Scientific Journal of Astana IT University
ISSN (P): 2707-9031
ISSN (E): 2707-904X

DOI: 10.37943/AITU.2020.91.98.002

Ye. Golenko
Doctoral Student
golenko.katerina@gmail.com, orcid.org/0000-0002-4643-4571
S. Seifullin Agrotechnical University, Kazakhstan

A. Ismailova
PhD, Senior Lecturer
a.ismailova@mail.ru, orcid.org/0000-0002-8958-1846
S. Seifullin Agrotechnical University, Kazakhstan

Ye. Rais
Master's Student
erbolrais@gmail.com, orcid.org/0000-0003-0097-8335
S. Seifullin Agrotechnical University, Kazakhstan

PROTEIN IDENTIFICATION USING SEQUENCE DATABASES

Abstract: The bottom-up proteomics approach (also known as the shotgun approach), based on the digestion of proteins in peptides and their sequencing using tandem mass spectrometry (MS/MS), has become widespread. The identification of peptides from the obtained MS/MS data is most often done using available sequence databases. This paper presents a detailed overview of the peptide identification workflow and a description of the main protein bioinformatics databases. Choosing the correct search parameters and the sequence database is essential to the success of this method, and we pay special attention to the practical aspects of searching for efficient analysis of MS/MS spectra. We also consider possible reasons why database search tools cannot find the correct sequence for some MS/MS spectra and highlight the misidentification issues that can significantly reduce the value of published data. To help assess the assignment of peptides to MS/MS spectra, we will look at the scoring algorithms that are used in the most popular database search tools. We also analyze statistical methods and computational tools for validating peptide compliance with MS/MS data. The final part describes the process of determining the identity of protein samples from a list of peptide identifications and discusses the limitations of bottom-up proteomics.

Keywords: Mass Spectrometry, MS/MS, Bioinformatics, Protein Identification, Proteomics, Databases, Protein Sequence

Introduction

The most widely used option for protein analysis is the bottom-up proteomics strategy [1-3]. It is a preliminary enzymatic cleavage of a protein (or large peptide) into smaller fragments, rapid purification of the obtained sample from low molecular weight impurities, and a mass spectrometric analysis of a mixture of proteolytic peptides. The introduction of the received data array into a computer protein search library (for example, Mascot) allows one to identify the initial protein and assess the reliability of its determination by the fraction of the total protein sequence characterized by the identified fragments (score). Proteolysis of several proteins can be carried out without prior separation. This significantly increases the complexity of the peptide set, but at the same time, significantly speeds up the process of protein identification with a slight decrease in the coverage of the sequence, i.e. reliability of identification. The sequence of procedures is shown in Fig. 1. The first step is using the desired protein complement. The proteins are then extracted from the mixture and cleaved by a protease to produce a mixture of peptides. The mixture of peptides is then loaded directly into a microcapillary column and the peptides are separated by hydrophobicity and charge. When peptides are eluted from the column, they are ionized and separated by m/z in the first step of tandem mass spectrometry. All tandem mass spectra obtained in the experiment, containing information on the m/z of the precursor ion, its retention time, as well as m/z product ions, are converted by the software. A Mass spectrometer in text form and already in the form of a large text file is loaded into the search engine. There are several search engine options for this search, the most popular of which are MS-Fit and Mascot (Peptide Mass Fingerprint). There are also sites for searching protein profiles of specific organisms.

