Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 1, pp. 151–159, DOI 10.12694/scpe.v26i1.3722

PEPTIDE SEQUENCE TAG EXTRACTION BY GRAPH CONVOLUTION NEURAL NETWORKS

XINYE BIAN[*], DONGMEI XIE[†], DI ZHANG[‡], XIAOYU XIE[§], YUYUE FENG[¶], PIYU ZHOU[‖], CHANGJIU HE[**], MINGMING LYU[††] AND HAIPENG WANG[‡‡]

Abstract. The peptide sequence tag extraction method plays a vital role in tandem mass spectrometry-based protein identification engines. This approach faces two significant challenges in practical applications: first, the issue of fixed tag lengths, where shorter tags lack sufficient specificity, leading to an excessive recall of non-target peptide sequences, and longer tags experience a reduction in precision as tag length increases, potentially failing to recall target peptide sequences; second, the sensitivity and precision of tag extraction remain relatively low. To address these issues, a variable-length peptide sequence tag extraction algorithm, TagEx, based on graph convolutional networks, is proposed. This method begins by training a de novo peptide sequencing scoring model utilizing graph convolutional networks. It then constructs a spectral peak connection graph from the mass spectrum, employing a depth-first search strategy to extract variable-length peptide sequence tags, with the trained graph convolutional network model scoring amino acid connections during the extraction process.Finally, tags are filtered based on length and scoring to obtain the final candidate peptide sequence tag set. To evaluate TagEx's performance, it was benchmarked against three representative tag extraction software tools: InsPect, PepNovo+, and DirecTag. The experimental results demonstrate that TagEx exhibits superior sensitivity, coverage, and precision, with improvements of 0.62-2.32, 3.22-11.14, and 3.29-8.31 percentage points, respectively, when retaining the top 100 tags.

Key words: proteomics,peptide sequence tag, graph convolutional neural network, de novo sequencing, tandem mass spectrometry

1. Introduction. With the continuous advancement of scientific and technological progress, the precision of mass spectrometry instruments has steadily improved, making tandem mass spectrometry analysis an indispensable key technology for protein identification. Tandem mass spectrometry analysis can generally be categorized into three primary methods, including database-based searching, de novo sequencing, and tag-based database searching method[1]. Among these, the protein database search method is the most commonly used for protein identification, and widely used database search software includes MS-GF+[2], pFind-Alioth[3], Comet[4]and others. The method mainly refers to existing mass spectrometry data and peptide sequences in protein databases; therefore, a major drawback of the database search method is that it is unable to explore proteins in unknown areas. Another common method for protein identification is de novo sequencing, which no longer depends on protein databases but infers peptide sequences directly from spectral information. Common de novo sequencing tools include SMSNet[5] , PointNovo[6], Casanovo[7], denovoGCN[8], and more. During the de novo sequencing process, the entire sequence corresponding to the spectrum needs to be predicted, and even if there are slight differences from the actual peptide sequence, it is considered an incorrect sequence. However, these incorrect sequences may contain some correct fragment sequences that still have value in peptide spectrum matching. Therefore, the sequence tag method is an approach that combines the strengths of both database searching and de novo sequencing methods.

[*]School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[†]School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[‡]School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[§]School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[¶]School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[‖]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China
[**]School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[††]School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China
[‡‡]School of Computer Science and Technology, Shandong University of Technology, Zibo 255000, China (Corresponding author, hpwang@sdut.edu.cn)

The peptide sequence tagging method is a technique utilized in proteomics mass spectrometry analysis for the identification of peptide sequences. This method initially infers partial fragments of peptides, namely peptide sequence tags, from tandem mass spectrometry data. Subsequently, it searches a protein database for candidate peptide sequences containing these sequence tags to achieve the final peptide spectrum matching results[9]. Within the field of proteomics research, this approach holds significant value for the rapid identification of peptide sequences, as well as the discovery and localization of unknown modifications.

The concept of peptide sequence tagging can be traced back to 1994, initially proposed by Mann et al[10]. They defined sequence tags as segments of consecutive fragment ions within tandem mass spectra and used these sequence tags as keywords to retrieve corresponding peptide sequences from protein sequence databases. In 2003, Tabb et al. published the first standalone sequence tag extraction software, GutenTag, which utilized statistical methods to generate a model of spectral peak relative intensity, thereby facilitating large-scale extraction of sequence tags. In the same year, Sunyaev et al. introduced MultiTag, employing a fault-tolerant sequence tagging approach to recall homologous proteins and perform statistical evaluations. In 2005, Tanner et al[11]. developed the InsPect protein search engine, primarily utilizing dynamic programming for peptide sequence tag extraction and leveraging these tags to filter retrieved proteins. Concurrently, Frank et al. published the de novo sequencing tool, PepNovo, adopting probabilistic network modeling combined with graph theory and dynamic programming to infer amino acid sequences from mass spectral data. In 2008[12], Tabb et al. introduced a new sequence tag extraction tool, DirecTag, which generated sequence tags through recursive enumeration and scored tags based on intensity rank, error variance, and ion complementarity, demonstrating improved performance in sequence tag extraction compared to GutenTag. In 2009, Frank et al[13]. enhanced PepNovo by introducing a new scoring function based on adaptive boosting, resulting in PepNovo+. However, since then, standalone sequence tag extraction tools have become relatively scarce, with most tools being integrated within protein search engines without offering an independent sequence tag output interface, such as MODa[14], Open-pFind[15], and MODplus[16], among others.

