Journal of Proteomics 125 (2015) 89–97

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

pFind–Alioth: A novel unrestricted database search algorithm to improve the interpretation of high-resolution MS/MS data

Hao Chi [a], Kun He [a], Bing Yang [b], Zhen Chen [c], Rui-Xiang Sun [a], Sheng-Bo Fan [a], Kun Zhang [a], Chao Liu [a], Zuo-Fei Yuan [a], Quan-Hui Wang [c], Si-Qi Liu [c], Meng-Qiu Dong [b], Si-Min He [a,*]

[a] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[b] National Institute of Biological Sciences, Beijing, Beijing 102206, China
[c] Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China

ARTICLE INFO

Article history:
Received 2 March 2015
Received in revised form 4 May 2015
Accepted 10 May 2015
Available online 12 May 2015

Keywords:
Unrestricted database search
Ion index
In-depth interpretation
High resolution MS/MS

ABSTRACT

Database search is the dominant approach in high-throughput proteomic analysis. However, the interpretation rate of MS/MS spectra is very low in such a restricted mode, which is mainly due to unexpected modifications and irregular digestion types. In this study, we developed a new algorithm called Alioth, to be integrated into the search engine of pFind, for fast and accurate unrestricted database search on high-resolution MS/MS data. An ion index is constructed for both peptide precursors and fragment ions, by which arbitrary digestions and a single site of any modifications and mutations can be searched efficiently. A new re-ranking algorithm is used to distinguish the correct peptide-spectrum matches from random ones. The algorithm is tested on several HCD datasets and the interpretation rate of MS/MS spectra using Alioth is as high as 60%–80%. Peptides from semi- and non-specific digestions, as well as those with unexpected modifications or mutations, can be effectively identified using Alioth and confidently validated using other search engines. The average processing speed of Alioth is 5–10 times faster than some other unrestricted search engines and is comparable to or even faster than the restricted search algorithms tested.

© 2015 Elsevier B.V. All rights reserved.

1. Introduction

The past decades have seen remarkable progress in proteomics [1], especially in peptide and protein identification technology, which is the bedrock of proteomics. Database search has long been the dominant approach to peptide and protein identification. In database search, the best matched candidate peptide can be retrieved from a specified protein sequence database for each spectrum. A few database search algorithms are used in the routine proteome analysis, such as Mascot [2], SEQUEST [3], X! Tandem [4,5], OMSSA [6], pFind [7,8] and Andromeda [9].

Generally speaking, a large number of peptide sequences are generated from the database even in a regular searching mode, where a few common modifications are allowed and usually the restriction of full protease specificity is applied. However, increasing demand for searching semi- or non-specifically digested peptides results in tens to hundreds of times more peptides than that in fully-specific digestion. For example, ~35 million fully tryptic peptides are expected from the SwissProt database v.56.2, compared to ~5.5 billion peptides (161-fold increase) if enzyme specificity is not considered [10]. Furthermore, the search space increases dramatically with each additional variable modification added. The database search will be very time-consuming if all possible forms of peptides are considered, so in most routine experiments, only considered are peptides with full protease specificity and a few modifications. However, such "restricted search" leads to low interpretation rates of MS/MS spectra. Elias et al. in 2005 reported that only 33.6% and 22.7% of the total CID spectra can be identified from LTQ and Q-TOF mass spectrometers, respectively [11]. Michalski et al. in 2011 reported that out of 21,906 HCD spectra from LTQ-Orbitrap Velos, 16,924 could be reliably recognized as peptide spectra and yet only 58% of them were identified at 1% FDR, bringing the total interpretation rate to 44.8% [12]. It has been shown that, because of the restricted search mode, peptides with unspecified modifications are not identified, which is one of the major factors underlying the low interpretation rate of MS/MS spectra [13,14].

To address this problem, many modification-tolerant database search algorithms were proposed in recent years. MS-Alignment uses a dynamic programming approach to compare a spectrum against the sequence database without any specified modification, and peptides with one or more unknown modifications can be discovered [15,16].

* Corresponding author.
E-mail address: smhe@ict.ac.cn (S.-M. He).

http://dx.doi.org/10.1016/j.jprot.2015.05.009
1874-3919/© 2015 Elsevier B.V. All rights reserved.

Interrogator constructs an index of *b*- and *y*-ions to speed up database search and to look for a single unspecified modification on a peptide [17]. There are a few other algorithms or tools aimed at unrestricted database search, such as Protein Prospector [13,18], PTMap [19], P-Mod [20] MassShiftFinder [21] and TwinPeaks [22]. In most of them, the mass of the precursor ions is no longer used to restrict the scope of candidate peptides, thus dramatically expanding the searching space for each spectrum. As such, it is less and less efficient as the database increases. However, the essence of this strategy, which is treating a modification as a mass shift in MS or MS/MS data, is widely used in the unrestricted database search algorithms in other forms. For example, in ModifiComb, the mass shift between the precursors of non-modified and potentially modified peptides and their retention time information are used for discovering unspecified modifications [23]. DeltAMT is an improved approach, which considers all spectrum pairs regardless of whether they are identified or not [24]. The spectral network approach proposed by Bandeira et al. takes advantage of modified and unmodified peptide pairs to reduce noise peaks and hence improves the efficiency of peptide identification [25,26]. Spectral library searching algorithms, such as pMatch [27] and QuickMod [28], are also based on finding spectrum pairs with mass shifts that can be interpreted as unknown modifications.

Another type of unrestricted database search tries to reduce the size of a protein database using an iterative strategy. For example, the error-tolerant search mode of Mascot enables users to search peptides with one unspecified modification against the protein list obtained from a regular search result [29]. X! Tandem also uses a similar strategy to enable more modifications in the refined search [4]. PeaksPTM is another algorithm that identifies single-site modified peptides by considering all modification types in the Unimod database [30], followed by another round of search looking for two or more common modifications per peptide among identified sequences [31]. Iterative searching effectively reduces the database in most cases, but the false discovery rate (FDR) may be underestimated in the second round. Hence, Bern et al. proposed a method that gives a conservative estimation of FDR in multi-stage database search [32]. Another problem of the iterative strategy is that the result depends on the quality of the first-round search; true positives may not be identified if the initial setting is not optimized.

Moreover, others attempt to reduce the search space by using information from MS/MS data, i.e., the continuity of fragment ions. A few algorithms use extracted sequence tags or full-length de novo reconstructions to filter candidate peptides. The tag-based approach was first proposed by Mann and Wilm [33], and improved fast in recent years [34–36]. Many tag-based database search tools are now available, such as InsPecT [14], Paragon [37], SPIDER [38], ByOnic [39], spectral dictionaries [40] and MOD[i] [41]. In addition, PEAKS DB incorporates the de novo sequencing results to improve the sensitivity and accuracy of the database search effectively [42]. This is a hybrid approach of de novo sequencing and database search and it is able to find unspecified modifications via partial sequence information, but it requires high-quality MS/MS data and accurate extraction of tags.