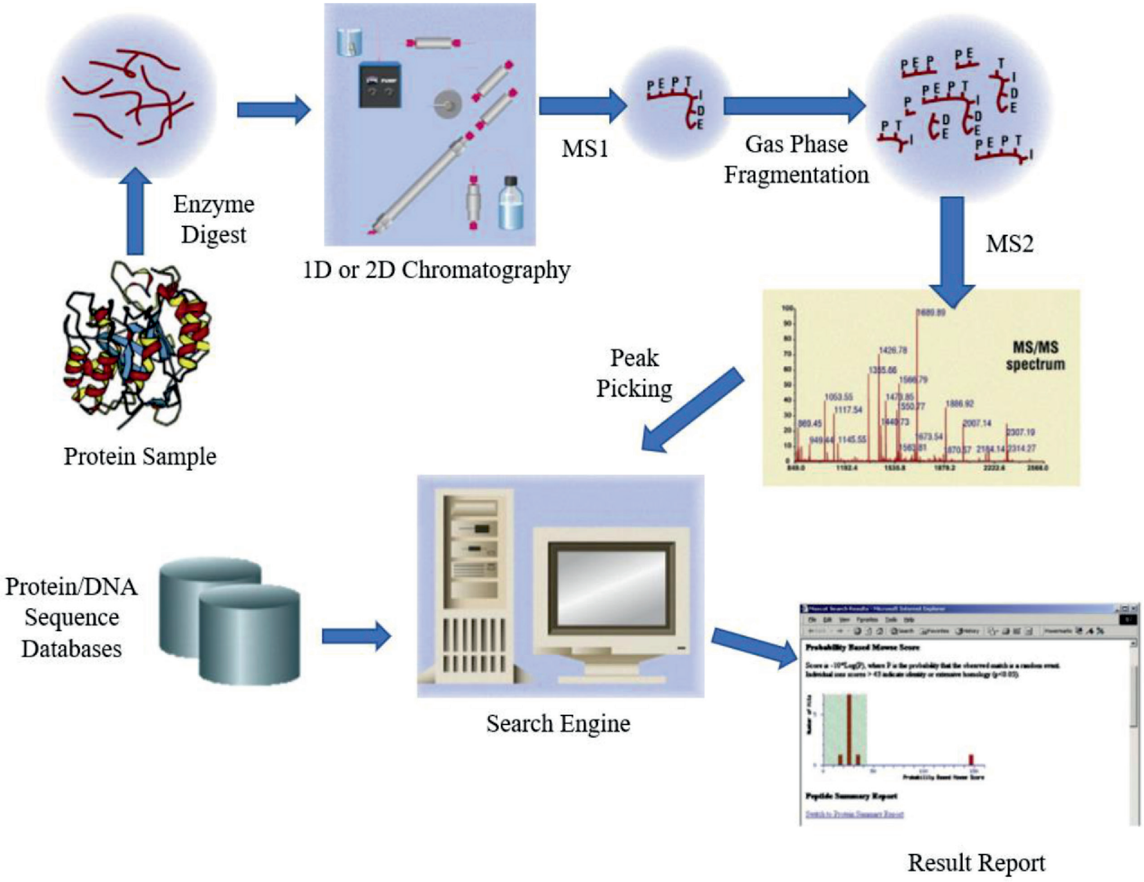

Fig. 1. A typical experimental workflow for protein identification and characterization using MS/MS data

After generating all possible structures of proteolytic peptides for each of them, the program calculates not only its mass, but also the theoretical tandem mass spectrum, taking into account the type of ions typical for the type of fragmentation initiation specified by the user. The search engine usually calculates c/z ion pairs, or b/y ion pairs, as well as the loss of $H_2O/CO/NH_3$ molecules from them. Theoretical MS/MS spectra are generated with a high probability of neutral emissions from the molecular ion. It is preferable to use high-resolution mass spectra, and in the case of low resolution, the useful information is provided by the MS/MS models and, in general, MS_n.

Since trypsin is used as a proteolytic enzyme in the vast majority of bottom-up studies, tryptic peptides dominate the databases. If we are talking about a sequence of non-tryptic peptides, even if they are already present in the database, the number of possible variants provided by the search engine increases significantly, which complicates the procedure for reliable identification. In this case, it is necessary to use the information of complementary methods of initiation of fragmentation or multistage tandem experiments.

When applying the above algorithm, it should be remembered that the reliability of the sequence loaded for the search is largely determined by the class of the mass spectrometer and is limited by the capabilities of the method itself. In particular, in the low-resolution spectra, it is impossible to unambiguously identify amino acid residues with the same integer mass Q and K, F and Mo), and the identification of isomeric Leu and Ile is possible only if the m/z values of satellite ions are present in the mass spectra [4]. If these issues are not resolved, it is advisable to load several sequences listing all possible variants of ambiguous amino acid residues.

Peptide Identification Methods

Three methods are used to identify proteins by mass spectra: database search, *de novo* sequencing, and hybrid approaches.

