increases as fragment ion mass tolerance increases. Edges of up to two amino acids are used.

Precision MS/MS data sharply reduce the complexity of spectrum graph, thereby increasing the efficiency of pNovo. As shown in Figure 4b, if the tolerance window is opened up to ±0.5 Da, equivalent to what is used for unit-resolution MS/MS data such as those generated in an ion trap instrument, the number of edges taken into consideration by the algorithm is nearly four times as many as at ±0.01 Da and the speed falls by ~90%.

Separation of Correct and Incorrect Identifications. We computed RnkScr and PnvScr (PepNovo scores), the ALC score (used by PEAKS), and *C*-score (designed in this work for pNovo) for correct and incorrect identifications made by PepNovo, PEAKS, and pNovo, respectively, using STD-208 as a test set. As shown in Figure 5, *C*-score achieved the best separation of correct and incorrect identifications. The average *C*-scores for correct and incorrect identifications are 0.70 and 0.39, respectively. The *C*-scores of correct and incorrect identifications overlap only slightly in the region between 0.40 and 0.65. No single spectrum is sequenced correctly with a *C*-score less than 0.40 or incorrectly with a *C*-score greater than 0.70. The Kolmogorov−Smirnov (KS) distance of the two *C*-score distributions is 0.864, which is much larger than that of PEAKS-ALC (0.359), PepNovo-RnkScr (0.387) and PepNovo-PnvScr (0.357). Furthermore, the average Similarity ratio of the incorrect identifications with a *C*-score greater than 0.6 is 0.87, while for