Although the methods discussed above are effective, the development of unrestricted search is still challenging. Firstly, unrestricted database search tools are usually very time-consuming and thus are seldom used in routine experiments. Besides, most of them are designed to identify peptides bearing perfect protease specificity, e.g. fully tryptic peptides. However, semi- and non-tryptic peptides or the like may account for a large proportion of MS/MS spectra. As a result, most of the unrestricted search strategies thus far have not succeeded in exploring the search space fully. On the other hand, open modification database search can lead to many high-scoring but incorrect identifications, and how to evaluate them remains a problem. If too many modifications are considered simultaneously, a true match may surrender to a false one, for example, a sequence with rare, improbable modifications might score higher than the plain, "ground truth" peptide. A few state-of-the-art evaluation algorithms, such as PeptideProphet [43,44], Percolator [45,46], SEPro [47] and iProphet [48], are not aimed at unrestricted search and have not taken full advantage of the occurrence frequency of different types of modifications.

In this paper, we present a novel algorithm, pFind–Alioth (hereinafter referred to as Alioth), to address the unrestricted database search problem with high resolution MS/MS data. A fragment ion index is constructed for theoretical precursors and their fragment ions. Different from the work of Tang et al. [17], in which a *b*- and *y*-ion table is constructed to speed up the database search, here we use a new and compact structure to store all peptides and their neutral fragment ions. For each spectrum, a list of queries is generated to retrieve sequence values through the index. All the peptides contained in a protein database, with or without protease specificity, and with one modification of any kind are all covered in the search space. To evaluate the quality of peptide–spectrum matches (PSMs), we used an iterative algorithm to re-rank the candidate peptides based on the estimated *p*-values of PSMs [8] and the occurrence probabilities of the candidate peptides. Compared with other existing tools, such as pFind (with KSDP scoring function and the traditional workflow [7,10]) and Mascot, Alioth shows superior performance and it is also better than the unrestricted search tools evaluated in our study, such as InsPecT (blind search mode) and Mascot (error tolerant search mode).

2. Methods

2.1. Constructing fragment ion index

A neutral fragment ion index table is constructed for a given protein database. For each protein sequence, all sub-sequences within a specified length and mass range are generated. For instance, given a peptide sequence AEHVAEADK, whose length is 9, the number of all its sub-sequences with length between 2 and 9 is 36. All sub-sequences generated from the protein database are sorted by their masses in an ascending order and stored in a datasheet. Then an index table is constructed in which the key is the mass of each peptide. Fig. 1 shows in details the construction of an ion index. In practice, the keys need not be stored due to their continuity. Each ion is represented by three integers: protein ID, start position and the amino acid (aa) length. Therefore, all ions are recorded in the datasheet with equal number of bytes. As shown in Fig. 1, given an explicit mass or mass range to be queried, the time complexity of finding the first valid position in the datasheet is $O(1)$.

Unlike the previous approaches, Alioth stores full peptide sequences and their neutral fragment ions in the same way. This greatly compresses the storage space, especially for peptides from non-specific in silico digestion. N- and C-terminal ions are uniformly stored in a single table rather than in different tables for specified ion types, therefore the storage space is further reduced and it is universal for searching different types of MS/MS data, e.g., CID and ETD. Given any two sequences in the database, s_1 and s_2, it is very easy to judge whether s_1 is the prefix or suffix of s_2 using the field values (protein ID, start position and length in aa) stored in the datasheet, which is essential in the querying step described below.

2.2. Generating queries

In the Alioth algorithm, each spectrum to be identified is transformed into a list of queries as follows. Firstly, isotopic and precursor-related peaks are removed and the detected monoisotopic peaks are transformed into singly-charged ions according to their charge states [49]. Secondly, k most intense peaks are picked out from the spectrum, where k is an empirical parameter and in this study it is set to 30. Each peak p can be represented using an $<m, i>$ tuple, in which m and i denote the singly charged m/z value and peak intensity of p, respectively. Thirdly, given the assumption that *b*- and *y*-ions are considered in the

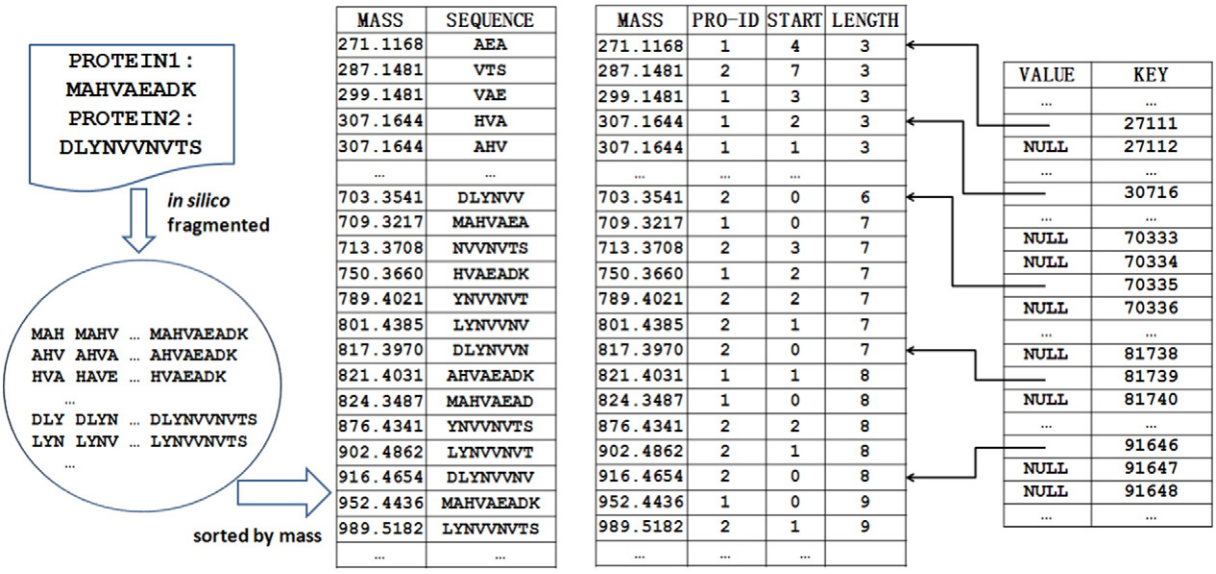

Fig. 1. Construction of the fragment ion index. Proteins are fragmented in silico into short subsequences, and then all subsequences are sorted by their masses in an ascending order and organized as a datasheet. At last, an index table is constructed, which is a list of (key, value) pairs. Each key denotes a valid mass that has been transformed into an integer, and the corresponding value to each key is a pointer to the first valid position in the datasheet.

MS/MS data, each selected peak p: <m, i> is transformed into four queries:

Q_1 : <$m - PROTON, i, Nterm$>,
Q_2 : <$m - H_2O - PROTON, i, Cterm$>,
Q_3 : <$comp(m) - PROTON, i, Nterm$ > and
Q_4 : < $comp(m) - H_2O - PROTON, i, Cterm$ > .

In the representation of the queries above, H_2O and $PROTON$ denote the mass of a water molecule and that of a proton, respectively, and $comp(m)$ denotes the m/z value of the complementary peak of p. That is, if the singly charged precursor mass is M, then $comp(m)$ equals to the value of $(M + PROTON - m)$. If a spectrum can be interpreted as a peptide with up to one modification on an arbitrary site, for every cleavage site in the peptide sequence, either the N- or the C-terminal ion, or both, are unmodified. For example, if the peak is a b-ion with an unknown modification, then its complementary peak, regardless of whether it exists in the real spectrum or not, is a non-modified y-ion, which is the basis of the generation of Q_4. Therefore, the four queries guarantee that for any b- or y-ion carrying up to one modification, its unmodified neutral form from the same cleavage is surely to be generated and later retrieved through the index.

