Scientific Research Publishing

Journal of Computer and Communications, 2022, 10, 14-23
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A New Tag Index Scheme Enables Fast Peptide Retrieval for Protein Identification

Piyu Zhou, Xinhang Hou, Haipeng Wang*

School of Computer Science and Technology, Shandong University of Technology, Zibo, China
Email: *hpwang@sdut.edu.cn

How to cite this paper: Zhou, P., Hou, X. and Wang, H. (2022) A New Tag Index Scheme Enables Fast Peptide Retrieval for Protein Identification. *Journal of Computer and Communications*, **10**, 14-23.
https://doi.org/10.4236/jcc.2022.104002

Received: March 11, 2022
Accepted: April 8, 2022
Published: April 11, 2022

Abstract

Sequence tag index in the field of computational proteomics can be used to facilitate faster open-search-based identification of modified peptides and in-depth analysis of mass spectrometry data. In protein-identification search engines, sequence tag index are playing a prominent role in recent ten years due to fast searching speed. However, in pursuit of less index space consumption, some protein search engines design excessively concise index schemes which lead to higher computational burden. We proposed a new tag index scheme named TIIP with a better balance between space and time complexity. TIIP has a unique two-level hierarchical index structure which allows rapid retrieval of all peptide sequences and their corresponding masses. Theoretically, the index space consumption of TIIP is not much higher compared to the typical tag index schemes, but the time complexity of sequence retrieval can be reduced to O(1), and practically, TIIP has about one million fold improvement in searching speed compared with brute force approach.

Keywords

Proteomics, Mass Spectrometry, Sequence Tag, Inverted Index, Tag Index, Open Search

1. Introduction

1.1. Sequence Tag

The concept of sequence tag was introduced by Mann *et al.* in 1994 [1]. It refers to the partial sequence of amino acids derived from a series of continuous fragment ions as shown in **Figure 1** [2] [3] [4] [5].

As an analysis strategy of mass spectrometry (MS) data in proteomics, it is an intermediate method between database search [6]-[19] and de novo sequencing

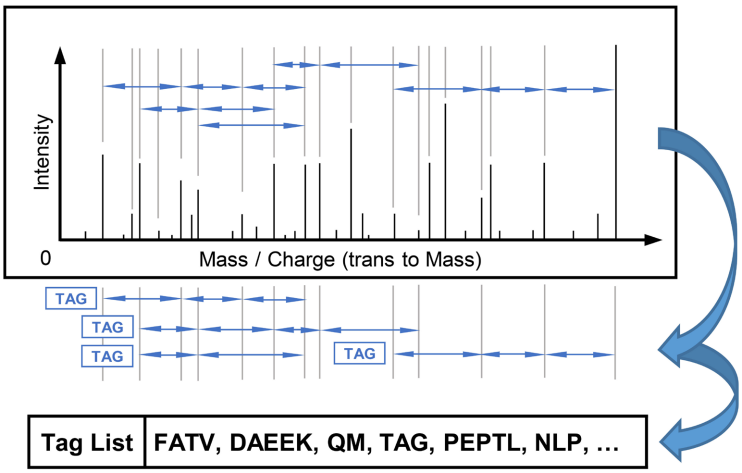

Figure 1. Sequence tags inferred from MS/MS spectra.

[2] [3] [4] [5] [10]. The sequence tag method can be used to prune exponentially larger search space in discovery proteomics to realize the effective filtering of candidate peptide sequences. Because of the characteristics above, search engines embrace the possibility of open search (expanding the search scope) without taking too much time [7] [9] [10] [13] [18]. In recent years, sequence tag method has been adopted by many modern protein search engines, such as Open-pFind, MODplus, and TagGraph, as an essential speeding up technique and become a fundamental peptide identification algorithms [7] [13] [18].

1.2. Tag Index Design

Sequence tag index, or tag index for short, is a hash table that can quickly search related peptide sequences [20] [21]. It is also an inverted index for search engines to accomplish rapid retrieval. The tag index, in essence, is a set of key-value pairs with a substring as the key, and the position of this substring in the original string as the value. In search engine technique, querying with inverted index is a commonly used approach. Compared to brute force with at least O(N) time complexity, inverted index method could transform a sequence tag to a hash code, which is related to an index entry and helps retrieve tag-containing sequences in O(1). As shown in **Figure 2**, exhaustively sequential substring matching can be avoided by using tag index, and search time of sequence tag can be reduced [7] [9] [10] [13] [18].