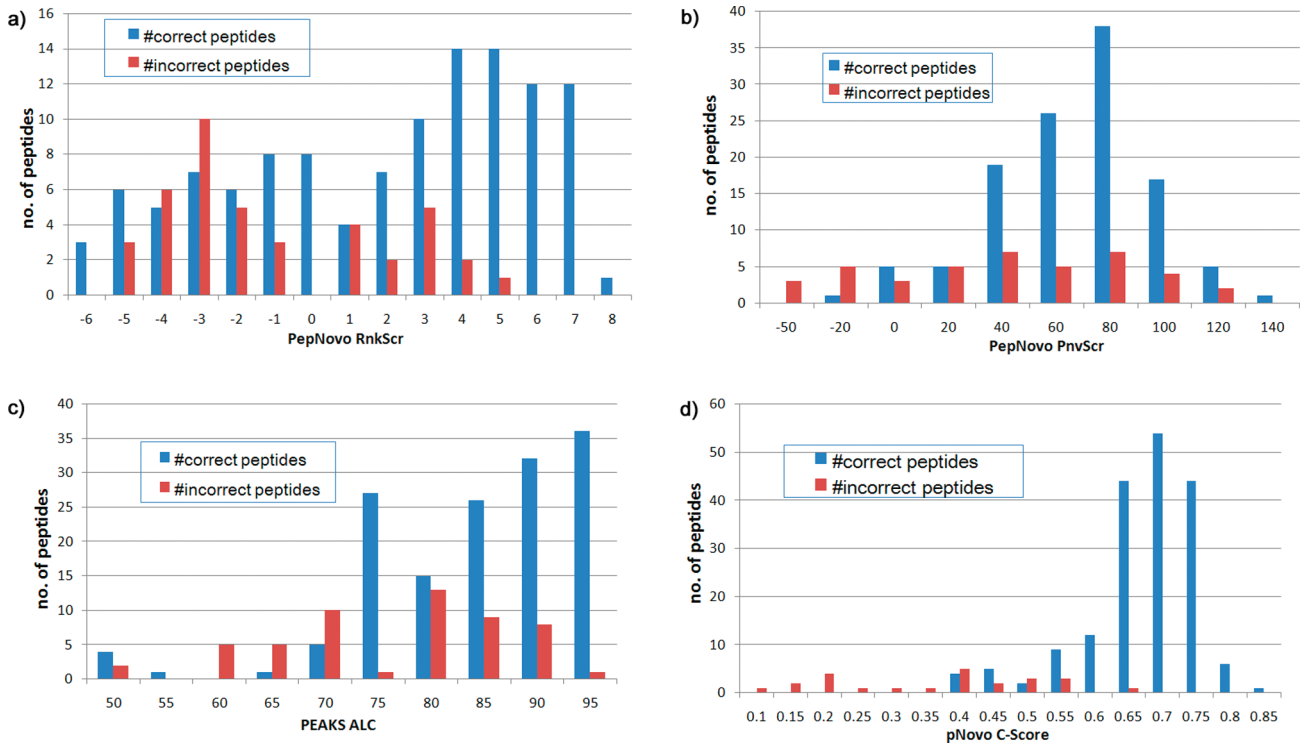

Figure 5. Distribution of scores given by different algorithms for all the identification results of STD-208.

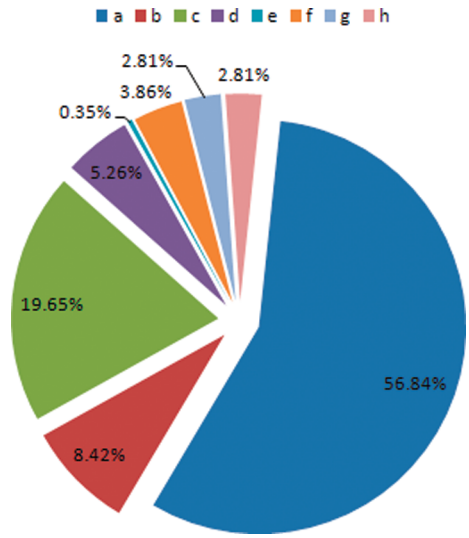

Figure 6. Breakdown of the spectra from STD-951 with accepted pNovo sequencing results. To each spectrum retained and represented in this chart, at least one peptide is given by pNovo with a *C-Score* greater than 0.6. (a) Spectra in STD-208, i.e. these spectra were also identified by Mascot and pFind using a 1% FDR cutoff. (b) Spectra for which pNovo identifications are less than seven-amino acid long. (c) These spectra were also identified by Mascot and pFind but were filtered out at 1% FDR cutoff. However, these identifications were likely correct as pNovo gave the same results. (d) From these spectra Mascot and pFind identified peptide sequences similar to what pNovo found with an average similarity ratio of 0.809. (e) Spectra from which pFind, but not Mascot, identified peptides and they were in complete agreement with the pNovo results. (f) Spectra from which the pNovo results were indicative of PTMs or amino acid mutations. (g) Spectra whose pNovo results contained partial sequence errors and perhaps PTMs or amino acids mutations. (h) For these spectra, the pNovo results were most likely wrong due to low spectral quality or other reasons.

the incorrect identifications with a *C*-score less than 0.6 it is 0.53. To sum it up, a proper threshold of *C*-score can effectively differentiate between the correct and incorrect identifications and to control the false positive rate.

Merging HCD Spectral Pairs. As mentioned above, a pair of HCD spectra (low-mass HCD and HCD) are acquired for each precursor, with the mass range of one starting from 50 m/z and the other from 100 m/z. For convenience, we call the former type low-mass HCD, and the latter normal HCD. A low-mass HCD spectrum lacks strong peaks above 500 m/z, but it is complementary to its cognate normal HCD spectrum, as it contains fragment ions from 50 to 100 m/z. Some amino acids such as Pro and Val tend to produce strong immonium ions that fall within this range. Besides, ions below 500 m/z are on average 1.5 times more intense in low-mass HCD than in normal HCD. As expected, the best result was obtained from the merged spectra which yielded 181 correct full-length peptide sequences, whereas the low-mass spectra yielded 122 and the normal spectra 176. Similarly, more accurate subsequences or sequence tags were found using the merged spectra. For example, the merged spectra yielded ~9.3% more correct 10-aa subsequences than the normal spectra. These results show that merging low-mass and normal HCD spectra improves *de novo* sequencing. We expect that the low-mass spectra will be more useful if the characteristics of low-mass internal fragment ions are investigated further.

