

UDC 543.51:577.112

A PRACTICAL WORKFLOW FOR MASCOT-BASED PROTEIN IDENTIFICATION AND POST-TRANSLATIONAL MODIFICATION ANALYSIS

S.Makhsatova^{ID}, A.Kurmanbay^{ID}, A.Kulyassov*^{ID}

National Center for Biotechnology, 13/5 Korgalzhyn Highway, Astana, 010000, Kazakhstan

*kulyassov@biocenter.kz

ABSTRACT

Modern biological research has become highly interdisciplinary, requiring mass spectrometry data generation to be combined with bioinformatics and data analysis skills to obtain meaningful results. However, many researchers who regularly conduct mass spectrometry experiments often lack practical experience in interpreting results and optimizing database searches. This practical article describes the use of the Mascot search engine for identifying peptides and proteins in shotgun proteomics. Using a local Mascot server (version 2.6.2), two datasets were analyzed: an in-house *Escherichia coli* tryptic lysate obtained on an Impact II QTOF mass spectrometer and a publicly available histone post-translational modifications dataset from the jPOST repository. The workflow includes optimizing search parameters, performing the search, interpreting results, and demonstrating the detection of post-translational modifications. The presented example of histones demonstrates the identification of known lysine modifications (including trimethylation) detected through search. This article serves as a practical guide for researchers wishing to improve their proteomic data analysis skills in Mascot and highlights the critical importance of selecting the right parameters to obtain accurate and reproducible results.

Keywords: Mascot, shotgun proteomics, database search, data analysis, post-translational modifications, peptide identification

INTRODUCTION

Omics-based methods are extensively employed in advanced biology to comprehend the structure, function, and dynamics of an organism. The genomics and transcriptomics offer useful insights into possible biological functions like gene expression, however they fail to encompass the complete overview of dynamic variations in protein levels, modifications, and activity. Studying intracellular processes at the protein level through proteomics techniques has become a crucial tool in modern molecular biology. Therefore, proteomic analysis is important for understanding biological regulation and disease mechanisms, because proteins directly represent the functional output of cellular systems.

Proteomics experiments are typically driven by two fundamental questions: which proteins are present in a given sample, and what is the relative abundance of each? Answers to these questions often serve as a springboard for deeper investigation into post-translational modifications, protein-protein interactions, and three-dimensional structures, all of which rely on the same core technological framework. The most widely adopted strategy for addressing these initial questions is shotgun (bottom-up) proteomics, in which proteins are enzymatically digested into peptides, analyzed by mass spectrometry, and identified by matching the generated spectra against protein databases using search engines [1, 2].

Shotgun proteomics is a high-throughput method that is used for large-scale protein analysis. This method emerged in the post-genomic era after the Human Genome Project revealed that DNA sequence alone is not enough to provide sufficient information about disease mechanisms and gene regulations. In the 1980s, Donald F. Hunt and colleagues pioneered tandem mass spectrometry (MS/MS) and in 1994 Eng, McCormack and Yates introduced the SEQUEST algorithm, which used genome-derived protein sequences to automatically interpret existing tandem mass spectra. This approach

laid the groundwork for proteogenomics, where proteomic data completes genome annotations [3, 4].

LC-MS/MS technology has advanced significantly over the past decade, with improvements in sensitivity, selectivity, and protein coverage, yet the raw data it produces does not directly identify proteins [2]. Instead, it generates complex spectra that represent fragmented peptide ions. Therefore, without computational interpretation, these spectra have no biological meaning [5].

Database search engines such as Mascot (Matrix Science) are required to translate raw MS/MS data into confident protein and peptide identifications [6]. The choice of database search parameters significantly impacts peptide and protein identification outcomes in any study [7]. Inaccurate parameter selection, such as using default values, can result in incomplete sequence coverage, higher false positive rates, poor detection of post-translational modifications, and missed peptide identifications, all of which can lead to biologically inaccurate findings [8].

In spite of the widespread use of the Mascot search engine, the detailed justification and descriptions of database search settings are often missing from publications or contain only a superficial description of their application [9]. The lack of methodological transparency limits reproducibility and comparability of proteomics datasets. It is essential to select parameters that are appropriate for the specific sample to ensure accurate classification of peptides [7]. This paper presents a practical tutorial and step-by-step protocol for protein identification using the Mascot search engine. This work is intended to help the research community who wants to increase identification certainty, transparency, and reproducibility while using database search engines.

To assist new users, we provide a brief glossary of key Mascot and mass spectrometry terms used throughout this practical workflow (Table 1).

Table 1. Key terms and definitions in Mascot-based proteomics.

Term	Definition
MS/MS (Tandem Mass Spectrometry)	An analytical technique in which selected precursor ions (peptides) are fragmented and the resulting fragment ions are analyzed to determine peptide sequence.
Precursor Ion (MS1)	The intact peptide ion measured before fragmentation.
Fragmented Ion (MS2)	Ions generated after breaking the precursor peptide (commonly b- and y-ions).
Peptide Spectrum Match (PSM)	The assignment of a peptide sequence to an experimental MS/MS spectrum.
Mascot Score	A probability-based score calculated as $-10 \times \log_{10}(P)$, where P is the probability that the observed match between the experimental spectrum and a peptide sequence in the database occurred by chance. Higher scores indicate more significant (less random) matches.
E-value	The number of matches with equal or better scores that are expected to occur by chance alone. The lower the expectation value, the more significant the score.
Peptide Tolerance (Precursor Mass Tolerance)	A filter for candidate sequences, instructing the software to only consider database peptides whose mass falls within this specified window (usually measured in Daltons, Da, or parts per million, ppm)
MS/MS tolerance (Fragment Mass Tolerance)	The mass window applied to the smaller pieces (fragment ions) that are generated <i>after</i> the precursor ion is broken apart.
Missed cleavage	Number of times trypsin failed to cleave at a possible site.
Fixed modification	A modification that is always present on specific amino acids (e.g., carbamidomethylation of cysteine).
Variable modification	A modification that may or may not occur (e.g., methionine oxidation).
Decoy Database	Reversed or randomized protein database used to estimate the False Discovery Rate (FDR).
False Discovery Rate (FDR)	Estimated percentage of false positive identifications among significant matches.
Spectral Library	Collection of experimentally observed and confidently identified MS/MS spectra used for direct spectrum matching
emPAI (Exponentially Modified Protein Abundance Index)	A label-free, relative quantitation of the proteins in a mixture based on protein coverage by the peptide matches in a database search result.