Currently, the majority of sequence tag extraction tools are configured with fixed tag extraction lengths. To reduce the difficulty of extracting correct tags and to ensure the recall rate of target peptide sequences, tag lengths are typically constrained to a range of three to five amino acids. However, setting a fixed tag length presents certain disadvantages. When the sequence tag length is set too short, the specificity of the tags may be insufficient, leading to the recall of numerous non-target peptide sequences; conversely, longer tags may decrease in precision as their length increases, potentially failing to recall target peptide sequences. At the same time, due to the outdated algorithms used for scoring tags, the sensitivity and precision of most current tag extraction tools remain low, significantly impacting the effectiveness of tag extraction methods in proteomics data analysis.

This article introduces a tag extraction method named TagEx, based on Graph Convolutional Neural Networks (GCN). This method initially involves training a peptide sequencing scoring model utilizing graph convolutional networks. Subsequently, a spectral peak connection graph is constructed based on the mass spectrum, employing a depth-first traversal strategy to extract variable-length peptide sequence tags. During the tag extraction process, the trained graph convolutional network model is utilized to score amino acids on connecting edges. Finally, tags are filtered based on their length and scores to obtain the final candidate peptide sequence tag set. By integrating a variable-length tag extraction algorithm with the GCN model, TagEx exhibits outstanding performance in terms of tag sensitivity, coverage, and precision.

2. Materials and methods.

2.1. Dataset description. TagEx employs high-precision data comprising 1528127 spectra from nine species, including V.mungo, M.musculus, M.mazei, C.endoloripes, S.lycopersicum, S.cerevisiae, A.mellifera, H.sapiens, Bacillus, as the dataset. The first eight species are utilized to train the de novo sequencing model based on graph convolutional neural networks. The original data for these species are sourced from the PRIDE database, specifically from PXD005025[17], PXD004948[18], PXD004325[19], PXD004536[20], PXD004947[21], PXD003868[22], PXD004467[23] and PXD004424[24]. The data from another species not included in the training set, PXD004565[25], is used as the test set. All datasets mentioned are uniformly exported as MGF files using the pParse+ software, with peptide sequence identification performed using PEAKS software to obtain peptide spectrum matching information. The data is filtered using a False Discovery Rate (FDR)[26],

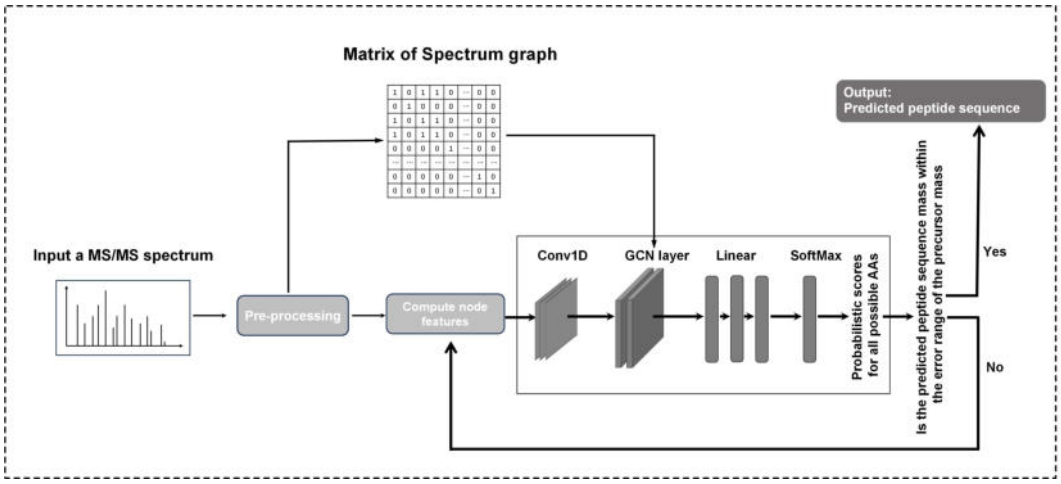

Fig. 2.1: De novo sequencing model training process.

the threshold of 0.01.