At last, the experimental precursor ion can also be transformed into a query:

Q_{pcr} : <$m_p - H_2O - PROTON, i_p, Nterm/Cterm$>,

where m_p denotes the mass of the singly charged precursor ion and i_p takes the intensity value of the base peak in the tandem mass spectrum. For short and unmodified peptides, this query is especially helpful to retrieved them successfully.

2.3. Retrieving candidate partial sequences

As described above, given k peaks selected from a spectrum, a query list containing $4k + 1$ queries is generated. Next, for each query in this list, all of the neutral fragments whose masses fall within a specified tolerance window are retrieved from the index table. At the same time, an N-terminal result list and a C-terminal result list are constructed to store the results retrieved by the queries. For the N-terminal result list, the retrieved neutral fragments are marked uniformly using their start aa positions. Those sharing the same start position in the same protein are gathered as an intermediate result and merged into a single item which takes the summed weight of all the constituents. Each item in the final N-terminal result list contains the start position and the summed weight. Fig. S1 shows the details; for example, three retrieved items from the N-terminal query 2, 3 and 12 share the same protein ID and start position, thus they are merged in the N-terminal result list. The construction of the C-terminal result list is similar except that the retrieved neutral fragments with the same end position in the same protein are gathered and merged. After all the queries in the list are searched against the index table, the top-ranked partial sequences in the two result lists are selected and scored further.

2.4. Generating full-length peptides and scoring

With each partial sequence, the full-length candidate peptides can be generated according to the specified modification list, e.g., the entire Unimod database and any amino acid substitutions. For example, for a partial sequence with a fixed N-terminus, we can enumerate all of its valid C-termini to keep the mass shift d between the mass of the peptide and that of the precursor within a specified range. Then the possible modifications can be looked up from a specified modification list that fits the mass of d. If d can be interpreted as at least one modification, we add the mass shift d as a modification to the peptide sequence and then score the PSM using the KSDP scoring function in pFind [7].

The PSM score is an important feature to measure the quality of matching. However, the PSM score alone cannot determine if a sequence is the correct answer for a spectrum. For example, given two peptide candidates: A: K.AEHVAEADCKG.T and B: K.AEHVAEADCK.G with a modification of carbamidomethylation on Cys, whose mass is absolutely equal to the residue mass of Gly, and then search engines are needed to distinguish them. Peptide B is correct; however, both peptides matched the spectrum with high scores and peptide A may score higher than B depending on which scoring function is used. Such error cannot be fully attributed to the design of the scoring function because the two peptides share the same sequence for the most part and there may be very little evidence in the spectrum to distinguish them. However, there was prior knowledge that the MS/MS data originated from a sample that, before being digested by trypsin, was reduced and alkylated with iodoacetamide, which would attach a carbamidomethyl group to most cysteines, so peptide B would be a more probable answer.

In other words, additional information (data about data, or metadata), rather than the PSM score alone, helps model the real problem and distinguish true positives. In this study, we present a new re-ranking algorithm which gives a more comprehensive scoring function to PSMs.

The new scoring function *EV-Score* is defined using the following equation:

$$EV\text{-}Score(\mathrm{PSM}_{pep,\ spec}) = \frac{p \text{ value of the PSM}}{P_{oc}(pep)},$$

where P_{OC} denotes the estimated occurrence probability of the peptide sequence in the PSM.

The *p* value of each PSM can be estimated using previously reported algorithms [8,50]. For $P_{OC}(pep)$, we can assume that the occurrence of a certain modification on a certain amino acid and the type of digestion are independent of one another, so the occurrence probability of modification and digestion specificity can be estimated based on their occurrence frequency in the dataset being analyzed, e.g. the occurrence probability of each amino acid, type of modification, and the specificity of protease digestion (fully-specific, semi-specific on the N-terminus or C-terminus, or non-specific).

The following procedure shows how to learn P_{OC} in an iterative fashion for every peptide in the database search result. In order to simplify the description, only the occurrence probability of each type of modification is taken into account.

1) Search against the combined target-decoy database and select every top-ranked PSM. All of the PSMs are sorted by their *EV-Score*. P_{OC} is set to 1 for each peptide initially.
2) Using the traditional target-decoy strategy to get a result set *D* under a specified FDR level, e.g., 1%.
3) For each type of modification mod_{aa} that occurs on the amino acid *aa*, the occurrence probability of the modification can be estimated as follows:

$$P_{OC}(mt_{aa}) = \frac{\text{Frequency of } mod_{aa} \text{ in } D}{\text{Frequency of } aa}.$$

The occurrence probability of regular amino acids without any modification can be similarly calculated using the equation above via a *null* modification type that occurs on all of the regular amino acids.
4) Assuming the independence between different types of modifications, the probability of a peptide can be calculated using the following equation:

$$P_{OC}(pep) = \prod_{i=1}^{\text{length of pep}} P_{OC}(mod_{pep(i)}),$$

where $pep(i)$ denotes the *i*th amino acid in the sequence of *pep*.
5) Re-calculate the *EV-Score* and then re-rank all PSMs according to their *EV-Scores*, and then go back to step 2. If there are no changes in the rank of all PSMs or the number of the iterations exceeds a specified parameter *t*, the procedure will be terminated.

Fig. S2 shows the change of the distribution of the target and decoy PSMs in different iterations.

2.5. Combining results from restricted database search

Although unspecified modifications can be automatically detected in Alioth, only one modification site is allowed for each candidate peptide. However, there are peptides that have two or more modifications. Thus Alioth is followed by a restricted database search using the traditional settings, i.e., fully-specific digestion and a few common modifications. The workflow of the restricted searching is similar to that shown in Ref. [10] but we made some slight changes. Firstly, the modifications specified in it are learned from Alioth, that is, *k* most abundant modifications are added into the following database search. Secondly, the preprocessing and re-ranking algorithms are the same as Alioth. Then the results from Alioth and the restricted database search are merged together. For each spectrum, the peptide with the best *EV-Score* is reported and further analyzed using the target-decoy strategy [11,51].

3. Experiment and result

3.1. Mass spectrometry and data sets

Two biological samples are used in our study. The first one is from a whole-cell lysate of *Thermoanaerobacter tengcongensis* which is digested by trypsin and then analyzed on a LTQ-Orbitrap Velos mass spectrometer. HCD (mass range 100–2000) is used for the generation of MS/MS spectra. *T. tengcongensis* cells were cultured at four different temperatures (55 °C, 65 °C, 75 °C and 80 °C) and a sample was generated at each temperature, leading to four datasets (TTE-55, TTE-65, TTE-75 and TTE-80, each containing 12 RAW files). The numbers of MS/MS spectra in these four datasets are 128,481, 113,531, 117,209 and 127,190, respectively. TTE-55 and TTE-65 are mainly used for the detailed analysis of the algorithm precision in this paper.