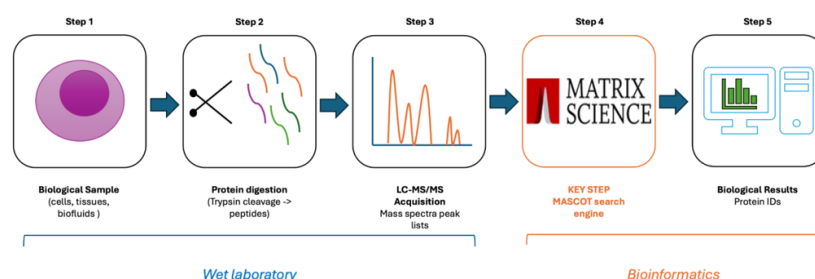


Figure 1. Overview of the shotgun proteomics workflow. Proteins extracted from a biological sample are enzymatically digested into peptides (Step 2) and analyzed by liquid chromatography coupled to tandem mass spectrometry (Step 3). The resulting MS/MS spectra are then processed by Mascot (Step 4), a probability-based database search engine primarily used for peptide and protein identification. This tutorial focuses on Step 4 - database searching and result interpretation using Mascot.

MATERIALS AND METHODS

2.1 Sample preparation

Escherichia coli DH5α cells were grown to mid-log phase, harvested by centrifugation, and washed with PBS. Cell pellets were resuspended in extraction buffer containing 4% SDS, 10 mM DTT in 100 mM Tris-HCl (pH 7.6), and sonicated on ice. Protein concentration was determined using the Pierce BCA Protein Assay Kit. Samples were denatured at 96°C, reduced, and alkylated with iodoacetamide. Proteins were precipitated using the chloroform/methanol method, resuspended

in 100 mM NH_4HCO_3 , and digested overnight with trypsin at a 1:100 ratio at 37°C. After digestion, samples were vacuum-dried and desalted prior to LC-MS/MS analysis [10] (Figure 1).

While this workflow demonstrates identification only, Mascot also supports relative quantification through additional tools. Mascot Distiller enables label-free quantification as well as analysis of isotopic labeling approaches (e.g., SILAC, TMT). Additionally, the emPAI (exponentially modified Protein Abundance Index) method provides a simple

semi-quantitative estimate of protein abundance based on the number of observed peptides per protein.

2.2 Mass spectrometry data acquisition

The selection and recording of peptide ions during analysis is determined by mass spectrometry data acquisition. Peptide samples were analyzed by Thermo Dionex Ultimate 3000 RSLCnano System coupled via CaptiveSpray nanoBooster ion source to an Impact II QTOF tandem mass spectrometer (Bruker Daltonics). Peptide separation and analysis were performed by liquid chromatography coupled with tandem mass spectrometry (LC-MS/MS) using an electrospray ionization source. To detect intact peptide ions, a full-scan mass spectrometric acquisition was first performed, followed by data-dependent acquisition, in which selected precursor ions were automatically isolated and fragmented. This fragmentation yielded sequence-specific ions, which were recorded as MS/MS spectra. The obtained spectra were used for additional database searches and peptide identification, coupled with the matching *m/z* values and charge states of the precursor ions.

2.3 Raw data processing and file conversion

Mass spectrometry output files change depending on the instrument vendor because each platform generates data in a specific format. These raw files are then converted to search engine-friendly forms using either vendor-specific software or universal conversion tools like MSConvert (ProteoWizard 3.0). Common raw data formats include .d files from Bruker and Agilent instruments. Thermo Scientific systems produce .raw files as well as .wiff files generated by AB Sciex platforms [11].

The resulting files, commonly generated in Mascot Generic Format (MGF), contain simplified spectral data suitable for downstream database searching.

2.4 Mascot database search and parameter selection

Protein identification using Mascot (Version: 2.6.2) is accomplished by defining search parameters before database analysis [6]. Search parameters are critical and should be defined properly. Curated databases, such as SwissProt, are commonly used due to their high annotation quality. Furthermore, providing the correct taxonomy and digestive enzyme reduces

the search space and increases identification accuracy [7] (see Table 2).

Database searches were performed against the target protein sequence database supplemented with a reversed decoy database. FDR was calculated by comparing each peptide sequence against its reversed counterpart in the decoy database, with matches to the decoy treated as false positive identifications. Peptide-spectrum matches were filtered to 1% FDR at the peptide level, and protein identifications were filtered to 1% FDR at the protein level.

Additionally, Mascot offers error-tolerant searching, which repeatedly searches for unanticipated variations such as non-specific cleavages, sequence variants, and unknown PTMs [26]. This feature is particularly useful when standard searching fails to identify a significant portion of spectra. However, it should be noted that error-tolerant searching considerably slows down the search engine due to the greatly expanded search space, and the increased number of false positive matches may make confirmation challenging.