Algorithmic Performance on CID Spectra. For analysis of the unit resolution ion trap spectra using pNovo, the penalty factor for mass accuracy and the use of internal ions are removed, and 100 peaks are kept in each spectrum. Although pNovo generated accurate subsequences up to 16-aa long, the average accuracy falls to 52.9%, which is lower by 6.7% than the average accuracy of PepNovo on the 8-protein STD data. Without the consideration of mass accuracy and internal ions, the scoring scheme of pNovo is simpler than PepNovo. This

Table 4. Nonredundant Identifications with Modifications or Mutations

peptides reported by pNovo	peptides in the database	comments
QPDIFKDIVNMIMHH**Q**R	QPDLFKDIVNMLMHH**D**R	D→Q
APNDFNIKDF**D**VGGYIQAVIQR	APNDFNLKDF**N**VGGYIQAVIQR	N→D
VI**T**SSAR	VL**A**SSAR	A→T
IIFDGVNSAFHIW**T**NGR	IIFDGVNSAFHLW**C**NGR	C→T
IE**D**GIIII**D**GKIPIIR	IE**N**GLLLL**N**GKPLLIR	N→D, N→D
YGDFGTAAQQPDGIAVVGVFIK**K**	YGDFGTAAQQPDGIAVVGVFIK**V**	V→K
DTDGSTDYGIIQI**D**SR	**N**TDGSTDYGILQI**N**SR	N→D

Table 5. Comparison of Protein Identification Results by *De novo* Sequencing (pNovo) vs Database Search (pFind and Mascot)

		database search								
		pFind			Mascot			*de novo* sequencing using pNovo		
protein	source of organism	#spec	#pep	% cov	#spec	#pep	% cov	#spec	#pep	% cov
Myosin	Rabbit	21	21	13.6	21	21	**14.2**	18	18	11.8
Glycogen phosphorylase	Rabbit	58	33	**45.2**	51	31	44.2	45	29	41.2
Serum albumin	Bovine	44	17	**33.3**	43	17	**33.3**	45	17	**33.3**
Beta-galactosidase	*E. coli*	26	16	**26.1**	18	16	**26.1**	17	11	15.3
Carbonic anhydrase	Bovine	28	9	**43.5**	27	9	**43.5**	19	9	42.7
Trypsin inhibitor	Soy bean	31	7	31.0	29	7	31.0	19	7	**34.7**
Ovalbumin	Chicken	12	7	**29.3**	11	7	**29.3**	7	6	22.8
Lysozyme	Chicken	30	6	39.5	28	6	39.5	34	8	**55.1**

again shows that the characteristics of HCD spectra, for example, high mass accuracy and abundant information of internal and immonium ions, are advantageous for *de novo* sequencing.

Algorithmic Performance on STD-951. A total of 285 identifications with *C-Scores* greater than 0.6 were kept and further analyzed (working on STD-208 at this threshold, 89% correct identifications were kept and the percentage of false identifications was ~0.6%). Of these, 241 are doubly charged peptides and the rest are triply charged.

As shown in Figure 6, identifications that also appear in STD-208 are the most dominant fraction (fraction a). These are surely correct identifications. The second largest fraction (fraction c in Figure 6) consists of sequence identifications that are probably right. This subset of spectra were identified by Mascot or pFind but were filtered out by the 1% FDR cutoff. However, these identifications were most likely correct as pNovo gave the same results as database search. Fraction d contains 15 pNovo identifications that are similar to both pFind and Mascot results but with small differences (the average similarity ratio is 0.809). From the spectra in Fraction f, pNovo identified peptides with modifications and mutations (see Table 4 for a complete list). For example, deamidation of Asn or Gln is a common post-translational modification and can also happen during sample preparation. It can be identified by database search at the cost of search time. In contrast, *de novo* peptide sequencing can handle post-translational modifications like deamidation and amino acid mutations almost without extra time cost.