The use of databases to identify proteins and peptides makes it possible to decode mass spectra of complex mixtures in a short time. Almost all currently known amino acid sequences of proteins and peptides are combined into publicly available databases on the Internet. To identify a polypeptide using databases, an intermediary program is needed that allows generating theoretical mass spectra of library polypeptides taking into account the input parameters and comparing them with experimental mass spectra. To load experimental mass spectra into the search engine, as a rule, built-in programs are used, which are included in the software packages of all leading manufacturers of mass spectrometric equipment: Biotools (Bruker), Biolink (Waters), and BioWorks (ThermoFischer). In this case, not only the sequence of the identified peptide is offered, but also information about the protein, of which this peptide may be a part. Thus, the search program allows identifying both peptides and proteins.

Determination of the amino acid sequence of peptides and proteins without using search programs and databases is called *de novo* sequencing. This approach is used to identify previously undescribed proteins, in the presence of unexplored mutations, post-translational modifications, etc. The applied *de novo* sequencing algorithms are based on various mathematical methods [5-9].

The first algorithms for determining the amino acid sequence consisted of an enumeration of all possible combinations of amino acids that make up the mass of the parent ion, the fragmentation of which was compared with the experimental mass spectrum. Another approach is to consider a small part of the sequence (tag), to which amino acids are added on both sides until the corresponding mass of the parent ion is reached.

Each of the described methods has its own advantages and disadvantages. The identification of peptides and proteins using database searches is the simplest and most common method for interpreting MS/MS data. First, this strategy is applicable only in the case of known proteins, the sequences of which are entered into databases. Second, with post-translational modifications, the search time can be significantly increased, and the likelihood of false results increases. *De novo* sequencing is irreplaceable when working with unknown peptides and proteins, but very high requirements are imposed on the quality of the obtained fragment spectra. Thus, a necessary condition is the presence of a complete set of fragment ions of the main series. This method shows the best results when using high-resolution mass spectrometers [10].

Protein Identification by MS/MS Database Searching

Each of the databases has its own data storage format, a different degree of redundancy, interconnection with related or similar databases. All databases can be categorized into five types.

The first type is archived databases in which information is added by researchers. These databases include (GenBank, EMBL, PDB). The second type is curated databases (the content of records is supervised by specialists), such as Swiss-Prot. The third type is automatic databases (records are generated by computer programs), such as TrEMBL. The fourth type is derived databases, which are replenished by processing data from databases of the first two types (SCOP, PFAM, GO, etc.). The fifth type is integrated databases that combine information from various databases (ENTREZ). Table 1 contains data on some of the databases [11].

Table 1. Overview of some protein bioinformatics databases

Category	DB short name	DB name	URLs
Sequence databases	CCDS	The consensus CDS protein set database	https://www.ncbi.nlm.nih.gov/CCDS/CcdsBrowse.cgi
	DDBJ	DNA Data Bank of Japan	http://www.ddbj.nig.ac.jp/
	ENA	European nucleotide archive	http://www.ebi.ac.uk/ena
	GenBank	GenBank nucleotide sequence database	https://www.ncbi.nlm.nih.gov/genbank/
	RefSeq	NCBI reference sequence database	https://www.ncbi.nlm.nih.gov/refseq/
	UniGene	Database of computationally identifies transcripts from the same locus	https://www.ncbi.nlm.nih.gov/unigene
	UniProtKB	Universal Protein resource (UniProt)	http://www.uniprot.org/
2D gel databases	COMPLUYEAST-2DPAGE	2-DE database at Universidad Complutense de Madrid, Spain	http://compluyeast2dpage.dacya.ucm.es/
	REPRODUCTION-2DPAGE	2-DE database at Nanjing Medical University, China	http://reprod.njmu.edu.cn/cgi-bin/2d/2d.cgi
	SWISS-2DPAGE	2-DE database at Swiss Institute of Bioinformatics, Switzerland	http://world-2dpage.expasy.org/swiss-2dpage/
	World-2DPAGE	The World-2DPAGE database	http://world-2dpage.expasy.org/repository/
3D structure databases	DisProt	Database of protein disorder	http://www.disprot.org/
	MobiDB	Database of intrinsically disordered and mobile proteins	http://mobidb.bio.unipd.it/
	ModBase	Database of comparative protein structure models	http://modbase.compbio.ucsf.edu/modbasecgi/index.cgi
	PDBe	Protein Data Bank at Europe	http://www.ebi.ac.uk/pdbe/
	PDBj	Protein Data Bank at Japan	http://pdbj.org/
	PDBsum	Pictorial database of 3D structures in the Protein Data Bank	http://www.ebi.ac.uk/pdbsum/
	ProteinModelPortal	Protein model portal of the PSI-Nature structural biology knowledgebase	http://www.proteinmodelportal.org/
	RCSB-PDB	Protein Data Bank at RCSB	http://www.pdb.org/
	SMR	Database of annotated 3D protein structure models	http://swissmodel.expasy.org/repository/
Proteomic databases	MaxQB	The MaxQuant DataBase	http://maxqb.biochem.mpg.de/mxdb/
	PaxDb	Protein Abundance Across Organisms	http://pax-db.org
	PeptideAtlas	PeptideAtlas	http://www.peptideatlas.org
	PRIDE	PRoteomics IDEntifications database	http://www.ebi.ac.uk/pride
	ProMEX	Protein Mass spectra EXtraction	http://promex.pph.univie.ac.at/promex/

