Abstract

Proteomics is key to understanding biological systems, providing insights into the temporal, spatial, and functional dynamics of proteins. This study introduces a Python 3-based computational pipeline for analysing bacterial genome-derived protein sequences from FASTA files. The pipeline predicts open reading frames (ORFS), digests them using four commonly used enzymes- Trypsin, V8 proteinase, Endoproteinase Lys-C and Endoproteinase Arg-C- while accounting for any missed cleavages. It calculates mass-to-charge (m/z) values for each peptide and evaluates their distribution within the 1000 to 1500 Da range.

Results demonstrate that Trypsin produced 5912 unique peptides, significantly outperforming other enzymes. This finding aligns with its established effectiveness in mass spectrometry-based proteomics, highlighting its high cleavage specificity. This pipeline facilitates accurate enzyme selection, saving wet-lab time and resources and enhancing protein identification for bacterial proteomic studies.

Introduction

Proteomics, the study of the proteome, provides critical insights into cellular function, disease mechanisms, and drug development by understanding when, where, and how proteins are expressed, modified and transported [1]. Advances in bioinformatics and computational tools have enabled efficient proteomic analysis of complex datasets, improved protein identification and providing deeper biological insights [2]. A key step in early proteomic analysis is identifying proteins from mass spectrometry data by generating diagnostic peptide fragments. Selecting the optimal enzyme for protein digestion is critical, as improper choices can negatively affect peptide yield, m/z distribution, and protein identification accuracy [3]. Computational approaches to enzyme selection streamline experimental design by optimising time and resources.

This study developed a Python-based command-line pipeline to identify the enzyme—Trypsin, V8 protease, Endoproteinase Lys-C, or Endoproteinase Arg-C—that maximizes peptide yield and diagnostic value within the 1000–1500 m/z range. The pipeline processes nucleotide sequences from an unknown bacterial genome to identify ORFs, digest proteins with enzyme-specific rules, calculate peptide m/z ratios, and generate statistical analyses using fixed-range and sliding-window histograms. The user can select various modes within the pipeline to tailor it to their specific needs. By identifying the most effective enzyme for this proteomic study through computational approaches, this pipeline facilitates integration with subsequent wet-lab studies, optimising time and resources.

Methods

Software and Libraries used: Python 3, standard libraries; argparse. Pipeline Link: Github

Parameters: Input FASTA file (.fasta), minimum ORF length, output file name (.txt), enzymes (Trypsin, V8 proteinase, Endoproteinase Lys-C, Endoproteinase Arg-C), missed cleavage tolerance, m/z range.

Command line input includes a FASTA file, minimum ORF length and output file. The fastaread() function extracts nucleotide sequences then translated them using the translate() function, excluding ambiguous nucleotides (“N”) or proteins (“X”). The reading_frames() function generates three forward and three reverse reading frames for each sequence; reverse reading frames are obtained using the reverse_complement() function. ORFs are identified by “ATG” start codons and “TAA”, “TAG” or “TGA” stop codons using locate_orfs(); this also prevents overlapping ORFs. Outputs including the ORF sequence, frame, start position and length are outputted.

Protein sequences are digested using enzyme-specific cleavage rules, with up to one missed cleavage allowed. Command-line arguments specify the input file, enzyme type, minimum peptide length, missed cleavages, and output file. The peptide_maker() function avoided cleavage before Proline (P) and extracted peptides, while metadata (e.g., enzyme type, missed cleavages) was added using digester_output_maker() in a FASTA format.

Peptide m/z values were calculated using the pep2mass() function to sum amino acid masses, accounting for water and proton mass for both monoisotopic and average mass types. The fastaread() function was used to read and validate sequences and metadata from the FASTA file. Calculated m/z values accounted for the different peptide masses and charges; outputs were saved to an output file. FASTA files, mass type (monoisotopic or average), output file naming, and charges were handled using argparse.

Various statistics were created, analysing peptide counts, distributions and identifications within the 1000-1500 Da range. Operation of the different modes was executed via command-line options using argparse. The script accepts arguments to calculate the total number of peptides within a user-defined m/z range, generates histograms based on a user-defined bin size to evaluate peptide distribution, apply a sliding window approach with a window size of 0.6 Da and a step size of 0.5 Da to examine peptide distribution density changes and calculate the number of unambiguously identified proteins, defined as those with at least one peptide with a unique m/z value in the proteome.

Results

ORFs were successfully identified across the three forward and three reverse frames of the inputted FASTA file, and results were outputted to a .txt file. These ORFS were digested into individual peptides, annotated by the enzyme used for digestion, the total number of digested fragments, and any missed cleavages. The mass analysis provided the m/z values for each peptide sequence, accounting for digestion by Trypsin, Endoproteinase Lys-C, Endoproteinase Arg-C, and V8 protease (Glu-C). Trypsin digestion generated 5912 unique peptides, compared to 1571 peptides for Arg-C, 1479 peptides for Lys-C, and 1519 peptides for V8 protease. Thus, Trypsin was identified as the most effective enzyme for peptide digestion, due to its superior cleavage efficiency.

