Journal of Computer and Communications, 2023, 11, 99-109
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

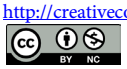
Classification of B and Y Ions in Peptide MS/MS Spectra Based on Machine Learning

Xinming Li

School of Computer Science and Technology, Shandong University of Technology, Zibo, China
Email: 1183307341@qq.com

How to cite this paper: Li, X.M. (2023) Classification of B and Y Ions in Peptide MS/MS Spectra Based on Machine Learning. *Journal of Computer and Communications*, **11**, 99-109.
https://doi.org/10.4236/jcc.2023.113008

Received: February 20, 2023
Accepted: March 28, 2023
Published: March 31, 2023



Abstract

In proteomics, b and y ions serve as the backbone ions for peptide sequencing in tandem mass spectrometry. Leveraging the existing ion recognition and separation methods, this article proposes a novel ion classification approach that combines machine learning with graph theory. By incorporating graph features, the method achieves higher accuracy and efficiency in ion type recognition, with the graph features playing a critical role in the classification process. Specifically, the method achieves a recall rate of nearly 90% for b and y ions, demonstrating its effectiveness in pre-processing de novo sequencing and improving its accuracy. The proposed method represents advancement in ion classification and has the potential to improve the accuracy and efficiency of de novo sequencing.

Keywords

Ion-Type Classification, Machine Learning, LightGBM, Proteomics, Tandem Mass Spectrometry

1. Introduction

As biotechnology advances, proteomics has gained increasing attention, and using computer technology to address proteomics problems is an important direction of development. The origin of computational proteomics lies in how to extract useful information from peptide and protein sequences.

After protein digestion, a mixture of peptides is obtained, which is then separated and enters the mass spectrometer. The mass spectrometer measures the mass-to-charge ratio of ions using the principle that differently charged ions of different masses move differently in an electromagnetic field, forming a mass spectrum. The horizontal axis of the mass spectrum represents the mass-to-charge ratio of the detected ions, and the vertical axis represents the intensity of the de-

tected ions.

The mass spectrometer first detects the charged ions corresponding to the peptide segments, forming a spectrum, where each peak corresponds to a peptide. The mass spectrometer then continues to select the peaks with higher intensities in the primary spectrum for fragmentation, breaking the peptide ions into charged fragment ions, forming a tandem mass spectrum [1] [2] [3] [4], as shown in Figure 1.

In secondary mass spectrometry, each peak corresponds to a fragment ion. Different types of fragment ions are produced at different positions, including N-terminal a, b, c ions and C-terminal x, y, z ions, with b and y ions commonly used for determining peptide sequences [5] [6] [7], as shown in Figure 2.

In the protein sequencing technology based on tandem mass spectrometry, database searching and de novo sequencing are two main identification methods. The database searching method searches for peptide sequences that match the tandem mass spectra in protein databases, while the de novo sequencing method infers peptide sequences directly from the mass spectra. The database searching method usually has higher accuracy, but it is limited by the protein database availability. For proteins that are difficult to obtain in databases, database searching method cannot be used for sequencing.

With de novo sequencing method, it is not relying on protein databases, but inferring the sequence based on the mass difference between peaks in the mass spectrum. Therefore, the quality of the data directly determines the performance of de novo sequencing method. When the ion coverage in the mass spectrum is

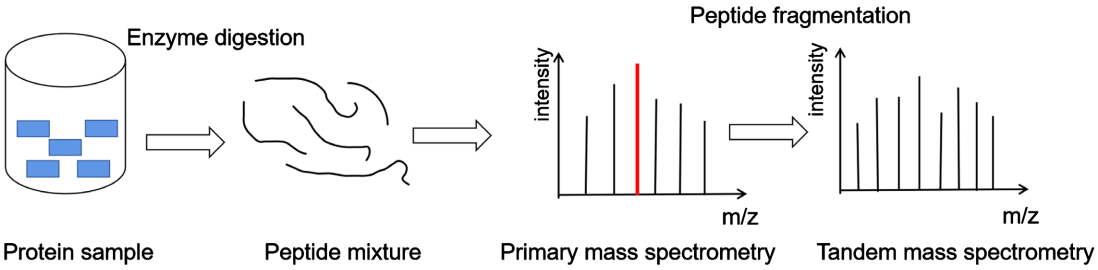


Figure 1. Mass spectrum data acquisition process.