Fig. 2 illustrates the MS/MS spectrum as input and compares it to theoretical fragmentation patterns plotted for peptides from a search database to find a match [12].

Typically, the user specifies the search restriction criteria in order to obtain a limited subset of peptides to be searched for these criteria including mass tolerance, types of post-translational modifications allowed, restriction of proteolytic enzymes, etc. First, the masses of peptides from the database are compared with the experimental data of the masses of peptides, taking into account the specified error. The score is then calculated for each match. The sum of the score of peptides gives the score for the protein. Also, for each of the candidates, species identification is indicated, which can be decisive in interpretation, and links to personal pages (final result) containing comprehensive information about a potential protein (values of its molecular weight and isoelectric point, decoding of the sequence of tryptic peptides, number of matches, percent of coverage of the complete amino acid sequence of the protein by the identified peptides, etc).

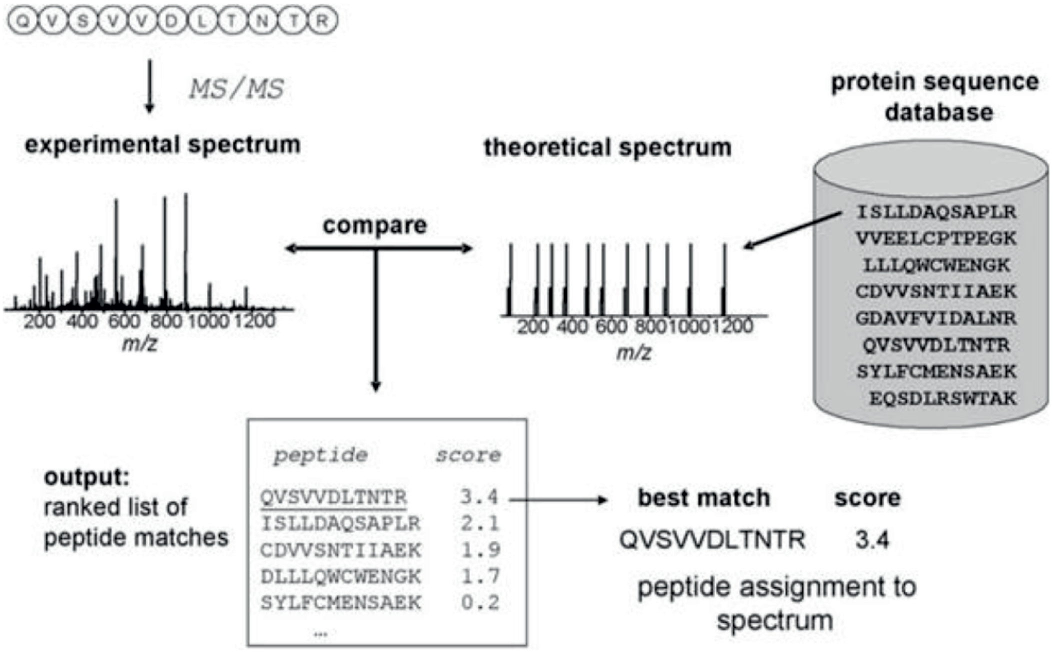

Fig. 2. *MS/MS* Database Searching

Table 2 shows the various search parameters that are used when searching the databases.

