doi 10.18699/bgrs2024-11.3-09

Functional cross-talks between transcription, alternative splicing, translation and novel post-translational modifications in Alzheimer disease: fill missings using proteomics

Tarasova I.[1]*, Brazhnikov M.[1, 2], Kopeykina A.[1, 3], Emekeeva D.[1, 4]

[1] *V.L. Talrose Institute for Energy Problems of Chemical Physics, N.N. Semenov Federal Research Center of Chemical Physics, RAS, Moscow, Russia*

[2] *Osipyan Institute of Solid State Physics, RAS, Chernogolovka, Russia*

[3] *Pirogov Russian National Research Medical University, Moscow, Russia*

[4] *Moscow Institute of Physics and Technology (State University), Dolgoprudny, Russia*

* *iatarasova@yandex.ru*

Key words: Alzheimer disease; post translational modifications; amino acid substitutions; alternative splicing; molecular mechanisms

Motivation and Aim: Alzheimer's disease (AD) is the most common dementia in clinical practice. According to WHO, the number of AD cases worldwide was 55 million in 2022, and the forecasts promise an increase by 10 million every year (https://www.who.int/news-room/fact-sheets/detail/dementia). Despite the research community contributing significant efforts in studying pathogenesis, understanding the molecular and structural changes in the cells remains incomplete. For example, alternative splicing (AS) represents the powerful mechanism of regulating protein function. However, the diversity of alternatively spliced proteins and their role in pathogenesis of Alzheimer disease are not investigated. Detection of AS at protein level is a challenging task, consequently, the contribution of alternatively spliced proteins to the cell life cycle is basically unknown and represents a matter of contradictory statements from no impact on proteome diversity to evidences that most frame-preserving alternatively spliced isoforms are translated. There are three major factors complicating measurement of alternative splicing by standard methods of mass spectrometry based proteomics: (1) low protein sequence coverage (30 % in average); (2) low proteome coverage (products from the best case of 1/2 to the regular 1/6 of human protein coding genes); (3) approximately 1/2 of known isoforms can be identified by the only one tryptic peptide from the alternatively spliced part of protein sequence. Next level of complexity is identification and characterization of post translational modifications (PTM). Most approaches assume characterization of a targeted PTM enriched by a specific experimental protocol. However, post translational modifications trigger other PTMs and together they provide the regulation of protein function. Investigation of such events requires simultaneous measurements of different types of PTMs. We argue that big proteomics data analysis is of high potential to make a first step to PTM interplay discoveries as well as for discovery of new PTMs affecting protein functions in pathology development. Our motivation behind the study is to develop mass spectrometry based bioinformatic approaches for discovery of the alternative splicing and amino acid substitutions translated into proteins in Alzheimer's disease as well as novel post translational modifications including PTMs of the alternatively spliced regions and characterization of their functional implications.

Methods and Algorithms: Public proteomics data was used in the study [1]. Proteomics data collected by the Mount Sinai Brain Bank (MSBB) (https://doi.org/10.7303/syn5759470) and the Banner Sun Health Research Institute (Banner) (https://doi.org/10.7303/syn7170616) were downloaded from AD Knowledge Portal at http://www.synapse.org within the controlled-access data use certificate. In total, the *post mortem* brain tissues (prefrontal cortex) from 300 AD patients and 200 healthy individuals were analyzed. Protein identification was performed using different strategies and their combinations: database search (IdentiPy [2], MSFragger [3]) within the strict and loose mass tolerances [4], and database search with MSFragger search engine supplemented by *de novo* peptide sequence assembling using Novor software [5]. Database search was performed against the canonical and isoform UniProt TrEMBL protein database (taxon ID: 9604) concatenated with a decoy database created by reversing the protein sequences from the UniProt TrEMBL database. Post search validation was performed using Scavager software [6]. Positions of post translational modifications were established using AA_stat [4]. Differential expression of splice protein isoforms was assessed based on the normalized intensities of peptide ions in the mass spectra reported by TMTcrunch software developed in this study and calculation of standard mean deviations (SMD). Statistical analysis was performed using SciPy [7] and program code developed in house. Gene ontology enrichment analysis was performed using STRING [8]. To validate variant peptides, the chromatographic retention time (RT) prediction with deep machine learning was used. Peptides with ΔRT value within $\pm\sigma$ of the median value were considered true. Pathogenicity of the identified amino acid polymorphisms was evaluated using VarSome [9]. UniProt and InterPro databases were used to match alternative splicing events and post translational modifications falling into the protein functional domains and binding sites.
Results: Three pipelines for identification of alternatively spliced proteins, amino acid substitutions and post translational modifications were developed. First pipeline allowed us to report translation of approximately 1,300 splice isoforms presented in the SwissProt database, unambiguously identified by isoform-unique peptide features. We observed translation of alternative splicing affecting functional domains and binding sites for ASPH, ATP2B2, DLGAP1, DNM1L, MAPT and other AD-related protein coding genes. Second pipeline implements identification of tryptic peptides by the *de novo* sequencing from mass spectrometry data, database search against the SwissProt canonical database, and comparison of peptides assembled *de novo* and peptides identified in database search to find out the peptide sequences differing by amino acid substitution, insertion or deletion. The pipeline provides a competitive search between variant and canonical peptides and produces tables of reliable peptides with amino acid substitutions, their positions in the protein, the corresponding canonical peptides and protein names. Using the developed workflows we identified a representative group of amino acid substitutions annotated in the Unimod database as misacylation of tRNA. Such events result in tRNA-dependent mistranslation that is earlier reported in microorganisms and human cells [10]. In our results, we observed tRNA mistranslation for the dihydropyrimidinase-like (DPYSL) proteins, the Na,K-ATPases ATP1A1/2/3, calcium/calmodulin-dependent kinase 2 and other proteins involved in neuro-developmental disorders. Our next findings were amino acid substitutions corresponding to single nucleotide polymorphisms that were predicted and/or annotated as pathogenic. In particular, we found proteomics evidences for a number of known pathogenic substitutions: HBB Q128E, GFAP N77D, NEFL N98D, etc. Using the developed

