Journal of proteome research

Article

pSite: Amino Acid Confidence Evaluation for Quality Control of De Novo Peptide Sequencing and Modification Site Localization

Hao Yang, Hao Chi, Wen-Jing Zhou, Wen-Feng Zeng, Chao Liu, Rui-Min Wang, Zhao-Wei Wang, Xiu-Nan Niu, Zhen-Lin Chen, and Si-Min He

J. Proteome Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.jproteome.7b00428 • Publication Date (Web): 13 Nov 2017

Downloaded from http://pubs.acs.org on November 15, 2017

ACS Publications

Journal of Proteome Research is published by the American Chemical Society. 1155 Sixteenth Street N.W., Washington, DC 20036

pSite: Amino Acid Confidence Evaluation for Quality Control of *De Novo* Peptide Sequencing and Modification Site Localization

Hao Yang[1,2§], Hao Chi[1§*], Wen-Jing Zhou[1,2], Wen-Feng Zeng[1,2], Chao Liu[1], Rui-Min Wang[1,2], Zhao-Wei Wang[1,2], Xiu-Nan Niu[1,2], Zhen-Lin Chen[1,2] and Si-Min He[1,2*]

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

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

[§] These authors contributed equally to this work.

[*] To whom correspondence authors should be addressed:

Hao Chi：

E-mail: chihao@ict.ac.cn

Si-Min He：

E-mail: smhe@ict.ac.cn

Abstract

MS-based *de novo* peptide sequencing has been improved remarkably with significant development of mass spectrometry and computational approaches, but still lacks quality control methods. Here we proposed a novel algorithm pSite to evaluate the confidence of each amino acid rather than the full-length peptides obtained by *de novo* peptide sequencing. A semi-supervised learning approach was used to discriminate correct amino acids from random ones and then an expectation-maximization algorithm was used to adaptively control the false amino-acid rate (FAR). On three test data sets, pSite recalled 86% more amino acids on average than PEAKS at the FAR of 5%. pSite also performed superiorly on the modification site localization problem, which is essentially a special case of amino acid confidence evaluation. On three phosphopeptide data sets, at the false localization rate of 1%, the average recall of pSite was 91% while those of Ascore and phosphoRS were 64% and 63%, respectively. pSite covered 98% of Ascore and phosphoRS results and contributed 21% more phosphorylation sites. Further analyses show that the use of distinct fragmentation features in high-resolution MS/MS spectra, such as neutral loss ions, played an important role in improving the precision of pSite. In summary, the effective and universal model together with the extensive use of spectral information makes pSite an excellent quality control tool for both *de novo* peptide sequencing and modification site localization.

Keywords: tandem mass spectrometry, *de novo* peptide sequencing, modification site localization, phosphorylation

Introduction

Mass spectrometry has become the technology of choice for protein identification[1]. Database search is regarded as the dominant method, but its performance relies heavily on the quality of proteome databases. An alternative method is *de novo* peptide sequencing, which can deduce the peptide sequences directly from tandem mass spectrometry (MS/MS) data without using any databases. Therefore, *de novo* peptide sequencing can be used to find novel peptides, including peptides with mutations and unexpected modifications[2]. Until now, a number of *de novo* peptide sequencing algorithms have been reported, such as SHERENGA[3], PEAKS[4], PepNovo[5,6], pNovo[2,7,8], UniNovo[9] and Novor[10].

However, it still remains challenging to control the false discovery rate (FDR)[11-13] in *de novo* peptide sequencing. As mentioned in the previous study, the precision of the full-length peptides yielded by *de novo* peptide sequencing cannot reach 70% even for the peptides with very high scores, no matter which algorithm is used[2]. The low precision of *de novo* peptide sequencing is not surprising because there is no reference database to restrict the generation of peptide candidates, which results in a space of ten orders of magnitude larger than that of database search[2].

As an alternative to controlling the FDR of the full-length peptides, controlling the false amino-acid rate (FAR) is more practical and also important in *de novo* peptide sequencing. Similar to FDR, which is defined by the number of the false full-length peptides divided by the number of all reported peptides, FAR is firstly defined in this paper by the number of the false amino acids divided by the number of all reported amino acids. In order to control FAR, the most pivotal procedure is to evaluate the confidence of each amino acid, based on which the credible amino acids can be reported for further studies, *e.g.*, generating reliable tags for database search[14-16] or assembling protein sequences based on the highly confident peptide sequences from different enzymatic digestions[17,18]. Unfortunately, until now few studies are concerned with FAR control in *de novo* peptide sequencing algorithms. PEAKS[4] and Novor[10] can report the confidence of each amino acid, but the details of scoring function were not shown in the related publications. Furthermore, no FAR control strategies have been proposed in all *de novo* peptide sequencing algorithms including PEAKS and Novor, according to our survey.

In addition, the amino acid confidence evaluation problem described above is very important because it is essentially a generalized model for another problem, modification site localization, which has always been an important issue in proteomics[19]. In other words, evaluating the confidence of an amino acid in *de novo* peptide sequencing is similar to evaluating the confidence of the localization for a modification. In 2006, Beausoleil *et al.*[20] firstly proposed a binomial distribution model to evaluate the confidence of phosphorylation sites and developed the Ascore algorithm, and then more software tools gradually appeared, most of which are based on two basic strategies: peak probability or difference score. A peak probability-based strategy calculates probabilities of the real matched peaks and random ones and these real matched peaks can determine modification sites, *e.g.*, Ascore, PTM Score[21], PLS in InsPecT[22], SLoMo[23], Phosphinator[24] and phosphoRS[25]. A difference score-based strategy calculates a search engine score difference between the best two candidate phosphorylation sites, *e.g.*, Mascot Delta

Score[26], SLIP score in Protein Prospector[27], VML score in Spectrum Mill[19] and LuciPHOr[28,29]. However, most of them only focused on the localization of phosphorylation but neglected other modification types in algorithm design or software development[30]. In addition, some features in MS/MS data, *e.g.*, the preference of modifications and the neutral losses, are crucial in determining the correct modification sites but were not well considered in these algorithms. Therefore, a more universal and accurate modification site localization algorithm will be of great value especially for those widely studied modifications, *e.g.*, phosphorylation, acetylation, glycosylation and ubiquitination. As mentioned above, such an algorithm can be designed by naturally generalizing the algorithm of the amino acid confidence evaluation.

In this paper, we firstly proposed a novel algorithm, pSite, to solve these two problems, *i.e.*, amino acid confidence evaluation and modification site localization, based on a unified approach. A support vector machine (SVM)[31,32] was used in our study in view of the fact that Percolator[33] adopted SVM to improve the identifications in database search and PTMFinder[34] applied this method to identify post-translational modifications. An SVM model was constructed to evaluate the confidence of each amino acid with or without modifications, and then a Bayesian model was used to evaluate the FAR at any given threshold. For amino acid confidence evaluation of *de novo* peptide sequencing, pSite was compared with PEAKS[4] and Novor[10], because they are the only two available software tools that can report the amino acid confidence as far as we know. On the three data sets, the recall of pSite was over 86% higher than those of the other two algorithms at the FAR of 5%. For the modification site localization problem, pSite was compared with Ascore[20], the first modification site localization algorithm, and phosphoRS[25], which performed well on the data sets containing more than 100,000 synthetic phosphopeptides[35]. On three phosphopeptide data sets, when the false localization rate (FLR)[27] was controlled at 1%, the average recall of pSite was 91% while the average recalls of Ascore and phosphoRS were 64% and 63%, respectively. In addition, pSite covered 98% of Ascore and phosphoRS results, and obtained 21% more phosphorylation sites.

Methods

Generally, a correct amino acid α in a *de novo* sequenced peptide P should match the corresponding region of a spectrum better or at least equally well when compared with the competitive ones enumerated according to the mass of P, as shown in Figure 1. Therefore, the score difference between the original and the competitive sequence can be used to estimate the confidence of the given amino acid, which is the basic principle of pSite.

pSite has five steps to compute the confidence of each amino acid on a peptide based on the peptide-spectrum match quality: 1) preprocessing MS/MS data, 2) enumerating the competitive sequences, 3) extracting features for each amino acid site, 4) estimating the confidence of each amino acid site, and 5) controlling the FAR of the reported amino acids. These five steps are introduced as follows.

