International Journal of Engineering Research and Science & Technology

www.ijerst.org

ISSN : 2319-5991

Vol. 19 No. 4 (2023)

ijerst.editor@gmail.com

editor@ijerst.com

International Journal of Engineering Research and Science & Technology

ISSN 2319-5991 www.ijerst.com
Vol. 19, No. 4, 2023

Research Paper

MULTI-LEVEL CHARACTERIZATION OF PROTEOFORM DIVERSITY AND POST-TRANSLATIONAL MODIFICATIONS VIA COMPLEMENTARY MASS SPECTROMETRIC STRATEGIES AND UNIPROT ANNOTATION

Veeramreddy S Ratna Devi[1], Dr. Pawan Kumar Soni[2]

[1]Research Scholar, Department of Biochemistry, J. S. University, Shikohabad, UP

[2]Associate Professor, Department of Biochemistry, J. S. University, Shikohabad, UP

Abstract :

The fusion of top-down and bottom-up mass spectrometry (MS) in tandem has gradually become a predominant method for revealing the whole picture in the case of proteoforms and post-translational modifications (PTM). Top-down MS is analyzing whole intact proteins which gives complete information about the different forms the material might be actually existing while bottom-up MS tears the proteins apart into peptides thus making the MS sensitive enough for the analysis along with specific site localization of the PTMs. The recent research has been showing that the combination of these two complementary approaches not only increases the coverage of the proteome but also improves the accuracy of PTM and validation through proteoforms. Tools such as PTM-TBA and ProteoCombiner make the integration of the two types of data from MS along with running the protein knowledge bases like UniProt for identifying, localizing, and validating the PTMs with even more confidence. The reviews by Allen and Eyers (2023) and others discuss the drawbacks of top-down workflows due to fragmentation efficiency, data complexity, and computational interpretation, while bottom-up methods are still not able to produce intact proteoforms. Hybrid approaches, as demonstrated by Wu et al. and later on through the development of computational pipelines, are able to deal with these problems since they make use of the strengths coming from both strategies. All together, these breakthroughs are proving the combination of top-down and bottom-up MS along with database annotation to be the most reliable method for precise characterization of proteoforms and thus deeper understanding of protein structure, regulation, and cellular function.

Keywords: *post-translational modification, top-down mass spectrometry, bottom-up mass spectrometry.*

Received: 22-10-2023 Accepted: 01-12-2023 Published: 07-12-2023

I. INTRODUCTION

Proteins can be found in different molecular forms, which are called proteoforms, and these are produced as a result of genetic variation, alternative splicing, and post-translational modifications (PTMs). The latter are the most important factors in controlling the entire range of cellular functions since they regulate the characteristics of proteins such as their activity, place of action, duration of existence, and interaction with other proteins. To properly identify and describe PTMs is essential for gaining insights into the regulation of proteins in both normal and diseased states. Mass spectrometry (MS) has turned out to be the technology that underlies all proteomic studies, with its unmatched

sensitivity and specificity being the main advantages for protein analysis. Among the different MS-based techniques, the two major ones, top-down and bottom-up MS, have gained the widest acceptance as they are complementary to each other with respective strengths and weaknesses. In the top-down approach, the mass spectrometer analyses intact proteins, and thus the complete set of PTMs is retained, making it possible to carry out full proteoform characterization. The processing of proteins, on the other hand, in the bottom-up approach, is done by congressing them into single peptides through an enzyme digest, that permits high-throughput protein identification and accurate PTM site localization. Not that both ways are equally limited: top-down loses without a doubt the sensitivity and more sophisticated interpretation of the data is the problem, while bottom-up loses its advantage of getting all the PTMs occurring next to each other on the same proteoform disconnected. There are integrative approaches that combine the best of both methods to the rescue. Recent computational tools like PTM-TBA and ProteoCombiner, not only merge top-down and bottom-up MS data with curated protein databases like UniProt but also improve significantly PTM detection and validation. The collaboration of the analytical techniques provides a wider view of the diversity of proteoforms thereby pushing the proteomics field towards the total characterization of proteins in terms of their structure and function.