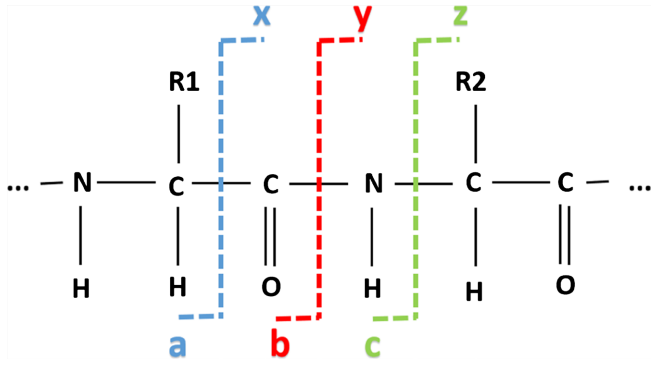


Figure 2. Different types of fragment ions formed by fragmentation of the main position of the peptide segment.

insufficient, the local sequence cannot be inferred correctly; and when the method is interfered by noise peaks, erroneous local correlations may be established. Therefore, preprocessing can be used to select b and y ion peaks, reduce the interference of noise peaks, and improve the accuracy of de novo sequencing, shorten processing time. Therefore, identifying and separating b and y ions is a common task in mass spectrometry analysis. In summary, how to improve the accuracy of b and y ion identification, and thus improve the accuracy of de novo sequencing, is an area worth further research.

In the process of identifying the development of b and y ions, there are two typical methods. In 2005, Bo Yan *et al.* proposed a graph theory method to solve the separation of b and y ions. Taking each spectral peak as a node, the first type of edge connects two peaks that may be of the same ion type, and the second type of edge connects two peaks that may be of different ion types, turning the ion separation problem into a graph partitioning problem. In this paper, a dynamic programming algorithm is developed to strictly solve the graph partitioning problem. Through a large number of simulated mass spectrometry and 19 sets of high-quality experimental tandem mass spectrometry tests, the separation accuracy of b and y ions reaches 90%. In 2013, James P Cleveland *et al.* proposed a neural network method to identify b and y ions, and generate multiple specific feature vectors for each spectral peak according to the characteristics of the data. This method improves other preprocessing techniques without detailed description of the peptide fragmentation process, which reduces the search space of candidate peptides without sacrificing the quality of candidate peptides. After preprocessing by this method, the accuracy and recall rate of de novo sequencing are higher than those of PepNovo + and pNovo [8] [9] [10] [11] [12].

In summary, based on the existing ion type identification and separation methods, combined with graph theory and neural network methods to extract commonalities, new ion type identification methods can be developed. This can be used for effective preprocessing in de novo sequencing to improve the accuracy of de novo sequencing.

2. Model and Feature Engineering

2.1. Model

GBDT (Gradient Boosting Decision Tree) is a popular machine learning model, which utilizes decision trees to iteratively train and obtain the optimal model. The model has advantages such as good training effects and being less prone to overfitting. LightGBM (Light Gradient Boosting Machine) is a framework that implements the GBDT algorithm. It supports efficient parallel training, and can achieve faster training speed, lower memory consumption, and better accuracy [13] [14] [15]. The machine learning-based LGB (LightGBM) model can be used for the classification and identification of b and y ions. As an efficient decision tree algorithm, it can handle large-scale and high-dimensional data, and there-

fore has wide applications in mass spectrometry analysis.

In this article, we employ the LightGBM machine learning model to achieve classification of ions. By tuning the model's parameters, including the depth of decision tree, learning rate, and regularization parameters, we enhance the model's generalization ability and classification performance while maintaining its accuracy. Furthermore, we utilize the feature_importance method and conduct comparative experiments to identify the crucial features that significantly contribute to the model. These features are then retained to further improve the model's performance. The approach significantly reduces the training time of the model without compromising its accuracy.