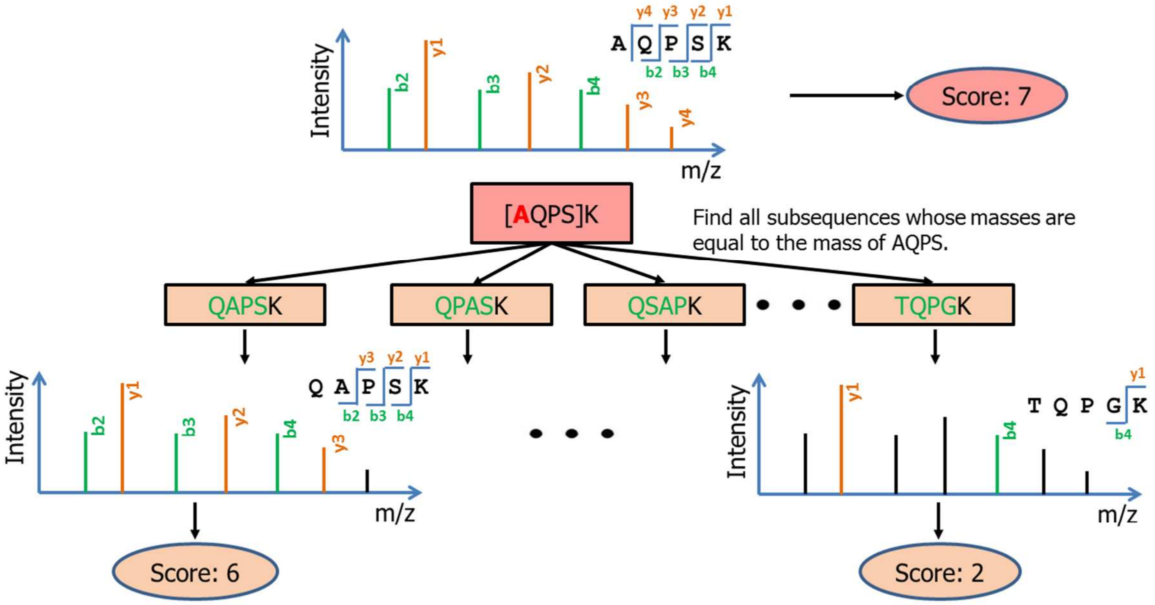


Figure 1. An example of confidence evaluation of amino acids by enumerating competitive partial sequences. Assuming that the correct peptide sequence is AQPSK and the confidence of the first residue A in AQPSK is to be evaluated. All subsequences whose masses are equal to the mass (with a given mass tolerance, *e.g.*, 20 ppm) of AQPS (383.18 Da) are enumerated, *e.g.*, QAPS, QPAS, QSAP, …, TQPG. Any enumerated subsequence whose summed mass of prefix residues and the amino acid to be evaluated are both the same as those of the original sequence should be removed, such as **A**PQS, **A**PSQ and **A**SPQ. Note that lengths of all subsequences do not have to be equal to 4 (the length of the original subsequence AQPS). For example, GATGP and GAAPS are also valid subsequences because their masses are also equal to 383.18 Da. Then the original subsequence is replaced by these enumerated subsequences to generate competitive peptides: QAPSK, QPASK, QSAPK, … , TQPGK. The score, *i.e.*, the number of matched peaks is 7 for the original sequence AQPSK and varies from 2 to 6 for the competitive sequences.

Preprocessing MS/MS data. Firstly, the raw files were extracted using pXtract and then the precursor ion masses were calibrated with pParse[36]. Secondly, for each spectrum, the charges of all peaks should be determined because a theoretical ion of charge 2+ matching with a peak of charge 1+ should be treated as a random match. For a spectrum, if the charge state of its precursor ion is *c*, the charge state of the peak is assigned by finding the best-fitting isotopic cluster from *c*, *c*−1, …, and 1. Also, peaks corresponding to the precursor ion and its neutral losses (*e.g.*, loss of water or ammonia) should be removed because they are often highly abundant and can easily be mismatched.

Enumerating competitive sequences. For a spectrum and its corresponding sequence $a_1 a_2 \ldots a_l$, if the confidence of a_i is to be evaluated, then all regions of *k*-long sequence tags containing a_i are to be considered, *e.g.*, $a_{i-3}a_{i-2}a_{i-1}a_i$, $a_{i-2}a_{i-1}a_i a_{i+1}$, $a_{i-1}a_i a_{i+1}a_{i+2}$ and $a_i a_{i+1}a_{i+2}a_{i+3}$ when *k* equals 4. The mass *m* of each tag is computed and all valid sequence tags are exhaustively generated whose masses are equal to *m* within a given tolerance. Some sequence tags whose prefix residue masses and amino acids to be evaluated are both the same to the original one should be removed (explained by one example in Figure 1 legend). Then the original sequence $a_1 a_2 \ldots a_l$ can be partially replaced by each of the generated tags, *e.g.*, $a_1 \ldots a_{i-4}S_1 a_{i+1} \ldots a_l$, $a_1 \ldots a_{i-3}S_2 a_{i+2} \ldots a_l$, $a_1 \ldots a_{i-2}S_3 a_{i+3} \ldots a_l$ and $a_1 \ldots a_{i-1}S_4 a_{i+4} \ldots a_l$, in which S_1, S_2, S_3 and S_4 denote four sets respectively and each set contains several tags generated from the original one. As shown in Table S1, the values of *k* are tested from 2 to 5 and 4 is the best value for the balance between the recall and the time efficiency, which is used for the following processing.

Extracting features for each amino acid site. For each amino acid site, the original sequence and all competitive sequences generated in the previous step are scored against the same spectrum, and then three types of features are concomitantly extracted as follows (Table 1):

a) *Scores related to the quality of the peptide-spectrum match*, including three features: 1) the original score reported by the *de novo* peptide sequencing algorithm, *e.g.*, pNovo+ in this study, 2) the ratio of the summed intensity of matched peaks to that of all peaks in the spectrum, and 3) the ratio of the number of matched ions to that of all theoretical ions from the peptide. The discriminations of these features are shown in Figures S1a–c. All these three score formulas are shown in Table S2.

b) *Score differences between two sequence candidates*, including three features: the differences of the three scores in a) between the original sequence and the competitive sequence with the maximum score, respectively. These features are more important to discriminate between the correct amino acids and the wrong ones (Figures S1d–f), which were also considered in the previous studies[30]. If the amino acid from the original sequence is correct and hence supported by a large number of peaks, its score should be much higher than that of any competitive sequence, otherwise there may be hardly any differences or the difference score can be even negative. All these three score formulas are also shown in Table S2.

c) *The peptide information*, including three types of features: the position of the amino acid (from 1 to the length of the peptide), the amino acid type and the length of the peptide, as shown in Figures S1g–i. For example, the N-terminal amino acids reported by *de novo* peptide sequencing algorithms are more challenging to sequence correctly because of the absence of b_1 ions for HCD spectra[37,38], hence the position information is important especially for the amino acids close to the N-termini of peptides.

Table 1. Nine features considered in pSite.

Class a: original score	1: *de novo* score
	2: intensity ratio[a]
	3: ion ratio[b]
Class b: different score	4: Δ *de novo* score
	5: Δ intensity ratio
	6: Δ ion ratio
Class c: peptide information	7: amino acid position
	8: amino acid type
	9: peptide length

[a] The ratio of the summed intensity of matched peaks to that of all peaks in the spectrum.

[b] The ratio of the number of matched ions to that of all theoretical ions from the peptide.