Mascot database searches were initially performed using precursor mass tolerance (PMT) of 1.2 Da and fragment mass tolerance (FMT) of 0.6 Da, as these values represent common default settings for many Q-TOF instruments and have been successfully applied in previous studies [10]. To evaluate the robustness of the results, additional searches were conducted with tighter tolerances (PMT = 0.02 Da, FMT = 0.05 Da). As shown in Table 3, protein scores were higher with the tighter mass tolerances, as expected due to increased matching specificity. Nevertheless, the overall number of identified proteins and protein families differed by only ~1% between the two conditions. This high consistency demonstrates that the wider, default tolerances provide reliable and comparable results while remaining practical for routine analysis. Optimal mass tolerance settings should always be chosen according to the type of mass spectrometer used.

2.5 Post-translational modification analysis (PTMs, coverage, biology)

A protein is initially synthesized as a primary amino acid chain but can then be modified. The ribosome creates the polypeptide chain; however, the final structure and function

Table 2. Mascot search parameters for LC-MS/MS proteomics analysis applied for processing *Escherichia coli* sample data

Parameter	Settings	Description
Database	SwissProt (AA)	Curated database used for peptide matching
Taxonomy	<i>Escherichia coli</i>	Target organism or group specified for database searching
Enzyme	Trypsin/P	Peptide digestion specificity
Fixed modifications	Carbamidomethyl (C)	Expected modification on all occurrences of an amino acid
Variable modifications	Oxidation (M) Acetyl (Protein N-term)	Modifications that may occur on some peptides
Missed cleavages	2	Allows incomplete enzymatic digestion
Peptide tolerance	1.2 Da	Precursor ion mass accuracy
Fragment mass tolerance	0.6 Da	MS/MS fragment ion mass accuracy (suitable for ion trap or low-resolution fragmentation)
Decoy	Yes	Reversed or randomized protein sequences used to estimate false discovery rates
False Discovery Rate (FDR)	Up to 1%	Threshold controlling the proportion of false-positive identifications

Table 3. Mascot protein scores obtained with different precursor mass tolerance (PMT) and fragment mass tolerance (FMT) settings using the *Escherichia coli* sample dataset.

Protein Accession	Default Tolerances (PMT - 1.2 Da; FMT - 0.6 Da)	Tight Tolerances (PMT - 0.02 Da; FMT - 0.05 Da)	Increase (%)
RBSB_ECOLI	20495	23549	+ 14.9
EFTU1_ECO24	18842	22617	+ 30.6
TNAA_ECOBW	10532	13027	+ 20
MDH_ECOBW		12073	+ 22.2
DCEB_ECOLI		11562	+ 18.4

of the protein are formed later by enzymes that add or remove chemical groups. These processes are called post-translational modifications (PTMs). PTMs regulate protein activity, signaling pathways, protein localization, stability, and disease [12, 13]. To demonstrate PTM detection capabilities of Mascot, variable modifications were specified in the search parameters using a publicly available histone modification dataset (JPST003119, jPOST repository), as described in Section 4.2. The modifications applied were adopted directly from those specified in the original dataset deposition: Acetyl (Protein N-term), Acetyl (K), Methyl (K), Dimethyl (K), Trimethyl (K), and GG (K) as variable modifications, with Carbamidomethyl (C) as a fixed modification. These reflect the known histone H3t PTM landscape of the original study.

Peptide Identification: Sequence Database vs. Spectral Library Searching

Peptide identification in proteomics is commonly performed using two complementary approaches: sequence database searching and spectral library searching.

Sequence database searching compares experimental tandem mass spectra with theoretical spectra generated from protein sequence databases. Search engines, such as Mascot, assign the best-matching peptide sequence as the identification. This method is particularly useful for discovering novel peptides predicted from genomic data, provided that relevant post-translational modifications are correctly specified in the search parameters.

Spectral library searching, in contrast, matches experimental spectra directly against previously acquired and confidently identified experimental spectra. This approach offers higher sensitivity and better handling of complex fragmentation patterns (including peak intensities and non-canonical ions), but is limited to peptides that have been previously observed and included in the library.

The advantages and limitations of both methods are summarized in the literature [14–21]. In this work, we combine both strategies within the Mascot platform. Figure 2 illustrates the integrated workflow used in this study: starting from high-

