RPML: A Learning-based Approach for Reranking Protein-Spectrum Matches

Qiong Duan[1], Hao Liang[1], Jun Wu[3], Bo Xu[*,1] and Zengyou He[*,1,2]

[1]School of Software, Dalian University of Technology, Dalian, China
[2]Key Laboratory for Ubiquitous Network and Service Software of Liaoning, Dalian, China
[3]School of Information Engineering, Zunyi Normal University, Zunyi, China
{boxu,zyhe}@dlut.edu.cn

Abstract. Searching top-down spectra against a protein database has been a mainstream method for intact protein identification. Ranking true Protein-Spectrum Matches (PrSMs) over their false counterparts is a feasible method for improving protein identification results. In this paper, we propose a novel model called RPML (Rerank PrSMs based on Machine Learning) to rerank PrSMs in top-down proteomics. The experimental results on real data sets show that RPML can distinguish more correct PrSMs from incorrect ones. The source codes of algorithm are available at `https://github.com/dqiong/spectra_protein_match_rerank`.

Keywords: protein identification, protein-spectrum matches, machine learning, rerank method

1 Introduction

Proteomics aims at studying the complete set of proteins expressed in a given cell, tissue, or organism [22]. One important problem of proteomics is to identify all proteins in a complex mixture effectively and sensitively within a limited time. Protein identification is a fundamental task in proteomics, which is the foundation to test further biological questions such as the relationship between protein abundance and disease states [23].

To date, there are two complementary technologies for protein identification: the bottom-up approach and the top-down approach [4]. In the bottom-up approach, proteins of interest are digested with a protease to produce a mixture of peptides. These peptides are subjected to liquid chromatography and then isolated and fragmented to generate tandem mass spectra (MS/MS) via mass spectrometer. Then, peptides can be identified by searching the spectra against a protein database. Finally, the identified peptides are assembled to infer proteins possibly contained in the samples.

The bottom-up approach have been widely used for identifying proteins, but it still deserves certain drawbacks [16, 27]. Since some peptides cannot be efficiently captured by the mass spectrometer to generate the corresponding MS/MS

spectra, proteins that contain these peptides may be missed in the final identification list. Furthermore, such a peptide-centric identification strategy is unable to identify proteoforms with the complex combinations of post-translational modifications (PTMs) [19]. In addition, the protein inference problem of generating a confident list of proteins from the set of identified peptides has not been completely resolved yet [13].

The current revolution in mass spectrometry instruments has made it possible to obtain high-resolution spectra of intact proteins. The top-down approach involves the direct analysis of intact proteins without the use of proteolytic digestion. A distinct advantage of this method over the bottom-up approach is that the abundance of the proteoforms can be determined directly, as intact proteins are less susceptible to instrumental biases than their short peptides. More importantly, the top-down approach is able to retain the correlation information between multiple PTMs. Therefore, the top-down approach has become a promising and complementary technique to the bottom-up method [31].

Over the last decade, some effective algorithms have been proposed to solve the top-down protein identification problem [2, 8, 10, 17, 15, 18, 21, 24, 28, 33]. ProSight [8, 17, 33] is the first search engine designed for identifying intact proteins, which adopts a shotgun annotation strategy to generate a virtual database including all proteoforms, but the size of virtual database will become to be very huge for complex samples. MS-TopDown [10] is based on the spectral alignment algorithm for scoring protein-spectrum matches (PrSMs). MS-Align+ [21] proposes some effective pruning strategies to accelerate the spectral alignment process, leading to a significant improvement in terms of running efficiency over MS-TopDown. pTop [28] is a recent proposed software package for identifying intact proteins from top-down spectra, which is much faster than other existing tools. In addition, there are a few other tools for top-down protein identification, such as TopPIC [15], MASHsuite [2], ProteinGoggle [18] and InformedProteomics [24].