The second one is from a whole-cell lysate of *Caenorhabditis elegans* [52]. Two enzymes, Trypsin and Asp-N, are used separately to digest proteins. Then the two digests are analyzed on a LTQ-Orbitrap XL mass spectrometer using a 6-step *Mu*lti-*d*imensional *P*rotein *I*dentification *T*echnology (MudPIT) [53]. In this experiment, the two most intense precursor ions from each full scan were isolated to generate five MS/MS spectra for each: low-mass HCD (mass range 50–2000), HCD (mass range 100–2000), CID detected in LTQ, ETD detected in LTQ and ETD detected in Orbitrap. Two HCD spectra from the same scan are merged together for further analysis. Two datasets, 12,488 HCD spectra from the trypsin digest and 11,288 HCD spectra from the Asp-N digest, are named as WORM-TRYP and WORM-ASPN respectively.

3.2. Database search

The Alioth algorithm, as well as Mascot (version 2.2) and pFind (version 2.6, in which KSDP scoring function is used [7]), was tested on the datasets described above. MS/MS spectra were extracted using an in-house tool of pFind Studio named pXtract. The database search parameters are shown in Table S1. For the *T. tengcongensis* data, we search against the original proteome database plus the six-frame translation database, which aimed at discovering novel genes [54,55]. For each target database, two shuffled decoy databases are generated, one for learning peptide occurrence probability and the other for FDR estimation. In the whole workflow of Alioth, the protein database is split into a few parts first and the total length of proteins in each part is no more than 1,000,000. Then for each partial database, the fragment ion index is constructed in memory and all of the MS/MS spectra to be identified are searched against the index. No more than 30 peaks are selected to generate queries. Up to 100 intermediate partial sequences are kept for each spectrum to generate full-length candidate peptides. We chose all of the modifications recorded in the Unimod database and all possible substitutions between any two different amino acids. After removing redundant items, a list containing 1362 modifications and amino acid substitutions are generated for Alioth.

The result from our algorithm was compared with Mascot and pFind in three search modes: regular, complex mode I, and complex mode II. In the regular search mode, full enzyme specificity and a few common modifications are specified, which is to simulate a routine proteomic analysis. In complex mode I, peptides due to semi- and non-specific digestion are first identified by Alioth+DB and then integrated into the original database, so these sequences are included in the Mascot and pFind searches when a fully specific digestion mode is applied. In complex mode II, the nine (the maximum allowed number of variable modifications in Mascot) most abundant modifications identified by Alioth are specified as variable modifications, which is the same setting

used in the restricted database search following Alioth search. For Mascot, we use Mascot Percolator [56] to extract true positive identifications and calculate *q*-values in the restricted searching mode. We did not show the comparison between Alioth and others in a real semi- or non-specific digestion mode because the running time of both Mascot and pFind was too long and the Mascot program finally became unresponsive in this condition.

3.3. Analysis of performance on the T. tengcongensis data

Shown in Fig. 2 is the comparison of Mascot, pFind and Alioth search results on the datasets TTE-55 and TTE-65. At 1% FDR, Alioth reported 94,099 PSMs from TTE-55, with a further increase to 98,283 PSMs when Alioth is combined with a follow-up restricted database search (Alioth+DB). This is 89.4% and 87.6% more than pFind (52,681) and Mascot (53,197) in the regular searching mode, respectively (Fig. 2a). On TTE-65, Alioth+DB also yielded the most identifications: 134.7% more than pFind and 125.2% more than Mascot (Fig. 2b). The spectrum interpretation ratio was ~41% using either Mascot or pFind, similar to what is reported recently on HCD data from LTQ Orbitrap Velos [12]. In contrast, 77.7% spectra in TTE-55 and 62.5% spectra in TTE-65 can be interpreted using Alioth+DB at the same 1% FDR cutoff, showing that the algorithm is an effective approach to improving spectrum interpretation rate.

To validate Alioth search results, we further compared them to the research results of pFind and Mascot in two complex search modes (Fig. 2c and d). Both pFind and Mascot identified many more PSMs in complex search modes than in the regular search mode. In complex search mode I, semi-specific peptides identified by Alioth-DB are added to the protein database used by pFind and Mascot. In complex search mode II, the nine most abundant modifications are set as variable modifications, based on the Alioth search results. The result of Mascot is slightly higher than that of pFind in both complex search modes. Alioth+DB still yielded ~10% and 3% more identifications on TTE-55 and TTE-65, respectively, compared to the Mascot result in complex search mode II. In addition, a vast majority of the reported PSMs by Alioth+DB are supported by the search results of other search engines (Fig. 2e and f). For example, 79.5% of the PSMs are identified from TTE-55 by all three algorithms, and 83.3% of the Alioth results are supported by at least one other search engine.

We compared the proportion of fully-, semi- and non-specific peptides between TTE-55 and TTE-65. As shown in Fig. S4a and b, semi- and non-specific peptides accounted for 38.3% of the identified spectra from TTE-65, more than double the percentage of semi- and non-specific peptides (17.3%) in TTE-55. As such, the presence of many more peptides from irregular digestion is likely the reason why the spectrum interpretation ratio achieved by Alioth is a bit lower on TTE-65 (62.5%) than on TTE-55 (77.7%). For Mascot and pFind, the peptides from irregular digestion are assumed to be from fully specific digestions and added to the original protein database. Thus, the PSM scores of these peptides are not influenced by their digestion specificity. However, the scores of peptides from semi- and non-specific digestion are penalized compared with the fully specific ones in Alioth. This is probably why the numbers of reported PSMs by pFind or Mascot in the complex search modes are closer to those by Alioth and Alioth+DB on TTE-65 than TTE-55. On the other hand, the modifications detected in these two datasets are almost the same (Tables S2 and S3).

The result of Alioth is also compared with the other two unrestricted database search algorithms, the error tolerant search mode of Mascot and the blind search mode of InsPecT, which is shown in Fig. S3. On both datasets, the number of the target PSMs reported by Alioth is remarkably greater than the other two engines. Compared with the result of Mascot, Alioth reported 7.0% and 29.5% more target PSMs at 1% FDR on TTE-55 and TTE-65, respectively. In addition, the result of Mascot in the error tolerant mode is 3.4% and 23.4% less than that in complex mode II on TTE-55 and TTE-65, respectively (shown in Fig. 2c and d), which indicates that the huge searching space in the unrestricted search contains much more peptide candidates and probably influences the performance of Mascot. Therefore, it is important to consider more features such as modifications and digestion forms, rather than the