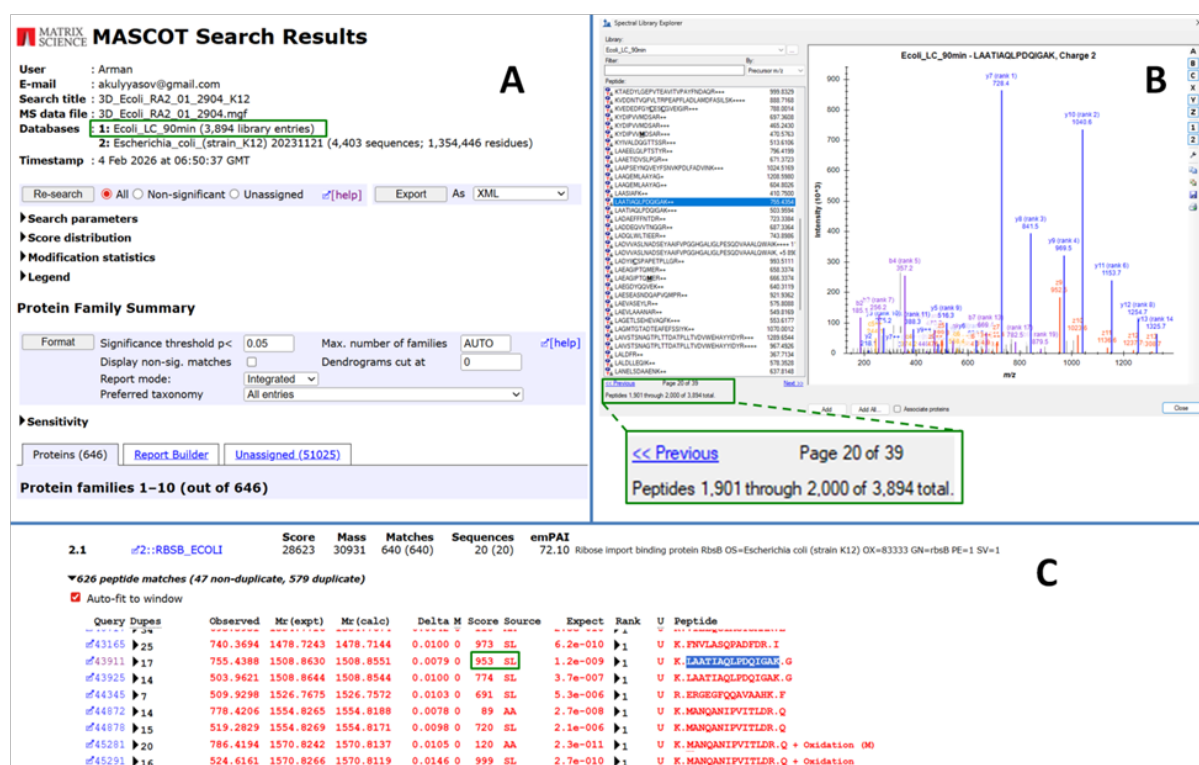


Figure 2. Demonstration of using both Spectral Library and Fasta database for Exploration of *Escherichia coli* Proteome on MASCOT Search Server.

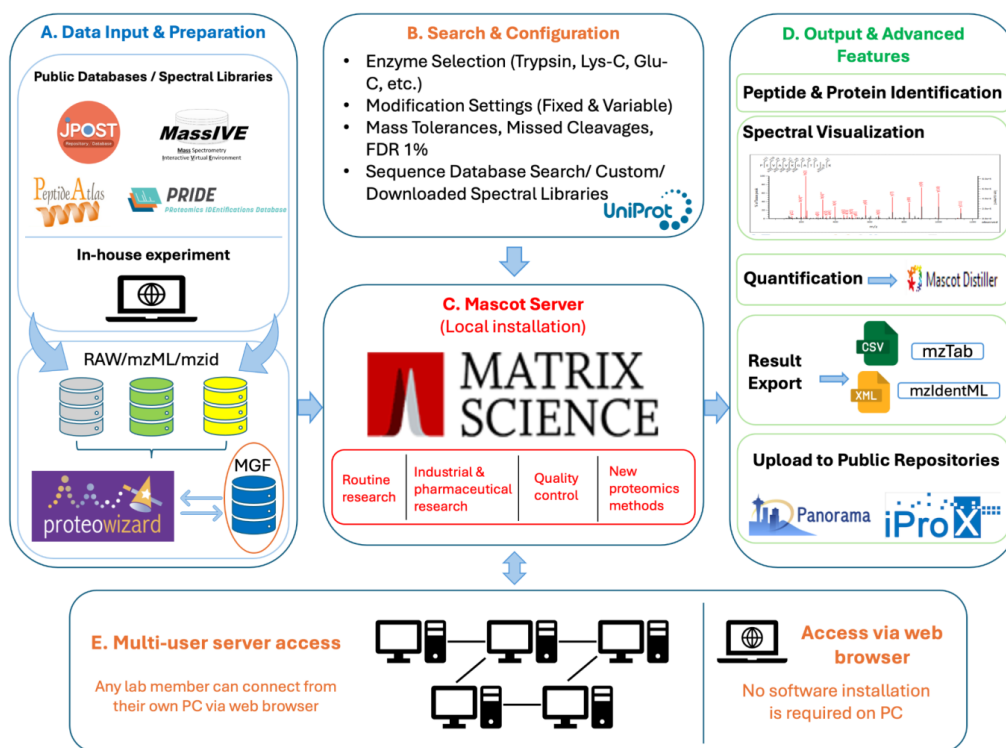


Figure 3. Comprehensive workflow and additional capabilities of the Mascot server platform. (A) Raw mass spectrometry data from public repositories (e.g., PRIDE, MassIVE) or in-house experiments are converted into Mascot-compatible formats (MGF) using tools such as ProteoWizard.

(B) Database search parameters are defined, including enzyme specificity, post-translational modifications, mass tolerances, and false discovery rate ($FDR \leq 1\%$), using reference protein databases (e.g., UniProt) or spectral libraries.

(C) Peptide-spectrum matching and protein identification are performed using the Mascot search engine deployed on a local server.

(D) Downstream analysis includes spectral visualization, peptide/protein identification, quantification, and export into standard formats (e.g., .csv, mzIdentML), with optional deposition to public repositories (e.g., Panorama, iProX).

(E) The system supports multi-user access via web-based interfaces without requiring local software installation.

level search parameters, proceeding through protein family summaries, and culminating in the detailed validation of individual peptide-spectrum matches using spectral library information. This combined approach enhances identification confidence and sensitivity in the analysis of the *Escherichia coli* proteome.

This figure presents a comprehensive overview of protein identification through MASCOT search and subsequent spectral library validation. Panel A displays the summary of the MASCOT search results, detailing the search parameters from the MS data file (3D_Ecoli_RA2_01_2904.mgf) and two databases: Spectral library (SL) database Ecoli_LC_90min and amino acid (AA) sequence database *Escherichia coli* strain K12. It also provides an overview of protein family identification, indicating 646 proteins identified. Panel B illustrates a representative MS/MS spectrum from the Spectral Library Explorer for the peptide «LAATIAQLPDQIGAK» with a charge of 2, originating from *E. coli* which was visualized by Skyline tool (26.1.0.057). The spectrum shows fragment ion peaks (e.g., y-ions and b-ions) used for peptide sequencing and confirmation. A total number of peptides in the Ecoli_LC_90min library is 3,894 (green frame in A panel and zoomed inset in B panel). Panel C provides a detailed table of peptide matches for the identified protein «Ribose import binding protein RbsB» (*Escherichia coli*), showing 626 pep-

tide matches. Note that scores are higher for matches derived from spectral library (SL) than the fasta databases (AA).