A key component in existing identification approaches is the scoring function used to evaluate the quality of protein-spectrum matches. However, current PrSM scoring algorithms are far from satisfactory since incorrect PrSMs are frequently observed in the identification results. These incorrect matches can be attributed to many factors, such as the poor quality of spectra, the existence of post-translational modifications and unavoidable random matches due to the increasing number of candidate PrSMs. Hence, it is quite challenging to design a PrSM scoring algorithm that is capable of distinguishing correct identifications from incorrect ones in a perfect manner. One alternative strategy is to develop new post-processing techniques by reranking PrSMs such that more true identifications are re-arranged before false identifications. As a result, we will obtain more true identifications at the same cut-off criteria (e.g., false discovery rate) threshold.

To date, some reranking methods have been introduced for improving the peptide identification results in bottom-up proteomics [9, 11, 12, 14, 25, 32]. The empirical results in these papers demonstrated that the reranking strategy is

capable of yielding a remarkable performance improvement for peptide identification. The success of the reranking methods in bottom-up proteomics motivates us to investigate if it is also feasible to rerank PrSMs so as to generate an improved top-down identification results. To the best of our knowledge, this is the first attempt that studies the PrSM reranking problem in the context of top-down protein identification.

In this paper, we propose a novel model called RPML based on machine learning to rerank PrSMs. Our method is tested on four public data sets. The experimental results show that RPML can distinguish more correct identifications from incorrect ones.

The rest of this article is organized as follows. In Section 2, we describe our method in detail. Section 3 presents the experimental results and Section 4 concludes the article.

2 Methods

2.1 Problem Definition

Let $\mathbb{C} = \{(s_1, p_1), (s_2, p_2), ..., (s_n, p_n)\}$ be a set of PrSMs, where s_i is a MS/MS spectra and p_i is a protein sequence. The set $\mathbb{C}$ is associated with a vector of initial ranking scores $X = (x_1, x_2, ..., x_n)^T$ provided by a standard top-down protein identification approach, where x_i is the score of (s_i, p_i). The type of score varies according to the baseline identification methods (e.g., E-value, the number of matching peaks). The goal of the reranking model is to assign (s_i, p_i) a new socre y_i so that the new vector $Y = (y_1, y_2, ..., y_n)^T$ can improve the

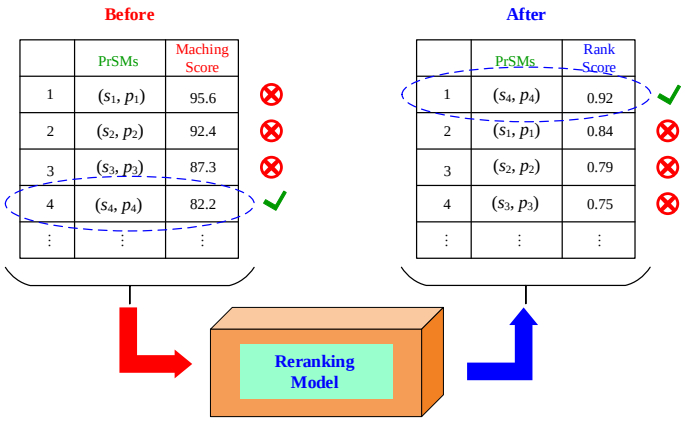

Fig. 1. An example on the PrSMs reranking problem. The correct PrSM (s_4, p_4) has a matching score 82.2 in the initial ranking list , which is arranged behind other incorrect PrSMs. The reranking model generates a new probability score vector for the set of PrSMs such that (s_4, p_4) is assigned with the highest score 0.92.

ranking results. Fig. 1 shows an example.

Note that it is impossible to have $(s_i, p_i) = (s_j, p_j)$ or $s_i = s_j$ for $i \neq j$ in the set $\mathbb{C}$, that is, spectra are different from each other. But it can be $p_i = p_j$ owing to the fact that a protein potentially may produce several spectra.

2.2 General Workflow of RPML

RPML consists of three main procedures: feature extraction, model construction and score integration, whose workflow is described in Fig. 2.

