Journal of proteome research

Article

pubs.acs.org/jpr

Open-pNovo: De Novo Peptide Sequencing with Thousands of Protein Modifications

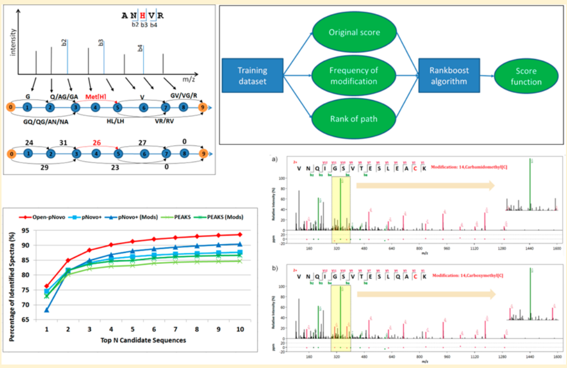
Hao Yang,[†,‡] Hao Chi,*,[†] Wen-Jing Zhou,[†,‡] Wen-Feng Zeng,[†,‡] Kun He,[†,‡] Chao Liu,[†] Rui-Xiang Sun,[†] and Si-Min He*,[†,‡]

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

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

Ⓢ *Supporting Information*

ABSTRACT: De novo peptide sequencing has improved remarkably, but sequencing full-length peptides with unexpected modifications is still a challenging problem. Here we present an open de novo sequencing tool, Open-pNovo, for de novo sequencing of peptides with arbitrary types of modifications. Although the search space increases by ~300 times, Open-pNovo is close to or even ~10-times faster than the other three proposed algorithms. Furthermore, considering top-1 candidates on three MS/MS data sets, Open-pNovo can recall over 90% of the results obtained by any one traditional algorithm and report 5−87% more peptides, including 14−250% more modified peptides. On a high-quality simulated data set, ~85% peptides with arbitrary modifications can be recalled by Open-pNovo, while hardly any results can be recalled by others. In summary, Open-pNovo is an excellent tool for open de novo sequencing and has great potential for discovering unexpected modifications in the real biological applications.



KEYWORDS: *tandem mass spectrometry, de novo peptide sequencing, dynamic programming, unexpected modifications*

■ INTRODUCTION

The past decades have seen remarkable progress in proteomics.[1] Researchers often use the mass spectrometry technology to analyze biological samples, in which peptide and protein identification has become the critical process. Database search has long been the dominant approach to peptide and protein identification. Many database search algorithms are used in the routine proteome analysis such as SEQUEST,[2] Mascot,[3] X! Tandem,[4,5] Andromeda,[6] pFind,[7,8] MS-GF+,[9] PEAKS DB,[10] and ByOnic.[11] Generally, the essence of these methods is retrieving all candidate peptides from a specified database for each spectrum, which also means that a protein database is indispensable.

An alternative method is de novo peptide sequencing, which deduces peptide sequences directly from MS/MS data without any databases. Whole peptide sequences are generated based on the mass difference of consecutive experimental MS/MS peaks. If there is no protein database available for the sample to be studied, de novo sequencing becomes an essential approach to analyzing MS/MS data. Multiple de novo peptide sequencing algorithms have been reported in recent years such as Lutefisk,[12] SHERENGA,[13] PEAKS,[14] NovoHMM,[15] PepNovo,[16,17] pNovo,[18,19] UniNovo,[20] and Novor.[21]

With the development of high resolution mass spectrometry, there has been an increasing emphasis on improving the identification rate of MS/MS data. More interpreted spectra are of great help to the identifications of peptides and proteins as well as the discovery of novel genes in proteogenomics.[22,23] A few studies showed that mutations and unexpected modifications are the principal factors leading to the unassigned mass spectra, while a potential advantage of de novo sequencing is to solve such problems, that is, discovering mutations and unexpected modifications.[22,24−26] Mutations are naturally considered in de novo sequencing algorithms, but detecting unexpected modifications is more challenging.

In previous studies, a few tag-based approaches have been proposed to identify peptides with unexpected modifications. Sequence tags or full-length de novo reconstructions are extracted first and then the intact peptide sequences are identified by expanding sequence tags or recovering the de novo reconstructions based on the protein databases. Mann et al. proposed a tag-based method in 1994,[27] and a few similar approaches are now available such as GutenTag,[28] MultiTag,[29] InsPecT,[30] MODi,[31] Paragon,[32] DirecTag,[33] and PEAKS DB.[10]

However, detecting peptides with unexpected modifications only via de novo sequencing is still an immense challenge. First, as shown in Figure 1, if all thousands of modifications in

Received: August 5, 2016
Published: December 26, 2016

DOI: 10.1021/acs.jproteome.6b00716

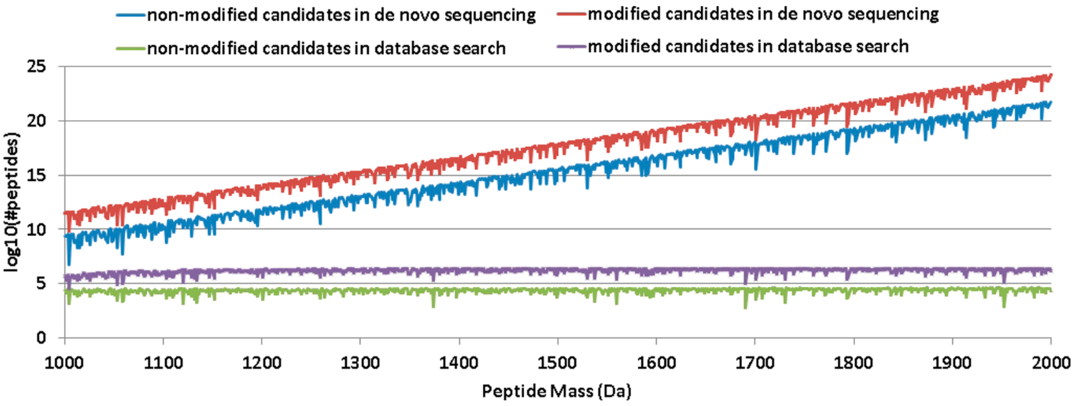
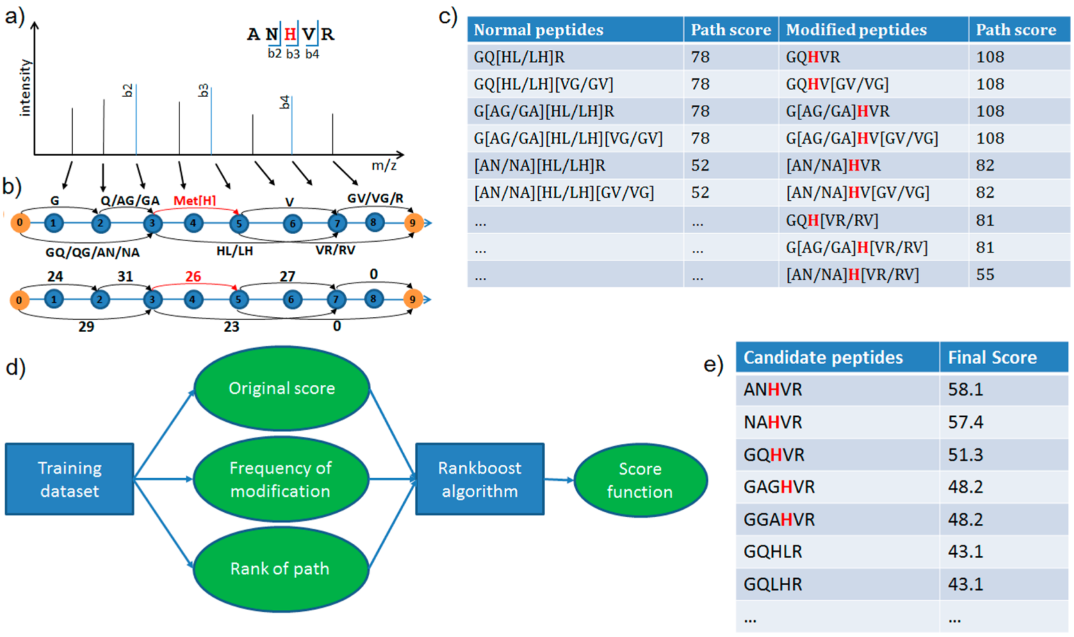