Table 2. Database Search Parameters

Parameter	Description
Types of fragment ions	General rules for peptide dissociation are used to calculate fragmentation patterns for each peptide [13].
Monoisotopic vs average mass	Monoisotopic mass and average mass are essential parameters in mass spectrometry. These values refer to the atoms of specific chemical elements. The key difference between monoisotopic mass and average mass is that monoisotopic mass is calculated based on one isotope, while average mass is calculated based on all common isotopes of a particular chemical element.
Peptide ion charge state	In order to determine the mass of a peptide from mass spectrometry data, it is necessary to know the charge state of the peptide ion. The pattern of isotope distribution in the MS spectrum helps to determine it with a fairly high probability [14].
Parent ion mass tolerance	For comparison with the experimental spectrum, only those peptides are added that have a calculated mass within a certain range. The choice of mass tolerance parameter depends on the type of mass spectrometer.
Enzymatic digestion constraint	Many proteolytic enzymes used to degrade proteins into peptides are specifically cleaved after certain residues in the protein sequence.
Chemical or posttranslational modifications	A database search can be performed with one or more static or variable modifications.

Reasons of Failure to Assign Correct Peptide Sequences

All MS/MS database search tools work in a similar way: they search the database for the most suitable peptide for each input spectrum unless there are no candidate peptides in the search database that match the search parameters specified by the user. However, the maximum match between the presented sample and the sequence from the database will not always be correct. The reasons the database search tools fail to assign the correct peptide sequences to so many experimental MS / MS spectra are listed below.

1. *Deficiencies of the scoring scheme*. The imperfection of the selected scoring system leads to a situation when, if there is a correct peptide sequence in the database, the search results give the wrong peptide with a higher score. Most of the assessment systems are based on a very simplified scheme for representing the process of peptide ion fragmentation [15] [16].
2. *Low MS/MS spectrum quality*. The efficiency of using programs for the interpretation of mass spectra and library search is determined, first of all, by the quality of mass spectral information. The first problem of this nature may be noise pollution. In addition, some MS / MS spectra were obtained not on peptides, but on various contaminants added to the sample during sample preparation.
3. *Fragmentation of multiple peptide ions*. Some MS / MS spectra, which can represent a significant percentage when analyzing mixtures of complex peptides, are the result of the simultaneous fragmentation of two more different peptide ions with similar m / z values. Since most database search tools operate on the assumption that the spectrum is from a single precursor ion, they often cannot assign any peptide sequence to the spectrum.
4. *Presence of homologous peptides*. Another important issue is the homologous but different peptides in the database used. This problem becomes very serious in the case of higher eukaryotes [17].
5. *Incorrectly determined charge state or peptide mass*. Usually, the spectrum assumes a double scan (first +2, and then +3), if the charge state of a multiply charged ion cannot be determined. If the charge state is +4 or higher, or the singly charged spectrum was incorrectly classified as multiply accused by the program, the correct peptide will not be found [18].
6. Restricted database search. Since a search taking into account, all post-translational modifications can accept an extended, users often refuse to consider modifications or consider only the most frequent changes.
7. *Sequence variants and novel peptides*. It should also be noted that peptides not previously entered into the database cannot be identified. To solve this problem, a global search across all databases can be performed, or other methods of peptide identification can be used, for example, *de novo* sequencing.

Scoring and Results Evaluation

As mentioned above, search tools do not always select the correct peptides for many MS/MS spectra, which means that the user must evaluate the performance of the search tool and remove incorrect results.

A database search score is a score calculated according to some scoring function that measures the degree of similarity between the experimental spectrum and theoretical fragmentation patterns of candidate peptides.

Different database search tools can use one or more different scoring schemes. There are a large number of schemes that are detailed in research. We propose to briefly review the schematics of the most commonly used and well-known search tools that are generally available and are currently used in many proteomic laboratories.

The Mascot search engine is based on the MOWSE (Molecular Weight Search) algorithm, proposed in 1993. This algorithm uses a mass peptide fingerprint search. First, the masses of peptides from the database are compared with the experimental data of the masses of peptides, taking into account the specified error. Then, for each match, the Score value (confidence level value) is calculated in accordance with (1):

$$Score = \frac{50000}{M_{prot} \times \Pi_n m_{i,j}} \tag{1}$$

where M_{prot} is the molecular weight of each matched protein, Π is the product, which is calculated from the $Mowse$-matrix of weights M for each coincidence of experimental data and peptide masses calculated from records in the genomic database [19].

This algorithm can be used for MS/MS search. In this case, a peptide plays the role of a protein in the $Score$ formula, and a fragment plays the role of a peptide. The sum of the $Score$ of peptides gives the $Score$ for the protein.