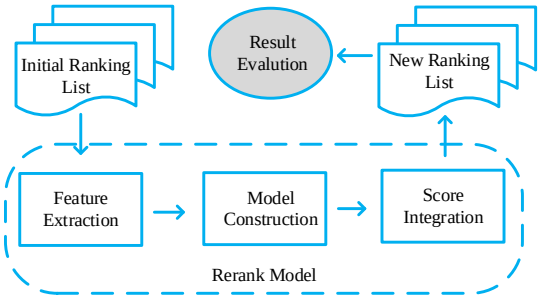

Fig. 2. The workflow of RPML. It is composed of three major steps: feature extraction, model construction and score integration. In feature extraction, we extract eleven features from both the initial PrSM and the corresponding spectrum as well as the protein sequence. In model construction, classification models are built to predict the probabilities of all PrSMs. In score integration, RPML aggregates the prediction results from multiple classifiers to generate a consensus probability for each PrSM.

2.3 Feature Extraction

In feature extraction, each PrSM is represented with eleven features listed in Table 1. To make the feature values comparable, the first 8 numeric features in Table 1 are transformed into [0,1] with the min-max normalization method:

$$z = \frac{f_i - min(f_i)}{max(f_i) - min(f_i)}\ , \tag{1}$$

where $max(f_i)$ and $min(f_i)$ represent the maximal value and the minimal value in a feature vector, respectively.

Other three features take binary values: 0 or 1. The *unexpected modification* and *residue* can be naturally marked as 0 or 1 as being specified in Table 1. The *charge state* is set to be 1 if the spectrum of a PrSM contains at most 15 charges, otherwise it will be 0.

Here we emphasize that RPML uses one feature (e.g.*E-value*) as the target feature and the remaining 10 features as dependent features. The choice of target

Table 1. Features used to represent PrSM.

No.	Feature	Description
1	mass difference	the absolute mass difference between the precursor mass of spectrum and protein mass
2	matched peaks	the number of matched peaks in the experimental spectrum
3	matched fragment ions	the proportion of matched fragment ions in the theoretical spectrum
4	P-value	the P-value reported by top-down approach
5	E-value	the E-value reported by top-down approach
6	protein length	the length of the matched protein
7	expected modifications	the number of PTMs in a set of known specific modifications
8	intensity of matched peaks	the sum of the intensity for matched peaks
9	charge state	is the number of charges smaller than 15 ?
10	unexpected modification	does the protein have unknown PTMs?
11	residue	is the difference between the first residue and the last residue greater than 50?

feature changes in accordance with the baseline identification method. We believe there would be other features that might be included, while these eleven features are sufficient to demonstrate the feasibility of our algorithm.

2.4 Model Construction

Since we already have an initial ranking list that is positively correlated with the set of true identifications, we can first construct the training data set based on the features that are used in the baseline identification methods. Then, we build the prediction model in parallel on multiple training data sets to predict probabilities for all PrSMs. Fig. 3 illustrates this process.

The PrSMs have no labels in the initial ranking list, we have to set up the labels artificially. Generally, the baseline identification method reports an initial ranking score rs (this score is the feature value from one of the eleven extracted features) for each PrSM. There will be a positive correlation between rs and the correctness of identifications. Therefore, we first sort all PrSMs based on rs. Then, we select a portion of top-ranked PrSMs as the candidate positive training set and the same portion of PrSMs at the end of the sorted list as the candidate negative training set. Finally, we randomly take a subset from both positive and negative candidate training sets to construct the training set used for model construction. Thus, we can generate multiple training data sets in parallel to build the prediction model so as to increase the stability and accuracy. It is difficult to determine the optimal size of both candidate training data and actual training data automatically. RPML select top 20% of PrSMs at the begining of sorted list to construct the candidate positive training data. Similarly, those 20% of PrSMs with lowest initial ranks are used as the candidate negative training