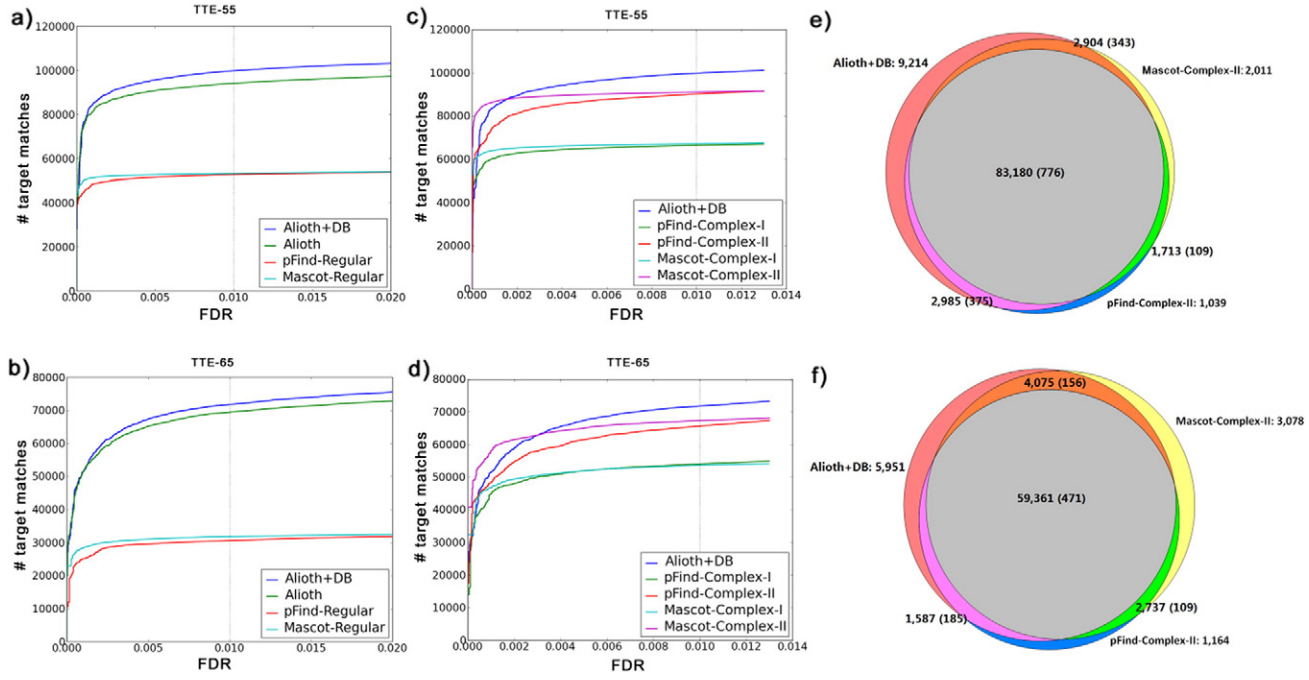

Fig. 2. Result comparison of Mascot, pFind and Alioth on the *T. tengcongensis* data. a) The number of correct PSMs as a function of FDR obtained by Alioth and Alioth+DB, as well as Mascot and pFind in the regular search mode on the dataset of TTE-55. b) Same as a) but on the dataset of TTE-65. c) The number of correct PSMs as a function of FDR obtained by Alioth+DB, as well as Mascot and pFind in the two complex search modes on the dataset of TTE-55. d) Same as c) but on the dataset of TTE-65. e) A Venn diagram that shows the consensus of the results confidently identified by Alioth+DB, Mascot and pFind in the complex search mode II on the dataset of TTE-55. The numbers in the parentheses indicate how many results are identified inconsistently. f) Same as e) but on the dataset of TTE-65.

quality of PSMs only, in result validation, especially for the unrestricted database search.

3.4. Analysis of the performance of Alioth on the C. elegans data

The comparison of Mascot, pFind and Alioth results is shown in Fig. 3. On the WORM-TRYP dataset, all of the algorithms can identify more than 60% of the total spectra. Alioth+DB reported 9523 PSMs, which is 12.6% more than pFind and 14.3% more than Mascot. The overall interpretation rate of the spectra is 76.3%. On the dataset of WORM-ASPN, the spectrum interpretation ratio of either pFind (43.4%) or Mascot (49.8%) is much reduced, while Alioth+DB still reported 7621 PSMs, which is 67.4% of the total spectra. On the WORM-TRYP dataset, pFind and Mascot identified over 65% of the spectra even in the regular search mode, closer to the Alioth and Alioth+DB results than on TTE-55 and TTE-65. In the WORM-TRYP dataset, 96.2% of the identified spectra are fully tryptic (Fig. S4c), and 76.1% of them are interpreted as peptides without modifications. Therefore, even in the regular search mode, most of the correct peptides are within the searching space. On the other hand, although the digestion and modification types are both unrestricted in Alioth, which increases the search space dramatically, Alioth still shows better performance than pFind and Mascot.

Curiously, fewer PSMs are reported by pFind in the complex search mode than pFind itself in the regular mode on both worm datasets (Fig. 3a and b). This is opposite to what is observed on the TTE-55 and TTE-65 datasets (compare Fig. 2a to c, and b to d). It may be partly explained by the fact that more than 95% of the worm peptides show perfect digestion specificity (Fig. S4c and d) and a very high percentage of the reported PSMs, 76.1% from WORM-TRYP and 84.8% from WORM-ASPN, are interpreted as unmodified peptides. So, although a few peptides with specified modifications or digestion specificity are gained using the complex search mode, a great number of theoretically possible but actually incorrect candidate sequences result in a net loss of PSMs at the same FDR cutoff. On the contrary, Mascot reported a comparative number of positive PSMs in the complex search mode because the information of modifications is taken into accounted in the Mascot percolator algorithm. In Alioth, the calculation of peptide occurrence probability also implies the consideration of different forms of digestions, as well as the different numbers and types of modifications. This is the reason why Alioth reported the most positive PSMs even when searching a huge space (i.e. arbitrary digestion forms, the Unimod database and all possible amino acid substitutions are all considered simultaneously).

Aside from the HCD spectra, we also acquired in the *C. elegans* datasets high resolution ETD data, and normal resolution ETD and CID data from the same precursors. Therefore, they can be used to validate the reliability of the Alioth+DB search results on HCD. As shown in Fig. 4, we compared the result of Alioth+DB with five other results: the result of Mascot on the HCD data and the result of pFind on the HCD, CID, ETD-LTQ and ETD-Orbitrap data. The consensus between the Alioth result and five other result groups is shown in Fig. 4a and c. As shown in Fig. 4a, 96.5% of the results can be validated by at least one result group on the WORM-TRYP data. The *q*-values of the remaining 3.5% target PSMs are distributed closer to zero than the *q*-values of the decoy PSMs (Fig. 4b, compare 0 and 6). In addition, 41.9% of the peptides in these target PSMs are also reported in the validated results (data not shown). From the WORM-ASPN dataset, only 82.7% of the results can be validated by at least one result group (Fig. 4c). However, as shown in Fig. 4d, the distribution of the *q*-values of the remaining 17.3% PSMs not validated by other result groups is much closer to zero than the decoy PSMs, just like the *q*-value distributions of the validated PSMs (Fig. 4d). These results suggest that most of these PSMs have much higher scores than decoy PSMs and are likely reliable.