Figure 1. Comparison of numbers of peptide candidates in de novo sequencing and database search. For each approach, nonmodified peptides, as well as peptides with at most one modification from Unimod, are counted, respectively. One-thousand precursor ions are uniformly sampled from 1000−2000 Da in a HeLa data set of Mann lab (the M-DS1 data set as described in the Results section). Peptide candidates in de novo sequencing are arbitrary amino acid sequences whose masses differ from the corresponding precursor ions within a tolerance window from −20 ppm to 20 ppm, while peptide candidates in database search are counted from a human database downloaded from UniProt using nonspecific enzyme digestion. The average number of nonmodified candidates is 1.35E20 in de novo sequencing and 2.66E4 in database search, and the average number of modified candidates is 3.58E22 in de novo sequencing and 1.70E6 in database search.



Normal peptides	Path score	Modified peptides	Path score
GQ[HL/LH]R	78	GQHVR	108
GQ[HL/LH][VG/GV]	78	GQHV[GV/VG]	108
G[AG/GA][HL/LH]R	78	G[AG/GA]HVR	108
G[AG/GA][HL/LH][VG/GV]	78	G[AG/GA]HV[GV/VG]	108
[AN/NA][HL/LH]R	52	[AN/NA]HVR	82
[AN/NA][HL/LH][GV/VG]	52	[AN/NA]HV[GV/VG]	82
...	...	GQH[VR/RV]	81
...	...	G[AG/GA]H[VR/RV]	81
...	...	[AN/NA]H[VR/RV]	55

Candidate peptides	Final Score
ANHVR	58.1
NAHVR	57.4
GQHVR	51.3
GAGHVR	48.2
GGAHVR	48.2
GQHLR	43.1
GQLHR	43.1
...	...

Figure 2. Workflow of Open-pNovo. (a) Example of an original spectrum from a peptide AN(Met[H])VR, where Met[H] denotes the methylation of Histidine residue. (b) Spectrum graph (DAG) for the original spectrum. The black edges denote the normal edges, while the red ones denote the modification ones. If we do not consider the red edge, the correct peptide cannot be obtained. (c) Normal peptides and the modified peptides are obtained by finding *k* longest paths in DAG. (d) Score function is trained by RankBoost using three features. (e) Candidate peptides are output and sorted by the final scores, which are obtained by RankBoost. Note that when translating the original spectrum a to the spectrum graph b, each peak is translated into two vertices rather than one vertex in panel b.

Unimod[34] are considered in de novo peptide sequencing, the search space will increase by two orders of magnitude so that few proposed de novo sequencing algorithms can deal with such a challenge. It is also shown in Figure 1 that if peptide candidates are restricted to a protein database, the search space of database search is smaller by ∼15 orders of magnitude than that of de novo sequencing. Second, compared with the database search approach, correct peptides in de novo sequencing are more difficult to be distinguished from other similar candidates because of the remarkable difference of search space. Therefore, designing score functions for de novo sequencing, especially for open de novo sequencing with thousands of unexpected modifications, is far more challenging.

In this paper, we present a novel method named Open-pNovo to address the problem of de novo peptide sequencing with thousands of protein modifications in Unimod. On the basis of our previous work of pNovo and pNovo+,[18,19] we propose a new dynamic programming method to detect modification sites and then output optimal paths. In addition, a RankBoost-based scoring function[35] is designed to distinguish correct PSMs from incorrect ones effectively. Open-pNovo is tested on three real MS/MS data sets and three simulated ones, and performs favorably compared with the latest versions of pNovo+, PEAKS, and Novor. In most cases, considering the top-10 results, more than 90% of all correct peptides can be recalled by Open-pNovo, while the speed is comparable to pNovo+ and even ~10-times faster than PEAKS, although the search space is ~300-times larger than the conventional de novo sequencing algorithms.

■ METHODS

Open-pNovo Workflow

The approach taken in Open-pNovo can be summarized into four steps: (1) preprocessing MS/MS data, (2) constructing a directed acyclic graph (DAG) for each spectrum, (3) finding the k longest paths using a dynamic programming method, and (4) scoring each peptide-spectrum match. The workflow of Open-pNovo is shown in Figure 2.

Preprocessing MS/MS Data

In the first step, monoisotopic peaks are recognized and then transformed to charge +1 according to their original charge states, and peaks corresponding to the precursor ions and the neutral loss ions, such as the losses of ammonia and water, are all removed. The details of the first step were shown in the previous study,[19] and the remaining three steps will be introduced in the following sections.

Constructing a DAG for Each Spectrum

First, peaks in each spectrum are transformed to vertices with nominal masses and weights. Given a peak p whose mass is m and the mass of the precursor ion is M, if only b and y ions are considered, then two vertices are generated, whose masses are $m - 1$ and $M - m$, respectively (all peaks are singly charged after the preprocessing step). The weights of these two vertices are both set as the natural logarithm of the intensity of the original peak p. After all peaks are converted, a source vertex and a destination vertex are added in the spectrum graph, whose masses are set as zero and $M - w$, respectively, where w denotes the summed mass of a water molecule and a proton, and the weights of both vertices are set as zero.

Second, the vertices are connected by two types of edges. For a pair of vertices u and v (assuming the mass of u is less than that of v), if the mass difference is equal to the mass of one or two amino acid residues, a directed edge is added from u to v. Such edges are called "normal" edges. On the other hand, if the mass difference is equal to the mass of an amino acid residue with a modification (e.g., the mass of a methylation of Histidine residue is ~151 Da), then the other type of directed edge, called "modified" edges, is added from u to v. In this study, all modified edges are generated based on a predefined modification list, for example, all modifications from Unimod,[34] which contains 1356 types of modifications until June 2016.

Figure 2b shows a spectrum graph containing two types of edges. Modified edges, denoted by the red ones, are not considered by the conventional de novo sequencing method, so the correct peptide AN(Met[H])VR where Met[H] denotes the methylation of Histidine residue cannot be obtained in routine de novo sequencing analysis. However, it can be sequenced if modified edges are considered in this study.