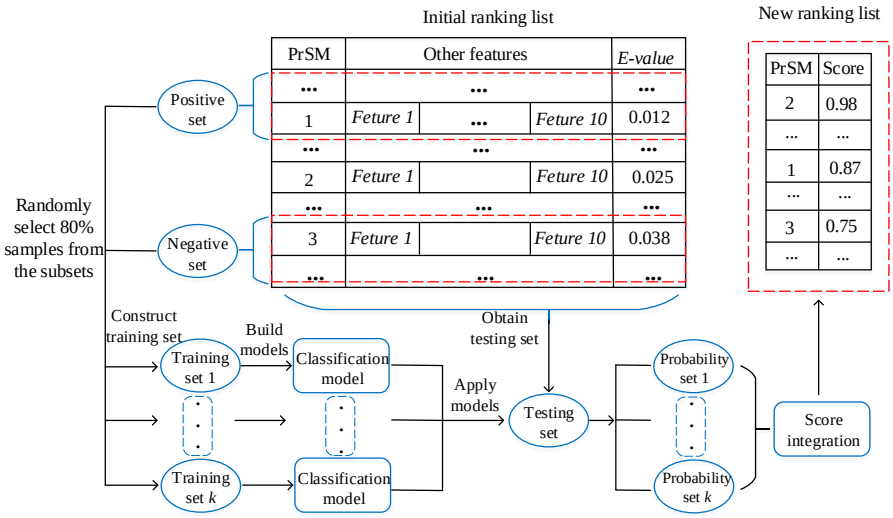

Fig. 3. Model construction process. We first sort the initial ranking list to generate a set of candidate training data. Then, we generate multiple training data sets by randomly sampling 80% of candidate training data. Finally, each training data set is used to construct a prediction model, which will assign each PrSM a new probability score.

data. In addition, we use a parameter c to control the size of actually selected training data. In our experiments, we set $c = 80\%$.

In this paper, we choose two widely used classifiers: xgboost (eXtreme Gradient Boosting) [5] and LR (Logistic Regression) [1] to construct prediction models on the training set to predict the probabilities for all PrSMs.

2.5 Score Integration

After model training and prediction, each classifier generates k scores for each PrSM, where k is the number of training data sets. For instance, xgboost will produce a probability score vector $SC_{xgb(i)} = (sc^1_{xgb(i)}, sc^2_{xgb(i)}, ..., sc^n_{xgb(i)})^T$ for the set of PrSMs based on the ith training data set, where n is the total number of PrSMs. As a result, there will be k new score vectors for the set of n PrSMs. In this paper, we use a simple approach to get the consensus score of each PrSM via calculating the arithmetic mean of its k scores. This means that each training data set is assigned with the same weight when considering its contribution to the final score for one classification model. For the xgboost method, the score for the ith PrSM is:

$$sc^i_{xgb} = \frac{sc^i_{xgb(1)} + sc^i_{xgb(2)} + ... + sc^i_{xgb(k)}}{k} . \quad (2)$$

Similarly, we can get the score of the ith PrSM for LR:

$$sc^i_{lr} = \frac{sc^i_{lr(1)} + sc^i_{lr(2)} + ... + sc^i_{lr(k)}}{k} . \quad (3)$$

To integrate the results from xgboost and LR, we set a parameter α to control the weights of two classifiers. The final consensus score of the ith PrSM is:

$$sc^i_{final} = \alpha sc^i_{xgb} + (1 - \alpha) sc^i_{lr} , \quad (4)$$

where α is set to be 0.5 in our experiments. After the score integration step, we can order the PrSMs according to the new ranking vector SC_{final}.

3 Experiments

We implemented the RPML algorithm in Python and compared its performance with MS-Align+ [21] on four data sets. All experiments were conducted on a computer with Intel Xeon E5607 2.3GHz CPU.

3.1 Data sets

In the experiments, we used four data sets to test the performance of RPML: Ecoli [3], H2A [29], ST [30] and Autopilot [6]. Each MS/MS spectrum of the four data sets was converted to a deconvoluted spectrum using MS-Deconv [20]. Only the spectra that have a precursor mass $\geq$ 2500 Da and at least 10 fragment peaks were retained for protein identification. The corresponding protein databases were downloaded from NCBI (`https://www.ncbi.nlm.nih.gov/`). The target-decoy strategy [7] was used for performance evaluation, where each decoy protein was generated by shuffling the corresponding protein sequence in the target databases. MS-Align+ was used as the baseline method with the following parameter settings: *searchType*=target+decoy, *shiftNumber*=5, *errorTolerance*=15, *cutoff*=0.01, and default values are used for other parameters. Meanwhile, MS-Align+ reported only the PrSM with the best matching score for each spectrum. Table 2 shows the number of spectra and the number of target and decoy PrSMs in the initial identification results of MS-Align+.