2.2. De novo sequencing model training based on graph convolutional networks. The Graph Convolutional Neural Network (GCN) was proposed by Kipf and Welling[27] in 2017 and swiftly became a seminal model within the domain of graph neural networks. Its foundational principle lies in conducting convolution operations on nodes and their adjacent nodes, thereby facilitating the extraction from local to global features. This feature enables GCNs to not only preserve graph structural information but also significantly enhance performance in various graph data tasks. The excellence of GCNs in tasks such as node classification and graph classification can be attributed to their convolution operations.

For the convenience of model performance testing, the training set is divided into an 8:2 ratio, with 80% allocated for training purposes and the remaining 20% serving as the validation set. The specific training process follows that of denovoGCN, as proposed by our research group. The initial step involves preprocessing the spectra, including the addition of four virtual peaks, converting the m/z ratio to neutral mass, and preserving spectral peaks among other operations. Subsequently, a mass spectrum connection matrix is constructed, and spectral peak features are developed based on the m/z ratio and intensity information of the currently predicted peptide sequence. Finally, the features and the mass spectrum connection matrix are input into the model to obtain probability estimates for all possible identities of the next amino acid, with the amino acid having the highest probability added to the predicted sequence. The sequence is output as the final de novo sequencing result once the mass of the predicted peptide sequence falls within the error range of the parent ion mass. The model training process is depicted in Figure 2.1.

2.3. Tag Extraction. Upon successful training of the model, it is utilized for tag extraction through the following specific process: 1) Spectrum preprocessing: conversion of spectral peak intensities to relative intensities; addition of four virtual peaks to the spectrum with m/z ratios equivalent to the mass of a proton, the mass of a singly charged water molecule, the mass of a singly charged parent ion, and the mass of a dehydrated singly charged parent ion, each with an intensity of 1; retention of the top 200 peaks with the highest intensities. 2) Construction of the spectral peak connection graph: starting from the first spectral peak in the preprocessed spectrum, all peaks are traversed. The mass difference between the current peak and other peaks with higher m/z ratios is calculated, and if the mass difference matches the mass of an amino acid, these two peaks are connected. 3) Amino acid edge scoring: during the traversal of spectral peaks, the mass of the traversed peak is input into the model to obtain scores for amino acids on connecting edges starting from the current peak. This continues until traversal is complete, resulting in a spectral peak connection graph with scored connecting amino acids. 4) Tag extraction: initially, a buffer is established to store the extracted amino

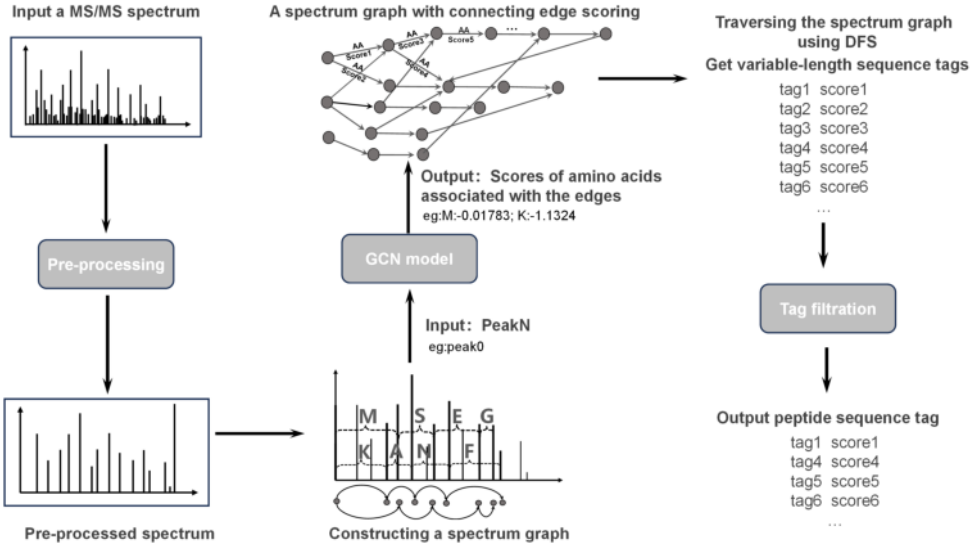

Fig. 2.2: Peptide sequence tag extraction flowchart.

acid sequence information. Then, all nodes with an indegree of zero are traversed, performing a depth-first traversal of the path from that node. During the depth-first traversal, the average of the scores of the amino acids on the traversed edges and those in the buffer is calculated to determine if this average is within the set scoring threshold t. If higher than the threshold, the amino acid and its score corresponding to the current edge are added to the buffer. If lower than the threshold, the length of the amino acid sequence in the buffer is assessed; if greater than 3, it is saved as a peptide sequence tag and the average of the amino acid scores is taken as the tag's score, then the tag and its score are saved before clearing the buffer; otherwise, the buffer is cleared directly. When traversal reaches a node with an outdegree of 0, the amino acid sequence in the buffer is saved as a tag and backtracking occurs. 5) Tag filtering: tags of varying lengths are filtered using different empirical scoring thresholds based on tag length, and the filtered results constitute the final candidate peptide sequence tag set. The peptide sequence tag extraction process is illustrated in Figure 2.2.