For simplicity, the open de novo sequencing problem is shown in Figure 2b with only one additional modified edge. However, the practical problem is far more challenging because of the significant growth of edges, especially for the modified ones, in the spectrum graph. If only ten modifications are considered, the average percentage of modified edges in each spectrum graph is only 25% (112/456), while the corresponding figures grows to 75% (1043/1387) if all 1356 modifications in Unimod are considered. However, there is only one unexpected modification on each peptide in most cases,[36] which means that among the thousands modified edges in one spectrum graph, only one is correct. Therefore, to distinguish the correct modified edges from thousands of modified edges is a very challenging task.

In Open-pNovo, the frequencies of modifications, which can be learnt automatically by iteratively running Open-pNovo or found by database search, are considered by Open-pNovo to distinguish the correct modification type from the wrong ones. The weight of a normal edge (u,ν) is assigned by the weight of v, while the weight of a modified edge (u,ν) is assigned by the weight of v multiplied by the frequency of the corresponding modification, as shown in eq 1. The frequency of a modification is assigned by the number of this modification divided by the number of all detected modifications. When there are more than one modification with similar masses in one modified edge, the frequency is assigned by the maximum one of all of these modifications:

$$w_{(u,v)} = \begin{cases} w_v & \text{if edge}(u,\ v) \text{ is a normal edge} \\ w_v \times \max_{m \in M(u,v)} f(m) & \text{if edge } (u,\ v) \text{ is a modified edge} \end{cases} \quad (1)$$

In eq 1, $w_{(u,\nu)}$ is the weight of edge (u,ν), w_ν is the weight of vertex v, $M(\mu,\nu)$ is the modification set of modified edge (u,ν), and $f(m)$ is the frequency of modification m and is between 0 and 1.

Finding the *k* Longest Paths

The k longest paths from the source to the destination should be found in the graph. Two types of paths are defined as valid solutions: one is called normal path if it consists of only normal edges, and the other is called modified path if it contains one modified edge. In principle, multiple modifications can also be supported, but only one modified edge is allowed in finding paths in this study. First, very few spectra contain multiple unexpected modifications, which is the reason why most open database search algorithms also allow at most one unexpected modification.[10,23] Second, if two or more modifications are considered, the error rate will increase significantly.[36,37] Despite all this, the search space of Open-pNovo also involves peptides with a few common modifications, that is, carbamidomethylation of cysteine and oxidation of methionine, and with one another unexpected modification, where the common modifications can be treated as regular amino acids.

The weight of a path is defined as the sum of its edge weights shown in eq 2:

$$w_{p(v_0,v_1,\ldots,v_n)} = \begin{cases} \sum_{i=1}^{n} w_{v_i} & \text{if } p \text{ is a normal path} \\ \sum_{i=1}^{j-1} w_{v_i} + w_{v_j} \times \max_{m \in M(v_{j-1},v_j)} f(m) + \sum_{i=j+1}^{n} w_{v_i} & \text{if } p \text{ is a modified path in which } (v_{j-1},\, v_j) \text{ is a modified edge} \end{cases} \quad (2)$$

The k longest normal paths and the k longest modified paths are to be found in Open-pNovo. It is easy to prove that if a path is one of the top-k longest paths from the source vertex s to the destination vertex t, then the subpath from s to every other vertex v must be one of the top-k longest paths from s to v, which is shown in eqs 3 and 4:

$$d_i(v) = \max_{u \in InvE1(v)}\{d_{u_j}(u) + w_{(u,v)}\} \quad (3)$$

$$d_i'(v) = \max\{\max_{u \in InvE2(v)}\{d_{u_j}(u) + w_{(u,v)}\},\ \max_{u \in InvE1(v)}\{d_{u_j}'(u) + w_{(u,v)}\}\} \quad (4)$$

In eqs 3 and 4, $d_i(v)$ and $d_i'(v)$ are the weights of the ith longest normal path and the ith longest modified path from source vertex to v, respectively. $InvE1(v)$ and $InvE2(v)$ denote two sets of all preceding vertices whose edges (u,v) are normal edges and modified edges, respectively, and u_j is the jth index of the vertex u where $i = 1 + \sum_j (u_j - 1)$. Therefore, when the longest paths from s to each vertex are computed in the graph, the top-ranked paths to all preceding vertices starting from s can be precomputed and stored, and then a dynamic programming method can be used to solve the problem. The details of the dynamic programming method are shown in the following section.

Dynamic Programming Method To Find *k* Longest Paths

First of all, all vertices are sorted by mass in ascending order. For each vertex v, the k longest normal and modified paths can be computed by the paths of all its preceding vertices. For each preceding vertex u, if (u,v) is a normal edge, then each of the k longest normal paths to u appended by (u,v) is a candidate of the k longest normal paths to v, and each of the k longest modified paths to u appended by (u,v) is a candidate of the k longest modified paths to v. On the other hand, if (u,v) is a modified edge, then only each of the k longest normal paths to u appended by (u,v) is a possible solution to the k longest modified paths to v. After all vertices are transversed in the graph, the longest paths are stored at the destination vertex. At last, backtrack all vertices on the optimal paths iteratively from the destination vertex to the source one. In the process, each vertex v is visited by $d(v)$ times where $d(v)$ is the degree of v. Before visiting a vertex, all the k longest path candidates of the preceding vertices, both the normal and the modified ones, have been computed earlier because of the ascending order of the masses of the vertices, so that no correct paths can possibly be omitted. This algorithm is called pDAG-I. An example explaining how the algorithm works is shown in the Supporting Information.

Antisymmetry Restriction

Algorithm pDAG-I is efficient to find peptides with one unexpected modification from a relatively small modification set. However, if a large modification set is used, pDAG-I is not accurate enough. The reason is that two vertices are easily to be randomly connected by one modified edge if more modifications are considered, so that high-weight vertices generated from the same peak are more likely to be appeared in one path. However, such conditions can probably lead to wrong results. When a spectrum graph is constructed, each peak is converted to two vertices (called a cognate vertex pair) because the ion type (e.g., b or y) of the peak is indeterminate. Nevertheless, at most one vertex in each pair is correct in most cases, which is equivalent to that one peak matches with at most one ion from a peptide. This is why an antisymmetry-path-finding problem is modeled in earlier studies.[13,38] The antisymmetry restriction means that only paths without any cognate vertex pairs are treated as valid solutions.

Chi et al.[19] suggested that the antisymmetry restriction can be removed in high resolution data with little loss of accuracy but with great speedup; however, when considering all modifications in Unimod,[34] the graph becomes much more complex and the antisymmetry restriction should be reconsidered. According to our statistics in all three real data sets, 15.5% of the total paths contain at least one cognate vertex pair, while the figure of the normal paths is 6.6%; however, there are only 7.0% of the spectra containing a peak that matches more than one types of ions in the real data sets. If the antisymmetry restriction is considered, the average rank of the correct paths in 15.0% of the spectra moved up from 73 to 29 and 8.1% of the correct peptides for these spectra can only be recalled under the antisymmetry restriction. Figure S1 shows that distributions of normal paths and modified paths containing at least one cognate vertex pair. As a result, the antisymmetry restriction is essential when unexpected modifications are considered in de novo sequencing.