2.2. Label Selection

The quality can be calculated based on the known sequence and modification information. The specific formula for calculating the mass of the neutral parent ion ($Mass(p)$) is as follows: given the recorded mass-to-charge ratio (m) and charge (zp) of the parent ion, $Mass(p)$ can be calculated. If the amino acid composition of a known prefix or suffix subsequence of the peptide is known, the theoretical mass-to-charge ratios of other related fragment ions can be obtained [16] [17] [18]. Let $P = \{AAi\}$ be a subsequence of the peptide, and let z be the charge of the fragment ion and massAa be the molecular weight of a certain amino acid residue. When P is a prefix sequence, the mass-to-charge ratio of the b ion can be calculated using formula (2.1), and then the corresponding mass-to-charge ratio of the y ion can be calculated using formula (2.2).

$$mz_b = \left(\sum_{i=1}^{len} mass_{AAi} + z * mass_{proton}\right) \Big/ 2 \tag{2.1}$$

$$mz_y = \left[Mass(p) - mz_b * z + 2 * z * mass_{proton}\right] \Big/ 2 \tag{2.2}$$

Actual matching of b and y ions: According to the theoretical mass-to-charge ratio of y ions, the mass-to-charge ratio of the actual fragment ions in each spectrum is matched in turn. The actual fragment ions that are matched within the error range and have the smallest mass-to-charge ratio error with the theoretical y ions are the actual matching y ions, which are also the labels required for this experiment.

2.3. Feature Engineering

In the process of b, y ion classification based on the Light GBM model, it is necessary to select features with discriminative power for training and classification. Therefore, selecting appropriate features has a crucial impact on the training and classification performance of the model. In general, feature design needs to be considered comprehensively based on the features of the mass spectrometry data and experimental data. Different feature combinations should be selected for different datasets and problems to obtain better classification performance.

For the handcrafted features extracted from the data, we chose a set of qualita-

tive features related to the secondary spectrum matching. Based on each secondary spectrum and its number of peaks, peak matching continuity, peak matching intensity, and peak matching mass deviation, we searched and calculated features such as fragment ion mass-to-charge ratio peaks, relative intensity peaks, intensity ratio peaks, isotope peaks, and mass difference peaks. Table 1 lists all the features that were calculated, and in the following text, we provide a detailed explanation of the most important representative features.

2.3.1. Features of mz

To extract comprehensive information of the mass-to-charge ratio and differentiate between different ion types, statistical features including mean, standard deviation, maximum value, and minimum value are calculated for each peak's mass-to-charge ratio data.

2.3.2. Features of Intensity

Normalization of intensity values is necessary because different ions have different intensity values, which need to be normalized for comparison. For each spectrum, the maximum intensity value is found, and normalization is performed on a spectrum-by-spectrum basis. After normalization, the intensity becomes a relative value, highlighting the b and y ions with significant intensity in the spectrum, making them distinguishable from other ions.

2.3.3. Features of Graph

Firstly, graph structures can be used to describe the relationships, feature extraction, and anomaly detection among ions in ion classification, which can improve the accuracy and robustness of ion classification, and also help analyze and visualize the relationships between ions. The spectral peak quality connection graph is an important feature that clearly displays the correlations between fragment ions. The construction method of mass spectrometry connection graphs can model the topological structure between ions in different ways, such as adjacency matrices and adjacency lists, to obtain more accurate topological information.

Therefore, this graph structure feature, which combines the principle information of the mass spectrum connection graph and establishes the association between

Table 1. Pattern features for model.

Feature	Value
mz	N, D
Intensity	N, D
Relative-intensity-ratio	N, D
Isotopologue	D
Gragh-node	D
Gragh-Acid mass	D

N specifies a normalized quantity, D specifies a discretized quantity.

peaks, has become a strong feature that can find more accurate candidate peaks and improve the overall performance of the model, thereby improving the recall rate of b and y ion identification.

Following the graph theory method of constructing graphs, if the difference between the distance between two peaks and the molecular weight of one amino acid residue falls within the set error range, an edge is created between the two peaks.

Let spectra be $S = \{(m_k, i_k)\}_{k=1}^{n_{peaks}}$ (m_k represents the mass of the kth peak and i_k represents the intensity of the kth peak, n_{peaks} refers to the number of peaks in a spectrum.), SA be the matrix of differences between peaks, $MASS(AA_k) = \{AA_k\}_{k=1}^{n=23}$, n = 23 represent the set of amino acid residues, including 20 standard residues and 3 modified residues. The process of calculating the adjacency matrix is expressed using Equations (2.3)-(2.6):