Table 2. Data sets used in experiments and the distribution of target and decoy PrSMs in the ranking list generated from MS-Align+.

Data set	#Spectra	#Target PrSMs	#Decoy PrSMs
Ecoli	1848	925	943
H2A	3529	2992	537
ST	4460	1725	2735
Autopilot	12102	7698	4404

3.2 Performance Evaluation

To compare the performance different methods, we plot a curve for each competing approach. In such a curve, the number of true positives (target PrSMs) is used as the y-axis and the q-value [26] is used as the x-axis. If we specify a score threshold t and refer to PrSMs with scores that are better than t as accepted PrSMs, the false discovery rate (FDR) is defined as the percentage of accepted PrSMs that are incorrect. The q-value is defined as the minimal FDR threshold at which a given PrSM is accepted. Here the target-decoy strategy is utilized to calculate the q-value. We use $h_1, h_2, ..., h_{mh}$ to denote the scores of target PrSMs and $d_1, d_2, ..., d_{md}$ to represent the scores of decoy PrSMs. The FDR at the rejection threshold t is defined as:

$$FDR(t) = \frac{\pi_0 \frac{mh}{md} |\{d_i > t; i = 1, ..., md\}|}{|\{d_i > t; i = 1, ..., md\}| + |\{h_i > t; i = 1, ..., mh\}|} , \tag{5}$$

where π_0 is the estimated proportion of target PrSMs that are incorrect (this value can be specified to be 1). The q-value for a given score gs is:

$$q(gs) = \min_{gs' \leq gs} \{FDR(gs')\} . \tag{6}$$

3.3 Results

We firstly compared the RPML algorithm with MS-Align+. Our goal is to correctly identify as many target PrSMs as possible for a given q-value. Therefore, in Fig. 4, we plot the number of identified target PrSMs as a function of q-value threshold. Note that RPML ensembles xgboost and LR to construct a prediction model, we also included these two classifiers in the performance comparison. That is, RPML equals xgboost and LR when the papameter $\alpha = 1$ and 0, respectively.

From the comparison results in Fig. 4, we can see that RPML has better performance than MS-Align+ on all data sets. This indicates that our method is capable of improving the initial ranking list of PrSMs generated from MS-Align+. For xgboost and LR, there is no method can always achieve better performance than the other. LR has better performance on the Ecoli and H2A data set, while xgboost exhibits good performance on the ST and Autopilot data set. This means that the results produced by different methods are very diverse. In contrast, RPML is able to obtain the best performance on the H2A, ST and Autopilot data set when the number of target PrSMs is apparently different from that of the decoy PrSMs. It is no doubt that the ensemble strategy is one key factor to the success of RPML.

In fact, the good performance of RPML can be partially attributed to the unbalanced distribution of target and decoy PrSMs. For instance, more than 80% PrSMs are target PrSMs in the H2A data set. As a result, the positive training data will be mainly composed of target PrSMs, while most of decoy PrSMs will be included in the negative training data. Such a training data set is extremely

helpful to build an accurate prediction model. Note that RPML can still achieve good performance when the target and decoy PrSMs have a similar proportion, as shown in Fig. 4(a).