Bit Vector Approach

As shown in the previous study,[38] the time complexity of finding the longest antisymmetric paths is $O(|V||E|)$. However, pDAG-I can be modified to satisfy the antisymmetry restriction by removing the invalid paths in real time during the iterative process. Because correct paths still often rank better than most random ones, the algorithm can store a relatively larger number of intermediate results, and finally the correct peptides can probably be recalled. When the paths to vertex v are computed, it can be checked whether each path p to the preceding vertices of v already contains the cognate vertex of v; if so, p will not be considered as a valid longest path to v. Because of the limited number of peaks (generally not greater than 300 after the preprocessing procedure), a bit vector approach can be used to record whether each peak has been used in each path as shown in Figure S2. The time complexity of judging if a cognate vertex has been visited is only O (1), while only ∼13 MB of memory are adequate.

Loser Tree to Speedup

A further improvement is using a loser tree[39] to effectively generate the k longest paths to v, which is based on the fact that the k longest paths to all of the preceding vertices of vertex v are sorted. In short, assuming that the preceding vertices of v are in S = $\{u_1, u_2, \ldots, u_d\}$, d is the in-degree of v, and k longest paths $\{p_{i_1}, p_{i_2},$

..., p_{i_k}} to each u_i in S are sorted, then the longest path to ν can be generated from $P = \{p_{1_1}, p_{2_1}, ..., p_{d_1}\}$. If the path is from the vertex u_j, then P is updated to be $P - \{p_{j_1}\} + \{p_{j_2}\}$, and the second path of ν should be generated in the updated P. If P is maintained as a loser tree, the time complexity of finding k paths to vertex v is $O(k\log d)$, where d is the in-degree of v. The improved algorithm is called pDAG-II. The pseudo codes of pDAG-I and pDAG-II are shown in the Supporting Information.

Analysis of the Time Complexity

The time complexity analysis of pDAG-II is $O(k|V| + |E| + k|V| \log \overline{d})$, where E is the edge set of the graph, V is the vertex set of the graph, k is the number of longest paths, and $\overline{d}$ is the average in-degree. The proof is shown in the Supporting Information.

Selection of the Number of Longest Paths

Experimental results show that the run time is with a linear increase when k becomes larger, while the recall rate becomes stable when k is no less than 150. Therefore, correct peptides can scarcely be omitted if a proper value of k is chosen in the algorithm. In this study, k is set as 150 to make a balance between the recall rate and the run time (shown in Table S10).

Refined Scoring by the RankBoost Algorithm

The k longest normal paths are converted to nonmodified candidate peptides, and the k longest modified paths are converted to candidate peptides containing one unexpected modification. Then a scoring function previously proposed in pNovo+ is used to evaluate the peptide-spectrum matches.[19] Furthermore, to better distinguish nonmodified peptide from modified ones, the frequencies of modifications detected in the data are used. These frequencies can be calculated initially by the de novo sequencing results with high original scores. In general cases, peptides without any modifications or with common ones are more credible than those with rare modifications.[23] We use the RankBoost algorithm (a machine learning scoring method[35]) to score these candidate peptides, in which three features are used as shown in Figure 2d. (1) The original score of the peptide-spectrum match. (2) The frequencies of the modifications. All values are between 0 and 1, and frequencies of nonmodified peptides are set as 1. (3) The rank of the path corresponding to the peptide (the range of this value is integers between 1 and k, where k is the number of paths). A scoring model was trained on the M-DS1 data set (shown in the following section), and the weights of each feature were sorted automatically by the RankBoost algorithm. Specifically, after learning from the training set, the importance of the features are sorted as follows: feature 1 > feature 2 > feature 3, which means that the original score is the most important feature, the frequency of the modification is the second important one. Some other features are also tested, that is, the precursor mass deviation, but the effect is negligible so that these features are not involved into the final scoring model of Open-pNovo.

This scoring model is used in Open-pNovo to evaluate all peptide-spectrum matches and obtain the final score shown in eq 5:

$$Score = \sum_{i=1}^{n} f_i(s_i) \tag{5}$$

In eq 5, n is the number of features, s_i is the value of ith feature, and f_i is a function of the ith feature. Specifically, f_i is the step function about s_i trained by RankBoost. The effect of the three features are shown in Figure S3.

RESULTS

Data Sets

The performance of Open-pNovo was tested on six data sets. The first two data sets are from HeLa cells, which are generated on LTQ Orbitrap Velos and Q Exactive, respectively. The third data set is a much larger one from budding yeast (*Saccharomyces cerevisiae*) generated on Q Exactive. All of the three data sets are provided by Matthias Mann's laboratory.[40,41] The open search mode of pFind[23] and PEAKS DB[10] are used to analyze the three data sets. The first two data sets are searched against UniProt human database (released in 2014−11), and the third data set is searched against UniProt yeast database (released in 2015−01). Both databases are appended with 286 common contaminant protein sequences. The parameters are shown in Table S1. Peptides with no modification or with one of the ten most abundant modifications were kept. The abundance of one modification was determined by the frequency of the modification. False discovery rate (FDR) was controlled at 1% at the peptide level for the quality assessment of the peptide−spectrum matches. In addition, inconsistent results of the two engines were removed, and three following criteria were used to generate test data sets. (1) The length of the peptide sequence is between 6 and 20 (the distribution of the peptide lengths is shown in Figure S4); (b) the maximum length of the gap in the matched ion series must be less than 2; and (c) the summed intensity of matched peaks is greater than 20% of the total in one spectrum. Finally, three data sets (referred to as M-DS1, M-DS2, and M-DS3) were generated, which consist of 8600, 6727, and 45 450 spectra, respectively. All these three real data sets are high resolution HCD data sets.

Besides the three real experimental data sets described above, another three simulated data sets were also used in this study. The data sets were generated in the following way. First, peptides were randomly generated whose lengths were between 6 and 25, and then one modification from Unimod[34] was selected randomly and then added to an arbitrary legal position on the peptides. For example, deamidation can be added only on N, Q, R, or F according to the record in Unimod. Second, theoretical spectra with full *b*- and *y*-series were created according to the peptides and then split into three subsets. For each spectrum, 10%, 15%, and 20% of the total peaks were randomly removed in three subsets, respectively, which was done to simulate the different level of fragment ion losses in the real condition. Third, for each data set, noise peaks were randomly added to each spectrum, whose intensity was 0.1-times the correct peak intensity and whose number was 0-, 1-, or 2-times the peaks in the original spectrum with equal probability of $^1/_3$. For instance, noise peaks whose number was 0-times the original peaks mean that there were no noise peaks, and there were one-third of such spectra without any noise peaks in each of the three subsets. Finally, three simulated MS/MS data sets, S-DS1, S-DS2, and S-DS3, were produced, whose sizes were 7761, 7372, and 8233, respectively. The simulated data sets seem fairly ideal because they were designed to explore the capability and boundary of finding unexpected modifications by Open-pNovo.