Peptide masses were calculated using both monoisotopic and average mass methods. A fixed histogram analysis determined the number of peptides within different m/z bins (Figure 1. A, B). Both methods produced an even peptide distribution across bins, but monoisotopic mass calculations exhibited noticeable fluctuations, while average mass calculations showed a gradual decline. Sliding window histogram analysis revealed a progressive decrease in peptide numbers as m/z values increased (Figure 1. C, D). Peptides calculated using monoisotopic mass displayed significant oscillations after approximately 1250 m/z reflecting ion complexity, whereas those based on average mass exhibited only slight fluctuations beyond this range.

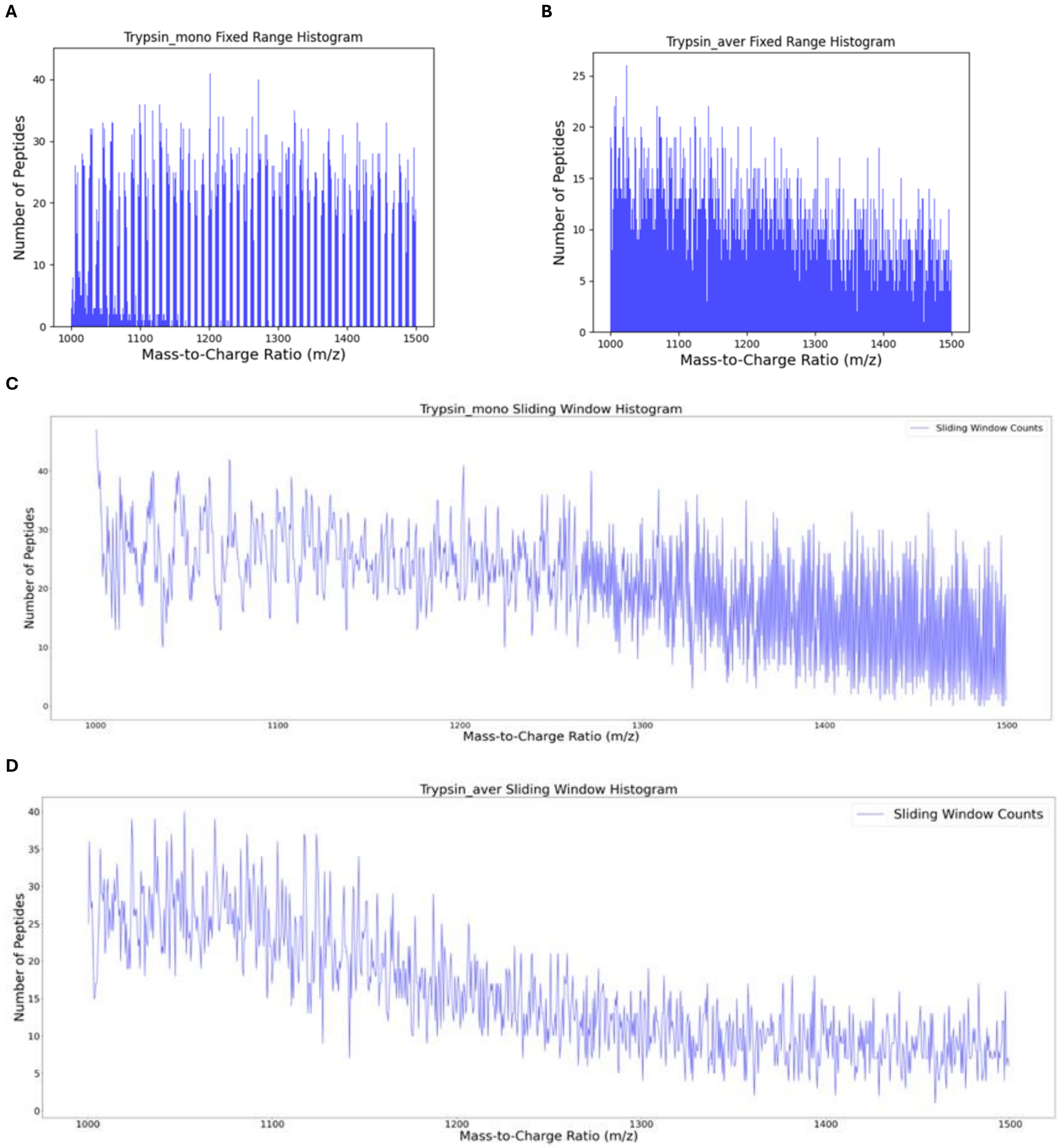

Figure 1. Relationship between Trypsin digested peptides and mass to charge ratios. A) Fixed range histogram showcasing the relationship between monoisotopic mass and mass to charge ratios for Trypsin digested peptides. B) Fixed range histogram showcasing the relationship between the average mass and mass to charge ratios for Trypsin digested peptides. C) Sliding Window Histogram for the monoisotopic mass showing changes in peptide distribution over ranges 1000-1500 D) Sliding Window Histogram for the average mass showing changes in peptide distribution over ranges 1000-1500