Practical application of Mascot database

Although Mascot has a range of additional capabilities, its primary use as a database search engine for peptide and protein identification (Figure 3). These include statistical validation via false discovery rate estimation, control over search parameters and post-translational changes, and adjustable database and spectrum library management inside the core search engine. A different data processing program called Mascot Distiller is needed for protein quantification, both label-based and label-free. Instead of being a stand-alone search engine, Mascot Server has a server-based architecture that facilitates multi-user access and collaborative data processing.

The Database Manager allows users to maintain up-to-date protein sequence databases such as Swiss-Prot/UniProt and build or import spectral libraries from their own experiments or publicly available data. Users have full control over all search parameters, including the choice of proteolytic enzyme, precursor and fragment ion mass tolerances, and fixed and variable modifications. For example, carbamidomethylation of cysteine is typically set as a fixed modification since it is introduced during sample preparation. Correctly defining these parameters is essential for reliable peptide identification and accurate spectral matching. Label-free and isobaric quan-

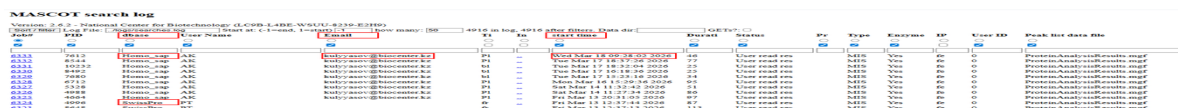


Figure 4. Annotated screenshot of Mascot server interface highlighted with key search parameters. Mascot's server mode allows for convenient searching, viewing results in real time, and filtering them by various criteria.



Figure 5. Adjustment of the false discovery rate (FDR) to 1% in the Mascot search parameters. This threshold is commonly used to ensure reliable peptide and protein identifications. Controlling FDR at 1% is a widely accepted standard in shotgun proteomics to minimize false positive results while maintaining good sensitivity.

tification is supported by the integrated Mascot Distiller module, and Mascot offers easy spectrum viewing in Peptide View after searching. All results can be easily exported to common formats and published in public repositories. Thus, a laboratory studying a rare disease or non-model organism can utilize thousands of validated spectra from published studies worldwide, significantly increasing the efficiency and depth of proteomic analysis.

Mascot Server is designed around a client-server architecture in which authorized users submit searches and retrieve results through a standard web browser interface.

4.1 Dataset description

To demonstrate the operation of the Mascot search engine and the features of its server interface, we reanalyzed previously obtained LC-MS/MS data (Figure 4). A tryptic lysate of *Escherichia coli*, previously analyzed in our laboratory, was used as a test set. The dataset consisted of the file 3D_Ecoli_RA2_01_2904.mgf containing 68,504 MS/MS spectra. These data were searched against the SwissProt database (version 2026_02, containing 575,503 sequences) restricted to *Escherichia coli* taxonomy (23,291 sequences). Search parameters are described in Table 2. After filtering at 1% False Discovery Rate (FDR) at the PSM level, the search yielded 12,080 PSMs and 521 protein groups (Figure 5).

This reanalysis allowed us to clearly demonstrate the Mascot web interface in server mode, the process of setting up search parameters, and the filtering and visualization of results.

4.2 Working with jPOST data

To demonstrate the identification of post-translational modifications, an independent public dataset was retrieved from the jPOST repository (JPST003119; Comprehensive post-translational modifications in the testis-specific histone

variant H3t protein validated in tagged knock-in mice) [24]. This external dataset was used to compare search results and demonstrate the reproducibility of the analysis on independent data. Raw files were downloaded and converted to MGF format using MSConvert (ProteoWizard) prior to uploading to the local Mascot server [25].

Using publicly available data allowed us to demonstrate the reproducibility of the analysis and the ability to work with external datasets.

4.3 Mascot search

The search was conducted through the web interface of the local Mascot server. The downloaded mgf files were loaded into the program, after which the search parameters presented in Table 2 were applied according to the experimental data provided in the repository. The selected fixed and variable modifications, along with other key settings, are highlighted in Figure 6.

4.4 Demonstration of post translational modifications identification

Mascot performs a conventional database search by matching experimental MS/MS spectra to theoretical peptide sequences extracted from the protein database. In this instance, it will only take into account the chosen enzyme, the particular modifications (phosphorylation, oxidation, etc.), and the sequences present in the database. Mascot operates strictly according to the parameters defined by the user. However, in reality, samples do not follow these rules, as a number of unknown processes occur before and during sample preparation. For example, unknown post-translational modifications (PTMs) and an amino acid substitution caused by Single Nucleotide Polymorphisms (SNPs). Non-specific cleavages can also prevent accurate identification during database searches. For example, trypsin cleaves at lysine and arginine residues,



Figure 6. Mascot database search and identification of histone H3.1 peptides from *Mus musculus*. (A) Database search parameters used for Mascot analysis. (B) Protein identification results. The two entries belong to the same protein group (sameset) as they share multiple peptides. (C) Peptide matches showing sequences present in both protein families and unique peptides. (D) MS/MS spectrum of the trimethylated peptide K.QLATKAAR.K (K5 +Trimethyl, Query 6145)

but sometimes peptides can be produced during incomplete digestion and/or in the presence of other proteases. These peptides do not match the specified parameters for the trypsin enzyme, meaning the Mascot ignores such sequences.

Mass spectrometry analysis successfully identified multiple post-translational modifications on histone H3.1. As shown in Figure 6, several lysine methylation states were detected, including trimethylation at lysine 5 on the peptide K.QLATKAAR.K (Panel D). The Mascot search results revealed both peptides shared between histone H3 protein families and unique peptides (Panel C), confirming reliable protein and peptide identification.