In order to facilitate the implementation of string matching algorithm, TagGraph constructs a database index structure by suffix array, in which the structured protein sequence information is called FM-indexed protein [18]. The other two search engines, Open-pFind and MODplus, adopt the scheme of k-length tag index, so they can find the corresponding peptide sequence information according to the k-length tags. In the index entries of Open-pFind, it only records the protein ID and starting position of tag in protein sequence to compress the memory space [7]. MODplus goes further in this direction. In MODplus, all protein sequences are concatenated into one linearized sequence delimited with

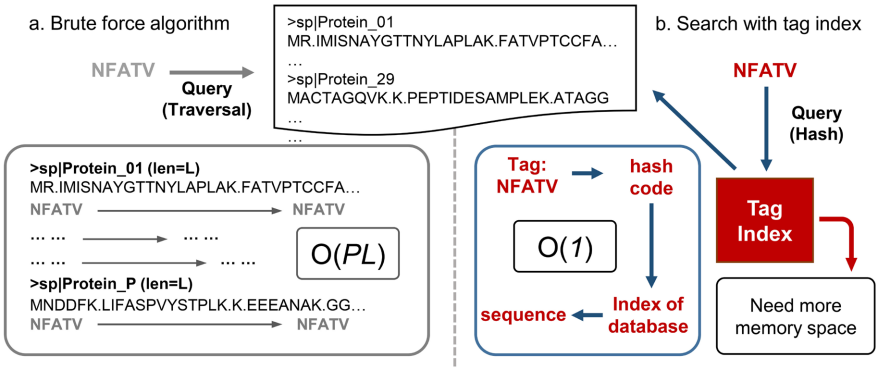

Figure 2. Brute force v.s. tag-index-based search.

the character "$". Therefore, tag index of MODplus only needs to store tag's start position at the linearized sequence string, so as to further compress the memory space occupied by inverted index [13]. The offset addresses of tags in Open-pFind and MODplus are encoded by dictionary order.

Both Open-pFind and MODplus compress the index information as much as possible in order to save memory space and complete the rapid search of tags. However, when retrieving sequences containing any of extracted tags, they all have unnecessary operations of amino acid traversal and mass calculation because of the design problem of tag index, which undoubtedly increases the cost of time. Nowadays, with the advances of computer hardware technology, memory usage of ordinary mass spectrometry data analysis is no longer a problem even for personal computer. Therefore, we should focus on how to retrieve the sequence more quickly.

Here, we introduce the new scheme named TIIP (Tag Index of Intact Proteins), a novel tag index design scheme. TIIP has a unique two-level index structure, which makes tag index and protein database cooperate effectively in search. Based on the design of TIIP, we can rapidly retrieve all peptide sequences that satisfy user-specified parameters, and get the theoretical masses information of peptide sequences quite fast.

2. Design of TIIP

2.1. Analysis of Sequence Retrieval Approach

Index design in search engine will affect the workflow design and efficiency of peptide sequence retrieval. The open search strategy in Open-pFind is to retrieve all the peptide sequences containing at least one extracted tag by extending this tag sequences from its both ends. This strategy can be compatible with semi-specific and non-specific searches.

The search strategy of MODplus will change with user-specified parameters. When searching for non-specific peptide sequences with MODplus, the search space of sequence is similar to the one considered by Open-pFind.

However, it should be noted that, as shown in **Figure 3**, it is very uneconomic to directly retrieve a large number of non-specific peptide sequences at the beginning

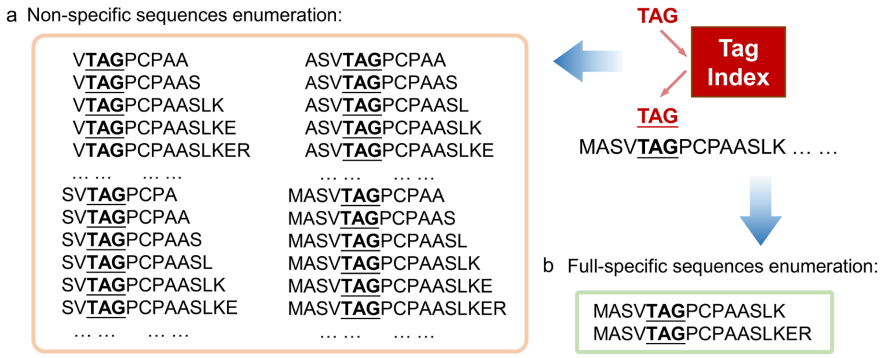

Figure 3. Different sequence retrieval approach.