3.5. Comparison of running time

Table 1 shows the running time of Mascot, InsPecT, pFind and Alioth in different search modes. Both pFind and Mascot in the regular search mode are faster than Alioth and Alioth+DB. In the complex search mode, pFind is 5–10 times faster than Mascot and still 3–4 times faster than Alioth+DB. We would like to point out that the search space explored by Alioth+DB is much larger than that by Mascot or pFind even in complex searching mode II. Non-specific digestion results in ~100 times more peptides, while the number of modifications and mutations considered in our algorithm is ~150 times more than that considered in pFind or Mascot. Under such condition, Alioth+DB is at least 10% faster on the worm datasets compared to Mascot in complex search mode II and even three times as fast as Mascot on the TTE datasets. If the real semi-specific digestion mode and nine most abundant modifications are specified, both pFind and Mascot become significantly slow and the reported target PSMs at 1% FDR are less than in complex mode II. In addition, the running time of our algorithm is compared with two unrestricted search algorithms: the blind search mode of InsPecT and the error tolerant search mode of Mascot. Up to one modification site is allowed per peptide in all of these algorithms. As shown

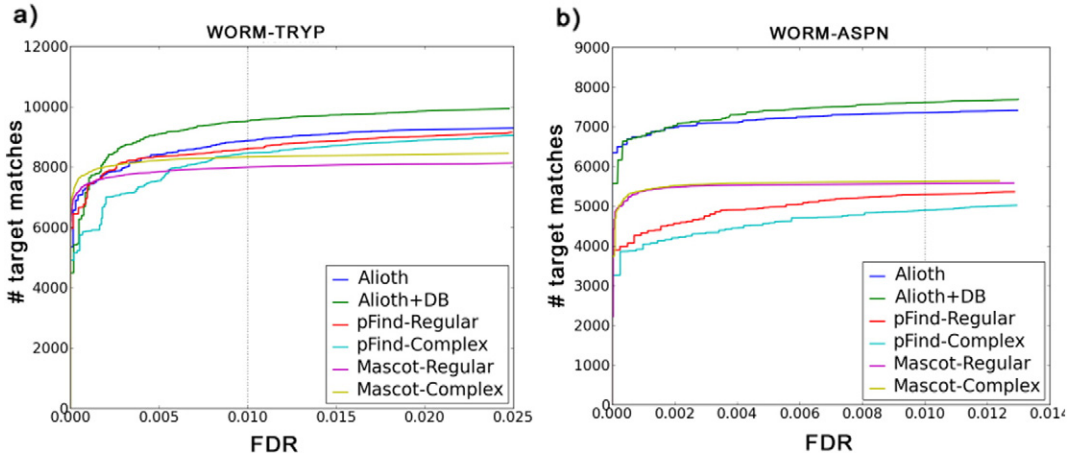

Fig. 3. Result comparison of Mascot, pFind and Alioth on the *C. elegans* data. a) The number of correct PSMs as a function of FDR obtained by Alioth and Alioth+DB, as well as Mascot and pFind in both the regular and complex search modes on the dataset of WORM-TRYP. The complex search mode is the same as the complex search mode II mentioned in Fig. 2. b) Same as a) but on the dataset of WORM-ASPN.

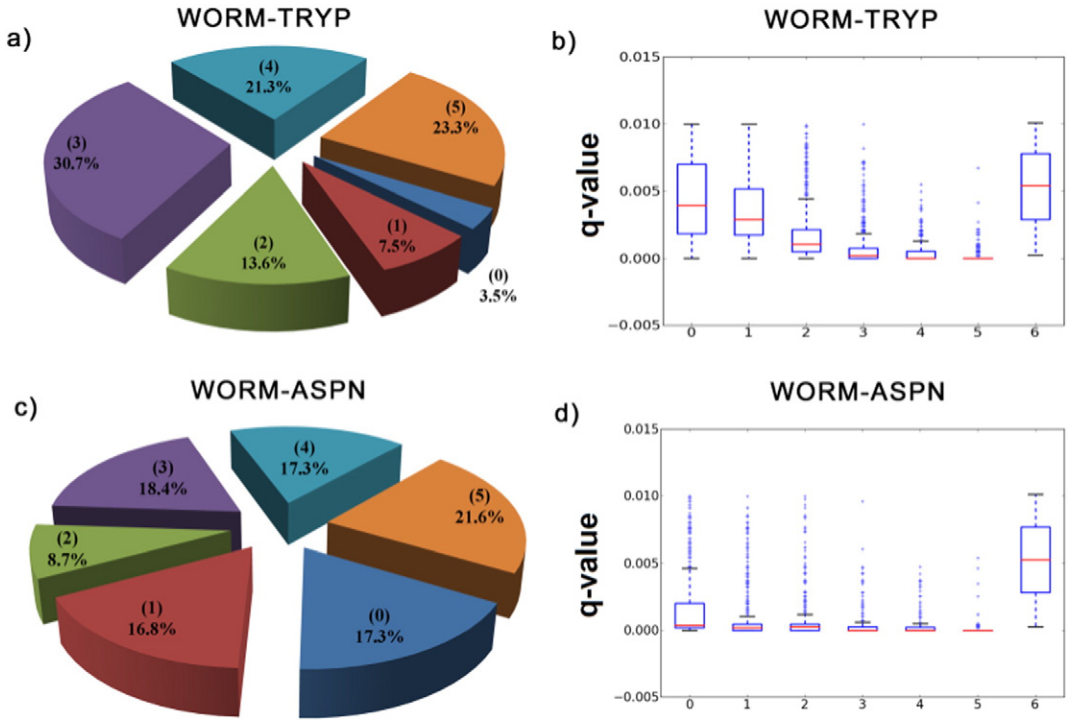

Fig. 4. Validation of the result of Alioth+DB on the *C. elegans* data. Five sets of results are used for validation: the search result of pFind on the HCD-FTMS, CID-ITMS, ETD-ITMS and ETD-FTMS data, as well as the search result of Mascot on the HCD-FTMS data. Then for each PSM identified by Alioth+DB, it is counted how many times this identification is supported by other result groups, that is, whether the same peptide is identified by other search engines from the same spectrum or its cognate spectra from the same precursor. a) The proportions of results with different levels of reliability. The numbers in the parentheses indicate how many result groups support the PSMs. For example, 23.3% of the total PSMs are supported by all of the five result sets. b) Box plots that show the distributions of *q*-values from different sections corresponding to a). The rightmost box plot shows the distribution of *q*-value from the decoy PSMs. c) Same as a) but on the dataset of WORM-ASPN. d) Same as b) but on the dataset of WORM-ASPN.

in Table 1, the running time of Alioth is about 5 and 10 times faster than the other two algorithms.

4. Discussion

In this paper, we describe a novel unrestricted database search algorithm, Alioth, which features a fragment ion indexing technique and an efficient retrieval approach to speed up database search with arbitrary forms of enzymatic digestion and thousands of modification types and mutations. In addition, an iterative strategy is used to learn the occurrence probability of candidate peptides. Alioth can be combined with a regular database search, which further improved the overall performance. Compared with a few other restricted and unrestricted database search tools, Alioth performs favorably on multiple test datasets containing ~500,000 MS/MS spectra in total.

It should be noted that Alioth is not a blind database search algorithm but rather based on what is recorded in the Unimod database, similar to PeaksPTM [31]. However, only a slight modification is needed for Alioth to fit the demand of blind search: in the step of candidate generation, the mass shifts in Unimod, as well as unknown mass shifts, can all be taken into account as potential modifications. However, the time cost will surely increase. Actually, the modifications in Unimod are sufficient for proteomic analysis in most cases. On the other hand, irregular digestion, especially the semi-specific digestion, is not uncommon in sample preparation and can lead to a sharp increase of search space. However, specifying digestion or modification types is sometimes risky because such information is unknown before analyzing the MS/MS data. In this case, Alioth can provide the ability to search a much larger space and, more importantly, view the MS/MS data from a global perspective and gather sample specific information such as digestion and modifications.

We have also tested the algorithm with normal mass resolution MS/MS data, but there is no obvious improvement compared with the traditional database search engines. The main reason is that because of reduced precision and accuracy, a peak (query) in a normal resolution MS/MS spectrum retrieves too many theoretical fragment ions. Therefore, it is much more difficult to extract the correct peptide for each spectrum. Besides, the running time and memory demand increase sharply. However, in high mass accuracy and high resolution MS/MS data, the correct peptides are easier to be distinguished even in unrestricted database search. Because high resolution and high mass accuracy mass spectrometers are widely used nowadays, we believe that fast, accurate, and unrestricted database search will be frequently used or even become routine in future high-throughput proteome analyses.