Comparison between Open-pNovo and Other Algorithms

Open-pNovo is compared with pNovo+,[19] PEAKS[14] (version 7.5), and Novor[21] on the six data sets described above. Two different sequencing modes of pNovo+, PEAKS, and Novor are tested in this study. The first one is called no-modification mode, in which only carbamidomethylation of cysteine for the fixed modification and oxidation of methionine for the variable

Table 1. Comparing Successful De Novo Peptide Sequencing Results between Open-pNovo and Other Algorithms When Considering Top-1 Results

data sets	open-pNovo	pNovo+	pNovo+ (Mods[a])	PEAKS	PEAKS (Mods)	Novor	Novor (Mods)
M-DS1 (8600)	77.9% (6703)	71.6% (6159)	71.7% (6170)	67.4% (5798)	70.4% (6053)	37.7% (3243)	34.2% (2940)
M-DS2 (6727)	74.6% (5020)	59.3% (3992)	62.5% (4203)	56.9% (3825)	64.5% (4341)	34.7% (2335)	33.5% (2256)
M-DS3 (45 450)	76.3% (34 659)	74.5% (33 879)	68.2% (31 019)	73.1% (33 226)	72.8% (33 080)	47.4% (21 527)	43.2% (19 616)
S-DS1 (7761)	85.6% (6641)	0.6% (45)	9.1% (704)	0.4% (34)		0.2% (17)	
S-DS2 (7372)	78.1% (5756)	0.7% (48)	8.5% (625)	0.5% (36)		0.2% (18)	
S-DS3 (8233)	69.7% (5740)	0.6% (51)	7.5% (616)	0.5% (38)		0.2% (15)	

[a]The second search mode in which more variable modifications is specified in pNovo+, PEAKS, and Novor.

Table 2. Comparing the Recall Rate of De Novo Sequencing Results between Open-pNovo and Other Algorithms on the PSMs Only with Modifications When Considering Top-1 Results

data sets	Open-pNovo	pNovo+	pNovo+ (Mods[a])	PEAKS	PEAKS (Mods)	Novor	Novor (Mods)
M-DS1 (2440)	65.0% (1587)	33.9% (828)	49.4% (1205)	34.9% (851)	51.1% (1247)	13.3% (325)	14.3% (350)
M-DS2 (2616)	67.0% (1753)	22.0% (576)	47.7% (1248)	21.9% (574)	46.0% (1204)	9.6% (251)	13.2% (345)
M-DS3 (10 047)	52.8% (5302)	38.0% (3813)	45.1% (4536)	39.3% (3945)	51.1% (5132)	18.3% (1841)	17.9% (1801)

[a]The second search mode in which more variable modifications is specified in pNovo+, PEAKS, and Novor.

modification is considered. This mode is to simulate the most popular usage of the traditional de novo sequencing tools. The second one is called modification mode, in which more variable modifications are specified according to the different characterizations of the data sets. For M-DS1 and M-DS2, six variable modifications including oxidation of methionine, carboxymethylation of cysteine, acetylation of N-terminus, carbamylation of N-terminus, pyro-glu of N-terminal Q, and pyro-glu of N-terminal E are specified, which are among the ten most abundant modifications according to the result of both pFind and PEAKS DB and cover 86% of all results. For M-DS3, four modifications including oxidation of methionine, acetylation of N-terminus, carbamylation of N-terminus, and pyro-glu of N-terminal Q are specified, which cover 91% of all results. For each of the simulated data sets, ten most abundant modifications are specified, which cover 9.7%, 10.3%, and 10.0% of all results. PEAKS and Novor cannot support so many rare variable modifications, so they are not tested in the modification mode in simulated data sets. Parameters for the modification search mode on the simulated data sets are shown in Table S2.

A peptide is correctly recalled if all of its residues, both the common and the modified ones, are correct according to the annotation of the data sets. In addition, if the mass of a residue reported by the algorithm is identical with that in the annotation, for example, Q and deamidated N, then the peptide is also considered to be correctly recalled.

The comparison of Open-pNovo and the two modes of other algorithms on all six data sets are shown in Table 1. Open-pNovo performed favorably in terms of the recall rate on all of the six data sets. On the real MS/MS data sets, the average recall rate of Open-pNovo is 76.3%, which is 5.3% more than that of the no-modification mode of pNovo+, the best algorithm in the test of the no-modification mode. In terms of the modification search mode on the real MS/MS data sets, PEAKS performs the best but Open-pNovo still reported 6.7% more than PEAKS. The result of pNovo+ and Novor in the modification mode is slightly less than the no-modification mode because they do not allow setting only one modification on each peptide, so peptides with multiple modifications interfered in the search space of these two algorithms.

The tools performed a little differently on the simulated data sets. For Open-pNovo, the percentages of the sequenced peptides were even higher than those on three real MS/MS data sets, although the simulated data sets contains far more complex modifications. By contrast, the recall rate of the no-modification mode of pNovo+ was less than 1% of the total, which is reasonable since all spectra are corresponding to the randomly modified peptides. However, there were still a few peptides recalled by the no-modification mode of pNovo+ because the masses of some residues with modifications are equal to some other amino acids, for example, the masses of both Glu and deamidated Gln residues are around 129 Da. Even if ten modifications were specified in pNovo+ and the percentages of the results containing these ten modifications are only 9.7%, 10.3%, and 10.0% in S-DS1, S-DS2, and S-DS3, respectively, the search space was yet too incomplete so that the recall rate was still less than 10%. PEAKS and Novor also reported a few correct peptides with the no-modification sequencing mode, and the modification mode is not tested because such large number of rare modifications is not supported by these two algorithms. However, it can be reasonably inferred that hardly any result can be reported for all traditional de novo sequencing algorithms due to the extreme incompleteness of the search space.

When only considering the modified results on the real data sets, Open-pNovo also performs the best as shown in Table 2. The recall rate of top-1 results of Open-pNovo is ∼62%, while those of pNovo+ and PEAKS are only ∼38%, and Novor is only ∼14% in the modified data sets. However, when only considering the unmodified results, as shown in Tables 1 and 2, Open-pNovo identified 5116 (6703−1587), 3267, and 29 357 in M-DS1, M-DS2, and M-DS3, while the figures of pNovo+ are 5331, 3416, and 30 066, and those of PEAKS are 4947, 3251, and 29 281. The performance of Open-pNovo is still better than PEAKS but slightly inferior to pNovo+ because there are more similar modified peptides to interfere with the correct unmodified peptide in open de novo sequencing.

Figure 3 shows the cumulative curves of the number of correct sequences from top-1 to top-10 candidate sequences on the three real data sets. Open-pNovo still performs the best regardless of how many top-ranked peptides are considered in the results. Because the search space of Open-pNovo is hundreds of times larger than the common de novo sequencing methods, correct peptides may be easily interfered with by other similar competitors, so that designing a scoring function to distinguish