of search, which will lead to a large amount of time wasted on enumerating incorrect peptide sequences and iterating over amino acids. Theoretically, assuming that average sequence length is L, the time complexity of retrieving all sequences is $O(L^2)$ when performing non-specific sequences, while the time complexity of retrieving only fully specific sequences is O(1). But practically the implementation of index structures in Open-pFind and MODplus didn't use cleavage information. Therefore, to find the cleavage sites in such index design schemes, even if only retrieving full-specifically digested sequences would take as much time as enumerating non-specific sequences.

The design and implementation of index structure and search flow will affect each other, so considering the needs of search can help index design. If retrieving all peptide sequences within the mass tolerance, lots of non-specific peptide sequences would reduce the efficiency of identification. From another point of view, the difference between fully specific and non-specific sequence from one protein sequence is several terminal amino acids, so we could regard terminal mass shifts as special-mass modifications, whose masses equal to the cumulative masses of terminal amino acids. Based on that, we can consider retrieving only fully specific sequences and dividing the search flow into two stages. The first one is using tag index to obtain candidate fully specific peptide sequence set and score to filter, while the second one includes precursor mass difference calculation, semi-specific/non-specific sequence detection and other operations for candidate peptide sequences. In view of previous analysis, we propose that TIIP is designed as a scheme for fast retrieving full-specifically digested peptide sequences, and moving the semi-specific and non-specific detection to the later stage, transferring the identification pressure of semi-specific and non-specific digested peptide sequences from the sequence retrieval stage to the subsequent stage.

2.2. Tag Index of Intact Proteins

In a word, the tag index and protein database structure we designed can rapid retrieve peptide sequences and fast calculate the difference between monoisotopic masses of precursor ions and theoretical masses of peptide sequences.

In order to compress the index space consumption, we use a strategy that does not need explicitly generate peptide sequences for database searching, just like the designs of Open-pFind and MODplus. The corresponding cleavage sites are recorded as a part of tag index structure. Consequently, we can retrieve fully specific peptide sequences immediately. Additionally, in order to calculate the theoretical masses of peptide sequences fast, the theoretical sequence masses are recorded in the tag index entries. Therefore, each index entry formed in this way needs three integers and one floating number: protein ID, left cleavage site, right cleavage site and theoretical mass of related peptide sequence. As shown in **Figure 4**, each tag should store the corresponding protein and peptide sequence information in related index entry in turn that matches the setting of missing cleavage sites number.

2.3. Refinement of TIIP

After finishing the initial design of tag index, further refinement is required. As shown in **Figure 5**, we consider refining the format of sequence records as follows, including cleavage sites, theoretical masses etc.

Firstly, the information of cleavage sites is stored in protein database, and each protein sequence corresponds to a list of all cleavage sites. Among them, in order to acquire peptide sequences, we need to store the offset addresses, which include zero and length of protein sequence, into the lists of cleavage sites. This refinement makes checking missing sites number become convenient, because any two closest sites of one protein sequence are adjacent in one list so their offset addresses are also adjacent.

Secondly, according to the list of cleavage sites belonging to each protein sequence, we calculate the cumulative mass of amino acid residues between two adjacent cleavage sites, and similarly store this masses information as a list so that we can calculate theoretical sequence masses so fast. Obviously, the length of cumulative mass list is equal to the length of related cleavage sites list minus one.

Finally, we only record protein ID, and offset addresses of the cleavage sites at the N-terminus and C-terminus of a peptide, and these two sites are the closest two to the tag in a protein sequence. Each index entry points to offset addresses of cleavage sites pair in list, so this design makes the tag index and protein database form a two-level index.

2.4. Sequence Retrieval of TIIP

As shown in **Figure 6**, when retrieving all possible peptide sequences, a sequence tag can be used to immediately acquire the shortest full-specifically digested sequence. Then, we extend cleavage sites of the shortest and enumerate all possible sequences with allowed number of missing cleavage sites. At the same time, we calculate the theoretical masses of the generated sequences and check if they are within the specified mass tolerance.