Estimating the confidence of each amino acid site. The machine learning method SVM[31,32] is used in this study with all of the features mentioned above. SVM is one of the most commonly used machine learning algorithm and the widely used Percolator algorithm[33] has adopted the SVM method to improve the identifications in database search. The results identified by database search with FDR ≤ 1% at the peptide level are used as the ground truth data. For each peptide from *de novo* peptide sequencing, if an amino acid is located at the correct position, *i.e.*, its N-terminal flanking mass is consistent with that in the database search results, then it is considered as a correct one; otherwise it is considered incorrect. For example, if the *de novo* sequenced peptide from a spectrum is AE***TP***K but the answer is AE***PT***K, then the correct amino acids are A, E and K and the incorrect ones are T and P. All amino acids labeled *correct* and *incorrect* are then used as the input of the radial-basis-function-based SVM model. Note that the original output of SVM model is a probability, so it is from 0 to 1. In order to be comparable with the scores of PEAKS and Novor, the probability is multiplied by 100 as the final output score of pSite (referred to as SVM-Score). The 5% of results in ground truth data on D-DS1 mentioned in the Results section were used to train the SVM model and all HCD results were used for testing the model performance. For ETD data, the 5% of results in ground truth data on trypsin ETD data set mentioned in the Results section were used to train the SVM model and ETD results of all enzymes were used for testing.

Controlling the FAR. On the data sets labeled by the results of database search, the FAR can be computed with a given SVM-Score threshold; however, on the real data sets, it is unknown that whether the amino acids are correct or not. More importantly, a fixed threshold cannot be used to accurately estimate the FAR since the scale of SVM-Score is quite different on different data sets: the cutoffs at precision 95% on D-DS1, D-DS2 and D-DS3 are 81, 89 and 79, respectively. Therefore, a method to estimate and control the FAR of the amino acids based on the distribution of the SVM-Score is needed. In this study, two Gamma distributions are used to fit the score distributions of the correct and incorrect amino acids respectively (Figure S2), similar to PeptideProphet[39]. Also we have compared the respective differences between the real distribution of SVM-Score and the Gamma or Gaussian distribution (Figure S3). These two Gamma distributions are used to compute the FAR as shown in Formula (1):

$$\mathrm{FAR} = \frac{p_w \times \varphi(X|\alpha_w, \beta_w)}{p_w \times \varphi(X|\alpha_w, \beta_w) + p_r \times \varphi(X|\alpha_r, \beta_r)}. \tag{1}$$

In this formula, p_w and p_r are the prior probabilities of the wrong and right results, respectively. X denotes the SVM-Score, α_w and β_w are the parameters of the Gamma distribution of the wrong results, and α_r and β_r are the parameters of the Gamma distribution of the right results. $\varphi(X|\alpha_w, \beta_w)$ and $\varphi(X|\alpha_r, \beta_r)$ respectively denote the probabilities of the SVM-Score exceeding X according to the

Gamma distributions of the wrong and right results. All these parameters (p_w, p_r, α_w, α_r, β_w and β_r) can be computed by the expectation-maximization algorithm[40]. These parameters are online computed according to each data set. Finally, given any one SVM-Score as the threshold to filter the results, the FAR can be estimated by this formula. So controlling the FAR means to determine the threshold of SVM-Score to select confident amino acids, making sure that the estimated FAR is not higher than the user-set value.

Table S3 shows the comparison of the real and estimated FARs on the three data sets. The real FAR was computed by the number of amino acids which were inconsistent with the results identified by database search divided by the number of all amino acids whose scores were greater than the score threshold. On all the three data sets, the estimated values by using Formula (1) were very close to the real ones, especially for the low FAR values, *e.g.*, less than 5%, which were practically used in the real proteomics applications.

Using pSite in modification site localization. The algorithm described above can be easily extended for localizing modification sites. For example, given a peptide Q**pS**HTYK with phosphorylation on the serine residue identified by database search, regarding the phosphorylated serine **pS** as a novel amino acid and then its confidence (referred to as s_1) can be computed in a similar way. In addition, other modification site candidates are usually needed to be compared with the given modification site in the same sequence. For example, there are two other phosphorylation site candidates on the peptide, *i.e.*, QSH**pT**YK and QSHT**pY**K and their confidence values are referred to as s_2 and s_3, respectively. Then the posterior probability of each phosphorylated residue can be computed by the Bayesian formula as shown in Formula (2):

$$p(t_i = 1|s_1, \dots, s_n) = \frac{p_i \times s_i}{\sum_i (p_i \times s_i)}, \tag{2}$$

in which $t_i = 1$ means the *i*-th position of residues is phosphorylated, s_i and p_i mean the confidence and the prior probability of the *i*-th phosphorylated amino acid, respectively. The prior probability p_i of the phosphorylation on each type of amino acid can be estimated by the frequency of the phosphorylated residues in the results of database search.

In order to further improve the algorithm performance by eliminating the effect of the large amount of sites with lower scores, Formula (2) is simplified to Formula (3) as the final score in which only the best candidate phosphorylation site (except the original one) rather than all candidates is considered:

$$p(t_i = 1|s_1, \dots, s_n) = \frac{p_i \times s_i}{p_i \times s_i + p_j \times max_{j|j \neq i}\, s_j}. \tag{3}$$

Results

Evaluating the confidence of the amino acids at each site in *de novo* peptide sequencing.

Data sets description. Three data sets (referred to as D-DS1, D-DS2 and D-DS3) were analyzed (Table S4). The first two data sets were from HeLa cells, which were generated on an LTQ Orbitrap Velos and a Q Exactive[41], respectively. The third data set was from budding yeast (*Saccharomyces cerevisiae*) generated on a Q Exactive[42]. All these three data sets were high-resolution HCD data which were measured in an Orbitrap analyzer. pFind[43] and PEAKS DB[44] were used to process these three data sets. The first two data sets were searched against the human database (UniProt, released in 2014-11) and the third data set was searched against the yeast database (UniProt, released in 2015-01). Both databases were appended with 286 common contaminant protein sequences. The precursor ion tolerance was set as 20 ppm for both pFind and PEAKS DB and the fragment tolerance was set as 20 ppm for pFind and 0.02 Da for PEAKS DB. The FDR was controlled at 1% at the peptide level for pFind and 1% at the PSM (peptide-spectrum match) level for PEAKS DB. The inconsistent PSMs of the two search engines were removed. In addition, peptides containing modifications except carbamidomethylation of cysteine were also removed. Finally, three data sets were generated which consisted of 49,803, 70,752 and 194,831 PSMs, respectively.

pNovo+[8], PEAKS[4] (v7.5) and Novor[10] (v1.1) were then used to get the *de novo* peptide sequencing results of the same data sets. Table S5 shows the numbers of correct and incorrect amino acids reported by the three algorithms. The average ratios of the correct amino acids reported by the three algorithms to the total amino acids were 64%, 63% and 38% and the average error rates of these three algorithms were 30%, 25% and 55%, respectively. pNovo+ reported more amino acids than PEAKS but the error rate was also slightly higher. In general, the results reported by all of the three algorithms were not accurate enough to be practically used in high-precision MS/MS data analysis. Consequently, amino acid confidence evaluation methods are indispensable to report as many amino acids as possible while controlling a low FAR of the result.

Comparison among pSite, PEAKS and Novor. Peptides reported by pNovo+ were subsequently processed by pSite, while the confidence of the amino acids reported by PEAKS and Novor were based on their built-in methods. Figures 2a–c show the precision-recall (PR) curves of the results above the precision threshold of 90% and Figure S4 shows the full PR curves with the precision of 0–100%. Although the precision of the amino acids reported by pNovo+ was slightly lower than PEAKS (Table S5), pSite can still report more amino acids than PEAKS and Novor at the same precision level because of the highly discriminative scoring method. When the precision was fixed at 95%, the recall of pSite was 56.0% on average, significantly higher than that of PEAKS (32.7%). Novor was faster than pNovo+ and PEAKS (Table S6) but reported less results at such a high precision level. The reason might be that it was not trained on high-resolution HCD data, which were also confirmed by Thilo Muth *et al.*[45] and Ngoc Hieu Tran *et al*[46].