2.4. The performance evaluation indicators. TagEx utilizes three performance metrics to evaluate the tag extraction algorithm, including sensitivity[28], coverage[29], and precision. Initially, sequence tags extracted by various software are ranked based on their scoring, with the top-scored tags being selected sequentially. Sensitivity is defined as the proportion of spectra containing correct tags within the top n tags to the total number of spectra. The formula for calculating tag sensitivity can be represented by Equation (2.1), where s denotes the number of spectra from which correct tags can be extracted, and S represents the total number of spectra tested.

$$sensitivity = \frac{s}{S} \tag{2.1}$$

For the evaluation metric of coverage, consider a specific spectrum and its corresponding peptide sequence "PEPTIDESEQ" as an example. Suppose that the first tag extracted by the tag extraction algorithm is "PEPTI" and the second tag is "EPTID". In this scenario, the coverage for the top 2 tags of this spectrum is calculated as the number of covered amino acids, 6, divided by the total length of the peptide sequence, 10, resulting in a coverage value of $\frac{6}{10} = 0.6$ for the top 2 tags of this spectrum. Subsequently, by calculating the coverage for all spectra in the dataset and then averaging these values, the overall dataset coverage is determined. Furthermore, in the assessment of coverage, the distribution of coverage by extracted tags when selecting the top 100 tags in each spectrum is also considered. This implies that not only the coverage of individual tags is calculated, but a comprehensive evaluation is also performed on the amino acids covered by all tags within each spectrum,

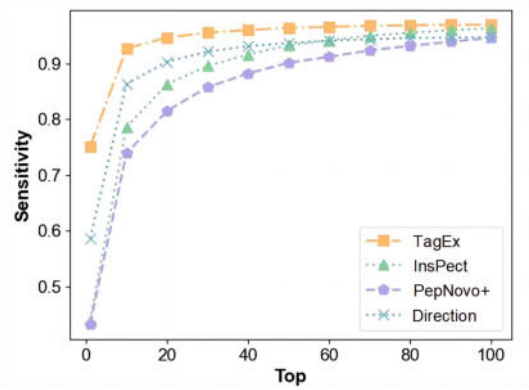

Fig. 3.1: Comparison of tag extraction sensitivity.

leading to a more holistic analysis of coverage. The formula for tag coverage can be denoted by Equation (2.2), where c represents the total coverage of all spectra in the dataset, and C denotes the number of spectra.

$$coverage = \frac{c}{C} \tag{2.2}$$

The precision of a spectrum's tags is measured by the proportion of correct tags within the top n selection. The precision across the test dataset is calculated as the mean precision for all spectra from which tags can be extracted. The formula for tag precision can be denoted by Equation (2.3), where p is the sum of precision for the spectra, and P is the count of spectra from which tags can be extracted.

$$precision = \frac{p}{P} \tag{2.3}$$

3. Experiment and Result Analysis.

3.1. Comparison Software Setup. To demonstrate the performance of TagEx in tag extraction, this article conducts a comparative analysis of TagEx with three other tag extraction tools—InsPecT, PepNovo+, and DirecTag—on the PXD004565 dataset. Since DirecTag requires input files in mzML format, it is necessary to first utilize the msConvert software to convert MGF files exported by the pParse+ software into the requisite mzML format. To ensure fairness in comparison, all tag extraction tools were configured to include only the fixed modification of Carbamidomethyl on amino acid C, and the amino acid mass error value was uniformly set to 0.02 Da. During the comparison process, the tag extraction lengths for the other three tools were set to 3 to enable optimal performance in tag extraction.

3.2. Sensitivity Comparison Experiment. On the PXD004565 dataset, the sensitivity performance of each tool is depicted in Figure 3.1. When selecting the top 1 tag from all software, TagEx achieves a sensitivity of 75.01%, which is respectively 16.38, 31.8, and 31.76 percentage points higher than DirecTag, InsPect, and PepNovo+. When selecting the top 100 tags, TagEx's sensitivity reaches 96.88%, which is respectively 2.23, 0.62, and 2.32 percentage points higher than DirecTag, InsPect, and PepNovo+. Throughout the entire process where sensitivity rises and eventually stabilizes, TagEx consistently maintains a leading advantage compared to the other tag extraction tools.