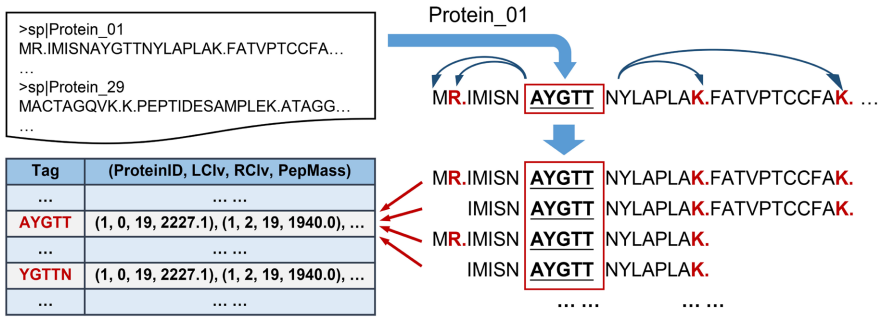

Figure 4. The initial design of TIIP scheme.

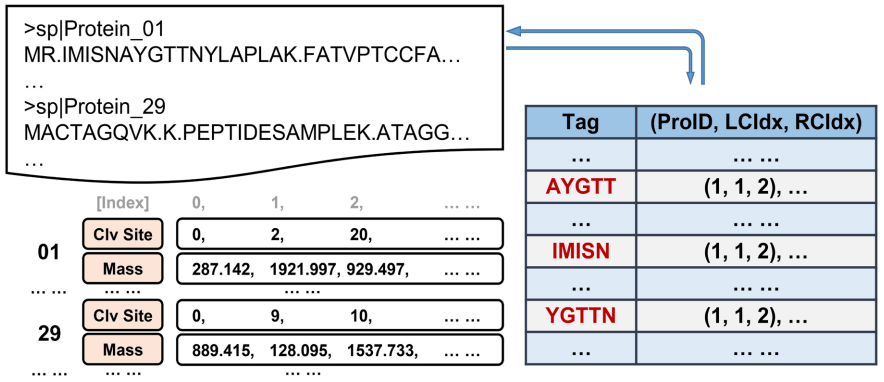

Figure 5. The refined design of TIIP.

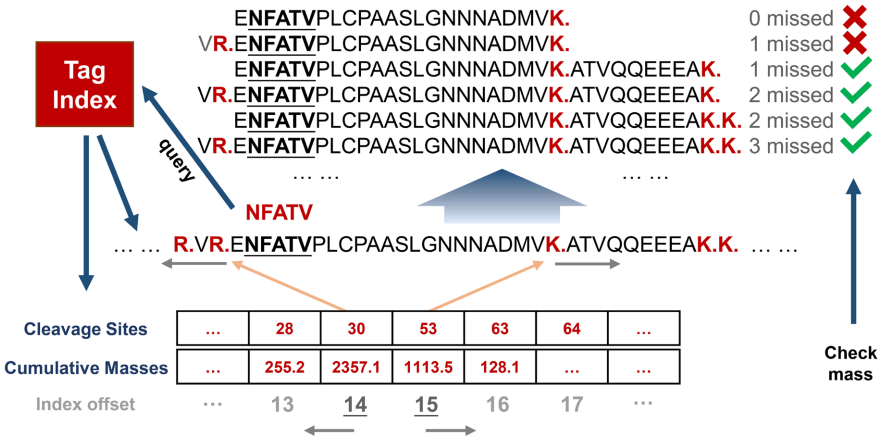

Figure 6. Overview of the sequence retrieval algorithm.

Theoretically, compared with the tag index schemes in Open-pFind and MODplus, TIIP scheme avoids unnecessary enumeration of amino acids when retrieving sequences. If the number of missing cleavage sites is fixed, the time complexity of retrieving sequence can be reduced to O(1), and the corresponding memory space consumption is acceptable.

3. Experiment and Result

3.1. Baseline, Dataset and Parameters

Because Open-pFind and MODplus do not open source code or independently

executable index components, it is inconvenient for us to conduct direct comparison test. Here, we only test and show the performances of TIIP scheme and brute force. The TIIP tested in this paper is implemented in Python.

The database in test with 20,350 human proteins was downloaded from UniProt on March 29, 2020. From the downloaded database, we randomly chosen 10,000 non-redundant peptide sequences and generated corresponding simulated mass spectra as the test dataset. Figure 7 shows an example of simulated spectrum.