II. LITERATURE SURVEY

Post-translational modifications (PTMs) are at the center of cellular activities as they are the main regulators of protein structure and function. Bischoff and Schlüter [1] drew attention to the chemical diversity of amino acids and the multitude of non-enzymatic modifications, pointing out their implications in the protein's lifetime and its interactions. Among them, glycosylation- an important PTM- has an impact on the very cellular mechanisms that are the basis of health and disease, since it plays a central role in cell signaling, immune response, and the whole tumor development [2]. In a study, Rikova et al. [3] performed a global investigation of phosphotyrosine signaling and pointed to the presence of oncogenic kinases in lung cancer, thus giving an example of the role of faulty PTMs in the development of the disease and their capacity to be treated with drugs. The accessibility of new proteomics tools has opened the door to the discovery of new PTMs, and it is particularly evident in cancer research where mapping out of specific modifications aids in revealing the disease mechanisms and the development of therapy [4]. Mass spectrometry is the method that is used most in the PTM identification process, as it provides the means to do a very detailed analysis of protein modifications and isoforms [5]. Eisenberg et al. [6] treated the topic of the post-genomic era where they pointed to the understanding of protein function via PTM characterization as a necessity for the unraveling of complex biological networks. Also, Zhang et al. [7] depicted the utilization of PTM dynamics in drug discovery, mainly within the kinase family, as the case where deep knowledge about PTMs can be an asset in the picking of therapeutic approaches.

Horizontally, these studies allampoosting very clearly the need for precise PTM identification and characterization, the impact of mass spectrometry on proteomics, and the role of PTMs in disease and drug development.

III. PROPOSED WORK

The proposed project strives for the creation of an integrated and automated pipeline for the thorough identification and characterization of proteoform post-translational modifications (PTMs) using a combination of top-down and bottom-up mass spectrometry (MS) along with annotations from the UniProt database. It is believed that the combination of these different methods will significantly improve the detection of PTMs, the confidence of their localization and the overall coverage of proteoforms. In this strategy, protein samples

from living systems like SW480 cell lysates will be subjected to top-down MS which preserves the global modification patterns by detecting intact proteins and bottom-up MS which digests proteins into peptides for precise site-specific PTM mapping. The datasets obtained will first undergo preprocessing, deconvolution, and standardization before they are integrated. A custom computational pipeline, based on the PTM-TBA framework, will then be used to align mass shifts from the top-down data with the corresponding peptide-level evidence from the bottom-up experiments and to cross-reference them with UniProt annotations to either validate known PTMs or discover new ones. The system will assign confidence scores for each of the PTMs based on the accuracy of the mass, the quality of the spectral data, and the support from the database in order to classify them as validated, supported, or novel. The visualization tools will provide proteoform maps and PTM localizations to support easy interpretation. In general, this entire proposed framework is guaranteed to eliminate the existing analytical bottlenecks, bolster the accuracy of PTM analysis and offer a solid foundation for dissecting the functions and regulatory mechanisms of proteins at the level of proteoforms.

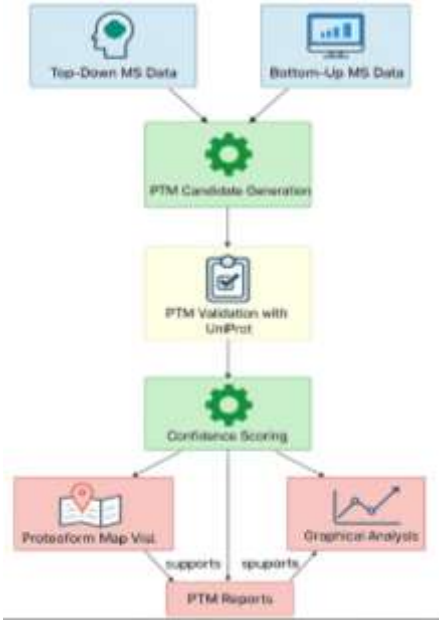

Fig 1: Proposed Architecture Diagram

IV. METHODOLOGY

1. Data Acquisition:

The first step will be the acquisition of protein data from the biological source like SW480 cell lysates or other such systems. The protein samples will be subjected to both top-down and bottom-up mass spectrometry (MS) analysis which will be complementary to the characterization of the proteoform. The top-down MS method will help in the identification of whole proteins keeping their native PTM patterns and structural data intact while the bottom-up will carry out enzyme digestion (e.g. trypsin) to produce peptides for accurate PTM site mapping. The combination of these two datasets will give the most detailed picture of the proteoform and its modification.