Table 5 shows that the sequence coverage of the eight standard proteins obtained by pFind, Mascot, or pNovo are at the same level except for Beta-galactosidase and Lysozyme. The sequence coverage of beta-galactosidase by pNovo (15.3%) is lower than that by pFind (26.1%) or Mascot (26.1%), whereas for Lysozyme it is the other way around (55.1% by pNovo and 39.5% by either pFind or Mascot). The better performance by pNovo on Lysozyme is due to the PTMs and amino acid mutations and can be explained by Table 4. In short, pNovo sequencing results are comparable with database search results, and pNovo has the advantage of discovering PTMs and amino acid mutations. We believe that for HCD spectra the *de novo* peptide sequencing approach is of great potential and will be very useful in proteomic research.

4. Discussions

There are several persistent obstacles in *de novo* sequencing. First of all, *de novo* peptide sequencing is hardly possible if fragment ion series contain too many gaps or if a gap is too big.[27] For this reason, CID spectra are especially troublesome because of loss of ions in the low mass region (often referred to as the "1/3 cutoff ").[41] Second, most algorithms cannot handle spectra of highly charged peptides.[40] Since highly charged peptides tend to be the long ones, they often give rise to a large number of fragments and these fragments can assume more than one charge state. If the charge states of the fragment ions cannot be determined, the resulting spectrum graph may become so complex that it overwhelms the algorithm. Lastly, top-ranked candidate peptides obtained by *de novo* sequencing are often very similar to each other, and it is extremely difficult to evaluate the candidates and determine which one is the most likely answer. Although a variety of validation models have been proposed and used in the database search engines,[48–50] only the one proposed by Kim et al. is suitable for *de novo* sequencing algorithms in a limited sense, for it slows down as the length of a peptide increases and cannot be readily applied to nonadditive scoring models.[51,52] Consequently, validation of *de novo* sequencing results mainly relies on database search results or manual interpretation, and this limits the application of *de novo* sequencing in proteomics.

Here we describe a *de novo* peptide sequencing algorithm called pNovo. pNovo is designed for HCD spectra, which as we have shown here have favorable features to help overcome the obstacles in *de novo* peptide sequencing. The features include the following: (1) a relatively wide mass range from 50 to 2000 m/z without low-mass cutoff, (2) more complete ion series than CID and ETD, (3) high resolution and high mass accuracy which translate into accurate determination of fragment ion mass and charge, and simplified spectrum graphs, (4) the presence of many internal and immonium ions which can be used to distinguish between sequences with minor

differences. Consideration of these features has made pNovo a successful algorithm. For example, as shown in the Table 5, pNovo made a similar number of peptide identifications as database search. This is a big improvement over previous *de novo* sequencing efforts. Below we discuss the next steps for *de novo* sequencing.

First, post-translational modifications can be analyzed by *de novo* sequencing from HCD spectra. For example, the mass difference of Phe (147.068) and oxidized Met (147.035) is only 0.033, but this difference is big enough to tell them apart in HCD spectra. Also, the presence or absence of the immonium ion of Phe or Met can lend further proof. PTM analysis by database search is time-consuming, especially if multiple PTMs are considered simultaneously because it will cause a serious combinatorial explosion of search space. It remains to be seen if *de novo* sequencing can find a way to analyze PTMs more efficiently, although an obvious advantage of *de novo* sequencing is that it can identify unexpected or unknown PTMs. Second, HCD spectra of +4 or higher charge-state peptides can be interpreted better. In this paper, only doubly and triply charged spectra are used in the experiments. However, longer peptides with higher charge state can also be obtained if more missed cleavages are considered or other enzymes such as Lys-N and Lys-C are used.[53] The mass accuracy of HCD spectra makes it possible to distinguish +4 or even +5 peaks based on isotopic peak clusters. The main problem may be how to determine the middle region of a peptide sequence, of which less information is expressed in the spectra. This problem may be alleviated by internal fragment ions observed in HCD spectra. Third, sequencing novel proteins remains a challenge. Most of the previous attempts were based on different enzymatic digestions to generate overlapping peptides.[54,55] We speculate that HCD coupled with multiple enzymatic digestions and other fragmentation methods such as ETD may be fruitful in automated protein sequencing.