$$A_k = \left| abs(SA) - MASS(AA_k) \right| \tag{2.3}$$

$$A_k = \begin{cases} 1, & a_{ij} \le \varepsilon \\ 0, & a_{ij} > \varepsilon \end{cases} \tag{2.4}$$

$$A_s = \sum A_k \tag{2.5}$$

$$\tilde{A} = A_s + E \tag{2.6}$$

Using formula (2.3), the absolute value of the peak difference matrix is calculated with the error matrix for each amino acid residue. If the error is within the given ε, the corresponding element is marked as 1; if it exceeds the range, it is marked as 0. Then, all the matrices are added to obtain the adjacency matrix of the current spectrum. The same dimension identity matrix E is added to incorporate the vertex self-information of the peak, which avoids the situation where the peak is isolated in the graph construction, *i.e.*, there is no adjacent ion peak of the same type in the spectrum due to the break of adjacent positions, and there is no edge connected to it.

Previously, we discussed how the B and Y ions in the spectrum form a sequence based on the mass difference of amino acid residues, as shown in **Figure 3**. Once the graph is constructed, Y and B ions in the spectrum are connected in a path based on the mass of the amino acid residues, forming a connected path as shown in **Figure 4**. The greater the abundance of B/Y ions in the spectrum, the more complete the connected path will be. To extract a feature from this, the mass constraint relationship is used to identify the longest path in the graph through depth-first search (DFS). The amino acid mass of the nodes on the longest path is then used as the node feature, while the other node features are left empty.

After constructing the connections between the peaks through graph construction, the information of the longest edge is extracted as a hand-crafted feature and fed into the existing model for learning. Essentially, this feature fuses

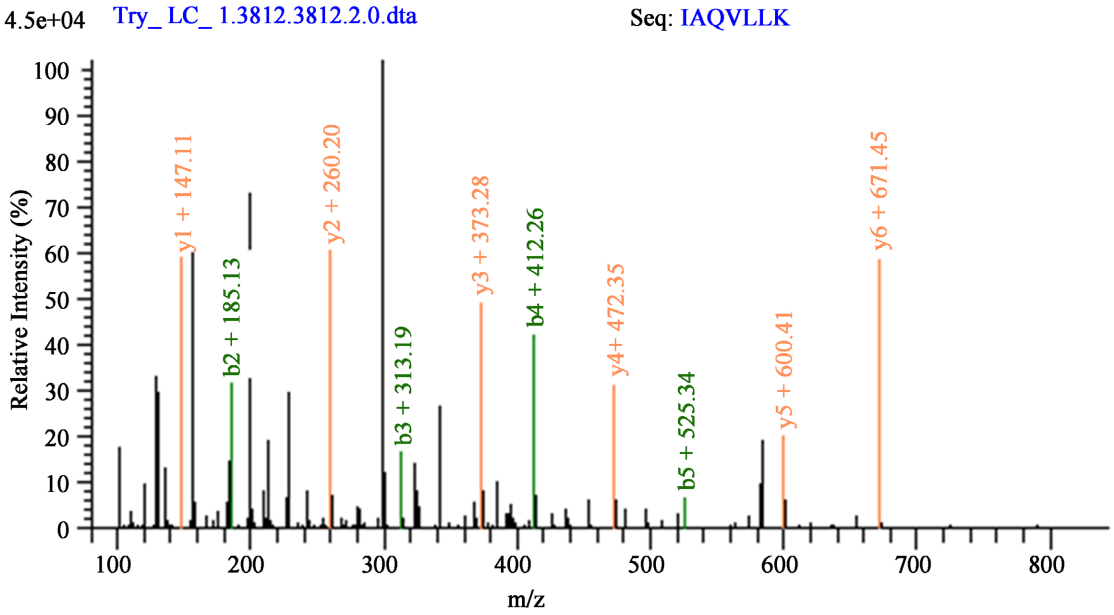


Figure 3. The labeled B-/Y-ions in tandem mass spectrometry.