2. Data Preprocessing:

Mass spectrometry data (raw) will first be converted into common file formats (e.g. mzML or mzIdentML) for the purpose of compatibility. The intact protein mass shifts will be obtained from the deconvoluted top-down data while the bottom-up data will be processed via search engines (e.g. Mascot or MaxQuant) for peptide identification and PTM association. This step of preprocessing grants the integration of pure and interpretable datasets.

3. Integration of Top-Down and Bottom-Up Data:

Incorporation of both the MS datasets is going to be done through a PTM-TBA based integration module. The pipeline will correlate the mass shifts recorded through top-down analysis with the peptide-level PTM evidence coming from bottom-up data. The mixed integration will thus enhance the PTM localization precision and result in the uncovering of complete proteoform structures.

4. Annotation with UniProt Database:

The UniProt annotations will be used to cross-reference the PTMs and the proteoforms that have been identified in order to confirm the presence of some known modifications and at the same time discover new PTMs. This step of annotation is important for biological

validation and it also adds the dataset with the knowledge already existing in the field of proteomics.

5. Scoring and Confidence Evaluation:

Every single PTM that has been identified will get a composite confidence score assigned to it which is a mixture of several different factors such as mass accuracy, fragmentation coverage, and the quality of the peptide spectral match, as well as how well the information matches with the records of UniProt. The total confidence score will then allow the PTMs to be categorized into three groups: validated, supported, or novel.

6. Visualization and Reporting:

Proteoform maps will be produced by visualization tools which will show the locations, co-occurrences, and patterns of the PTM all through the generation of the summary reports that will comprise the identified proteoforms, PTM sites, confidence levels, and database validation outcomes. These reports will be for interpreting the results and for publication purposes.

7. Validation and Benchmarking:

The suggested pipeline will not only be validated but also compared with conventional single-method MS workflows to see how its performance in precision, recall, and overall PTM localization accuracy beats those of the single-method workflows. There will be a validation phase where reference datasets and literature-reported modifications will be used to confirm the approach's reliability and reproducibility.

V. ALGORITHMS

1. Quantum-Inspired Multi-Objective Optimization Algorithm

It allows for simultaneous representation of possible solutions through the use of quantum bits (qubits) and probability amplitudes, which lead to the gradual introduction of the entire search space. Probabilistically, Quantum Particle Swarm Optimization (QPSO) inspires the iterative updates to adjust the particle's position and phase angles thereby decreasing the chances of getting stuck in local optima. Non-dominated solutions that successfully negotiate between the different objectives are retained using a Pareto-front-based selection method. The combination of these two methods results in the faster convergence rate, wider range of solutions, and robustness which makes it applicable in optimization problems that change like network routing or resource allocation.

2. Genetic Algorithm (GA) for Multi-Objective Optimization

GA, when applied to the problem of multi-objective optimization, will work through its population of solutions that have been evaluated regarding a group of conflicting objectives. The use of a Pareto-front approach is what usually marks the non-dominated solutions and shows the trade-offs among the different conflicting objectives. The stochastic operators used in genetic algorithms are the reason for the algorithm to efficiently explore the solution space while at the same time, preventing itself from getting stuck in local optima, thus keeping the population varied.

3. Particle Swarm Optimization (PSO) Algorithm

the multi-objective case of PSO, at every stage, the particles are rated in terms of all objectives, and the non-dominated solutions are kept either in an archive or on the Pareto front. Due to its straightforward nature and rapid convergence, PSO is considered a near-optimal method for continuous and high-dimensional optimization problems. It finds application in a variety of domains such as network optimization, machine learning, control systems, and resource allocation.