The Sequest search engine is based on a separate identification of each mass spectrum [20]. First, peptides corresponding to the mass of the parent ion of the studied peptide are selected from the protein database. For each candidate, a theoretical mass spectrum of fragmentation is generated and checked against experimental data [21]. Then a cross-correlation analysis of the spectra is carried out, which is reduced to calculating the integer function $R(\tau)$ according to (2):

$$R(\tau) = \sum_{i=0}^{n-1} x[i]y[i+\tau] \tag{2}$$

where n is the number of channels in the mass spectrum; $x[i]$ and $y[i]$ – the intensity of the mass spectrum signals on the i-th channel; τ is the shift of the calculated spectrum relative to the experimental one. This function is maximum at $\tau=0$ [22].

The X!Tandem search engine is the most advanced because it is open-source software [23]. In this algorithm, the calculated and experimental mass spectra are reduced to the form of a multidimensional vector of $n=m_{prt}/\Delta m$, where m_{prt} is the mass of the parent ion, and Δm is the maximum error in determining the mass of the daughter ion. The calculated mass spectrum includes the masses of the series ions and the masses of their ions with neutral losses ($NH3$ and $H2O$). To assess the coincidence of the calculated and experimental spectra, a rating is used, calculated by (3):

$$x = n_b!n_y!\sum_{i=0}^{n} I_i P_i \tag{3}$$

where n_b and n_y are the number of b- and y-series ions detected in the experimental mass spectrum, respectively; $\sum_{i=0}^{n} I_i P_i$ – a scalar product of vectors of experimental and calculated mass spectra.

To assess the reliability of protein identification, the protein rating E_{pro} is calculated by (4), based on the reliability e of each spectrum of the peptide of this protein:

$$E_{pro} = \sqrt{M} \prod_{i=1}^{n} e_i\left(x_i^*\right) \tag{4}$$

where N is the total number of spectra; n is the number of spectra associated with the protein [24]. X!Tandem also performs the identification of peptides with incomplete or nonspecific hydrolysis or in the presence of modifications in them in a relatively short time.

Peptide and protein identifications were compared using multiple search engines - Mascot, SEQUEST. For each search, common search parameters were used, including a sequence library.

Mascot provides the user with two scores: an ion score and an E score. The ion score is compared with two thresholds that are calculated independently for each peptide, a homology score and an identity score. Traditionally, the identity score is a reported threshold and is used in most laboratories. The homology score is usually lower than the identity score for longer peptides and higher for shorter peptides. A second criterion is shown using identity and homology scores as it is an effective way to remove short peptides that have a higher tendency to false positives than longer peptides. This results in higher confidence in the dataset with an increase in the estimated amount of detected proteins.

Sequest also offers two separate scores. The first score is actually a mixture of several scores: The cross-correlation score (Xcorr) is a measure of the cross-correlation of theoretical and experimental spectra. Typically, a different Xcorr threshold is used for each state of charge to reflect the difference in the chance of coincidence for each state. A score measuring the difference between the two best candidate spectra is also given in the form of a delta correlation value, where scores are normalized so that the highest score is set to 1 and the difference is taken. Less commonly used are a preliminary estimate (Sp) and a preliminary estimate (RSp), which are estimates calculated first as a filter to limit the number of spectra to be estimated for cross-correlation. The second Sequest score is the likelihood score offered in the most recent version of Bioworks software (Sequest p).

Differences between Mascot and SEQUEST.

For both LTQ and QSTAR runs, peptide populations confidently selected by only one algorithm were indistinguishable from peptides selected by another based on several measurements, including state of charge, residue composition, and peptide length (data not shown). Spectrum quality MS / MS is a possible explanation for the difference in performance between Mascot and SEQUEST. We have noticed that the apparent complexity and signal-to-noise ratios of MS/MS spectra have a profound and distinct influence on the magnitude of the ratings assigned by Mascot and SEQUEST.

Protein Inherence

It is important to note that the database search for MS/MS spectra identifies peptides rather than proteins. It is very difficult to determine, using peptide sequences, which proteins were present in the original sample. This is because many peptide sequences can be attributed to several proteins. The basic principle of protein deduction is to make a list with the minimum amount of proteins that can make up the required peptides.

We highlight several difficulties that complicate the process of assembling peptides into proteins.