These results demonstrate that the overall pattern of lysine methylation on histone H3.1 is consistent with modifications previously reported on canonical histone H3. This supports the finding that H3t-FLAG undergoes post-translational modifications in a manner largely comparable to that of canonical H3.1/H3.2, although differences in the relative abundance of specific marks (e.g., at K9 and K27) were observed between H3t-FLAG and bulk H3 [24]. Different areas of the

cell undergo modifications. For instance, phosphorylation typically occurs in the cytoplasm and nucleus, while oxidative alterations may result from reactive oxygen species generated during cellular metabolism. Every mutation alters the mass of the peptide, which causes the spectra to vary during LC-MS/MS analysis. Mascot must take these modifications into account or the peptide will not be detected in the database.

Mascot looks for peptides by comparing the database's theoretical peptides with the experimental spectra. Numerous modification sites may be present in a single peptide, and there may be a large number of modified versions. If there are more modifications, the number of variants grows exponentially, and Mascot must check all combinations, which consequently leads to an increase in the search space, a slowdown in the computational analysis, and increase in the random matches [7]. Therefore, it is important to select a small number of biologically sound modifications in Mascot to prevent the search space from expanding unnecessarily.

4.5 Mascot file formats

Comprehensive data, including MS1 and MS2 spectra, in-

Table 4. Classification of common Mascot file formats.

Category	File Format	Extension	Description	Purpose / When Used
Input / Peak List Files	Mascot Generic Format (MGF)	.mgf	Plain text format containing MS/MS peak lists (m/z and intensity values)	Submit spectra to Mascot for searching. Most common input format.
Input / Peak List Files	mzML	.mzML	Standard HUPO-PSI XML format for raw or processed mass spectra	Alternative to MGF (more structured and standardized).
Search Result Files	Mascot DAT File	.dat	Proprietary binary format containing complete search results (old format)	Native Mascot result file (used in versions ≤ 2.8). Still supported for backwards compatibility.
Search Result Files	Mascot Search Results (MSR)	.msr	SQLite relational database format (newer, since Mascot 3.0)	Modern internal result storage on the server.
Export / Interchange Formats	Mascot XML	.xml	Custom XML format with full search results	Machine-readable, good for databases or further processing.
Export / Interchange Formats	CSV	.csv	Tabular text format (same content as Mascot XML)	Easy to open in Excel or import into spreadsheets.
Export / Interchange Formats	mzIdentML	.mzid	Standard HUPO-PSI format for peptide/protein identification results	Community standard; best for sharing and depositing data (e.g., PRIDE).
Export / Interchange Formats	mzTab	.mzTab	Lightweight tab-delimited format for summary results	Simple reporting of identifications and quantification.
Export / Special	MGF (Peak List Export)	.mgf	Export of peak lists from search results	Useful to re-use spectra or export for other tools.

strumental noise, calibration data, and operating metadata, can be found in raw output files. However, these raw instrument files cannot be processed directly by database search engines like Mascot. As a result, database searching requires conversion to appropriate formats.

During the conversion process, relevant MS/MS spectra are extracted, a portion of the background noise is eliminated, and precursor ion m/z values and charge states are assigned, creating a simplified dataset appropriate for database searching later on.

Mascot uses different file formats depending on the stage of the analysis [11]. These can be classified into three main groups (Table 4):

Input/Peak List Files - include processed MS/MS spectra (mostly **MGF** and **mzML**) that are prepared for database searches.

Internal Search Result Files - Store the complete search output from Mascot (**DAT** for older versions and **MSR** for newer versions).

Export/Interchange Formats - Allow results to be exported for further analysis, visualization, or deposition in public repositories. The most important are **mzIdentML** (standard), **XML/CSV** (Mascot-specific), and **mzTab** (summary reporting).

In this study, MGF files downloaded from the jPOST repository were used as input.

5. CONCLUSION

This study showed that Mascot is a practical and reliable tool for protein identification in shotgun proteomics. We demonstrated that appropriate parameter selection including enzyme specificity, mass tolerances, and modification settings

is essential for precise and repeatable identifications by examining both in-house *Escherichia coli* LC-MS/MS data and a publicly accessible histone modification dataset from the jPOST repository.

The workflow highlighted two complementary search strategies available in Mascot: sequence database searching, which matches experimental spectra against theoretically predicted peptides derived from FASTA protein databases, and spectral library searching, which compares spectra directly using both *m/z* values and fragment ion intensities. Spectral library searching demonstrated higher sensitivity, though it is limited to previously observed peptides.

The successful detection of known post-translational modifications, such as trimethylation on histone H3 peptides, confirms that the search engine can identify biologically relevant PTMs when appropriate parameters are used. MGF was used as the input file format throughout, demonstrating its practicality as a universal format compatible with both local instrument data and publicly available datasets.

Mascot remains a reliable and accessible tool for proteomics data interpretation, particularly for research groups transitioning into mass spectrometry-based workflows. This practical workflow aims to lower the barrier for researchers who generate proteomics data but lack experience in database searching, ultimately contributing to greater methodological transparency in the field.

LIMITATIONS

This workflow was demonstrated using Mascot version 2.6.2. As newer versions of Mascot introduce updated features and interface changes, some details described here may differ in more recent releases. Additionally, the analysis was limited to data-dependent acquisition (DDA) using trypsin as the digestion enzyme, and the described workflow may not fully apply to data-independent acquisition (DIA)-based experiments or alternative enzymatic digestion strategies, both of which are becoming increasingly common in proteomics research [27, 28].