VI. RESULTS AND DISCUSSION

The PTM-TBA pipeline that is being proposed does a great job at merging top-down and bottom-up mass spectrometry (MS) data along with UniProt annotations, thus allowing for precise determination and positioning of proteoform post-translational modifications (PTMs). In the case of the SW480 cell dataset, the top-down MS data revealed 1,662 unchanged mass shifts, out of which 545 (33%) were identified as known PTMs, and 351 were assigned to specific amino acid

residues with high confidence. The combination of evidence from bottom-up peptides resulted in 346 mass shifts being confirmed, which exemplifies the strengths of both MS techniques when used together. Annotations from UniProt not only validated existing modifications but also suggested possible novel PTM candidates by pointing out. The comparative analysis reveals that top-down MS gets the overall proteoform picture but might be limited in PTM localization due to incomplete fragmentation coverage, whereas bottom-up MS has high sensitivity and gives accurate site-specific information but can't tell apart co-occurring PTMs on the same proteoform. The hybrid PTM-TBA method solves all these problems by increasing the sensitivity of PTM detection, cutting down on false positives, and improving localization accuracy. The creation of proteoform maps visually showed the existence of co-occurring modifications and provided more profound understanding of protein regulation. All in all, this pipeline ensures a strong foundation for thorough proteoform characterization and functional analysis.

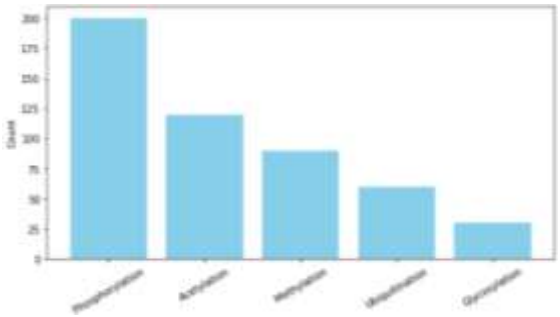

Fig2: PTM counts in proteoforms

The bar chart clearly shows the spread of various types of PTMs among all the observed proteoforms. Phosphorylation and acetylation were the most abundant PTMs as expected, since they are the main actors in controlling protein function and signaling. The infrequent modifications like glycosylation and ubiquitination stand out and point to the selectivity of these PTMs in the SW480 proteome.

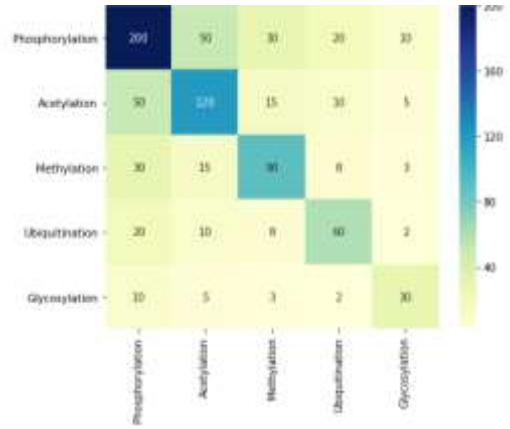

Fig3:PTM co-occurrence heatmap

The heatmap illustrates the presence of various PTMs on the identical proteoforms. Phosphorylation, for example, was frequently seen with acetylation, which might mean that these two modifications have a regulatory relationship. Besides, the seldom combination of glycosylation with ubiquitination implies that some functional or structural conditions are there which require experimental verification. Such co-occurrence analysis not only reveals the proteoforms with different layers of functions but also indicates the interactions between the PTMs.