3.3. Coverage Comparison Experiment. On the PXD004565 dataset, the coverage comparison curves of various tag extraction tools are illustrated in Figure 3.2. Since TagEx employs a variable-length tag extraction method, unlike other tools that utilize fixed-length tags, it is not feasible to fairly compare coverage at the top 1 tag; hence, the coverage for the top 1 tag is not recorded. When selecting the top 20 tags from all software, TagEx's coverage reaches 57.21%, which is respectively 8.58, 20.85, and 17.92 percentage points higher than DirecTag, InsPect, and PepNovo+. Upon selecting the top 100 tags from all software, TagEx's

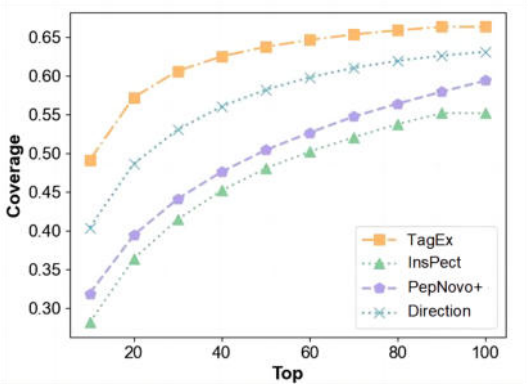

Fig. 3.2: Comparison of tag extraction coverage.

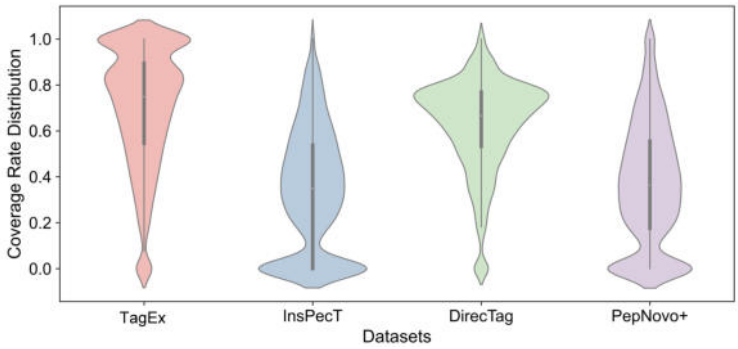

Fig. 3.3: Distribution of per-spectrum coverage in the Top100 tags by software.

coverage attains 66.33%, which is respectively 3.22, 11.14, and 6.94 percentage points higher than DirecTag, InsPect, and PepNovo+. Consequently, TagEx demonstrates superior performance in terms of coverage.

In the evaluation of tag coverage, the average value across all spectra is utilized, which may disadvantage tag extraction tools that exhibit low coverage in only a small subset of spectra. Therefore, after selecting the top 100 tags for each tag extraction tool, the distribution of coverage for each spectrum was analyzed, as shown in Figure 3.3. The coverage for each spectrum by TagEx is predominantly distributed around 100% and 80%, whereas the coverage for each spectrum by InsPect and PepNovo+ is mostly distributed within the 0-20% range, and the coverage for each spectrum by DirecTag is largely distributed within the 60-80% range. Hence, from the perspective of coverage per spectrum, TagEx also demonstrates an advantage.

3.4. Precision Comparison Experiment. Within the PXD004565 dataset, the precision of tags extracted by various tag extraction tools is depicted in Figure 3.4. When selecting the top 20 tags from each tag extraction tool, TagEx achieves a precision of 41.05%, which is respectively 10.91, 21.108, and 25.752 percentage points higher than DirecTag, InsPect, and PepNovo+. Upon selecting the top 100 tags from each tag extraction tool, the tag precision of TagEx is 14.56%, which is respectively 6.735, 3.287, and 8.313 percentage points higher than DirecTag, InsPect, and PepNovo+. Thus, TagEx demonstrates exceptional performance in terms of precision.

The precision of extracted tags decreases as the length of the tag increases. To further elucidate TagEx's performance in terms of precision, a comparison was made between the precision of sequence tags of lengths 3, 4, and 5 extracted by TagEx and those extracted by other tag extraction tools of corresponding lengths, with results depicted in Figure 3.5. Since DirecTag is only capable of extracting tags of lengths 3 and 4, its data

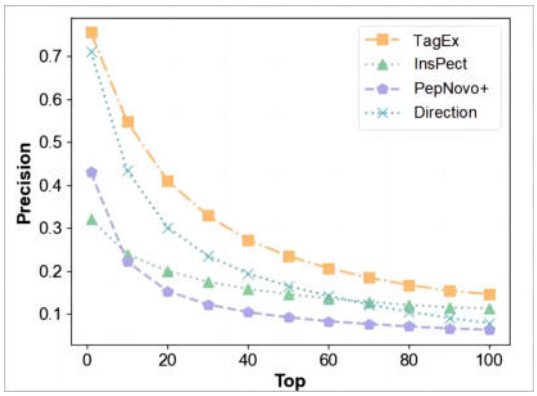

Fig. 3.4: Comparison of tag extraction precision.