Some parameter settings, software and hardware environment during the test are shown in Table 1 and Table 2.

3.2. Memory Space and Time Cost of Tag Index

In the environment and parameter settings mentioned above, we tested the space consumption of TIIP index design scheme and the time cost of index generation. The memory space consumption of 5-tag index (index with 5-length tag) is acceptable, while index generation has fast speed. The details are shown in Figure 8 and Figure 9.

Table 1. Key parameters of sequence retrieval.

Property	Value
Tag Length	5
Precursor Mass Window	[−500, 500]
Max Missed Cleavage Number	3

Table 2. Hardware and software environment.

Property	Value
CPU	Intel(R) Core(TM) i5-4210M CPU @ 2.60 GHz
Memory (RAM)	16.0 GB DDR3L 1600 MHz
Operating System	Windows 10 64 bit Professional

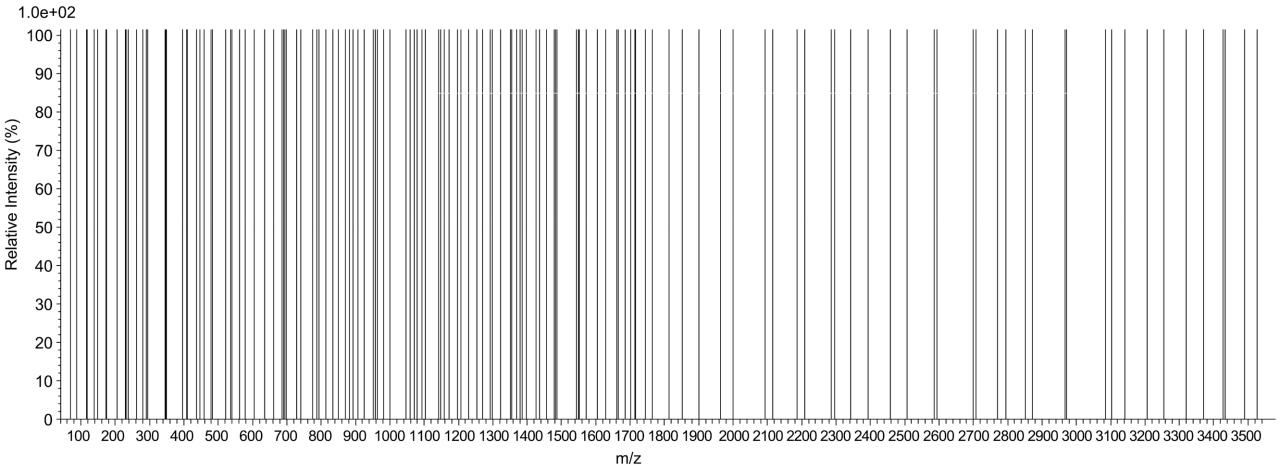

Figure 7. An example of simulated MS/MS spectrum.

3.3. Time Cost of Peptide Sequences Retrieval

This test is in full enumeration method which means that all full-specifically digested sequences are retrieved. We compared the time consumption of full traversal (brute force method) and TIIP scheme (searching with inverted index) using 5-length sequence tags. The result details are shown in Figure 10.

If using brute force method, the average time cost of only one spectrum is 2156.59 s. If we use TIIP scheme, the total time consumption for 10,000 spectra is only 60.73 s which includes the time (31.09 s) consumed in the process of database and index generation. The average time cost for a single spectrum is only 0.006 s. If the scale of spectra is larger, the database building time will be further diluted so that the average single spectrum time cost will be shorter.

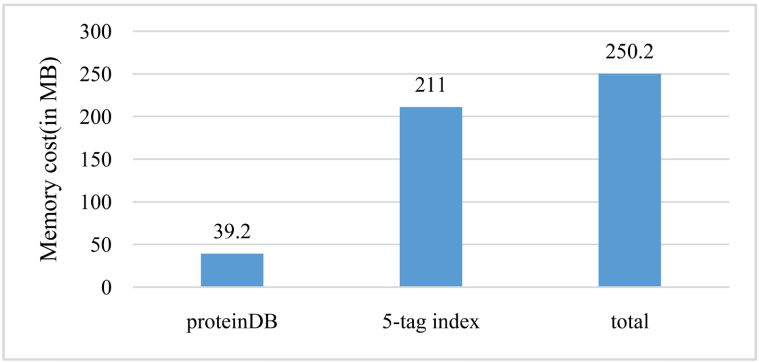

Figure 8. Memory cost of database construction and 5-tag index generation.