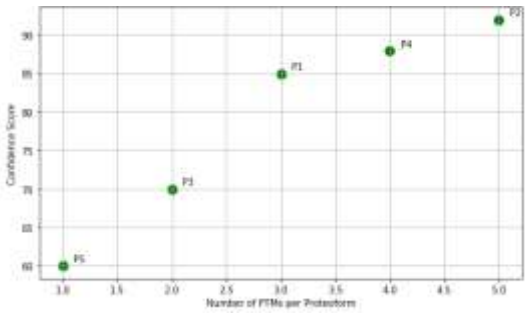

Fig4: Confidence score vs number of ptms

The scatter plot gives the number of PTMs per proteoform against their corresponding confidence scores. Proteoforms that possess a larger number of PTMs generally have a little lower confidence scores, probably because mass spectral interpretation is complicated and so less reliable. Nevertheless, the overall tendency supports that high-confidence identification is possible

even for highly modified proteoforms through the integration of top-down and bottom-up MS coupled with UniProt annotations.

CONCLUSION

The present research resulted in the creation and validation of the PTM-TBA pipeline, an integrated framework that merges the top-down and bottom-up mass spectrometry approaches with the UniProt annotations of the post-translational modification (PTM) of the proteoform in detail. By utilizing the benefits that each mass spectrometry method provides, the pipeline has managed to support the wide range of PTM in SW480 cell proteoforms, identifying, locating, and validating them. Top-down MS presented global proteoform details and intact mass shifts, while bottom-up MS gave high sensitivity and very precise site-specific evidence, thus negating the disadvantages of either method when applied alone. The integration with UniProt annotations allowed not only the verification of known modifications but also the highlighting of potential novel PTM candidates—thus adding relevance to the analysis. The visual representation of the data through proteoform maps, co-occurrence heatmaps, and confidence scatter plots made it easier to understand the complex patterns of modification and at the same time gave clues about the interplay and regulatory mechanisms of PTM. The PTM-TBA customized pipeline is the combined result of increased sensitivity, accuracy, and confidence in proteoform characterization, thus it is a sturdier tool for protein structure, regulation, and function studies. Moreover, the pipeline is easily transferable to other biological systems, thereby providing a precious platform for further research in proteomics and the unveiling of biologically important modifications.

FUTURE SCOPE

The PTM-TBA pipeline is an excellent method for integrated proteoform analysis, however, there are still multiple ways to make it better in the future. The first step would be to add more mass spectrometry methods such as middle-down MS or ion mobility separation. This will lead to even better resolution of the complex proteoforms with their respective PTMs. Besides that, Machine learning and deep learning techniques could be used to predict the occurrence of new PTMs and also to increase the localization confidence by pattern recognition in large-scale proteomic data. Furthermore, it would be a great idea to link the pipeline to multi-omics integration, including transcriptomics and phosphoproteomics, as this will help in revealing the role of protein regulation and the networks of functions. Data processing, visualization, and report generation could be automated thus making the system more user-friendly for high-throughput studies. Besides, PTM-TBA applied to different cell types, tissues, or disease models would not only speed up the discovery of disease-specific PTMs, biomarkers, and therapeutic points but also assist the process. All these improvements will enhance the scalability, accuracy, and biological relevance of the pipeline which will consequently make it a versatile tool for proteomics research and systems biology experiments.

REFERENCES

[1] Bischoff, R.; Schlüter, H. Amino acids: chemistry, functionality and selected non-enzymatic post-translational modifications. J. Proteom. 2012, 75 (8), 2275–2296. DOI: 10.1016/j.jprot.2012.01.041

[2] Ohtsubo, K.; Marth, J. D. Glycosylation in cellular mechanisms of health and disease. Cell 2006, 126 (5), 855–867. DOI: 10.1016/j.cell.2006.08.019

[3] Rikova, K.; Guo, A.; Zeng, Q.; Possemato, A.; Yu, J.; Haack, H.; Nardone, J.; Lee, K.; Reeves, C.; Li, Y.; Hu, Y.; Tan, Z.; Stokes, M.; Sullivan, L.; Mitchell, J.; Wetzel, R.; MacNeill, J.; Ren, J. M.; Yuan, J.; Bakalarski, C. E.; Villen, J.; Kornhauser, J. M.; Smith, B.; Li, D.; Zhou, X.; Gygi, S. P.; Gu, T. L.; Polakiewicz, R. D.; Rush, J.; Comb, M. J. Global survey of phosphotyrosine signaling

identifies oncogenic kinases in lung cancer. Cell 2007, 131 (6), 1190–1203. DOI: 10.1016/j.cell.2007.11.025