Acknowledgment. This work was supported by the National Key Basic Research & Development Program (973) of China under Grant Nos. 2010CB912701 and 2002CB713807, the National High Technology Research and Development Program (863) of China under Grant Nos. 2007AA02Z315 and 2008AA02Z309, the CAS Knowledge Innovation Program under Grant No. KGGX1-YW-13 and the National Natural Science Foundation of China under Grant No. 30900262. We thank Li-Yun Xiu and Kun Zhang for the support of MS/MS spectra labeling software pLabel, and Ding Ye, Wen-Ping Wang, Yan-Jie Wu, and Chen Zhou for valuable discussions.

Supporting Information Available: Supplementary Figure S1: Offset frequency functions for prefix/suffix of HCD spectra mentioned in the algorithm section. Supplementary Table S1: Probabilities of common ion types appeared in HCD data. Supplementary Table S2: Proteins used in the experiments. Supplementary Table S3: Parameters of database search used in the experiments. This material is available free of charge via the Internet at http://pubs.acs.org.

References

(1) Apweiler, R.; Bairoch, A.; Wu, C. H. Protein sequence databases. *Curr. Opin. Chem. Biol.* **2004**, *8* (1), 76–80.
(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), 3551–67.
(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), 976–89.
(4) Craig, R.; Beavis, R. C. A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2310–6.
(5) Craig, R.; Beavis, R. C. TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* **2004**, *20* (9), 1466–7.
(6) 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), 1948–54.
(7) Li, D.; Fu, Y.; Sun, R.; Ling, C. X.; Wei, Y.; Zhou, H.; Zeng, R.; Yang, Q.; He, S.; Gao, W. pFind: a novel database-searching software system for automated peptide and protein identification via tandem mass spectrometry. *Bioinformatics* **2005**, *21* (13), 3049–50.
(8) Wang, L. H.; Li, D. Q.; Fu, Y.; Wang, H. P.; Zhang, J. F.; Yuan, Z. F.; Sun, R. X.; Zeng, R.; He, S. M.; Gao, W. pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2007**, *21* (18), 2985–91.
(9) Colinge, J.; Masselot, A.; Cusin, I.; Mahe, E.; Niknejad, A.; Argoud-Puy, G.; Reffas, S.; Bederr, N.; Gleizes, A.; Rey, P. A.; Bougueleret, L. High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics. *Proteomics* **2004**, *4* (7), 1977–84.
(10) Colinge, J.; Masselot, A.; Giron, M.; Dessingy, T.; Magnin, J. OLAV: towards high-throughput tandem mass spectrometry data identification. *Proteomics* **2003**, *3* (8), 1454–63.
(11) 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), 958–64.
(12) Lu, B.; Chen, T. Algorithms for de novo peptide sequencing via tandem mass spectrometry. *Biosilico* **2004**, *2* (2), 85–90.
(13) 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), 327–42.
(14) Frank, A.; Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **2005**, *77* (4), 964–73.
(15) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun. Mass Spectrom.* **2003**, *17* (20), 2337–42.
(16) 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), 1067–75.
(17) 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), 1768–74.
(18) 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), 4870–8.
(19) 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), 1867–78.
(20) 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), 1694–9.
(21) 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), 3033–8.
(22) 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), 7265–73.
(23) 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), 325–37.