workflows we characterized the landscape of post translational modifications in precortex tissues of AD patients and healthy individuals. It was represented by deamidation of asparagine, glutamine and arginine, dehydrogenation, acetylation, phosphorylation, oxidation, glycosylation, methylation and many others. In total we identified few dozen of proteins with statistically significant changes in the level of post translationally modified proteins in Alzheimer's disease relative to the control group. Cross-validation using different bioinformatic approaches and comparison with deamidation and citrullination annotated in UniProt and dbPTM allowed us to propose three dozen of new deamidation sites enriched in AD and located on N, Q and R in protein products of MYO5A, ATP1A3, GATD3B, NEFM, and TKT genes. In analysis of amino acid motifs, the *N*G motif was enriched in peptides with modified asparagine.
Conclusion: The developed bioinformatic pipelines for processing mass spectrometry data represent powerful tools to follow genomic events translated to proteins. Using those workflows, we observed numerous molecular events pretending for reproducibility and novelty, and their analysis is to be continued. Our efforts are further focused on data integration, interpretation and systematization to reconstruct the molecular mechanisms.
Funding: The study is supported by the Russian Science Foundation (No. 23-45-00012).

References

1. Bereczki E., Branca R.M., Francis P.T. et al. Synaptic markers of cognitive decline in neurodegenerative diseases: a proteomic approach. *Brain.* 2018;141(2):582-595. doi 10.1093/brain/awx352
2. Levitsky L.I., Ivanov M.V., Lobas A.A. et al. IdentiPy: an extensible search engine for protein identification in shotgun proteomics. *J Proteome Res.* 2018;17(7):2249-2255. doi 10.1021/acs.jproteome.7b00640
3. 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
4. Levitsky L.I., Bubis J.A., Gorshkov M.V., Tarasova I.A. AA_stat: Intelligent profiling of *in vivo* and *in vitro* modifications from open search results. *J Proteomics.* 2021;248:104350. doi 10.1016/j.jprot.2021.104350
5. Ma B. Novor: Real-time peptide *de novo* sequencing software. *J Am Soc Mass Spectrom.* 2015;26(11):1885-1894. doi 10.1007/s13361-015-1204-0
6. Ivanov M.V., Levitsky L.I., Bubis J.A., Gorshkov M.V. Scavager: a versatile postsearch validation algorithm for shotgun proteomics based on gradient boosting. *Proteomics.* 2019;19(3):e1800280. doi 10.1002/pmic.201800280
7. Virtanen P., Gommers R., Oliphant T.E. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods.* 2020;17(3):261-272. doi 10.1038/s41592-019-0686-2
8. Szklarczyk D., Gable A.L., Nastou K.C. et al. The STRING database in 2021: customizable protein – protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* 2021;49(D1):D605-D612. doi 10.1093/nar/gkaa1074
9. Kopanos C., Tsiolkas V., Kouris A., Chapple C.E., Albarca Aguilera M., Meyer R., Massouras A. VarSome: the human genomic variant search engine. *Bioinformatics.* 2019;35(11):1978-1980. doi 10.1093/bioinformatics/bty897
10. Lant J.T., Berg M.D., Sze D.H.W. et al. Visualizing tRNA-dependent mistranslation in human cells. *RNA Biol.* 2018;15(4-5):567-575. doi 10.1080/15476286.2017.1379645