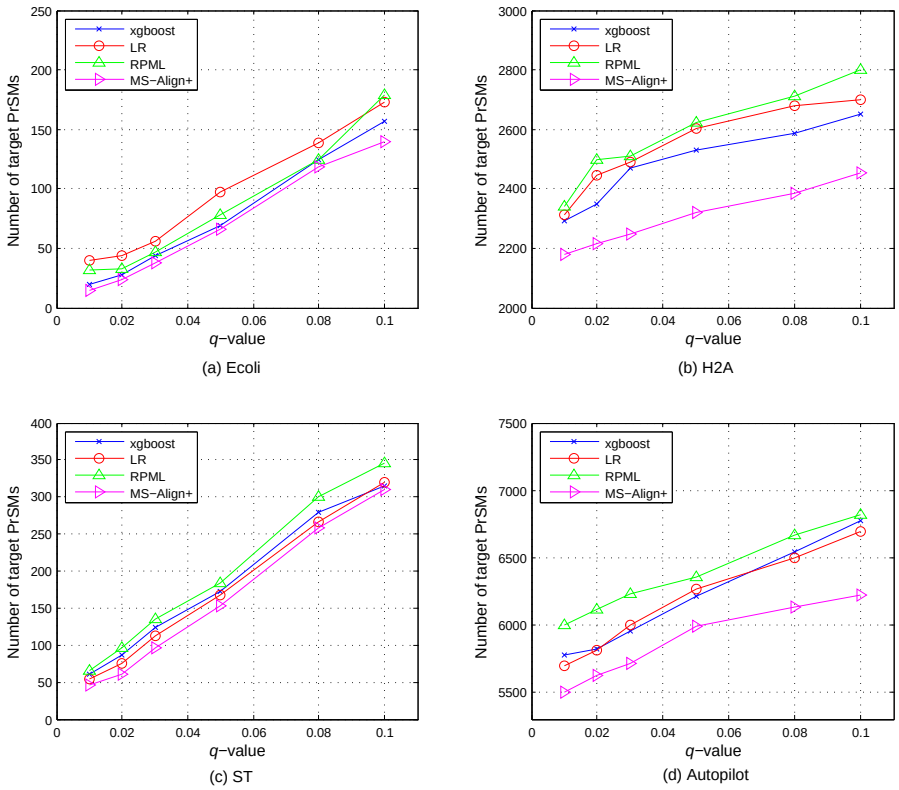

Fig. 4. Performance comparison among four algorithms. The parameters for RPML were specified as: $\alpha = 0.5$, $c = 80\%$ and $k = 100$.

3.4 Parameter sensitivity

In model construction, we generate multiple training data sets to increase the stability and accuracy. The number of training data sets k will influence the performance of RPML. To test the sensitivity of this parameter, we vary k from 10 to 200 to check its influence on the number of identified target PrSMs in Fig. 5. From Fig. 5, it is clearly visible that the identification performance will be improved when k is increased. In addition, it can also be observed that the increase on the number of target PrSMs is neglectable when $k > 100$. In practice, k can be set as a value in the range $[100, 110]$ so as to obtain a stable result within a reasonable time slot.

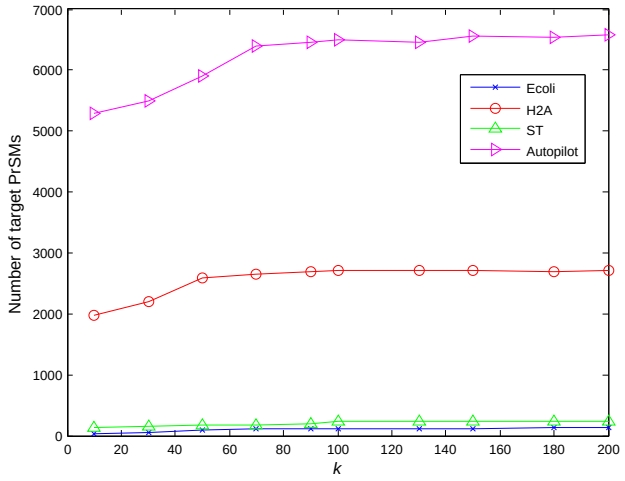

Fig. 5. The influence of parameter k on the final result of RPML when $\alpha = 0.5$, $c = 80\%$ and q-value$= 0.05$.

4 Conclusion

Ranking true Protein-Spectrum Matches (PrSMs) over their false counterparts is a feasible strategy to improve the protein identification performance in top-down proteomics. In this paper, we propose the first PrSM reranking algorithm in the literature. Our new algorithm is called RPML, which is based on the idea of supervised ensemble learning and a novel adaptive training data construction method. The experiments on four data sets show that the proposed algorithm is capable of distinguishing more correct identifications from incorrect ones. Actually, many research efforts in bottom-up proteomics have demonstrated that the use of different classifiers and different features may led to significantly different prediction performance (e.g. [9, 11, 12, 14, 25, 32]). Therefore, we will investigate the feasibility of incorporating different features into our prediction model according to different baseline methods in the future work. Moreover, new effective training data selection strategies will be further investigated as well.