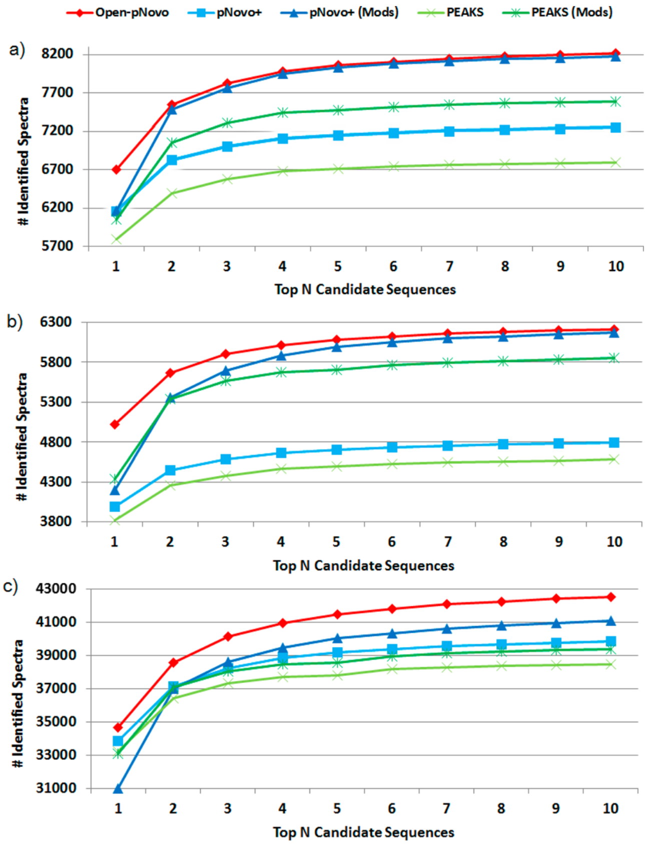


Figure 3. Cumulative curves of the number of correct sequences among the top-1 to top-10 candidates from all algorithms on (a) M-DS1, (b) M-DS2, and (c) M-DS3. In all three real data sets, the top-10 recall of Open-pNovo is 93.7%, while the corresponding figures for pNovo+ and PEAKS are 85.4% and 82.0%, respectively, and 91.2% and 86.9% for pNovo+ (Mods) and PEAKS (Mods), respectively. Only the top-1 results are reported by Novor: the recall of three real data sets are 37.7%, 34.7%, and 47.4% for the no-modification mode, and 34.2%, 33.5%, and 43.2% for the modification mode, respectively.

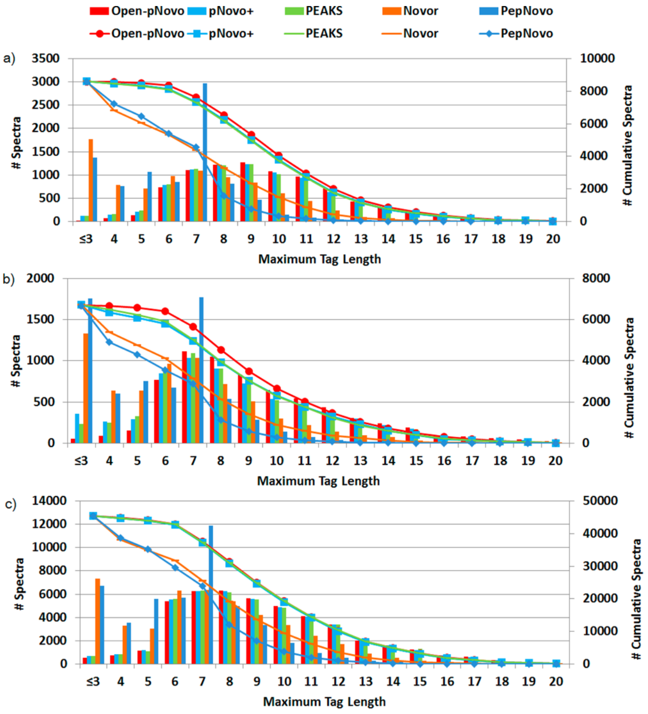


Figure 4. Comparison of identifications with the maximum correct sequence tags in the top-1 results of Open-pNovo, pNovo+, PEAKS, Novor, and PepNovo on (a) M-DS1, (b) M-DS2, and (c) M-DS3.

them is much more difficult. Take Figure 3a as an example, when considering the top-10 candidates, the identified spectra of pNovo+ (Mods) are almost as many as that of Open-pNovo; however, if only the top-1 candidates are considered, the result of pNovo+ (Mods) are significantly less than those of Open-pNovo and even slightly less than those of the no-modification mode of pNovo+. In terms of the modification mode of pNovo+ and PEAKS, the difference between top-1 and top-2 is much larger, which can be shown in the curves. However, the trends of the other three curves (Open-pNovo, pNovo+, and PEAKS) are quite consistent to each other, which shows that the RankBoost-based scoring function provides more powerful ability to distinguish correct PSMs from other random matches.

The RankBoost algorithm ranked more correct peptides, especially for the top-1 results: the use of the RankBoost algorithm yielded a relative improvement of 27.3% more PSMs in total. For PSMs with modified peptides only, the improvement is 12.9%. The details of the effect of the RankBoost algorithm are shown in Table S3.

Figure 4 shows the comparison of the maximum sequence tags in the top-1 results identified by Open-pNovo, pNovo+, PEAKS, Novor, and PepNovo.[16] The sequence tags identified by Open-pNovo are slightly longer than pNovo+ and PEAKS and much longer than Novor and PepNovo. The ratio of the sequence tags identified by PepNovo whose lengths are longer than 8 is low because PepNovo considers the gaps of N-terminus and C-terminus, and the percentages of the top-1 results with no gaps are only 29.7%, 34.3%, and 40.3% on M-DS1, M-DS2, and M-DS3.

Consistency Analysis

The comparison of the correct top-1 results of Open-pNovo, pNovo+, and PEAKS was shown in Figure 5. About 96% of the pNovo+ result and 90% of the PEAKS result can also be obtained by Open-pNovo. The result of pNovo+ is more consistent with that of Open-pNovo because they share the same scoring function (partially in Open-pNovo). We find that the other results identified only by Open-pNovo are all modified results, which can not be recalled by pNovo+ or PEAKS in the no-modification mode.

Modification Analysis

Figure 6 and Tables S4−S9 show the number of correct peptides with different modifications recalled in the top-10 candidates. In most cases, Open-pNovo gives more correct peptides than others, and few modifications can be detected by pNovo+, PEAKS, or Novor if no modifications are specified, except deamidation on Gln's and Asn's, which leads to the same masses of Glu and Asp, respectively. When more modifications were added, more correct PSMs with modifications can be reported, but still inferior to that of Open-pNovo because the scoring functions in the traditional de novo sequencing algorithms only aimed at peptides without unexpected modifications. In addition, modifications with similar masses can also be effectively distinguished in Open-pNovo. Figure S5 gives an example of two PSMs with very similar peptide sequences but different modifications. If algorithms only considered carbamidomethylation (one of the most common modifications), both pNovo+ and PEAKS gave a wrong peptide VNQLGSVTESLEAC(+57)K

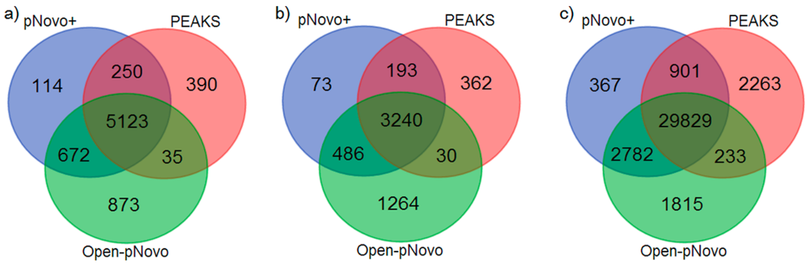


Figure 5. Comparison of the correct top-1 results of Open-pNovo, pNovo+, and PEAKS on (a) M-DS1, (b) M-DS2, and (c) M-DS3.