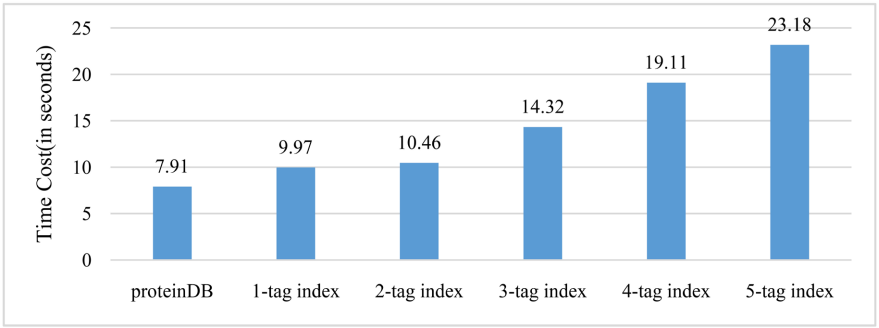

Figure 9. Time cost of database construction and *k*-tag index generation.

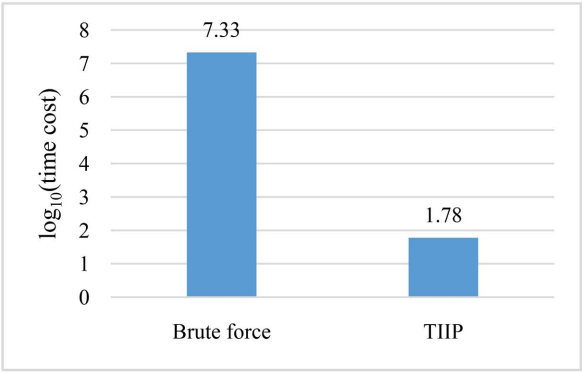

Figure 10. Time cost of peptide sequence retrieval.

It can be found that the search based on TIIP is actually very fast, which thanks to the enumeration of only full-specifically digested peptide sequences, and compared with brute force method, TIIP has about one million fold improvement in searching speed.

4. Conclusion

We designed TIIP scheme using two-level index structure with a tag hash table and a pre-digested protein database. When searching with TIIP, users can fast acquire candidate peptide sequences only restricted by the number of missed cleavage sites and open search mass window. Compared with the index design of Open-pFind and MODplus, TIIP scheme can avoid a large number of unnecessary enumeration of non-specifically digested peptide sequences, and thus save search time. TIIP scheme is more applicable to the cases of full-specifically digested peptides identification, while semi-specifically and non-specifically digested peptides are supported, too. For TIIP, after scoring and pruning, the candidate peptide sequences' number will be considerably reduced. Thus, the identification of semi-specific and non-specific peptide sequences only need to be determined by the flanking masses of hitting sequence tags. Therefore, the computational costs of finding semi-specific and non-specific peptides are reduced.

According to the current research progress, we will continue further development in tag index design for better performance of any protein search engines.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Mann, M. and Wilm, M. (1994) Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags. *Analytical Chemistry*, **66**, 4390-4399. https://doi.org/10.1021/ac00096a002

[2] Tabb, D.L., Saraf, A. and Yates, J.R. (2003) GutenTag: High-Throughput Sequence Tagging via an Empirically Derived Fragmentation Model. *Analytical Chemistry*, **75**, 6415-6421. https://doi.org/10.1021/ac0347462

[3] Aebersold, R. and Mann, M. (2003) Mass Spectrometry-Based Proteomics. *Nature*, **422**, 198-207. https://doi.org/10.1038/nature01511

[4] Sunyaev, S., Liska, A.J., Golod, A. Shevchenko, A. and Shevchenko, A. (2003) MultiTag: Multiple Error-Tolerant Sequence Tag Search for the Sequence-Similarity Identification of Proteins by Mass Spectrometry. *Analytical Chemistry*, **75**, 1307-1315. https://doi.org/10.1021/ac026199a

[5] Patterson, S.D. and Aebersold, R.H. (2003) Proteomics: The First Decade and Beyond. *Nature Genetics*, **33**, 311-323. https://doi.org/10.1038/ng1106

[6] Bern, M., Kil, Y.J. and Becker, C. (2012) Byonic: Advanced Peptide and Protein Identification Software. *Current Protocols in Bioinformatics*, **40**, 13.20.1-13.20.14. https://doi.org/10.1002/0471250953.bi1320s40

[7] Chi, H., *et al.* (2018) Comprehensive Identification of Peptides in Tandem Mass Spectra Using an Efficient Open Search Engine. *Nature Biotechnology*, **36**, 1059-1061.