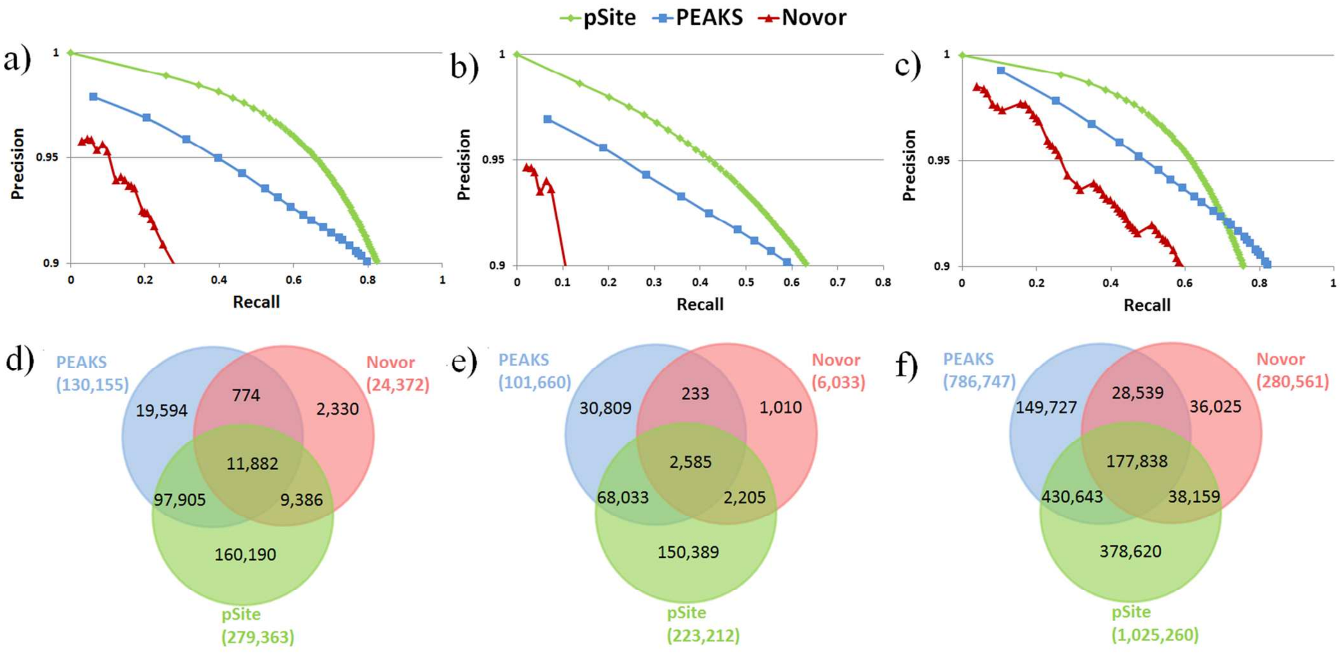


Figure 2. a–c) The precision-recall (PR) curves with precision greater than 90% for pSite, PEAKS and Novor on a) D-DS1, b) D-DS2 and c) D-DS3. d–f) The Venn diagrams of pSite, PEAKS and Novor on d) D-DS1, e) D-DS2 and f) D-DS3 at the 95% precision.

Figures 2d–f show the Venn diagrams of the correct amino acids (precision over 95%) reported by the three algorithms. pSite covered ~77.1% and ~81.2% of the results of PEAKS and Novor on average, respectively. As mentioned above, the input sequences between pSite and PEAKS were not the same: pSite used the results of pNovo+ as input while PEAKS evaluated the amino acid confidence based on the *de novo* peptide sequencing results of its own. For the same spectrum, pNovo+ and PEAKS may report two different sequences. Therefore, it meant that a few correct sequences only reported by PEAKS have no chance to be processed by pSite and vice versa. In order to make a more fair comparison, pSite also processed the peptides from PEAKS (referred to as pSite-P) and then compared again with PEAKS so that the input sequences were absolutely identical. As shown in Figure S5, pSite-P covered ~82.2% of the results of PEAKS, which was slightly higher than that shown in Figures 2d–f because of the same input peptide sequences. Furthermore, pSite-P can also individually report 40.1–69.2% results which were correct but below the 95% precision threshold of PEAKS. Also, pSite was still far better than PEAKS in evaluating the confidence of amino acids no matter which algorithm was used as the input sequences.

The score distributions of correct and incorrect amino acids reported by pSite-P and PEAKS were also analyzed on these three data sets (Figure S6), which demonstrated that the score function of pSite-P was much more discriminative than that of PEAKS. As shown in this figure, there were also many incorrect results with high scores in PEAKS's results, which may interfere with the correct ones. Almost 92% of the results with scores greater than 90 reported by pSite-P were correct while only 80% of the

PEAKS results were correct above the same score threshold. Therefore, the score function of PEAKS was less discriminative even for the results with high scores, which led to the low recall at a high precision level.

Performance of evaluating the confidence of the isomeric amino acids or amino acid combinations. There are a few combinations of amino acids whose summed masses were identical to some other amino acids or combinations (*e.g.*, AG=GA=Q, GG=N, …). For example, given a correct sequence **AG**CDLK while there are no b_1 and y_5 ions in this spectrum, *de novo* peptide sequencing algorithms can easily report an incorrect sequence like **Q**CDLK (AG=Q). It is interesting to analyze the performance of pSite on distinguishing the isomeric amino acids or combinations. Figure 3 shows the Q-Q plots of six different situations. For example, in Figure 3a, the blue curve displayed the relationship between the scores of the correct amino acid Q and the incorrect amino acids AG or GA. The red curve displays the line y=x. According to Figure 3, pSite can distinguish between the correct and incorrect amino acids for all of these six situations to some degree, especially for Q, AG, N and *xy* ("*xy*" denotes any two different amino acids).