(24) Lu, B.; Chen, T. A suboptimal algorithm for de novo peptide sequencing via tandem mass spectrometry. *J. Comput. Biol.* **2003**, *10* (1), 1–12.
(25) Ning, K.; Ye, N.; Leong, H. W. On preprocessing and antisymmetry in de novo peptide sequencing: improving efficiency and accuracy. *J. Bioinform. Comput. Biol.* **2008**, *6* (3), 467–92.
(26) 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), 114–23.
(27) Savitski, M. M.; Nielsen, M. L.; Kjeldsen, F.; Zubarev, R. A. Proteomics-grade de novo sequencing approach. *J. Proteome Res.* **2005**, *4* (6), 2348–54.
(28) 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), 703–14.
(29) 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), 753–61.
(30) 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), 10313–7.
(31) Datta, R.; Bern, M. Spectrum fusion: using multiple mass spectra for de novo Peptide sequencing. *J. Comput. Biol.* **2009**, *16* (8), 1169–82.
(32) Bringans, S.; Kendrick, T. S.; Lui, J.; Lipscombe, R. A comparative study of the accuracy of several de novo sequencing software packages for datasets derived by matrix-assisted laser desorption/ionisation and electrospray. *Rapid Commun. Mass Spectrom.* **2008**, *22* (21), 3450–4.
(33) Pevtsov, S.; Fedulova, I.; Mirzaei, H.; Buck, C.; Zhang, X. Performance evaluation of existing de novo sequencing algorithms. *J. Proteome Res.* **2006**, *5* (11), 3018–28.
(34) Pitzer, E.; Masselot, A.; Colinge, J. Assessing peptide de novo sequencing algorithms performance on large and diverse data sets. *Proteomics* **2007**, *7* (17), 3051–4.
(35) Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **1994**, *66* (24), 4390–9.
(36) Sunyaev, S.; Liska, A. J.; Golod, A.; Shevchenko, A.; Shevchenko, A. MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. *Anal. Chem.* **2003**, *75* (6), 1307–15.
(37) Tabb, D. L.; Saraf, A.; Yates, J. R. 3rd, GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal. Chem.* **2003**, *75* (23), 6415–21.
(38) 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), 4626–39.
(39) Shilov, I. V.; Seymour, S. L.; Patel, A. A.; Loboda, A.; Tang, W. H.; Keating, S. P.; Hunter, C. L.; Nuwaysir, L. M.; Schaeffer, D. A. The Paragon Algorithm, a next generation search engine that uses sequence temperature values and feature probabilities to identify peptides from tandem mass spectra. *Mol. Cell. Proteomics* **2007**, *6* (9), 1638–55.
(40) Kim, S.; Gupta, N.; Bandeira, N.; Pevzner, P. A. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteomics* **2009**, *8*, 53–69.
(41) Olsen, J. V.; Macek, B.; Lange, O.; Makarov, A.; Horning, S.; Mann, M. Higher-energy C-trap dissociation for peptide modification analysis. *Nat. Methods* **2007**, *4* (9), 709–12.
(42) Falick, A. M.; Hines, W. M.; Medzihradszky, K. F.; Baldwin, M. A.; Gibson, B. W. Low-mass ions produced from peptides by high-energy collision-induced dissociation in tandem mass-spectrometry. *J. Am. Soc. Mass Spectrom.* **1993**, *4* (11), 882–93.
(43) Fridman, T.; Razumovskaya, J.; Verberkmoes, N.; Hurst, G.; Protopopescu, V.; Xu, Y. The probability distribution for a random match between an experimental-theoretical spectral pair in tandem mass spectrometry. *J. Bioinform. Comput. Biol.* **2005**, *3* (2), 455–76.
(44) Zhang, J.; Ma, J.; Dou, L.; Wu, S.; Qian, X.; Xie, H.; Zhu, Y.; He, F. Bayesian nonparametric model for the validation of peptide identification in shotgun proteomics. *Mol. Cell. Proteomics* **2009**, *8* (3), 547–57.
(45) 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.
(46) Levenshtein, V. Binary codes capable of correcting deletions, insertions, and reversals. *Sov. Phys.-Dokl* **1966**, *10* (8), 707–710.
(47) Aho, A.; Corasick, M. Efficient string matching: an aid to bibliographic search. *Commun. ACM* **1975**, *18* (6), 333–40.
(48) Keller, A.; Nesvizhskii, A. I.; Kolker, E.; Aebersold, R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal. Chem.* **2002**, *74* (20), 5383–92.
(49) Fenyo, D.; Beavis, R. C. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Anal. Chem.* **2003**, *75* (4), 768–74.
(50) Elias, J. E.; Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **2007**, *4* (3), 207–14.
(51) Kim, S.; Gupta, N.; Pevzner, P. A. Spectral probabilities and generating functions of tandem mass spectra: a strike against decoy databases. *J. Proteome Res.* **2008**, *7* (8), 3354–63.
(52) Kim, S.; Bandeira, N.; Pevzner, P. A. Spectral profiles, a novel representation of tandem mass spectra and their applications for de novo peptide sequencing and identification. *Mol. Cell. Proteomics* **2009**, *8* (6), 1391–400.
(53) 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), 650–60.
(54) 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), 1123–34.
(55) 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), 2174–80.