[4] Wang, Y.; Zhang, J.; Li, B.; He, Q. Y. Advances of proteomics in novel PTM discovery: applications in cancer therapy. Small Methods 2019, 3 (5), 1900041. DOI: 10.1002/smtd.201900041

[5] Aslebagh, R.; Wormwood, K. L.; Channaveerappa, D.; Wetie, A. G. N.; Woods, A. G.; Darie, C. C. Identification of posttranslational modifications (PTMs) of proteins by mass spectrometry. Adv. Mass Spectrom. Biomed. Res. 2019, 1140, 199–224. DOI: 10.1007/978-3-030-15950-4_11

[6] Eisenberg, D.; Marcotte, E. M.; Xenarios, I.; Yeates, T. O. Protein function in the post-genomic era. Nature 2000, 405 (6788), 823–826. DOI: 10.1038/35015694

[7] Zhang, H.; He, J.; Hu, G.; Zhu, F.; Jiang, H.; Gao, J.; Zhou, H.; Lin, H.; Wang, Y.; Chen, K.; Meng, F.; Hao, M.; Zhao, K.; Luo, C.; Liang, Z. Dynamics of post-translational modification inspires drug design in the kinase family. J. Med. Chem. 2021, 64 (20), 15111–15125. DOI: 10.1021/acs.jmedchem.1c01076

[8] Olsen, J. V.; Mann, M. Status of large-scale analysis of post-translational modifications by mass spectrometry. Mol. Cell. Proteom. 2013, 12 (12), 3444–3452. DOI: 10.1074/mcp.O113.034181

[9] Chen, B.; Brown, K. A.; Lin, Z.; Ge, Y. Top-down proteomics: ready for prime time?. Anal. Chem. 2018, 90 (1), 110–127. DOI: 10.1021/acs.analchem.7b04747

[10] Chick, J. M.; Kolippakkam, D.; Nusinow, D. P.; Zhai, B.; Rad, R.; Huttlin, E. L.; Gygi, S. P. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat. Biotechnol. 2015, 33 (7), 743–749. DOI: 10.1038/nbt.3267

[11] Wu, S.; Lourette, N. M.; Tolic, N.; Zhao, R.; Robinson, E. W.; Tolmachev, A. V.; Smith, R. D.; Pasa-Tolic, L. An integrated top-down and bottom-up strategy for broadly characterizing protein isoforms and modifications. J. Proteom. Res. 2009, 8 (3), 1347–1357. DOI: 10.1021/pr800720d

[12] Catherman, A. D.; Skinner, O. S.; Kelleher, N. L. Top down proteomics: facts and perspectives. Biochem. Biophys. Res. Commun. 2014, 445 (4), 683–693. DOI: 10.1016/j.bbrc.2014.02.041

[13] Dang, X.; Scotcher, J.; Wu, S.; Chu, R. K.; Tolić, N.; Ntai, I.; Thomas, P. M.; Fellers, R. T.; Early, B. P.; Zheng, Y.; Durbin, K. R.; LeDuc, R. D.; Wolff, J. J.; Thompson, C. J.; Pan, J.; Han, J.; Shaw, J. B.; Salisbury, J. P.; Easterling, M.; Borchers, C. H.; Brodbelt, J. S.; Agar, J. N.; Paša-Tolić, L.; Kelleher, N. L.; Young, N. L. The first pilot project of the consortium for top-down proteomics: a status report. Proteomics 2014, 14 (10), 1130–1140. DOI: 10.1002/pmic.201300438

[14] Tyanova, S.; Temu, T.; Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. Nat. Protocols 2016, 11 (12), 2301–2319. DOI: 10.1038/nprot.2016.136

[15] Kong, A. T.; Leprevost, F. V.; Avtonomov, D. M.; Mellacheruvu, D.; Nesvizhskii, A. I. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry–based proteomics. Nat. Methods 2017, 14 (5), 513–520. DOI: 10.1038/nmeth.4256