ACKNOWLEDGMENTS

The authors would like to thank Patrick Emery (Matrix Science Ltd) for his valuable support and expert assistance throughout the data analysis and Mascot search optimization. His guidance and insights were instrumental in the successful completion of this work.

This study was carried out within the framework of the Scientific and Technical Program BR24992881, funded by the Science Committee of the Ministry of Science and Higher Education of the Republic of Kazakhstan.

AUTHOR CONTRIBUTIONS

S.M. conceived and wrote the manuscript, performed the literature analysis, and prepared the original draft. A.K. (Supervisor) supervised the work, contributed to manuscript editing, and authored the section “Advantages and Limitations of Spectral Libraries”. A.K. (co-author) contributed to the analytical aspects of the study. All authors reviewed and approved the final version of the manuscript.

REFERENCES

1. Shuken S. R. An Introduction to Mass Spectrometry-Based Proteomics // *Journal of Proteome Research*. – 2023. – Vol. 22, № 7. – P. 2151–2171.
2. Zhang Y. et al. Protein Analysis by Shotgun/Bottom-up Proteomics // *Chemical Reviews*. – 2013. – Vol. 113, № 4. – P. 2343–2394.
3. Yates J. R. The Revolution and Evolution of Shotgun Proteomics for Large-Scale Proteome Analysis // *Journal of the American Chemical Society*. – 2013. – Vol. 135, № 5. – P. 1629–1640.
4. Henzel W. J., Watanabe C., Stults J. T. Protein identification: The origins of peptide mass fingerprinting // *Journal of the American Society for Mass Spectrometry*. – 2003. – Vol. 14, № 9. – P. 931–942.
5. Cottrell J. S. Protein identification using MS/MS data // *Journal of Proteomics*. – 2011. – Vol. 74, № 10. – P. 1842–1851.
6. Perkins D. N. et al. Probability-based protein identification by searching sequence databases using mass spectrometry data // *Electrophoresis*. – 1999. – Vol. 20, № 18. – P. 3551–3567.
7. Nesvizhskii A. I., Aebersold R. Interpretation of Shotgun Proteomic Data // *Molecular & Cellular Proteomics*. – 2005. – Vol. 4, № 10. – P. 1419–1440.
8. Marquioni V., Nunes F. M. F., Novo-Mansur M. T. M. Protein Identification by Database Searching of Mass Spectrometry Data in the Teaching of Proteomics // *Journal of Chemical Education*. – 2021. – Vol. 98, № 3. – P. 812–823.
9. Calderón Celis F. et al. Exploring Proteomics Workflows: Hands-On Learning and Evaluation of Analytical Techniques in Proteomics // *Journal of Chemical Education*. – 2025. – Vol. 102, № 7. – P. 2732–2742.
10. Kulyyassov A., Makhsatova S., Kurmanbay A. Protein Identification Improvement in Complex Samples Using Higher Frequency MS Acquisition and PEAKS Software // *Applied Sciences*. – 2025. – Vol. 15, № 2. – P. 666.
11. Deutsch E. W. File Formats Commonly Used in Mass Spectrometry Proteomics // *Molecular & Cellular Proteomics*. – 2012. – Vol. 11, № 12. – P. 1612–1621.
12. Brandi J. et al. Advances in enrichment methods for mass spectrometry-based proteomics analysis of post-translational modifications // *Journal of Chromatography A*. – 2022. – Vol. 1678. – P. 463352.
13. Ridgeway N. H., Biggar K. K. Uncovering Enzyme-Specific Post-Translational Modifications: An Overview of Current Methods // *Proteomes*. – 2025. – Vol. 13, № 3. – P. 37.
14. Ahrens C. H. et al. Generating and navigating proteome maps using mass spectrometry // *Nature Reviews Molecular Cell Biology*. – 2010. – Vol. 11, № 11. – P. 789–801.
15. Craig R., Cortens J. P., Beavis R. C. The use of proteotypic peptide libraries for protein identification // *Rapid Communications in Mass Spectrometry*. – 2005. – Vol. 19, № 13. – P. 1844–1850.
16. Craig R. et al. Using Annotated Peptide Mass Spectrum Libraries for Protein Identification // *Journal of Proteome*

Research. – 2006. – Vol. 5, № 8. – P. 1843–1849.

17. Shao W., Lam H. Tandem mass spectral libraries of peptides and their roles in proteomics research // *Mass Spectrometry Reviews*. – 2017. – Vol. 36, № 5. – P. 634–648.

18. Eng J. K., McCormack A. L., Yates J. R. 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*. – 1994. – Vol. 5, № 11. – P. 976–989.

19. Ji C. et al. Extending the coverage of spectral libraries: A neighbor-based approach to predicting intensities of peptide fragmentation spectra // *PROTEOMICS*. – 2013. – Vol. 13, № 5. – P. 756–765.

20. Lam H., Aebersold R. Using Spectral Libraries for Peptide Identification from Tandem Mass Spectrometry (MS/MS) Data // *Current Protocols in Protein Science*. – 2010. – Vol. 60, № 1.

21. Matthiesen R. Methods, algorithms and tools in computational proteomics: A practical point of view // *PROTEOMICS*. – 2007. – Vol. 7, № 16. – P. 2815–2832.

22. Stein S. E., Scott D. R. Optimization and testing of mass spectral library search algorithms for compound identification // *Journal of the American Society for Mass Spec-*

trometry. – 1994. – Vol. 5, № 9. – P. 859–866.

23. MacLean B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments // *Bioinformatics*. – 2010. – Vol. 26, № 7. – P. 966–968.

24. Kawaguchi T. et al. Comprehensive posttranslational modifications in the testis-specific histone variant H3t protein validated in tagged knock-in mice // *Scientific Reports*. – 2024. – Vol. 14, № 1. – P. 21305.