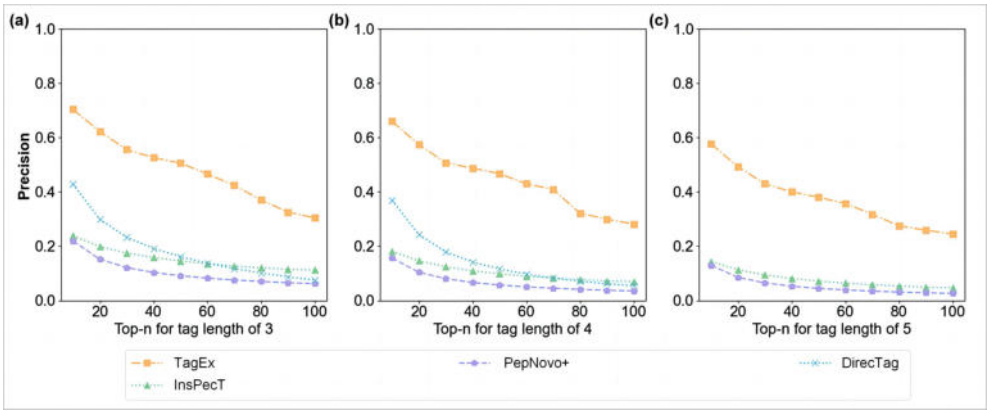

Fig. 3.5: (a) Precision comparison of tags of length 3. (b) Precision comparison of tags of length 4. (c) Precision comparison of tags of length 5.

is not included in the comparison for length 5 tags. Figure 3.5(a) displays the comparison results for length 3, where TagEx achieves a precision of 30.39% for tags of length 3 when selecting the top 100 tags, which is 19.1 to 24.14 percentage points higher than other software. Figure 3.5(b) shows the comparison results for length 4, with TagEx achieving a precision of 28.14% for tags of length 4, which is 17.06 to 21.99 percentage points higher than other software. Figure 3.5(c) presents the comparison results for length 5, where TagEx achieves a precision of 24.17% for tags of length 5, which is respectively 23.19 and 25.74 percentage points higher than InsPect and PepNovo+. Therefore, even when comparing tag precision segmented by tag length, TagEx still exhibits the best performance.

TagEx demonstrates outstanding precision performance, primarily due to our use of a de novo sequencing model based on graph convolutional networks as the scoring model for tags. This method more effectively ensures that the correct tags extracted by TagEx receive higher scores, thereby enhancing the overall precision of tag identification. The advantage of TagEx becomes more apparent when comparing precision based on tag length because the difficulty of determining tags increases with length, inevitably leading to a reduction in precision. Unlike other tag extraction tools that use fixed lengths of 3 to 5 amino acids, TagEx extracts tags of variable lengths, all of which are at least three amino acids long. Therefore, TagEx is at a comparative disadvantage when evaluated against tools that use a fixed tag length. However, when the tags extracted by TagEx are analyzed by length and compared for precision against other tools, their superiority becomes more evident.

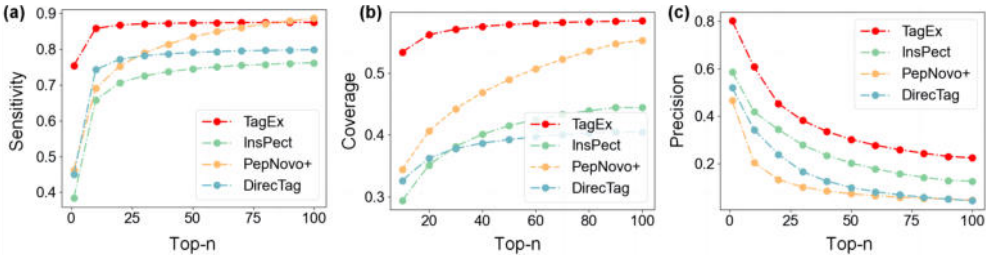

Fig. 3.6: (a) Comparison of sensitivity in the PXD009449 dataset. (b) Comparison of coverage in the PXD009449 dataset. (c) Comparison of precision in the PXD009449 dataset.

3.5. Software Performance Comparison in the PXD009449 Dataset. In the PXD009449 dataset, this study conducted a comparative analysis of the performance of various tag extraction software. The primary purpose of this analysis was to evaluate the capability of TagEx in extracting tags within complex modification environments and to ensure the algorithm's generalizability across different datasets, thus preventing its performance from being limited to specific datasets.

In terms of sensitivity, TagEx demonstrated significant superiority, with improvements of 12.19 to 38.86 percentage points and 8.57 to 32.25 percentage points over InsPect and DirecTag, respectively. Additionally, when selecting the Top80 tags, GCNTag showed an increase of 0.29 to 29.98 percentage points compared to PepNovo+. However, in the selection of Top90 to Top100 tags, PepNovo outperformed GCNTag, exhibiting higher sensitivity by 0.51 to 1.21 percentage points as the Figure 3.6(a). In terms of tag coverage, TagEx also performed better than InsPect and DirecTag, with coverage increases of 9.95 to 21.55 percentage points, 2.79 to 18.43 percentage points, and 13.91 to 19.01 percentage points as the Figure 3.6(b). Regarding precision, the improvements in TagEx were 22.23 to 35.15 percentage points, 27.14 to 45.11 percentage points, and 26.75 to 41.3 percentage points compared to InsPect and DirecTag as the Figure 3.6(c).