[8] Fu, Y., *et al.* (2004) Exploiting the Kernel Trick to Correlate Fragment Ions for Peptide Identification via Tandem Mass Spectrometry. *Bioinformatics*, **20**, 1948-1954.

[9] Na, S., Bandeira, N. and Paek, E. (2012) Fast Multi-Blind Modification Search through Tandem Mass Spectrometry. *Molecular & Cellular Proteomics*, **11**, M111.010199. https://doi.org/10.1074/mcp.M111.010199

[10] Tanner, S., *et al.* (2005) InsPecT: Identification of Posttranslationally Modified Peptides from Tandem Mass Spectra. *Analytical Chemistry*, **77**, 4626-4639. https://doi.org/10.1021/ac050102d

[11] Wang, X., Li, Y., Wu, Z., Wang, H., Tan, H. and Peng, J. (2014) JUMP: A Tag-Based Database Search Tool for Peptide Identification with High Sensitivity and Accuracy. *Molecular & Cellular Proteomics*, **13**, 3663-3673. https://doi.org/10.1074/mcp.O114.039586

[12] Yates, J.R., Eng, J.K., McCormack, A.L. and Schieltz, D. (1995) Method to Correlate Tandem Mass Spectra of Modified Peptides to Amino Acid Sequences in the Protein Database. *Analytical Chemistry*, **67**, 1426-1436. https://doi.org/10.1021/ac00104a020

[13] Na, S., Kim, J. and Paek, E. (2019) MODplus: Robust and Unrestrictive Identification of Post-Translational Modifications Using Mass Spectrometry. *Analytical Chemistry*, **91**, 11324-11333. https://doi.org/10.1021/acs.analchem.9b02445

[14] Kong, A.T., Leprevost, F.V., Avtonomov, D.M., Mellacheruvu, D. and Nesvizhskii, A.I. (2017) MSFragger: Ultrafast and Comprehensive Peptide Identification in Mass Spectrometry-Based Proteomics. *Nature Methods*, **14**, 513-520. https://doi.org/10.1038/nmeth.4256

[15] Wang, L.H., *et al.* (2007) pFind 2.0: A Software Package for Peptide and Protein Identification via Tandem Mass Spectrometry. *Rapid Communications in Mass Spectrometry*, **21**, 2985-2991. https://doi.org/10.1002/rcm.3173

[16] Li, D., *et al.* (2005) pFind: A Novel Database-Searching Software System for Automated Peptide and Protein Identification via Tandem Mass Spectrometry. *Bioinformatics*, **21**, 3049-3050. https://doi.org/10.1093/bioinformatics/bti439

[17] Chi, H., *et al.* (2015) pFind-Alioth: A Novel Unrestricted Database Search Algorithm to Improve the Interpretation of High-Resolution MS/MS Data. *Journal of Proteomics*, **125**, 89-97. https://doi.org/10.1016/j.jprot.2015.05.009

[18] Devabhaktuni, A., *et al.* (2019) TagGraph Reveals Vast Protein Modification Landscapes from Large Tandem Mass Spectrometry Datasets. *Nature Biotechnology*, **37**, 469-479. https://doi.org/10.1038/s41587-019-0067-5

[19] Craig, R. and Beavis, R.C. (2004) TANDEM: Matching Proteins with Tandem Mass Spectra. *Bioinformatics*, **20**, 1466-1467. https://doi.org/10.1093/bioinformatics/bth092

[20] Zhou, C., *et al.* (2010) Speeding up Tandem Mass Spectrometry-Based Database Searching by Longest Common Prefix. *BMC Bioinformatics*, **11**, Article No. 577. https://doi.org/10.1186/1471-2105-11-577

[21] Lu, B. and Chen, T. (2003) A Suffix Tree Approach to the Interpretation of Tandem Mass Spectra: Applications to Peptides of Non-Specific Digestion and Post-Translational Modifications. *Bioinformatics*, **19**, ii113-ii121. https://doi.org/10.1093/bioinformatics/btg1068