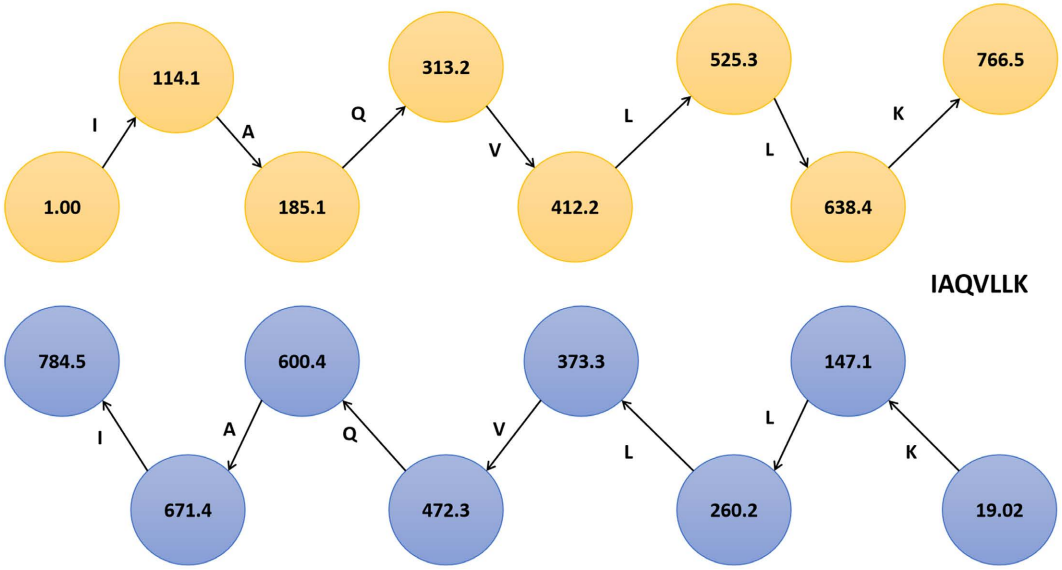


Figure 4. B-/Y-ion connection path constitutes a sequence.

the information from the mass spectrum connection graph and searches for the longest path, providing more accurate candidates for identifying b and y ions, thereby improving the recall rate of b and y ion identification.

After the graph is constructed, node degree can also be used as a feature to extract. For each node in the spectrum, its degree (*i.e.* the number of adjacent nodes) is calculated. For b and y ions, which connect adjacent amino acid residues, they typically connect to more nodes. Recording the degree and position of nodes as features can extract key nodes at critical positions. Therefore, by calculating the degree of nodes, it can better distinguish b, y ions from other ions.

2.3.4. Summary of Features

After considering all the feature designs, to further improve the identification of b and y ions, we need to return to the graph structure. We will add the previously designed features of mass-to-charge ratio peaks, fragment ion relative intensity peaks, isotope peaks, and mass difference peaks to the features of the nodes in the spectrum peak connection graph. By establishing the association between peaks, they become strong features. The relationships between nodes after connecting the edges will also be used as features. In summary, all features are included in the model for training, and the model outputs the classification of ion types.

This paper is based on machine learning and processes mass spectrometry data into graph data features. By processing mass spectrometry data into graph data, the most important information of fragment ions, which is the relationship between peaks, can be directly calculated and determined without the need to learn the relationship features between peaks through a deep learning model. This overall improves the performance of b and y ion identification and increases the accuracy of de novo sequencing.

3. Result

3.1. Dataset

The dataset was obtained by preprocessing the raw data using pParse to extract the tandem mass spectra, resulting in mgf files. Then, pFind was used for a targeted search, with the fasta file downloaded from the Uniprot database used as the search database, to parse the information of the spectrum peaks and their types. The dataset uses data from the yeast species, which has abundant b and y fragment ions and sufficient coverage of ions. The training set, test set a total of 30,000 spectra. Therefore, this dataset can be used for ion classification.

Preprocessing certain mass spectrometry data, such as denoising, normalization, and feature scaling, can improve the model's generalization ability and prediction accuracy. However, in mass spectrometry data, it is important to ensure sufficient fragment ion information to accurately classify ions. Therefore, in this dataset, no preprocessing was performed before classification.

The label of the number of b, y, and u ions (Unknown ions) calculated from the peak types of the tandem mass spectrum and the manually extracted features of b, y, and u ions were combined, and the training set was used to train the classifier. Training was stopped when the loss on the validation set was minimized, and the posterior probability estimate of b and y ions was output.

3.2. Evaluation Metrics and Results

Choosing appropriate metrics depends on the specific requirements of the task. In the ion classification task, both high accuracy and high recall are required, while controlling false positives and false negatives is also important. Therefore, precision and recall were chosen as evaluation metrics.