PR100182K

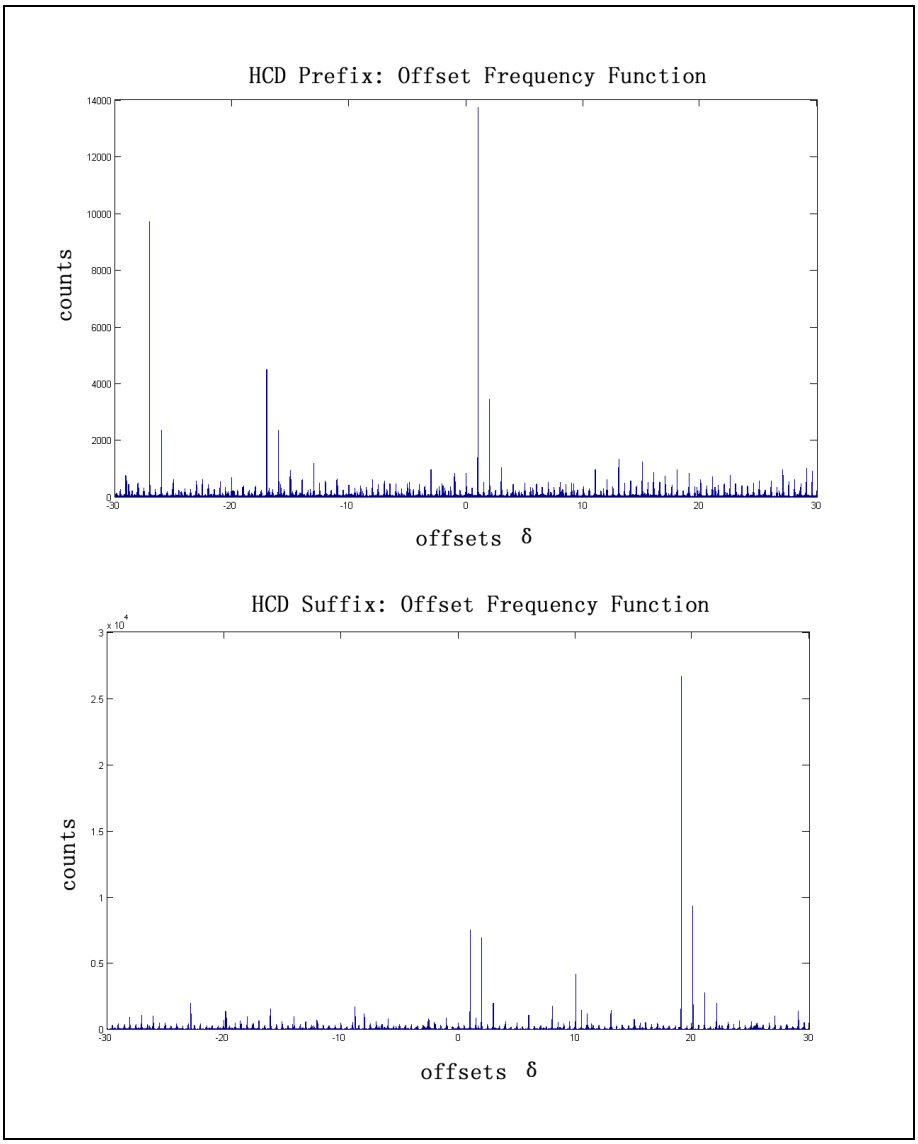

Figure S1. Plots of Offset Frequency Functions for prefix/suffix of HCD spectra mentioned in the algorithm section. In the plot of the prefix OFF, the 3 most intense peaks are located at 1.0061, -26.9885 and -17.0041, and we can confidently infer their corresponding ion types as *b, a, and b-*H_2O (Table 2). In the plot of the suffix OFF, the peak located at 19.0159 is undoubtedly attributed to *y*-ions, and the next two peaks at 20.0189 and 21.0217 are mainly due to isotopic peaks of *y*-ions. The average distance between two adjacent peaks in this cluster is 1.0029, which is very close to the theoretical value, 1.0034. The offsets at 1.9905 and 1.0065 correspond to *y-*NH_3 and *y-*H_2O ions. In addition, we also chose to consider the offset at 10.0106, which is due to the y^{2+} ions, in the pNovo algorithm. It appears in the triply charged spectra with a relatively higher frequency.

Table S1. Probabilities of common ion types appeared in HCD data, with the consideration of mass regions and the relations between these types of ions.

	total	**low**	**medium**	**high**
Prob (*a*)	0.200	0.464	0.097	0.051
Prob (*b*)	0.286	0.0.533	0.217	0.111
Prob(*y*)	0.668	0.661	0.783	0.539
$Prob(y^0)$	0.155	0.235	0.118	0.125
$Prob(y^*)$	0.177	0.273	0.156	0.111
$Prob(y^{2+})$	0.100	0.033	0.071	0.199