Therefore, based on the results of the comparative experiments, it is evident that GCNTag still exhibits a significant advantage in overall tag extraction performance on the PXD009449 dataset compared to other tag extraction tools.

4. Conclusions. This article introduces a variable-length tag extraction method to address the issues posed by fixed-length sequence tags and employs a graph neural network model to fit peptide spectrum matching patterns as a scoring tool to enhance the sensitivity and precision of the extracted tags. To evaluate TagEx's performance, it was benchmarked against three representative tag extraction software tools: InsPect, PepNovo+, and DirecTag. The experimental results demonstrate that TagEx exhibits superior sensitivity, coverage, and precision, with improvements of 0.62-2.32, 3.22-11.14, and 3.29-8.31 percentage points, respectively, when retaining the top 100 tags; the advantages are even more pronounced when only the top-ranked tag is retained, with sensitivity and precision increasing by 16.38-31.76 and 4.387-32.597 percentage points, respectively.

In summary, in comparison with three representative tag extraction software, the tags mentioned by TagEx exhibit advantages in sensitivity, coverage, and precision. Moreover, this variable-length peptide sequence tag extraction method provides a more flexible and accurate algorithmic foundation for subsequent rapid peptide sequence identification, discovery, and localization of unknown modifications.

REFERENCES

[1] J.Yuming,R.Devasahayam,S.Dina,N.Benjaminand,V.Norbert,M.Amanda,Peters-Clarke,T.M,E.Susan,K.Simion, *pFind–Alioth:A novel unrestricted database search algorithm to improve the interpretation of high-resolution MS/MS data*, Journal of Proteomics,125(2015),pp. 89-97.

[2] K.Sangtae and P.PavelA, *MS-GF+ makes progress towards a universal database search tool for proteomics*, Nature Communications,2023.

[3] C.Hao,H.Kun,Y.Bing,C.Zhen,S.Rui-Xiang,F.Sheng-Bo,Z.Kun,L.Chao,Y.Zuo-Fei,W.Quan-Hui, *Comprehensive Overview of Bottom-Up Proteomics using Mass Spectrometry*, ArXiv,125(2015),pp. 89-97.