Non-random grouping of peptides. This problem can be best explained using the illustration in Fig. 3. Peptides that are correctly identified tend to group into relatively few proteins. In contrast, misassignment of peptides can be described as coincidental overlaps with entries in a very large database of protein sequences, and almost every misassignment of a high-score peptide adds one new protein misidentification. As a result, even a small error rate of false identification at the peptide level can lead to a high error rate at the protein level [25].

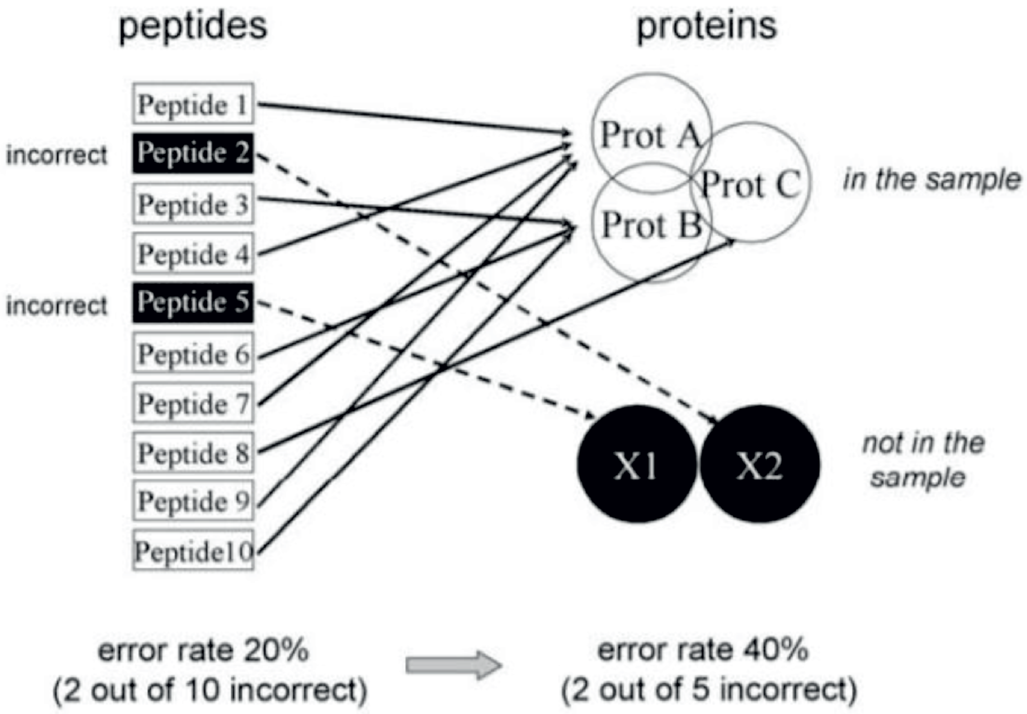

Fig. 3 Non-random grouping of peptides

Shared peptides. The identification of standard peptides, that is, peptides whose sequence is present in more than one entry in a protein sequence database, makes it difficult to determine the specific corresponding protein (or proteins) present in a sample. Such cases most often arise due to the presence of homologous proteins, splicing variants, or duplicate entries in the protein sequence database [25] [26]. This problem is severe in the case of higher eukaryotic organisms. As a result, it is often impossible to distinguish between different protein isoforms in bottom-up proteomics.

Conclusion

Thus, identification of peptides and proteins using database searches is the simplest and most common method for interpreting MS/MS data, but it is not without its drawbacks. First, this strategy is applicable only in the case of known proteins, the sequences of which are entered into databases. Secondly, with post-translational modifications, the search time can be significantly increased, and the likelihood of false results increases. In addition, a typical problem in the analysis of proteolytic mixtures of peptides is the high degree of homology among peptides. As a result, a plurality of sequences appears in the list of candidate sequences produced by the search program, which are assigned relative index values, and the program leaves only one sequence in the final list of identifications. Interpretation results are negatively affected by the presence of errors in the amino acid sequences of proteins in the databases. In order to eliminate this deficiency, work is constantly being carried out to correct and update the databases.

References

1. Link, A.J., Eng, J., Schieltz, D.M., Carmack, E., Mize, G.J., Morris, D.R., ... & and Yates, J.R. (3). rd. (1999). Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology, 17*, 676-82. doi: 10.1038/10890.
2. Gygi, S.P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., & Aebersold, R. (1999). Quantitative analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology, 17* (10), 994-999.