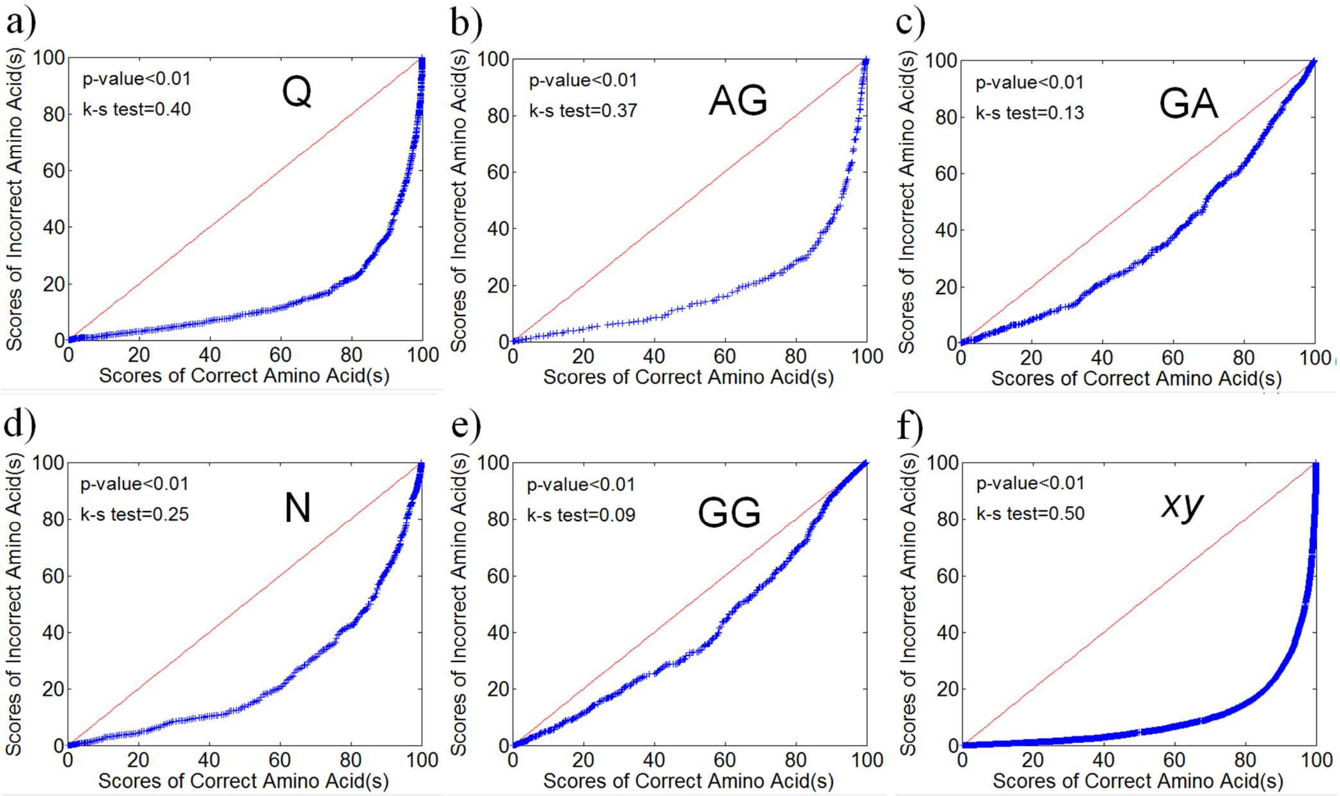


Figure 3. The Q-Q plot curves on D-DS1. The blue curve displays the distribution between the scores of the correct amino acid (*x* axis) and the incorrect amino acid (*y* axis). The red curve displays the line *y*=*x*. a) The correct amino acid is Q while the incorrect amino acids are AG and GA. b) The correct amino acid is AG while the incorrect amino acid(s) is Q or GA. c) The correct amino acids are GA while the incorrect amino acid(s) is Q or AG. d) The correct amino acids is N while the incorrect amino acids are GG. e) The correct amino acids are GG while the incorrect amino acid is N. f) The correct amino acids are *xy* while the incorrect amino acids are *yx*, in which both “*xy*” and “*yx*” denote any two different amino acids (*e.g.*, the correct amino acids are AC while the incorrect amino acids are CA). The p-values of these six distributions are all smaller than 0.01 based on the two-sample Kolmogorov-Smirnov test.

Performance of estimating the confidence of amino acids on ETD data sets. We also compared pSite and PEAKS on three ETD data sets with different enzymes (*e.g.*, Asp-N, Glu-C or trypsin)[8]. The PR curves of pSite and PEAKS on these data sets were shown in Figure S7, indicating that the performance of pSite was much better than PEAKS. When the precision threshold was fixed at 95%, the recall of pSite was 71% on average while that of PEAKS was 39%. Figure S8 shows the Venn diagrams of these two algorithms in which pSite covered ~80% of PEAKS results and can also report ~60% more results which cannot be identified by PEAKS. These results demonstrated that pSite also performed well on ETD data sets.

Performance on phosphorylation site localization.

Data sets description. Three phosphopeptide data sets (referred to as P-DS1, P-DS2 and P-DS3) were analyzed (Table S4). The first data set was generated on an Orbitrap Velos containing more than 100,000 synthetic phosphorylated peptides[35]. The other two data sets were phosphopeptide enrichment data sets: one was generated on a Q Exactive[47] and the other one was generated on a Q Exactive HF[48]. Both of these two data sets were generated from human samples. All of the three data sets were high-resolution HCD data which were measured in an Orbitrap analyzer. Whereafter, they were all analyzed by pFind[43] and PEAKS DB[44] and the parameters were shown in Table S7. The FDR was controlled at 1% at the peptide level for pFind and 1% at the PSM level for PEAKS DB. The inconsistent results of pFind and PEAKS DB were removed to make sure that the retained phosphopeptides including the modifications sites were more credible. P-DS1 was a synthetic data set that the phosphorylated sites were known, so the results which were not consistent with the known sites were removed. For P-DS2 and P-DS3, although neither pFind nor PEAKS DB can evaluate the precision of the modification site localization, the intersected results of these two engines were proved to be high confident with error rate of 1.1% by analyzing P-DS1 in the same way (Table S8). On all of the three data sets, peptides with only one phosphorylation but more than one candidate site were kept. Finally, there were 49,085, 7,743 and 9,608 PSMs in P-DS1, P-DS2 and P-DS3, respectively.

Comparison between pSite, Ascore and phosphoRS. pSite was compared with Ascore[20] and phosphoRS[25] (v3.1) on these three data sets. All of the three algorithms reported the confidence of not only the given phosphorylation site but also the most probable phosphorylation site, which may be different from the given one. For each spectrum, if the most probable phosphorylation site reported by an algorithm was consistent with the ground truth, *i.e.*, the recorded results in P-DS1 or the database search results in P-DS2 and P-DS3, then this site was considered correct, otherwise incorrect. All results were sorted by the confidence in the descending order. The FLR[19] can be estimated by #incorrect sites / (#incorrect sites + #correct sites).

Figures 4a–c show the FLR curves of pSite, Ascore and phosphoRS on the three data sets. No matter how large or small the FLR threshold was, the number of the correct phosphorylation sites reported by pSite was larger than those of the other two algorithms, which also implied that pSite had a better scoring method to discriminate between the correct and incorrect phosphorylation sites. Interestingly, the performance for phosphoRS seemed worse in Figure 4a compared with Figures 4b and c. We have manually checked the results and found that phosphoRS considered the neutral loss ions of the phosphorylation on Y but they were not often observed[49]. Also, we found that in P-DS1 data set, the percentage of phosphorylated Y was 49.7% while only 0.4% and 0.1% in other two data sets (P-DS2 and P-DS3), respectively. This was probably the reason why phosphoRS performed not well on P-DS1.

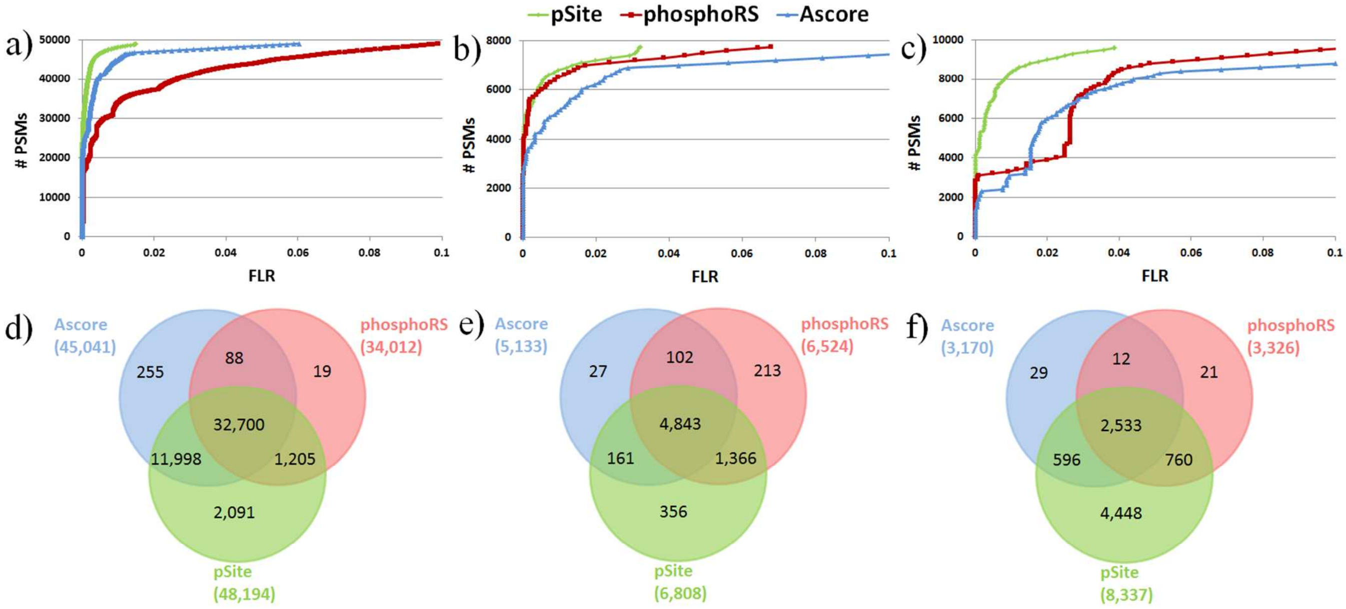


Figure 4. a–c) The FLR curves of pSite, Ascore and phosphoRS on a) P-DS1, b) P-DS2 and c) P-DS3. d–f) The comparison of the correct results of pSite, Ascore and phosphoRS on d) P-DS1, f) P-DS2 and f) P-DS3.

Consistency analysis. When the FLR was controlled at 1%, the recalls of pSite were 98.2%, 87.9% and 86.8%, while those of Ascore were 91.8%, 66.3% and 33.0% and phosphoRS were 69.3%, 84.3% and 34.6% on P-DS1, P-DS2 and P-DS3, respectively. Figures 4d–f show the comparison of the correct phosphorylation sites (FLR ≤ 1%) of these three algorithms. pSite covered ~98% results of Ascore and phosphoRS and can independently report ~21% more phosphorylation sites at the same FLR threshold.