25. Chambers M. C. et al. A cross-platform toolkit for mass spectrometry and proteomics // *Nature Biotechnology*. – 2012. – Vol. 30, № 10. – P. 918–920.

26. Mann M., Wilm M. Error-Tolerant Identification of Peptides in Sequence Databases by Peptide Sequence Tags // *Analytical Chemistry*. – 1994. – Vol. 66, № 24. – P. 4390–4399.

27. Guo T., Aebersold R. Recent advances of data-independent acquisition mass spectrometry-based proteomics // *PROTEOMICS*. – 2023. – Vol. 23, № 7–8. – P. 2200011.

28. Li J., Smith L. S., Zhu H.-J. Data-independent acquisition (DIA): An emerging proteomics technology for analysis of drug-metabolizing enzymes and transporters // *Drug Discovery Today: Technologies*. – 2021. – Vol. 39. – P. 49–56.

УДК 543.51:577.112

ПРАКТИЧЕСКИЙ ПРОТОКОЛ ИДЕНТИФИКАЦИИ БЕЛКОВ И АНАЛИЗА ПОСТТРАНСЛЯЦИОННЫХ МОДИФИКАЦИЙ С ИСПОЛЬЗОВАНИЕМ MASCOT

Махсатова С.А. <https://orcid.org/0009-0003-8772-086X>, Курманбай А.Б. <https://orcid.org/0009-0007-2131-0090>,
Кулыясов А.Т.* <https://orcid.org/0000-0002-7932-5689>

ТОО «Национальный центр биотехнологии», Кургальжинское шоссе, здание 13/5, Астана, 010000, Казахстан

*kulyyasov@biocenter.kz

АБСТРАКТ

Современные биологические исследования стали высоко междисциплинарными, где генерация данных масс-спектрометрии должна сочетаться с навыками биоинформатики и анализа данных для получения осмысленных результатов. Однако многие исследователи, регулярно проводящие масс-спектрометрические эксперименты, часто испытывают недостаток практических знаний в интерпретации результатов и оптимизации поиска в базах данных. Данная практическая статья описывает использование поискового движка Mascot для идентификации пептидов и белков в shotgun-протеомике. С помощью локального сервера Mascot (версия 2.6.2) были проанализированы два набора данных: собственный триптический лизат *Escherichia coli*, полученный на масс-спектрометре Impact II QTOF, и общедоступный датасет по посттрансляционным модификациям гистонов из репозитория jPOST. Рабочий процесс включает оптимизацию параметров поиска, выполнение поиска, интерпретацию результатов и демонстрацию выявления посттрансляционных модификаций. В представленном примере с гистонами показана идентификация уже известных модификаций лизина (включая триметилирование), обнаруженных в результате поиска. Статья служит практическим руководством для исследователей, желающих улучшить навыки анализа протеомных данных в Mascot, и подчеркивает критическую важность правильного выбора параметров для получения точных и воспроизводимых результатов.

Ключевые слова: Mascot, shotgun-протеомика, поиск в базе данных, анализ данных, посттрансляционные модификации, идентификация пептидов

ЭОЖ 543.51:577.112

MASCOT КӨМЕГІМЕН АҚУЫЗДАРДЫ ИДЕНТИФИКАЦИЯЛАУ ЖӘНЕ ПОСТТРАНСЛЯЦИЯЛЫҚ МОДИФИКАЦИЯЛАРДЫ ТАЛДАУДЫҢ ПРАКТИКАЛЫҚ ПРОТОКОЛЫ

Махсатова С.А. <https://orcid.org/0009-0003-8772-086X>, Құрманбай А.Б. <https://orcid.org/0009-0007-2131-0090>,
Кулыясов А.Т.* <https://orcid.org/0000-0002-7932-5689>

Ұлттық биотехнология орталығы, Қорғалжын тас жолы, 13/5, Астана, 010000, Қазақстан

*kulyyasov@biocenter.kz

ТҮЙІН

Қазіргі заманғы биологиялық зерттеулер пәнаралық сипатқа ие болды, бұл маңызды нәтижелерге қол жеткізу үшін масс-спектрометрия деректерін жасауды биоинформатика және деректерді талдау дағдыларымен біріктіруді талап етеді. Дегенмен, масс-спектрометрия тәжірибелерін үнемі жүргізетін көптеген зерттеушілерде нәтижелерді түсіндіру және деректер іздеулерін оңтайландыру бойынша практикалық тәжірибе жетіспейді. Бұл практикалық мақалада мылтық протеомикасындағы пептидтер мен ақуыздарды анықтау үшін Mascot іздеу жүйесін пайдалану сипатталған. Жергілікті Mascot серверін (2.6.2 нұсқасы) пайдаланып, екі деректер жиынтығы талданды: Impact II QTOF масс-спектрометрінде алынған *Escherichia coli* триптикалық лизаты және jPOST репозиторийінен жалпыға қолжетімді гистоннан кейінгі трансляциялық модификациялар деректер жиынтығы. Жұмыс процесіне іздеу параметрлерін оңтайландыру, іздеуді орындау, нәтижелерді түсіндіру және трансляциядан кейінгі модификацияларды анықтауды көрсету кіреді. Гистондардың ұсынылған мысалы іздеу арқылы анықталған белгілі лизин модификацияларын (триметилденуді қоса алғанда) анықтауды көрсетеді. Бұл мақала Mascot бағдарламасында протеомикалық деректерді талдау дағдыларын жетілдіргісі келетін зерттеушілер үшін практикалық нұсқаулық ретінде қызмет етеді және дәл және қайталанатын нәтижелерге қол жеткізу үшін дұрыс параметрлерді таңдаудың маңыздылығын атап көрсетеді.

Кілтті сөздер: Mascot, shotgun-протеомика, деректер базасы бойынша іздеу, деректерді талдау, посттрансляциялық модификациялар, пептидтерді идентификациялау