Table 1
Comparison of the running time among the database search algorithms in different search modes (in minutes)[a].

	Running time (in minutes)		
	WORM-TRYP (12,488)[b]	WORM-ASPN (11,288)	TTE (486,411)
Alioth	36	33	643
Alioth+DB	66	65	1449
pFind-regular	5	5	136
pFind-complex-II	17	16	440
Mascot-regular	7	5	148
Mascot-complex-II	74	86	4460
InsPecT (blind search)	–	–	6540
Mascot (error tolerant)	–	–	3552

[a] In the experiment reported in this the paper, we used a Dell PC, which has an Intel Core i5 CPU @ 2.90 GHz and 4 GB memory.
[b] The number in the parentheses indicates how many MS/MS spectra are contained in the corresponding dataset.

Now Alioth can be downloaded from http://pfind.ict.ac.cn/software/Alioth/index.html.

Acknowledgments

This work was supported by the National Key Basic Research and Development Program of China (973) under Grant Nos. 2013CB911203 and 2012CB910602 to R.-X.S. and 2010CB912701 to S.-M.H. 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 CAS Knowledge Innovation Program under Grant Nos. KGCX1-YW-13 and ICT-20126033. The authors thank Long Wu and Wen-Feng Zeng from the Institute of Computing Technology, Chinese Academy of Sciences, for their valuable discussions.

Appendix A. Supplementary data

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2015.05.009.

References

[1] R. Aebersold, M. Mann, Mass spectrometry-based proteomics, Nature 422 (6928) (2003) 198–207.
[2] D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell, Probability-based protein identification by searching sequence databases using mass spectrometry data, Electrophoresis 20 (18) (1999) 3551–3567.
[3] J. Eng, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database, J. Am. Soc. Mass Spectrom. 5 (11) (1994) 976–989.
[4] R. Craig, R.C. Beavis, A method for reducing the time required to match protein sequences with tandem mass spectra, Rapid Commun. Mass Spectrom. 17 (20) (2003) 2310–2316.
[5] R. Craig, R.C. Beavis, TANDEM: matching proteins with tandem mass spectra, Bioinformatics 20 (9) (2004) 1466–1467.
[6] L.Y. Geer, S.P. Markey, J.A. Kowalak, L. Wagner, M. Xu, D.M. Maynard, X. Yang, W. Shi, S.H. Bryant, Open mass spectrometry search algorithm, J. Proteome Res. 3 (5) (2004) 958–964.
[7] Y. Fu, Q. Yang, R. Sun, D. Li, R. Zeng, C.X. Ling, W. Gao, Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry, Bioinformatics 20 (12) (2004) 1948–1954.
[8] L.H. Wang, D.Q. Li, Y. Fu, H.P. Wang, J.F. Zhang, Z.F. Yuan, R.X. Sun, R. Zeng, S.M. He, W. Gao, pFind 2.0: a software package for peptide and protein identification via tandem mass spectrometry, Rapid Commun. Mass Spectrom. 21 (18) (2007) 2985–2991.
[9] J. Cox, N. Neuhauser, A. Michalski, R.A. Scheltema, J.V. Olsen, M. Mann, Andromeda: a peptide search engine integrated into the MaxQuant environment, J. Proteome Res. 10 (4) (2011) 1794–1805.
[10] Y. Li, H. Chi, L.H. Wang, H.P. Wang, Y. Fu, Z.F. Yuan, S.J. Li, Y.S. Liu, R.X. Sun, R. Zeng, et al., Speeding up tandem mass spectrometry based database searching by peptide and spectrum indexing, Rapid Commun. Mass Spectrom. 24 (6) (2010) 807–814.
[11] J.E. Elias, W. Haas, B.K. Faherty, S.P. Gygi, Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations, Nat. Methods 2 (9) (2005) 667–675.
[12] A. Michalski, J. Cox, M. Mann, More than 100,000 detectable peptide species elute in single shotgun proteomics runs but the majority is inaccessible to data-dependent LC–MS/MS, J. Proteome Res. 10 (4) (2011) 1785–1793.
[13] R.J. Chalkley, P.R. Baker, L. Huang, K.C. Hansen, N.P. Allen, M. Rexach, A.L. Burlingame, Comprehensive analysis of a multidimensional liquid chromatography mass spectrometry dataset acquired on a quadrupole selecting, quadrupole collision cell, time-of-flight mass spectrometer: II. New developments in Protein Prospector allow for reliable and comprehensive automatic analysis of large datasets, Mol. Cell. Proteomics 4 (8) (2005) 1194–1204.
[14] S. Tanner, H. Shu, A. Frank, L.C. Wang, E. Zandi, M. Mumby, P.A. Pevzner, V. Bafna, InsPecT: identification of posttranslationally modified peptides from tandem mass spectra, Anal. Chem. 77 (14) (2005) 4626–4639.
[15] D. Tsur, S. Tanner, E. Zandi, V. Bafna, P.A. Pevzner, Identification of post-translational modifications via blind search of mass-spectra, Proc. IEEE Comput. Syst. Bioinform. Conf. 2005, pp. 157–166.
[16] S. Tanner, P.A. Pevzner, V. Bafna, Unrestrictive identification of post-translational modifications through peptide mass spectrometry, Nat. Protoc. 1 (1) (2006) 67–72.
[17] W.H. Tang, B.R. Halpern, I.V. Shilov, S.L. Seymour, S.P. Keating, A. Loboda, A.A. Patel, D.A. Schaeffer, L.M. Nuwaysir, Discovering known and unanticipated protein modifications using MS/MS database searching, Anal. Chem. 77 (13) (2005) 3931–3946.
[18] R.J. Chalkley, P.R. Baker, K.F. Medzihradszky, A.J. Lynn, A.L. Burlingame, In-depth analysis of tandem mass spectrometry data from disparate instrument types, Mol. Cell. Proteomics 7 (12) (2008) 2386–2398.
[19] Y. Chen, W. Chen, M.H. Cobb, Y. Zhao, PTMap — a sequence alignment software for unrestricted, accurate, and full-spectrum identification of post-translational modification sites, Proc. Natl. Acad. Sci. U. S. A. 106 (3) (2009) 761–766.
[20] B.T. Hansen, S.W. Davey, A.J. Ham, D.C. Liebler, P-Mod: an algorithm and software to map modifications to peptide sequences using tandem MS data, J. Proteome Res. 4 (2) (2005) 358–368.
[21] H. Barsnes, S.O. Mikalsen, I. Eidhammer, Blind search for post-translational modifications and amino acid substitutions using peptide mass fingerprints from two proteases, BMC Res. Notes 1 (2008) 130.
[22] M. Havilio, A. Wool, Large-scale unrestricted identification of post-translation modifications using tandem mass spectrometry, Anal. Chem. 79 (4) (2007) 1362–1368.
[23] M.M. Savitski, M.L. Nielsen, R.A. Zubarev, ModifiComb, a new proteomic tool for mapping substoichiometric post-translational modifications, finding novel types of modifications, and fingerprinting complex protein mixtures, Mol. Cell. Proteomics 5 (5) (2006) 935–948.
[24] Y. Fu, L.Y. Xiu, W. Jia, D. Ye, R.X. Sun, X.H. Qian, S.M. He, DeltAMT: a statistical algorithm for fast detection of protein modifications from LC–MS/MS data, Mol. Cell. Proteomics 10 (5) (2011) (M110 000455).
[25] N. Bandeira, Spectral networks: a new approach to de novo discovery of protein sequences and posttranslational modifications, Biotechniques 42 (6) (2007) 687 (689, 691 passim).
[26] N. Bandeira, D. Tsur, A. Frank, P.A. Pevzner, Protein identification by spectral networks analysis, Proc. Natl. Acad. Sci. U. S. A. 104 (15) (2007) 6140–6145.
[27] D. Ye, Y. Fu, R.X. Sun, H.P. Wang, Z.F. Yuan, H. Chi, S.M. He, Open MS/MS spectral library search to identify unanticipated post-translational modifications and increase spectral identification rate, Bioinformatics 26 (12) (2010) i399–i406.
[28] E. Ahrne, F. Nikitin, F. Lisacek, M. Muller, QuickMod: a tool for open modification spectrum library searches, J. Proteome Res. 10 (7) (2011) 2913–2921.
[29] D.M. Creasy, J.S. Cottrell, Error tolerant searching of uninterpreted tandem mass spectrometry data, Proteomics 2 (10) (2002) 1426–1434.
[30] D.M. Creasy, J.S. Cottrell, Unimod: protein modifications for mass spectrometry, Proteomics 4 (6) (2004) 1534–1536.
[31] X. Han, L. He, L. Xin, B. Shan, B. Ma, PeaksPTM: mass spectrometry-based identification of peptides with unspecified modifications, J. Proteome Res. 10 (7) (2011) 2930–2936.
[32] M. Bern, B.S. Phinney, D. Goldberg, Reanalysis of *Tyrannosaurus rex* mass spectra, J. Proteome Res. 8 (9) (2009) 4328–4332.
[33] M. Mann, M. Wilm, Error-tolerant identification of peptides in sequence databases by peptide sequence tags, Anal. Chem. 66 (24) (1994) 4390–4399.
[34] D.L. Tabb, Z.Q. Ma, D.B. Martin, A.J. Ham, M.C. Chambers, DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring, J. Proteome Res. 7 (9) (2008) 3838–3846.
[35] D.L. Tabb, A. Saraf, J.R. Yates III, GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model, Anal. Chem. 75 (23) (2003) 6415–6421.
[36] S. Sunyaev, A.J. Liska, A. Golod, A. Shevchenko, A. Shevchenko, MultiTag: multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry, Anal. Chem. 75 (6) (2003) 1307–1315.
[37] I.V. Shilov, S.L. Seymour, A.A. Patel, A. Loboda, W.H. Tang, S.P. Keating, C.L. Hunter, L.M. Nuwaysir, D.A. Schaeffer, 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 6 (9) (2007) 1638–1655.
[38] Y. Han, B. Ma, K. Zhang, SPIDER: software for protein identification from sequence tags with de novo sequencing error, J. Bioinform. Comput. Biol. 3 (3) (2005) 697–716.
[39] M. Bern, Y. Cai, D. Goldberg, Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry, Anal. Chem. 79 (4) (2007) 1393–1400.
[40] S. Kim, N. Gupta, N. Bandeira, P.A. Pevzner, Spectral dictionaries: integrating de novo peptide sequencing with database search of tandem mass spectra, Mol. Cell. Proteomics 8 (1) (2008) 53–69.
[41] S. Kim, S. Na, J.W. Sim, H. Park, J. Jeong, H. Kim, Y. Seo, J. Seo, K.J. Lee, E. Paek, MODi: a powerful and convenient web server for identifying multiple post-translational peptide modifications from tandem mass spectra, Nucleic Acids Res. 34 (Web Server issue) (2006) W258–W263.
[42] J. Zhang, L. Xin, B. Shan, W. Chen, M. Xie, D. Yuen, W. Zhang, Z. Zhang, G.A. Lajoie, B. Ma, PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification, Mol. Cell. Proteomics 11 (4) (2012) (M111 010587).
[43] H. Choi, A.I. Nesvizhskii, Semisupervised model-based validation of peptide identifications in mass spectrometry-based proteomics, J. Proteome Res. 7 (1) (2008) 254–265.
[44] A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold, Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search, Anal. Chem. 74 (20) (2002) 5383–5392.
[45] M. Spivak, J. Weston, L. Bottou, L. Kall, W.S. Noble, Improvements to the percolator algorithm for Peptide identification from shotgun proteomics data sets, J. Proteome Res. 8 (7) (2009) 3737–3745.
[46] L. Kall, J.D. Canterbury, J. Weston, W.S. Noble, M.J. MacCoss, Semi-supervised learning for peptide identification from shotgun proteomics datasets, Nat. Methods 4 (11) (2007) 923–925.
[47] P.C. Carvalho, J.S. Fischer, T. Xu, D. Cociorva, T.S. Balbuena, R.H. Valente, J. Perales, J.R. Yates 3rd, V.C. Barbosa, Search engine processor: filtering and organizing peptide spectrum matches, Proteomics 12 (7) (2012) 944–949.
[48] D. Shteynberg, E.W. Deutsch, H. Lam, J.K. Eng, Z. Sun, N. Tasman, L. Mendoza, R.L. Moritz, R. Aebersold, A.I. Nesvizhskii, iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates, Mol. Cell. Proteomics 10 (12) (2011) M111 007690.