Results with high scores but only reported by pSite were further analyzed. For example, in Figure 5, the phosphorylation on S was more reasonable than on Y in this sequence because the peptide with phosphorylation on S (Figure 5a) contributed lots of neutral loss ions, which cannot be observed from peptides with phosphorylation of Y. Another possible evidence is that the frequencies of phosphorylation on S, T and Y were respectively 94.6%, 5.0% and 0.4% on this data set (P-DS2), in other words phosphorylation on S had a much higher prior probability. These two issues were both well considered in pSite but not in Ascore and phosphoRS. Another example was shown in Figure S9. The correct sequence was TLGHMVILDQTK in which the N-terminal T was phosphorylated because it was supported by the record in the synthetic data set (P-DS1). However, Ascore reported the T at the 11th position with a score of 14.9 (over 90% confidence) and phosphoRS also reported this position with 99.1% confidence. As shown in this figure, the correct peptide contributed many *y* ions with water losses, which was incorrectly assigned to the neutral losses of phosphorylated ions by Ascore and phosphoRS. Therefore, neutral loss ions, *e.g.*, water and ammonia loss ions, should be considered in the algorithm design of modification site localization. Table S9 shows what ion types were considered in different algorithms. Figure S10 shows the performance of pSite without considering ions with water and ammonia losses, referred to as pSite-NH on P-DS1. When the FLR was controlled at 1%, the recall of pSite-NH decreased from 98.2% to 90.5%, which shows the effect of considering the neutral loss ions.

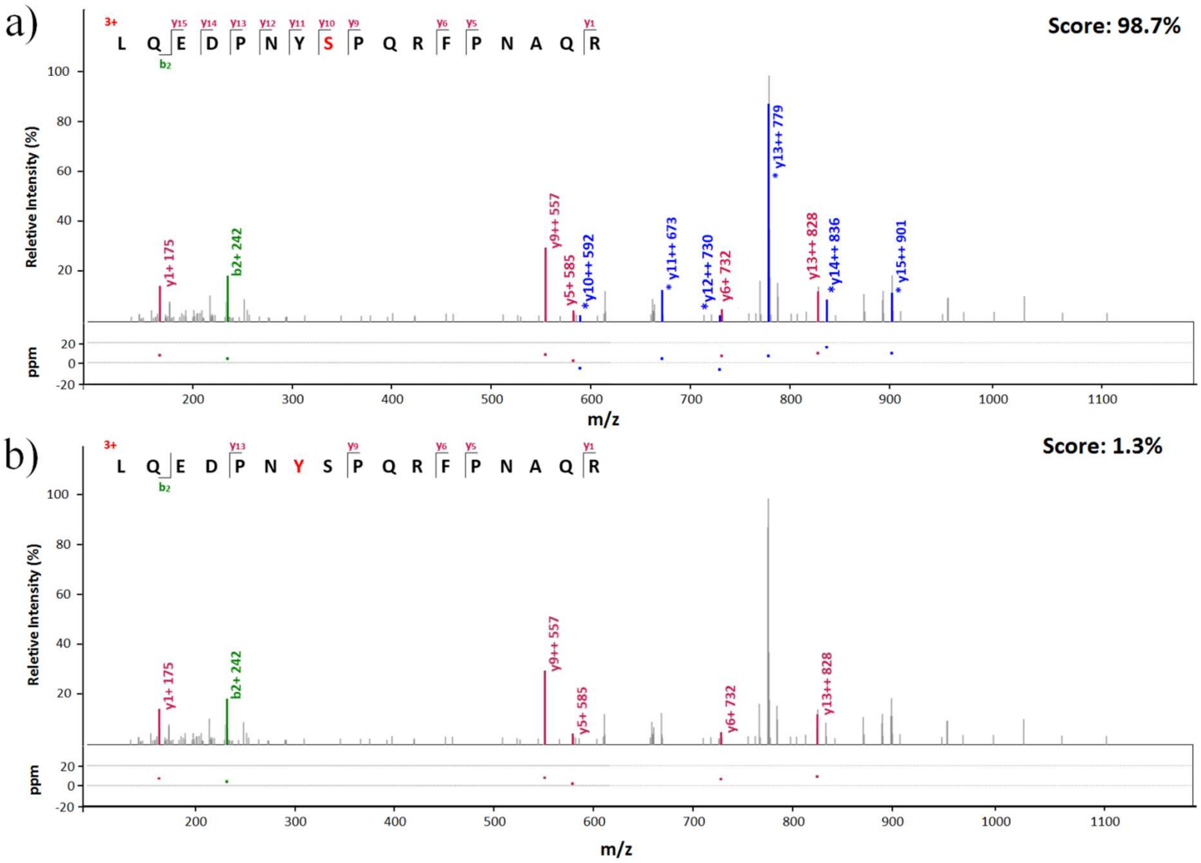


Figure 5. A peptide-spectrum match reported by pSite only. This spectrum was from P-DS2 and the title was 83031_PT3.6376.6376.3.dta. a) The match of the correct peptide LQEDPNYpSPQRFPNAQR, which was also reported by pSite. b) The match of the incorrect peptide LQEDPNpYSPQRFPNAQR reported by phosphoRS. pSite gave the confidences of these two sites 98.7% and 1.3%, respectively. The peptide with phosphorylation on S contributed lots of neutral loss ions, which cannot be observed from peptides with phosphorylation of Y. Each "*" in this figure denoted a neutral loss ion (−98 Da) of phosphorylation.

Performance on ETD data sets and multi-phosphorylated peptide data sets. We also compared pSite with Ascore and phosphoRS on other three data sets. One was an ETD data set containing 100,000 synthetic phosphorylated peptides, which was generated on an Orbitrap Velos[35]. The other two data sets were respectively generated on a Q Exactive[47] and a Q Exactive HF[48], which consisted of peptides with two or more phosphorylation sites. pFind[43] was used to build three benchmark data sets with the FDR ≤

1% at the peptide level. On these three data sets, there were 13,633, 988 and 11,151 PSMs respectively. Figure S11 shows the FLR curves of three algorithms on these three data sets. In general, the performance of pSite was still better than those of Ascore and phosphoRS, especially on the ETD data set. Figure S12 shows the comparison of the correct phosphorylation sites recalled by these three algorithms when the FLR was controlled at 1%. pSite covered 83.2% results of Ascore and 81.8% of phosphoRS on average while also reported 13.1–27.0% more PSMs which cannot be reported by Ascore and phosphoRS. When the FLR was controlled at 1%, the recalls of pSite were 76.3%, 32.4% and 26.2% on these three data sets while those of Ascore were 66.7%, 14.2% and 20.3% and those of phosphoRS were 55.0%, 23.3% and 28.1%, respectively. It revealed that correctly localizing all of the multiple phosphorylation sites on one peptide was much more difficult.

Discussion

In this study, we have proposed a novel method to evaluate the confidence of amino acids from either the results of *de novo* peptide sequencing or the modification sites identified by database search. Evaluation on three large-scale data sets for *de novo* peptide sequencing and other three data sets for phosphorylation site localization shows that pSite can report more correct amino acids and locate more correct phosphorylation sites at a high precision level using a universal model.

The confidence evaluation of amino acids at each site of a peptide rather than the full-length peptides is a new emerging field in computational proteomics. With the development of proteomics in recent years, full-length peptides with any types of modifications, as well as unexpected digestions, can be efficiently identified by some open search strategies[43,50,51]. However, the types of modifications are difficult to be determined and their sites are usually not able to be correctly localized. In other words, peptides are more likely to be partially correct, *i.e.*, neither exactly correct nor entirely incorrect, and the FDR of partially correct results cannot be modeled or estimated by the traditional target-decoy strategy[51]. Hence it is required to determine the confidence of the individual amino acids in each peptide, which aims at obtaining a more precise interpretation of the large-scale proteome data.