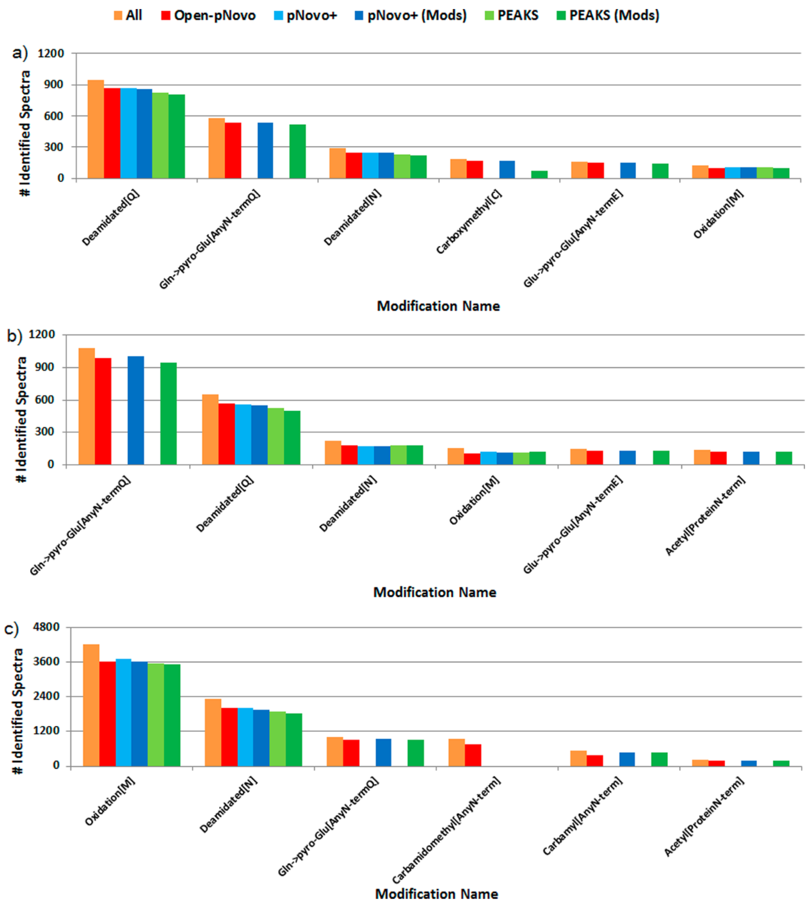


Figure 6. Distribution of correct PSMs on each type of modification obtained by Open-pNovo, pNovo+, and PEAKS on (a) M-DS1, (b) M-DS2, and (c) M-DS3.

Table 3. Run Time Comparison between Open-pNovo and Other Algorithms on Six Data Sets[a]

data sets	Open-pNovo	pNovo+	pNovo+ (Mods)	PEAKS	PEAKS (Mods)	Novor	Novor (Mods)
M-DS1	62.9	42.5	43.7	420.0	480.0	10.0	11.0
M-DS2	52.0	34.4	36.0	300.0	420.0	9.0	9.0
M-DS3	258.1	166.1	177.0	2280.0	2820.0	37.0	39.0
Avg.[b]	162.9	250.1	236.8	20.3	16.3	1085.3	1030.1
S-DS1	151.1	108.5	227.6	900.0		15.0	
S-DS2	127.1	90.2	185.1	960.0		14.0	
S-DS3	146.8	103.1	221.3	1020.0		14.0	
Avg.	55.0	77.4	36.9	8.1		543.4	

[a] All of the software packages were executed on the same PC (Dell Optiplex 9010, Intel(R) Core(TM) i7−4770 CPU at 3.40 GHz, 12GB Memory).
[b] The average number of spectra can be processed in one second.

(Figure S5a); however, Open-pNovo reported another peptide VNQIGSVTESLQAC(+58)K with a better peptide−spectrum match, which is identical with the result given by the two open database search algorithms, pFind and PEAKS DB (Figure S5b). Furthermore, when carboxymethylation is specified, the correct peptide can also be given by pNovo+ and PEAKS. This example shows that a more comprehensive search space is the basis of obtaining more correct results. If the search space is insufficient, a similar but incorrect result is more likely to be obtained. In addition, a more discriminating scoring function is also indispensable so that correct peptides can still be distinguished from random ones in a more comprehensive search space.

Run Time Analysis

The run time comparison of Open-pNovo, pNovo+, PEAKS, and Novor is shown in Table 3. Open-pNovo can process ∼163 spectra per second on the real data sets and ∼55 spectra per second on the simulated data sets, which means that Open-pNovo has potential for real time spectral analysis in shotgun proteomics. Although the search space is hundreds of times larger, Open-pNovo is a bit faster than pNovo+ and 8−10-times faster than PEAKS. Novor is the fastest one in our experiment, which is mainly due to that only the first candidate of each spectrum is reported. Actually, if only one temporary path is kept in the algorithm and only the first candidate of each spectrum is reported in Open-pNovo, it can process ∼1105 spectra per second on the real data set, which is still slightly faster than Novor. The recall rate of Open-pNovo in such condition is 52%, while the corresponding figure of Novor is 45%. It can also be inferred that the recall rate of Novor is lower than that of other algorithms because of the lower number of temporary results.

On the simulated data sets, the average in-degree of all vertices is only 4.1, while on the real data sets, it is up to 14.7 (Figure S6). As a result, the simulated spectra are fairly simpler than the real MS/MS data. However, all four algorithms run more slowly in the simulated spectra than the real ones, which is mainly due to their different peptide length distributions (the upper bounds of the lengths on the real and simulated data sets are 20 and 25, respectively). As shown in Figure S7, the average time per spectrum grows exponentially when the peptide length increases, and the time used on sequencing peptides with length greater than 20 is ∼64% of the total.

DISCUSSION

In this paper, we presented a new de novo sequencing tool called Open-pNovo, which can sequence peptides with any one of the thousands of modifications that are predefined in a database such as Unimod. On both the real and the simulated data sets, Open-pNovo performs favorably compared with two sequencing modes of pNovo+, PEAKS, and Novor. On the real data sets, the recall rate on the top-1 candidate sequences of Open-pNovo is ∼9% more than that of pNovo+, ∼7% more than that of PEAKS, and ∼79% more than that of Novor. On high-quality simulated data set, the recall rate on the top-1 candidate sequences of Open-pNovo is as high as ∼85%, while few results can be reported by other tested algorithms because that the common methods are not designed for the open de novo sequencing of peptides with thousands of modifications.

On the real data sets, the speed of Open-pNovo is comparable with that of the two modes of pNovo+ and even ∼10-times faster than PEAKS, although the search space is ∼300-times larger; on the simulated data sets, Open-pNovo is nearly twice as fast as the modification mode of pNovo+. A possible reason why pNovo+ is slower than Open-pNovo on the simulated data sets is that Open-pNovo can process long peptides more efficiently with the algorithm pDAG-II explained in the Methods section. De novo sequencing of longer peptides is essential because more valuable information tends to be carried.

The false discovery rates (FDRs) of Open-pNovo, pNovo+, PEAKS, and Novor are also analyzed on three complete real data sets.[40,41] Results identified by database search with FDR ≤ 1% at the peptide level are used to evaluate the FDR of de novo sequencing. If a PSM is consistent with the results of database search, it is considered correct, otherwise incorrect. The value of $\frac{\text{no. correct results}}{\text{no. correct results} + \text{no. incorrect results}}$ can be used to estimate the FDR of de novo sequencing. Figure S8 shows the FDR curves of four algorithms; the FDRs of Open-pNovo and PEAKS with high score results are ∼10%, while the FDRs of all four algorithms with whole results are ∼50%.