3. Washburn, M.P., Wolters, D., & Yates, J.R. (2001). Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology,* 19, 242-247.
4. Chung, T.W., & Tureček, F. (2010). Backbone and side-chain specific dissociations of *z* ions from non-tryptic peptides. *Journal of the American Society for Mass Spectrometry*, 21, 1279-1295
5. Dancik, V., Addona, T.A., Clauser, K.R., Vath, J.E., & Pevzner, P.A. (1999). De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology*, 6 (3-4), 327-342.
6. Taylor, J.A., & Johnson, R.S. (2001). Implementation and uses of automated de novo peptide sequencing by tandem mass spectrometry. *Analytical Chemistry*, 73 (11), 2594-2604.
7. Chen, T., Kao, M.Y., Tepel, M., Rush, J., & Church, G.M. (2001). A dynamic programming approach to de novo sequencing via tandem mass spectrometry, *Journal of Computational Biology*, 8 (3), 325-337.
8. Ma, B., Zhang, K., Hendrie, C., et al. (2003). PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry, 17,* 2337-2342.
9. Frank, A., & Pevzner, P. (2005). PepNovo: de novo peptide sequencing via probabilistic network modeling. *Analytical chemistry,* 77 (4), 964-973.
10. Berizovskaya, E.I., Ichalaynen A.A., Antochin A.M. (2015). Methods of processing mass spectrometry data for identification of the peptides and proteins. *Vestnik Moskovskogo universiteta*, 56, 266-278.
11. Chen, C., Huang, H., & Wu, C.H. (2017). Protein bioinformatics databases and resources. *Protein Bioinformatics,* 3-39.
12. Nesvizhskii, A.I. (2007). Protein Identification by Tandem Mass Spectrometry and Sequence Database Searching. *Mass Spectrometry Data Analysis in Proteomics*, 367, 87-119. doi: 10.1385/1-59745-275-0:87.
13. Aebersold, R. & Goodlett, D.R. (2001). Mass spectrometry in proteomics. *Chemical Reviews, 101,* 269-295.
14. Keller, A., Nesvizhskii, A.I., Kolker, E., & Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical Chemistry, 74,* 5383-5392.
15. Tabb, D.L., Smith, L.L., Breci, L.A., Wysocki, V.H., Lin, D., & Yates, J.R. (2003). Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. *Analytical Chemistry,* 75, 1155-1163.
16. Kapp, E.A., Schütz, F., Reid, G.E., Eddes, J.S., Moritz, R.L., O'Hair, R.A., ... & Simpson, R.J. (2003). Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation, *Analytical Chemistry,* 75, 6251-6254.
17. Resing, K.A., Meyer-Arendt, K., Mendoza, A.M., et al. (2004). Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics, *Analytical Chemistry,* 76, 3556-3568.
18. Keller, A., Nesvizhskii, A.I., Kolker, E., & Aebersold, R. (2002). 'Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Analytical chemistry,* 74 (20), 5383-5392.
19. Avtonomov D., Agron I., Kononikhin A., Nikolaev E. (2009). *Sozdaniye bazy dannykh tochnykh masso-vo-vremennykh metok dlya kachestvennogo i kolichestvennogo podkhoda v issledovanii proteoma mochi cheloveka s ispol'zovaniyem izotopnogo mecheniya*. Proceedings of Moscow Institute of Physics and Technology, 1 (1), 24-29.
20. Eng, J., McCormack, A., Yates, J. (1994*).* An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry, 5,* 976-989.
21. Yates, J.R., Eng, J.K., & McCormack, A.L. (1995). Mining Genomes: Correlating Tandem Mass Spectra of Modified and Unmodified Peptides to Sequences in Nucleotide Databases. *Analytical Chemistry*, 67 (18), 3202-3210. https://doi.org/10.1021/ac00114a016.
22. Lyutvinsky J. (2007). Method of recognition of amino acid sequences in mass spectra of peptides for proteomics problems. *Dissertation*.
23. Craig, R., Beavis, R.C. (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, 20 (9), 1466-1467.
24. Fenyö, D., Beavis, R.C., (2003). A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. *Analytical Chemistry,* 75, 768-774.
25. Nesvizhskii, A.I., Keller, A., Kolker, E., & Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Analytical Chemistry,* 75, 4646-4658.
26. Rappsilber, J., & Mann, M. (2002). What does it mean to identify a protein in proteomics? *Trends in Biochemical Sciences,* 27 (2), 74-78.