- Precision: the proportion of true positive samples among all samples classified as positive, *i.e.* TP/(TP + FP).
- Recall: The proportion of true positive samples that are correctly classified as positive, *i.e.*, TP/(TP + FN).

The LightGBM model, which has high accuracy and fast speed, was used for classification. Since each peak contains isotopic peaks and double peaks that are correlated before and after, the original order of the data was not disrupted. After training and saving the model, the test set was used for prediction, and metrics such as precision and recall were used to evaluate the model. After comparing the effects of various groups of results, insignificant features such as noise peaks, unknown ions, and intensity levels were discarded, while significant features that improved the model's quality were retained. Due to imbalanced data, the precision and recall of b and y ions were calculated separately to better understand the model's performance. Finally, based on the trained classifier, the maximum probability of output label is compared with the original label result. The calculated recall rate and accuracy rate is as follows Table 2.

We performed a table statistical analysis of the changes in the recall rate of Y ions before and after adding graph features, which play a crucial role in the features.

As shown in Figure 5, it can be seen that graph features are crucial for improving the accuracy of model classification, thus confirming that our approach of constructing mass spectral connectivity graphs to obtain features is correct, and ion classification must be based on graph structures.

Table 2. Classification results of b-/y-ions.

	Recall	Precision
Y	90.8%	85.3%
B	86.4%	82.5%

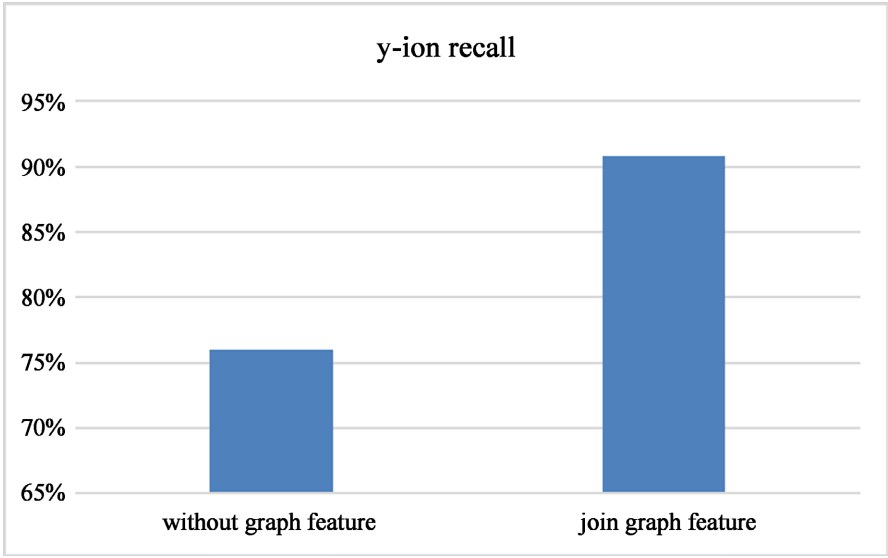


Figure 5. Comparison of y ion recall rate before and after adding graph feature.

4. Conclusions

In this paper, a method of ion classification of common mass spectrometry data under machine learning is proposed. Combined with the method of graph theory, a spectral peak connection diagram is constructed to identify and separate the main by ions to the greatest extent, and the pretreatment steps of protein sequencing are completed. The key ions are filtered out for sequencing, which lays a foundation for improving the accuracy of sequencing.

This article proposes a machine learning-based classification method that can effectively distinguish b, y ions from other ions, especially with the addition of new graph features, significantly improving the recall rate of ions and achieving good results in this experiment. However, this method also has certain limitations, as it may be affected to varying degrees by factors such as noise and mass spectrometry signal intensity in different mass spectrometry data. Therefore, the ion selection and improvement should be based on the specific situation of the ions in the spectrum. To improve the identification and classification of b and y ions using machine learning, it is essential to consider various factors such as data quality, feature selection, and model optimization, and to continuously iterate and optimize the method.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Yan, B., Pan, C., Olman, V., *et al.* (2005) A Graph-Theoretic Approach for the Separation of b and y Ions in Tandem Mass Spectra. *Bioinformatics*, **21**, 563-574. https://doi.org/10.1093/bioinformatics/bti044