Acknowledgements

This work was partially supported by the Natural Science Foundation of China (Nos. 61572094, 61502071), the Fundamental Research Funds for the Central Universities (Nos. DUT2017TB02, DUT14QY07) and the Science-Technology Foundation for Youth of Guizhou Province (No.KY[2017]250).

References

1. Bishop, C.: Pattern recognition and machine learning. Springer, New York (2007)

2. Cai, W., Guner, H., Gregorich, Z.R., Chen, A.J., Ayazguner, S., Peng, Y., Valeja, S.G., Liu, X., Ge, Y.: Mash suite pro: A comprehensive software tool for top-down proteomics. Molecular and Cellular Proteomics 15(2), 703–714 (2016)
3. Cannon, J.R., Cammarata, M., Robotham, S.A., Cotham, V.C., Shaw, J.B., Fellers, R.T., Early, B.P., Thomas, P.M., Kelleher, N.L., Brodbelt, J.S.: Ultraviolet photodissociation for characterization of whole proteins on a chromatographic time scale. Analytical Chemistry 86(4), 2185–2192 (2014)
4. Chait, B.T.: Mass spectrometry: Bottom-up or top-down? Science 314(5796), 65–66 (2006)
5. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. pp. 785–794. ACM (2016)
6. Durbin, K.R., Fellers, R.T., Ntai, I., Kelleher, N.L., Compton, P.D.: Autopilot: An online data acquisition control system for the enhanced high-throughput characterization of intact proteins. Analytical Chemistry 86(3), 1485–1492 (2014)
7. Elias, J.E., Gygi, S.P.: Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nature Methods 4(3), 207–214 (2007)
8. Fellers, R.T., Greer, J.B., Early, B.P., Yu, X., Leduc, R.D., Kelleher, N.L., Thomas, P.M.: Prosight lite: Graphical software to analyze top-down mass spectrometry data. Proteomics 15(7), 1235–1238 (2015)
9. Frank, A.: A ranking-based scoring function for peptide-spectrum matches. Journal of Proteome Research 8(5), 2241–2252 (2009)
10. Frank, A., Pesavento, J.J., Mizzen, C.A., Kelleher, N.L., Pevzner, P.A.: Interpreting top-down mass spectra using spectral alignment. Analytical Chemistry 80(7), 2499–2505 (2008)
11. He, Z., Yu, W.: Improving peptide identification with single-stage mass spectrum peaks. Bioinformatics 25(22), 2969–2974 (2009)
12. He, Z., Zhao, H., Yu, W.: Score regularization for peptide identification. asia pacific bioinformatics conference 12(1), 1–10 (2011)
13. Huang, T., Wang, J., Yu, W., He, Z.: Protein inference: a review. Briefings in Bioinformatics 13(5), 586–614 (2012)
14. Kall, L., Canterbury, J.D., Weston, J., Noble, W.S., Maccoss, M.J.: Semi-supervised learning for peptide identification from shotgun proteomics datasets. Nature Methods 4(11), 923–925 (2007)
15. Kou, Q., Xun, L., Liu, X.: Toppic: a software tool for top-down mass spectrometry-based proteoform identification and characterization. Bioinformatics 32(22), 3495–3497 (2016)
16. Lane, N.M., Gregorich, Z.R., Ge, Y.: Top-down proteomics. In: Manual of Cardiovascular Proteomics, pp. 187–212. Springer (2016)
17. LeDuc, R.D., Taylor, G.K., Kim, Y.B., Januszyk, T.E., Bynum, L.H., Sola, J.V., Garavelli, J.S., Kelleher, N.L.: Prosight ptm: an integrated environment for protein identification and characterization by top-down mass spectrometry. Nucleic acids research 32(suppl_2), W340–W345 (2004)
18. Li, L., Tian, Z.: Interpreting raw biological mass spectra using isotopic mass-to-charge ratio and envelope fingerprinting. Rapid Communications in Mass Spectrometry 27(11), 1267–1277 (2013)
19. Liu, X., Hengel, S.M., Wu, S., Tolic, N., Pasatolic, L., Pevzner, P.A.: Identification of ultramodified proteins using top-down tandem mass spectra. Journal of Proteome Research 12(12), 5830–5838 (2013)