As the confidence evaluation of amino acids is the groundwork for the peptide and protein identifications, it can also be used in several other fields. For example, in the cross-linking studies, the site of the cross-linker needs to be precisely determined, and the shorter peptide in each peptide pair is prone to be randomly matched[52]. Therefore, the confidence of the localization of the cross-linker and the match of the shorter peptides need to be separately estimated. Moreover, the amino acid confidence evaluation algorithm can also be used in top-down proteomics because of the huge number of modification combinations in which the correct proteoform is hard to be identified using the traditional top-down search engines[53,54].

The software can be downloaded on the following website:

http://pfind.ict.ac.cn/software/pNovo/pSite_v1.0.exe.

References

1. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347-355 (2016).
2. Yang, H., *et al.* Open-pNovo: De Novo Peptide Sequencing with Thousands of Protein Modifications. *Journal of Proteome Research* **16**, 645-654 (2017).
3. 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* **6**, 327-342 (1999).

4. Ma, B.*, et al.* PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry : RCM* **17**, 2337-2342 (2003).
5. Frank, A. & Pevzner, P. PepNovo: de novo peptide sequencing via probabilistic network modeling. *Anal Chem* **77**, 964-973 (2005).
6. 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* **6**, 114-123 (2007).
7. Chi, H.*, et al.* pNovo: de novo peptide sequencing and identification using HCD spectra. *J Proteome Res* **9**, 2713-2724 (2010).
8. Chi, H.*, et al.* pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra. *J Proteome Res* **12**, 615-625 (2013).
9. Jeong, K., Kim, S. & Pevzner, P.A. UniNovo: a universal tool for de novo peptide sequencing. *Bioinformatics* **29**, 1953-1962 (2013).
10. Ma, B. Novor: Real-Time Peptide de Novo Sequencing Software. *Journal of the American Society for Mass Spectrometry* **26**, 1885-1894 (2015).
11. Soric, B. Statistical Discoveries and Effect-Size Estimation. *Journal of the American Statistical Association* **84**, 608-610 (1989).
12. Benjamini, Y. & Hochberg, Y. Controlling the False Discovery Rate - a Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B-Methodological* **57**, 289-300 (1995).
13. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 9440-9445 (2003).
14. Tabb, D.L., Saraf, A. & Yates, J.R., 3rd. GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model. *Anal Chem* **75**, 6415-6421 (2003).
15. Shilov, I.V.*, et al.* 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**, 1638-1655 (2007).
16. Tanner, S.*, et al.* InsPecT: identification of posttranslationally modified peptides from tandem mass spectra. *Anal Chem* **77**, 4626-4639 (2005).
17. Bandeira, N., Pham, V., Pevzner, P., Arnott, D. & Lill, J.R. Automated de novo protein sequencing of monoclonal antibodies. *Nature biotechnology* **26**, 1336-1338 (2008).
18. Tran, N.H.*, et al.* Complete De Novo Assembly of Monoclonal Antibody Sequences. *Sci Rep* **6**, 31730 (2016).
19. Chalkley, R.J. & Clauser, K.R. Modification site localization scoring: strategies and performance. *Mol Cell Proteomics* **11**, 3-14 (2012).
20. Beausoleil, S.A., Villen, J., Gerber, S.A., Rush, J. & Gygi, S.P. A probability-based approach for high-throughput protein phosphorylation analysis and site localization. *Nature Biotechnology* **24**, 1285-1292 (2006).

21. Olsen, J.V., *et al.* Global, in vivo, and site-specific phosphorylation dynamics in signaling networks. *Cell* **127**, 635-648 (2006).

22. Albuquerque, C.P., *et al.* A multidimensional chromatography technology for in-depth phosphoproteome analysis. *Molecular & Cellular Proteomics* **7**, 1389-1396 (2008).

23. Bailey, C.M., *et al.* SLoMo: Automated Site Localization of Modifications from ETD/ECD Mass Spectra. *Journal of Proteome Research* **8**, 1965-1971 (2009).

24. Phanstiel, D.H., *et al.* Proteomic and phosphoproteomic comparison of human ES and iPS cells. *Nature Methods* **8**, 821-U884 (2011).

25. Taus, T., *et al.* Universal and Confident Phosphorylation Site Localization Using phosphoRS. *Journal of Proteome Research* **10**, 5354-5362 (2011).

26. Lemeer, S., *et al.* Phosphorylation site localization in peptides by MALDI MS/MS and the Mascot Delta Score. *Analytical and Bioanalytical Chemistry* **402**, 249-260 (2012).

27. Baker, P.R., Trinidad, J.C. & Chalkley, R.J. Modification Site Localization Scoring Integrated into a Search Engine. *Molecular & Cellular Proteomics* **10**(2011).

28. Fermin, D., Walmsley, S.J., Gingras, A.C., Choi, H. & Nesvizhskii, A.I. LuciPHOr: Algorithm for Phosphorylation Site Localization with False Localization Rate Estimation Using Modified Target-Decoy Approach. *Molecular & Cellular Proteomics* **12**, 3409-3419 (2013).

29. Fermin, D., Avtonomov, D., Choi, H. & Nesvizhskii, A.I. LuciPHOr2: site localization of generic post-translational modifications from tandem mass spectrometry data. *Bioinformatics* **31**, 1141-1143 (2015).

30. Mirzaei, H. & Carrasco, M. *Modern Proteomics–Sample Preparation, Analysis and Practical Applications*, (Springer, 2016).

31. Cortes, C. & Vapnik, V. Support-Vector Networks. *Machine Learning* **20**, 273-297 (1995).

32. Vapnik, V.N. An overview of statistical learning theory. *Ieee Transactions on Neural Networks* **10**, 988-999 (1999).

33. Kall, L., Canterbury, J.D., Weston, J., Noble, W.S. & MacCoss, M.J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat Methods* **4**, 923-925 (2007).

34. Tanner, S., *et al.* Accurate annotation of peptide modifications through unrestrictive database search. *Journal of Proteome Research* **7**, 170-181 (2008).

35. Marx, H., *et al.* A large synthetic peptide and phosphopeptide reference library for mass spectrometry-based proteomics. *Nat Biotechnol* **31**, 557-564 (2013).

36. Yuan, Z.F., *et al.* pParse: a method for accurate determination of monoisotopic peaks in high-resolution mass spectra. *Proteomics* **12**, 226-235 (2012).

37. Paizs, B. & Suhai, S. Combined quantum chemical and RRKM modeling of the main fragmentation pathways of protonated GGG. II. Formation of b(2), y(1), and y(2) ions. *Rapid communications in mass spectrometry : RCM* **16**, 375-389 (2002).

38. Paizs, B. & Suhai, S. Towards understanding some ion intensity relationships for the tandem mass spectra of protonated peptides. *Rapid Commun Mass Sp* **16**, 1699-1702 (2002).

39. Keller, A., Nesvizhskii, A.I., Kolker, E. & Aebersold, R. Empirical statistical model to estimate

the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem* **74**, 5383-5392 (2002).

40. Dempster, A.P., Laird, N.M. & Rubin, D.B. Maximum Likelihood from Incomplete Data Via Em Algorithm. *Journal of the Royal Statistical Society Series B-Methodological* **39**, 1-38 (1977).

41. Michalski, A.*, et al.* Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Molecular & Cellular Proteomics* **10**(2011).

42. 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* **11**, 319-U300 (2014).

43. Chi, H.*, et al.* pFind-Alioth: A novel unrestricted database search algorithm to improve the interpretation of high-resolution MS/MS data. *J Proteomics* **125**, 89-97 (2015).

44. Zhang, J.*, et al.* PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Molecular & cellular proteomics : MCP* **11**, M111 010587 (2012).

45. Muth, T. & Renard, B.Y. Evaluating de novo sequencing in proteomics: already an accurate alternative to database-driven peptide identification? *Briefings in bioinformatics* (2017).

46. Tran, N.H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences of the United States of America* (2017).

47. Huesgen, P.F.*, et al.* LysargiNase mirrors trypsin for protein C-terminal and methylation-site identification. *Nat Methods* **12**, 55-58 (2015).

48. Kelstrup, C.D.*, et al.* Rapid and deep proteomes by faster sequencing on a benchtop quadrupole ultra-high-field Orbitrap mass spectrometer. *J Proteome Res* **13**, 6187-6195 (2014).