[49] H. Chi, R.X. Sun, B. Yang, C.Q. Song, L.H. Wang, C. Liu, Y. Fu, Z.F. Yuan, H.P. Wang, S.M. He, et al., pNovo: de novo peptide sequencing and identification using HCD spectra, J. Proteome Res. 9 (5) (2010) 2713–2724.
[50] D. Fenyo, R.C. Beavis, A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes, Anal. Chem. 75 (4) (2003) 768–774.
[51] J.E. Elias, S.P. Gygi, Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry, Nat. Methods 4 (3) (2007) 207–214.
[52] H. Chi, H.F. Chen, K. He, L. Wu, B. Yang, R.X. Sun, J.Y. Liu, W.F. Zeng, C.Q. Song, S.M. He, et al., pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra, J. Proteome Res. 12 (2013) 615–625.
[53] M.P. Washburn, D. Wolters, J.R. Yates III, Large-scale analysis of the yeast proteome by multidimensional protein identification technology, Nat. Biotechnol. 19 (3) (2001) 242–247.
[54] K. Zhang, Identifying novel genes and gene refinements in *Thermoanaerobacter tengcongensis*, The 12th Korea–Japan–China Bioinformatics Training Course 2014, 2014.
[55] Q. Bao, Y. Tian, W. Li, Z. Xu, Z. Xuan, S. Hu, W. Dong, J. Yang, Y. Chen, Y. Xue, et al., A complete sequence of the *T. tengcongensis* genome, Genome Res. 12 (5) (2002) 689–700.
[56] M. Brosch, L. Yu, T. Hubbard, J. Choudhary, Accurate and sensitive peptide identification with Mascot Percolator, J. Proteome Res. 8 (6) (2009) 3176–3181.