20. Liu, X., Inbar, Y., Dorrestein, P.C., Wynne, C., Edwards, N., Souda, P., Whitelegge, J.P., Bafna, V., Pevzner, P.A.: Deconvolution and database search of complex tandem mass spectra of intact proteins a combinatorial approach. Molecular and Cellular Proteomics 9(12), 2772–2782 (2010)
21. Liu, X., Sirotkin, Y., Shen, Y., Anderson, G., Tsai, Y.S., Ying, S.T., Goodlett, D.R., Smith, R.D., Bafna, V., Pevzner, P.A.: Protein identification using top-down spectra. Molecular and Cellular Proteomics Mcp 11(6), M111.008524 (2012)
22. Nilsson, T., Mann, M., Aebersold, R., Yates, J.R., Bairoch, A.M., Bergeron, J.J.M.: Mass spectrometry in high-throughput proteomics: ready for the big time. Nature Methods 7(9), 681–685 (2010)
23. Noble, W.S., MacCoss, M.J.: Computational and statistical analysis of protein mass spectrometry data. PLoS computational biology 8(1), e1002296 (2012)
24. Park, J., Piehowski, P.D., Wilkins, C., Zhou, M., Mendoza, J., Fujimoto, G.M., Gibbons, B.C., Shaw, J.B., Shen, Y., Shukla, A.K.: Informed-proteomics: open-source software package for top-down proteomics. Nature Methods 14(9), 909 (2017)
25. Spivak, M., Weston, J., Bottou, L., Kall, L., Noble, W.S.: Improvements to the percolator algorithm for peptide identification from shotgun proteomics data sets. Journal of Proteome Research 8(7), 3737–3745 (2009)
26. Storey, J.D.: A direct approach to false discovery rates. Journal of The Royal Statistical Society Series B-statistical Methodology 64(3), 479–498 (2002)
27. Sun, R.X., Luo, L., Chi, H., Liu, C., He, S.M.: Top-down proteomics: The large-scale proteoform identification. Progress in Biochemistry and Biophysics 42(2), 101–114 (2015)
28. Sun, R., Luo, L., Wu, L., Wang, R., Zeng, W., Chi, H., Liu, C., He, S.: ptop 1.0: A high-accuracy and high-efficiency search engine for intact protein identification. Analytical Chemistry 88(6), 3082–3090 (2016)
29. Tian, Z., Tolic, N., Zhao, R., Moore, R.J., Hengel, S.M., Robinson, E.W., Stenoien, D.L., Wu, S., Smith, R.D., Pasatolic, L.: Enhanced top-down characterization of histone post-translational modifications. Genome Biology 13(10), 1–9 (2012)
30. Tsai, Y.S., Scherl, A., Shaw, J.L., Mackay, C.L., Shaffer, S.A., Langridgesmith, P.R.R., Goodlett, D.R.: Precursor ion independent algorithm for top-down shotgun proteomics. Journal of the American Society for Mass Spectrometry 20(11), 2154–2166 (2009)
31. Whitelegge, J.P.: Intact protein mass spectrometry and top-down proteomics. Expert Review of Proteomics 10(2), 127–129 (2013)
32. Yang, C., He, Z., Yang, C., Yu, W.: Peptide reranking with protein-peptide correspondence and precursor peak intensity information. IEEE/ACM Transactions on Computational Biology and Bioinformatics 9(4), 1212–1219 (2012)
33. Zamdborg, L., LeDuc, R.D., Glowacz, K.J., Kim, Y.B., Viswanathan, V., Spaulding, I.T., Early, B.P., Bluhm, E.J., Babai, S., Kelleher, N.L.: Prosight ptm 2.0: improved protein identification and characterization for top down mass spectrometry. Nucleic acids research 35(suppl_2), W701–W706 (2007)