Therefore, the error rate control of amino acids on a proteome-scale may be more realistic. The precision and recall rates of the amino acids identified by Open-pNovo and PEAKS are shown in Figure S9. When the recall rate is ∼50%, the precision rates of Open-pNovo and PEAKS are ∼95% and ∼90%, respectively.

In summary, Open-pNovo can be an efficient tool to de novo sequence the modified peptides, and it can be downloaded from the following Web site: http://pfind.ict.ac.cn/software/pNovo/Open-pNovo_v1.0.exe.

■ ASSOCIATED CONTENT

Ⓢ Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acs.jproteome.6b00716.

Pseudo code of pDAG-I/II; example of the algorithm pDAG-I; proof that pDAG-II is always not worse than pDAG-I; time complexity analysis of pDAG-II; figures and tables (PDF)

■ AUTHOR INFORMATION

Corresponding Authors

*E-mail: chihao@ict.ac.cn.
*E-mail: smhe@ict.ac.cn.

ORCID

Hao Yang: 0000-0002-1277-2628

Author Contributions

H.Y. designed the algorithms and performed the data analysis. H.C. wrote the manuscript, and S.-M.H. edited the manuscript. W.-J.Z. produced the simulated data sets. W.-F.Z. and K.H. suggested using a lose tree algorithm and proved the time complexity. C.L. and R.-X.S. modified the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

This work was supported by the National Key Research and Development Program of China (No. 2016YFA0501301) and the National High Technology Research and Development Program of China (863) under Grant Nos. 2014AA020902 to S.-M.H. and 2014AA020901 to H.C. This work was also supported by the National Natural Science Foundation of China (No. 31470805) and Youth Innovation Promotion Association CAS (No. 2014091). Finally we thank Prof. Gong Zhang's lab for valuable discussions.

■ REFERENCES

(1) Aebersold, R.; Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **2016**, *537* (7620), 347−355.

(2) Eng, J. K.; McCormack, A. L.; Yates, J. R. 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.

(3) 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.

(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), 2310−2316.

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

(6) Cox, J.; Neuhauser, N.; Michalski, A.; Scheltema, R. A.; Olsen, J. V.; Mann, M. Andromeda: A Peptide Search Engine Integrated into the MaxQuant Environment. *J. Proteome Res.* **2011**, *10* (4), 1794−1805.

(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) 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), 2985−2991.

(9) Kim, S.; Pevzner, P. A. MS-GF+ makes progress towards a universal database search tool for proteomics. *Nat. Commun.* **2014**, *5*, 5277.

(10) Zhang, J.; Xin, L.; Shan, B. Z.; Chen, W. W.; Xie, M. J.; Yuen, D.; Zhang, W. M.; Zhang, Z. F.; Lajoie, G. A.; Ma, B. PEAKS DB: De Novo Sequencing Assisted Database Search for Sensitive and Accurate Peptide Identification. *Mol. Cell. Proteomics* **2012**, *11* (4), M111.010587.

(11) Bern, M.; Kil, Y. J.; Becker, C. Byonic: advanced peptide and protein identification software. *Curr. Protoc Bioinformatics* **2012**, 20.

(12) 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.

(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), 327−342.

(14) 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), 2337−2342.

(15) 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.

(16) Frank, A.; Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal. Chem.* **2005**, *77* (4), 964−973.

(17) 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.

(18) 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.

(19) Chi, H.; Chen, H.; He, K.; Wu, L.; Yang, B.; Sun, R. X.; Liu, J.; Zeng, W. F.; Song, C. Q.; He, S. M.; Dong, M. Q. pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra. *J. Proteome Res.* **2013**, *12* (2), 615−625.

(20) Jeong, K.; Kim, S.; Pevzner, P. A. UniNovo: a universal tool for de novo peptide sequencing. *Bioinformatics* **2013**, *29* (16), 1953−1962.

(21) Ma, B. Novor: Real-Time Peptide de Novo Sequencing Software. *J. Am. Soc. Mass Spectrom.* **2015**, 1−10.

(22) Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* **2015**, *33* (7), 743−749.

(23) Chi, H.; He, K.; Yang, B.; Chen, Z.; Sun, R. X.; Fan, S. B.; Zhang, K.; Liu, C.; Yuan, Z. F.; Wang, Q. H.; Liu, S. Q.; Dong, M. Q.; He, S. M. pFind-Alioth: A novel unrestricted database search algorithm to improve the interpretation of high-resolution MS/MS data. *J. Proteomics* **2015**, *125*, 89−97.

(24) Lu, B.; Chen, T. Algorithms for *de novo* peptide sequencing using tandem mass spectrometry. *Drug Discovery Today: BIOSILICO* **2004**, *2* (2), 85−90.

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

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

(27) Mann, M.; Wilm, M. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* **1994**, *66* (24), 4390−4399.

(28) 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), 6415−6421.

(29) 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), 1307−1315.

(30) 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.

(31) Kim, S.; Na, S.; Sim, J. W.; Park, H.; Jeong, J.; Kim, H.; Seo, Y.; Seo, J.; Lee, K. J.; Paek, E. MODi: a powerful and convenient web server for identifying multiple post-translational peptide modifications from tandem mass spectra. *Nucleic Acids Res.* **2006**, *34* (Web Server), W258−263.

(32) 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), 1638−1655.

(33) Tabb, D. L.; Ma, Z. Q.; Martin, D. B.; Ham, A. J. L.; Chambers, M. C. DirecTag: Accurate sequence tags from peptide MS/MS through statistical scoring. *J. Proteome Res.* **2008**, *7* (9), 3838−3846.

(34) Creasy, D. M.; Cottrell, J. S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **2004**, *4* (6), 1534−1536.

(35) Freund, Y.; Iyer, R.; Schapire, R. E.; Singer, Y. An efficient boosting algorithm for combining preferences. *Journal of Machine Learning Research* **2004**, *4* (6), 933−969.

(36) Tsur, D.; Tanner, S.; Zandi, E.; Bafna, V.; Pevzner, P. A. Identification of post-translational modifications via blind search of mass-spectra. *Nat. Biotechnol.* **2005**, *23* (12), 1562−1567.

(37) Pevzner, P. A.; Mulyukov, Z.; Dancik, V.; Tang, C. L. Efficiency of database search for identification of mutated and modified proteins via mass spectrometry. *Genome Res.* **2001**, *11* (2), 290−299.

(38) 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.

(39) Knuth, D. E. The art of Computer Programming: Sorting and Searching. *Addison-Wesley Series in Computer Science and Information Processing*; Addison-Wesley, 1973; Vol. 3.

(40) Michalski, A.; Damoc, E.; Hauschild, J. P.; Lange, O.; Wieghaus, A.; Makarov, A.; Nagaraj, N.; Cox, J.; Mann, M.; Horning, S. Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Mol. Cell. Proteomics* **2011**, *10* (9), M111.011015.

(41) Kulak, N. A.; Pichler, G.; Paron, I.; Nagaraj, N.; Mann, M. Minimal, encapsulated proteomic-sample processing applied to copy-number estimation in eukaryotic cells. *Nat. Methods* **2014**, *11* (3), 319−U300.