[2] Cleveland, J.P. and Rose, J.R. (2013) Identification of b-/y-Ions in MS/MS Spectra Using a Two Stage Neural Network. *Proteome Science*, **11**, S4. https://doi.org/10.1186/1477-5956-11-S1-S4

[3] Alicia, L., *et al.* (2013) Neutron-Encoded Signatures Enable Product Ion Annotation from Tandem Mass Spectra. *Molecular & Cellular Proteomics*, **12**, 3812-3823. https://doi.org/10.1074/mcp.M113.028951

[4] Overmyer, K.A., Tyanova, S., Hebert, A.S., *et al.* (2018) Multiplexed Proteome Analysis with Neutron-Encoded Stable Isotope Labeling in Cells and Mice. *Nature Protocols*, **13**, 293-306. https://doi.org/10.1038/nprot.2017.121

[5] Rose, C.M., Merrill, A.E., Bailey, D.J., *et al.* (2013) Neutron Encoded Labeling for Peptide Identification. *Analytical Chemistry*, **85**, 5129-5137. https://doi.org/10.1021/ac400476w

[6] Potts, G.K., Voigt, E.A., Bailey, D.J., *et al.* (2016) Neucode Labels for Multiplexed, Absolute Protein Quantification. *Analytical Chemistry*, **88**, 3295-3303. https://doi.org/10.1021/acs.analchem.5b04773

[7] Tran, N.H., Zhang, X., *et al.* (2018) De Novo Peptide Sequencing by Deep Learning. *PNAS*, **114**, 8247-8252.

[8] Frank, A. and Pevzner, P. (2005) PepNovo: De Novo Peptide Sequencing via Proba-

bilistic Network Modeling. *Analytical Chemistry*, **77**, 964-973. https://doi.org/10.1021/ac048788h

[9] Renard, B.Y., *et al.* (2010) When Less Can Yield More—Computational Preprocessing of MS/MS Spectra for Peptide Identification. *Proteomics*, **9**, 4978-4984.

[10] Mo, L., Dutta, D., Wan, Y., *et al.* (2007) MSNovo: A Dynamic Programming Algorithm for de Novo Peptide Sequencing via Tandem Mass Spectrometry. *Analytical Chemistry*, **79**, 4870-4878. https://doi.org/10.1021/ac070039n

[11] Dimaggio, P.A. and Floudas, C.A. (2007) De Novo Peptide Identification via Tandem Mass Spectrometry and Integer Linear Optimization. *Analytical Chemistry*, **79**, 1433-1446. https://doi.org/10.1021/ac0618425

[12] Chi, H., Sun, R.X., Yang, B., *et al.* (2010) PNovo: De Novo Peptide Sequencing and Identification Using HCD Spectra. *Journal of Proteome Research*, **9**, 2713-2724. https://doi.org/10.1021/pr100182k

[13] Wessels, H., Bloemberg, T.G., Dael, M., *et al.* (2012) A Comprehensive Full Factorial LC-MS/MS Proteomics Benchmark Data Set. *Proteomics*, **12**, 2276-2281. https://doi.org/10.1002/pmic.201100284

[14] Zhou, X.X., Zen, W.F., Chi, H., *et al.* (2017) pDeep: Predicting MS/MS Spectra of Peptides with Deep Learning. *Analytical Chemistry*, **89**, 12690-12697. https://doi.org/10.1021/acs.analchem.7b02566

[15] Yan, D., Chen, H., Cheng, J., *et al.* (2018) Scalable De Novo Genome Assembly Using Pregel. 2018 *IEEE* 34*th International Conference on Data Engineering* (*ICDE*), Paris, 16-19 April 2018, 1216-1219.

[16] Ma, B., Zhang, K. and Liang, C. (2005) An Effective Algorithm for Peptide de Novo Sequencing from MS/MS Spectra. *Journal of Computer and System Sciences*, **70**, 418-430. https://doi.org/10.1016/j.jcss.2004.12.001

[17] Yang, H., *et al.* (2019) Precision De Novo Peptide Sequencing Using Mirror Proteases of Ac-LysargiNase and Trypsin for Large-Scale Proteomics. *Molecular & Cellular Proteomics*: *MCP*, **18**, 773-785.

[18] Nguyen, M.N. and Vien, N.A. (2019) Scalable and Interpretable One-Class SVMs with Deep Learning and Random Fourier Features: Recognizing Outstanding. Ph.D. Research.