[4] B.Navratan,B.Elena,C.Enrique,L.AnaVictoria,M.Spiros,T.Marco,E.Iakes,J.Manuel, M.Ricardo,L.Ana, *Comprehensive quantification of the modified proteome reveals oxidative heart damage in mitochondrial heteroplasmy*, Cell Reports,23(2018),pp. 3685-3697.
[5] W.Jinwei,Z.Junjie,Y.Qilin,L.Xiangyang,Z.Yuhui,S.Yun-Qing,J.Sunil, *SmsNet: A New Deep Convolutional Neural Network Model for Adversarial Example Detection*, IEEE Transactions on Multimedia,24(2021),pp. 230-244.
[6] Q.Rui,T.Ngoc-Hieu,X.Lei,C.Xin,L.Ming,S.Baozhen,G.Ali, *Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices*, Nature Machine Intelligence,3(2021),pp. 420-425.
[7] Y.Melih,F.William,B.Wout,O.Sewoong,N.William, *De novo mass spectrometry peptide sequencing with a transformer model*, International Conference on Machine Learning,(2022),pp. 25514-25522.
[8] W.Ruitao,Z.Xiang,W.Runtao,W.Haipeng, *Denovo-GCN: De Novo Peptide Sequencing by Graph Convolutional Neural Networks*, Applied Sciences,13(2023),p. 4604.
[9] V.Kira, *Validation of de novo peptide sequences with bottom-up tag convolution*, Proteomes,10(2021),p. 1.
[10] M.Matthias,W.Matthias, *Error-tolerant identification of peptides in sequence databases by peptide sequence tags*, Analytical Chemistry,24(2021),pp. 4390–4399.
[11] T.Stephen,S.Hongjun,F.Ari,W.LingChi,Z.Ebrahim,M. Marc,P.Pavel,B.Vineet, *InsPecT: identification of posttranslationally modified peptides from tandem mass spectra*, Analytical Chemistry,77(2005),pp. 4626–4639.
[12] T.David,M.Ze-Qiang,M.Daniel,H.Amy-Joan,C.Matthew, *DirecTag: accurate sequence tags from peptide MS/MS through statistical scoring*, Journal of Proteome Research,9(2008),pp. 3838–3846.
[13] F.Ari, *A ranking-based scoring function for peptide- spectrum matches*, Journal of Proteome Research,8(2009),pp. 2241–2252.
[14] N.Seungjin,B.Nuno,P.Eunok, *Fast multi-blind modification search through tandem mass spectrometry*, Molecular & Cellular Proteomics,11(2012),p. 012087.
[15] C.Hao,L.Chao,Y.Hao,Z.Wen-Feng,W.Long,Z.Wen-Jing,W.Rui-Min,N.Xiu-Nan,D.Yue-He,Z.Yao, *Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine*, Nature biotechnology,11(2018),pp. 1059–1061.
[16] N.Seungjin,K.Jihyung,P.Eunok, *MODplus: robust and unrestrictive identification of post-translational modifications using mass spectrometry*, Analytical Chemistry,91(2019),pp. 11324–11333.
[17] P.Ana-LS,O.Jose-TA,S.Gustavo,V.Ilka, *Label-free proteomic reveals that cowpea severe mosaic virus transiently suppresses the host leaf protein accumulation during the compatible interaction with cowpea*, Journal of Proteome Research,15(2016),pp. 4208-4220.
[18] N-Nathalie,T.Lucie,C.Cerina,A.Zuzanna,L.Joanna,G.Francois,B.Anne,E. Aleksander,A.Corinne,G.Ida-Chiara, *Impact of cystinosin glycosylation on protein stability by differential dynamic stable isotope labeling by amino acids in cell culture (SILAC)*, Molecular & Cellular Proteomics,16(2017),pp. 457-468.
[19] C.Liam,P.Daniela,L.Dennis,S.Ruth,T.Andreas, *Combination of bottom-up 2D-LC-MS and semi-top-down GelFree-LC-MS enhances coverage of proteome and low molecular weight short open reading frame encoded peptides of the archaeon Methanosarcina mazei*, Journal of Proteome Research,15(2017),pp. 3773–3783.
[20] P.Jillian,K.Anna,G.Harald,C.Ulisse,V.Matthijs,K.Manuel,B.Silvia,M.Marc, H.Craig,S.Brandon, *Chemosynthetic symbionts of marine invertebrate animals are capable of nitrogen fixation*, Nature Microbiology,2(2016),pp. 1-11.
[21] M.Clara,F.Bertrand,H.Maarten-LATM,P.Harriet,D.MichaelJ,L.KathrynS,N.BartM, *In-depth characterization of the tomato fruit pericarp proteome*, Proteomics,17(2017),p. 1600406.
[22] S.Gunnar,Me.David,S.Nesli-Ece,G.Anika,K.Sylvia,A.Georg, *Quantitative global proteomics of yeast PBP1 deletion mutants and their stress responses identifies glucose metabolism, mitochondrial, and stress granule changes*, Journal of Proteome Research,16(2017),pp. 504-515.
[23] H-Han,B.Kaspar,W.Jakob,Z.Fred,H.Yue,F.Mao,H.Bin,F.Yu,W.Abebe-Jenberie,L.Jianke, *Proteome analysis of the hemolymph, mushroom body, and antenna provides novel insight into honeybee resistance against varroa infestation*, Journal of Proteome Research,15(2016),pp. 2841-2854.
[24] C.Wojciech,L.Martina,P.Anne,N.Tuula,M.Sampsa, *Proteomic and bioinformatic characterization of extracellular vesicles released from human macrophages upon influenza A virus infection*, Journal of proteome research,16(2017),pp. 217-227.
[25] R.Daniel,A.Josef,M.Ulrike,R.Hermann,I.Till,a.Praveen-Kumar,T.Andrea,G.Cyprien,N.Pierre,S.Leif, *Large-scale reduction of the Bacillus subtilis genome: consequences for the transcriptional network, resource allocation, and metabolism*, Genome research,27(2017),pp. 298-299.
[26] C.Hao,L.Chao,Y.Hao,Z.Wen-Feng,W.Long,Z.Wen-Jing,W.Rui-Min,N.Xiu-Nan,D.Yue-He,Z.Yao, *Comprehensive identification of peptides in tandem mass spectra using an efficient open search engine*, Nature biotechnology,36(2018),pp. 1059-1061.
[27] K.ThomasN,W.Max, *Semi-supervised classification with graph convolutional networks*, ArXiv,(2016).
[28] F.Zhengcong,W.Kaifei,C.Hao, *GameTag: A New Sequence Tag Generation Algorithm Based on Cooperative Game Theory*, Proteomics,20(2022),pp. 21-22.
[29] F.Zhengcong, *Novel Peptide Sequencing With Deep Reinforcement Learning*, 2020 IEEE International Conference on Multimedia and Expo (ICME),79(2022),pp. 1-6.

Edited by: Jingsha He
Special issue on: Efficient Scalable Computing based on IoT and Cloud Computing
Received: Apr 10, 2024
Accepted: May 27, 2024