Discussion

This study identified Trypsin as the most effective enzyme for proteomic digestion, due to its high cleavage specificity and superior peptide yield, generating nearly four times as many peptides as the other enzymes. It generates peptides that are well-suited for separation by HPLC and subsequent analysis via mass spectrometry, consistent with its established role in proteomics [4,5].

Monoisotopic mass analysis provided greater precision, critical for distinguishing peptides with close m/z values but exhibited significant local fluctuations across the mass distribution. For larger and more complex peptides, which may occur in non-bacterial proteomes in future applications it can introduce ambiguity. By contrast, average mass analysis provides smoother trends due to its inherent averaging approach, mitigating the effect of outliers but sacrificing specificity, making it less suitable for distinguishing peptides with closely related masses [6,7]. Both analyses confirmed that Trypsin produced the greatest number of peptides (Figure 1), reinforcing its effectiveness in bacterial proteomics [8].

Future pipeline improvements include introducing support for processing multiple FASTA files simultaneously and incorporating additional proteolytic enzymes which are commercially available. Addressing the exclusion of post-translational modifications in this pipeline, which are common in bacterial proteomes would enhance the accuracy and reliability of the pipeline but would also introduce considerable complexity to the code [9].

In conclusion, this pipeline demonstrates Trypsin's effectiveness for bacterial proteomic digestion due to its high specificity and ability to generate the largest number of diagnostic peptides of the four enzymes measured, backing up its established role in the field. This was validated using monoisotopic and average mass analyses and informs future proteomic studies into this sequence. Expanding enzyme options and incorporating post-translational modification handling would further enhance this pipelines utility, supporting future *in-silico* and wet-lab studies.

References

1. Al-Amrani, S., Al-Jabri, Z., Al-Zaabi, A., Alshekaili, J., & Al-Khabori, M. (2021). Proteomics: Concepts and applications in human medicine. World journal of biological chemistry, 12(5), 57–69. https://doi.org/10.4331/wjbc.v12.i5.57

2. Schmidt, A., Forne, I. & Imhof, A. Bioinformatic analysis of proteomics data. BMC Syst Biol 8 (Suppl 2), S3 (2014). https://doi.org/10.1186/1752-0509-8-S2-S3

3. Cravatt, B. F., Wright, A. T., & Kozarich, J. W. (2008). Activity-based protein profiling: from enzyme chemistry to proteomic chemistry. Annual review of biochemistry, 77, 383–414. https://doi.org/10.1146/annurev.biochem.75.101304.124125

4. Chen, Q., Yan, G. and Zhang, X. (2015) 'Applying multiple proteases to direct digestion of hundred-scale cell samples for proteome analysis', *Rapid Communications in Mass Spectrometry*, 29(15), pp. 1389–1394. Available at: https://doi.org/10.1002/rcm.7230.

5. Mansuri, M.S. *et al.* (2024) 'Optimal conditions for carrying out trypsin digestions on complex proteomes: From bulk samples to single cells', *Journal of proteomics*, 297, pp. 105109–105109. Available at: https://doi.org/10.1016/j.jprot.2024.105109.

6. Radziński, P. *et al.* (2022) 'Envemind: Accurate Monoisotopic Mass Determination Based On Isotopic Envelope', *Journal of the American Society for Mass Spectrometry*, 33(11). Available at: https://doi.org/10.1021/jasms.2c00176.

7. Claesen, J. *et al.* (2015) 'Differences in the Elemental Isotope Definition May Lead to Errors in Modern Mass-Spectrometry-Based Proteomics', *Analytical Chemistry*, 87(21), pp. 10747–10754. Available at: https://doi.org/10.1021/acs.analchem.5b01165.

8. Cardoza, J. D., Parikh, J. R., Ficarro, S. B., & Marto, J. A. (2012). Mass spectrometry-based proteomics: qualitative identification to activity-based protein profiling. *Wiley interdisciplinary reviews. Systems biology and medicine*, *4*(2), 141–162. https://doi.org/10.1002/wsbm.166

9. Macek, B., Forchhammer, K., Hardouin, J. *et al.* Protein post-translational modifications in bacteria. *Nat Rev Microbiol* 17, 651–664 (2019). https://doi.org/10.1038/s41579-019-0243-0