49. Creasy, D.M. & Cottrell, J.S. Unimod: Protein modifications for mass spectrometry. *Proteomics* **4**, 1534-1536 (2004).

50. Chick, J.M.*, et al.* A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nature biotechnology* **33**, 743-749 (2015).

51. Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D. & Nesvizhskii, A.I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat Methods* **14**, 513-520 (2017).

52. Trnka, M.J., Baker, P.R., Robinson, P.J., Burlingame, A.L. & Chalkley, R.J. Matching cross-linked peptide spectra: only as good as the worse identification. *Mol Cell Proteomics* **13**, 420-434 (2014).

53. Smith, L.M., Kelleher, N.L. & Consortium for Top Down, P. Proteoform: a single term describing protein complexity. *Nat Methods* **10**, 186-187 (2013).

54. Sun, R.X.*, et al.* pTop 1.0: A High-Accuracy and High-Efficiency Search Engine for Intact Protein Identification. *Anal Chem* **88**, 3082-3090 (2016).

Supporting Information

The following files are available free of charge at ACS website http://pubs.acs.org:

1) 12 figures

2) 9 tables

3) An example to explain how to extract the nine features

Acknowledgments

This work was supported by the National Key Research and Development Program of China (No. 2016YFA0501301), the National Natural Science Foundation of China (No. 31470805), the Youth Innovation Promotion Association CAS (No. 2014091), the National High Technology Research and Development Program of China (863) (Grant Nos. 2014AA020902 to S.-M.H. and 2014AA020901 to H.C.).

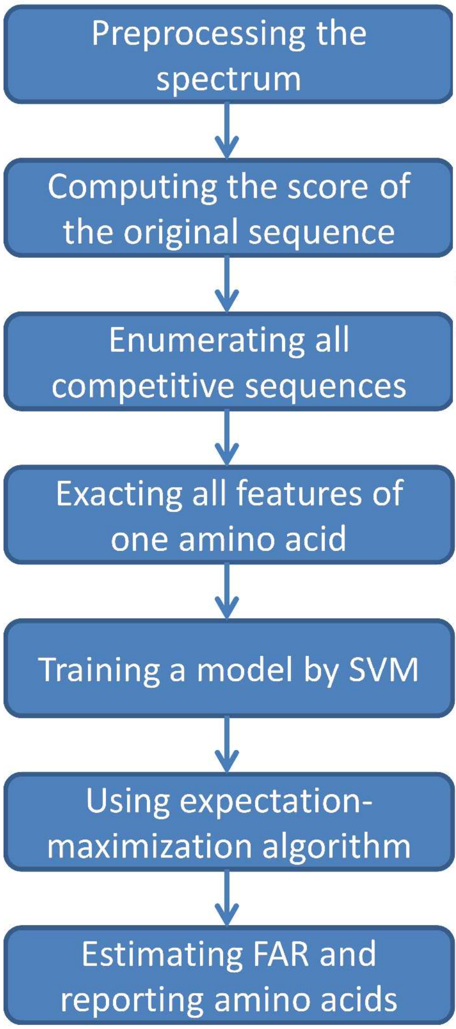
Author contributions

Hao Yang designed the algorithms, performed the data analysis and wrote the manuscript. Hao Chi wrote the manuscript and modified the figures and tables. Si-Min He edited the manuscript and put forward several key suggestions. Wen-Jing Zhou downloaded the synthetic phosphorylated data set. Wen-Feng Zeng implemented the EM algorithm based on two Gamma distributions. Rui-Min Wang proposed the suggestion of using SVM model. Wen-Jing Zhou, Wen-Feng Zeng, Chao Liu, Zhao-Wei Wang, Xiu-Nan Niu and Zhen-Lin Chen modified the manuscript.

Competing financial interests

The authors declare no competing financial interests.

For TOC only



Preprocessing the spectrum

Computing the score of the original sequence

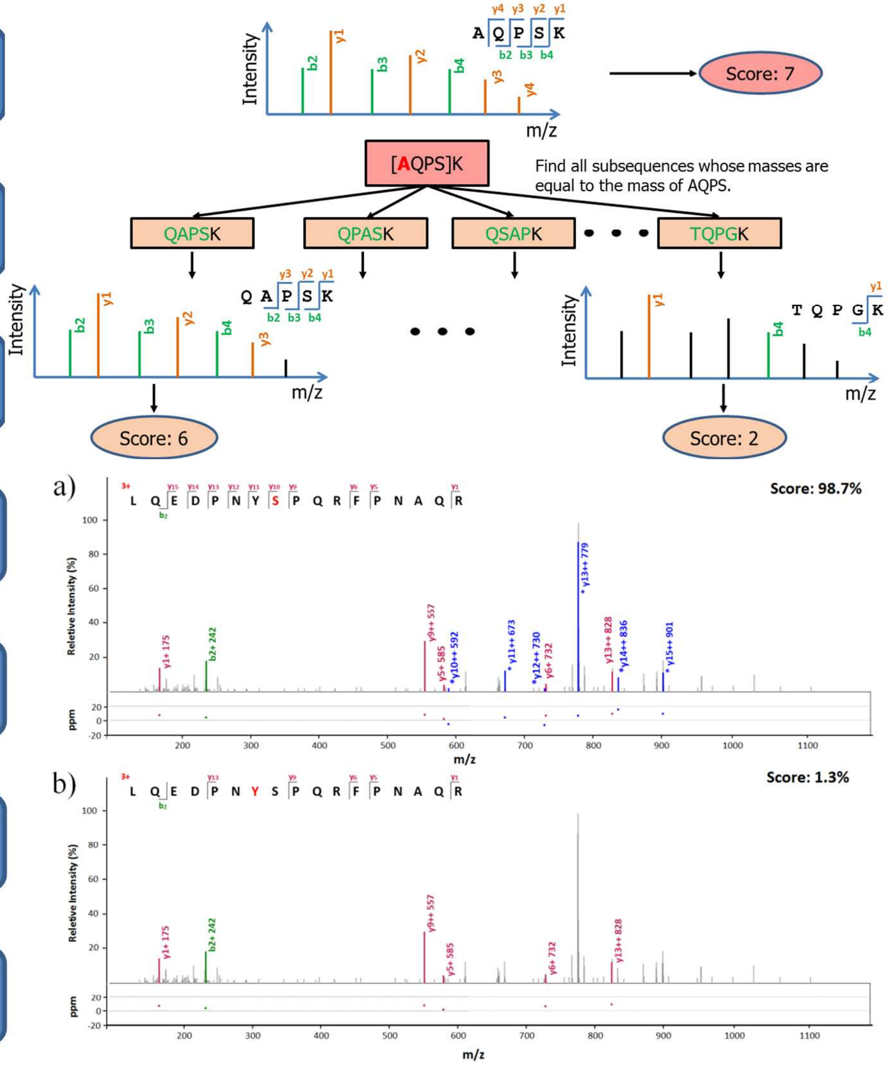
Enumerating all competitive sequences

Exacting all features of one amino acid

Training a model by SVM

Using expectation-maximization algorithm

Estimating FAR and reporting amino acids



Intensity
m/z
y4 y3 y2 y1
A Q P S K
b2 b3 b4
b2 y1 b3 y2 b4 y3 y4
Score: 7
[AQPS]K
Find all subsequences whose masses are equal to the mass of AQPS.
QAPSK
QPASK
QSAPK
TQPGK
y3 y2 y1
Q A P S K
b2 b3 b4
b2 y1 b3 y2 b4 y3
Score: 6
T Q P G K
y1
b4
Score: 2
a)
L Q E D P N Y S P Q R F P N A Q R
Score: 98.7%
Reletive Intensity (%)
ppm
y1+ 175
b2+ 242
y9++ 557
y5+ 585
y10++ 592
y11++ 673
y12++ 730
y6+ 732
y13++ 779
y13++ 828
y14++ 836
y15++ 901
m/z
b)
L Q E D P N Y S P Q R F P N A Q R
Score: 1.3%
Reletive Intensity (%)
ppm
y1+ 175
b2+ 242
y9++ 557
y5+ 585
y6+ 732
y13++ 828
m/z