$Prob(y^{2+})^{\#}$	0.388	0.089	0.413	0.634
$Prob(y\|y^*)$	0.932	0.924	0.948	0.922
$Prob(y\|y^0)$	0.970	0.951	0.989	0.982
$Prob(y^0\|y)$	0.225	0.338	0.149	0.228
$Prob(y^*\|y)$	0.247	0.382	0.190	0.189
Prob(*b*|*a*)	0.739	0.691	0.877	0.867
Prob (*a*|*b*)	0.518	0.602	0.392	0.398
Prob (*b*|*y*)	0.328	0.159	0.254	0.649
Prob (*y*|*b*)	0.766	0.657	0.917	0.952
$Prob(y\|y^{2+})^{\#}$	0.520	0.977	0.889	0.188

only triply charged spectra are considered.

Note: y^0 and y* denote y-ions with a neutral loss of water and ammonia, respectively. Three regions, low, medium and high ones, are computed by evenly split the range between 0 and the value of the peptide molecular weight plus a Proton.

Table S2. Proteins used in the Experiments and their corresponding IDs in Swiss-Prot database (v.56.2)

Protein	**ID in Swiss-Prot database**
Myosin	Q28641
Glycogen phosphorylase	P00489
Serum albumin	P02769
Beta-galactosidase	P00722
Carbonic anhydrase	P00921
Trypsin inhibitor	P01070
Ovalbumin	P01012
Lysozyme	P00698

Table S3. Parameters of database search.

Item	**Setting in pFind and Mascot**
Database	Target-reversed strategy is used and the target database consists of the proteins in Table S2.
Enzyme	Trypsin
Maximum missed cleavage sites	2
Precursor tolerance	±10 ppm
Fragment tolerance	±0.01Da
Fixed Modifications	Carbamidomethylation (C)