DOI 10.18699/BGRSSB-2018-091

# The size of the Human Proteome: how many protein species are detectable today?

E. Ponomarenko*, A. Kopylov, V. Zgoda, E. Poverennaya, E. Ilgisonis, A. Lisitsa, S. Naryzhny, E. Petrenko, S. Radko, A. Archakov
*Institute of Biomedical Chemistry, Moscow, Russia*
** e-mail: 2463731@gmail.com*

**Key words:** Human Proteome Project, proteoforms, proteome size, human chromosome 18

*Motivation and Aim*: The size of the human proteome, which is determined by the number of protein species («*width*») and the number of copies of an individual proteoform in a biosample («*depth*»), is still unknown [1, 2]. Given the limitations of sensitivity of analytical methods and the absence of amplification reaction in proteomics, it remains a challenge to define the part of the proteome that can be observed experimentally.
*Methods and Algorithms*: Here, meta-analysis of neXtProt knowledge base is proposed for theoretical prediction of the number of different proteoforms arising from alternative splicing, single amino acid polymorphisms and post-translational modifications. Experimental part was performed using targeted MS for chromosome 18 encoded proteins, along with the estimation of copy numbers in plasma, liver, and HepG2 cell line. The proposed approaches for estimation of proteome *width* and *depth* were validated using UPS1 and UPS2 protein calibration standards provided by Sigma Aldrich and consisting of 48 proteins present in same (UPS1) and different (UPS2) concentration.
*Results*: A range of 0.55-7.14 millions of protein species (proteoforms) in the human body was estimated using different methods of calculation based on the average number of variations per gene from neXtProt. In particular, 275 protein-coding genes predicted for human chromosome 18 could potentially encode about 8 to 18 thousand of proteoforms. In total, proteins were detected and measured for only approximately 30 % of the predicted protein-coding genes in selected types of biomaterials. When using UPS1 and UPS2 standards we found that shotgun LC-MS/MS analysis allows to identify only a half of proteins presented in the sample in a pure solution. In comparison, there was just four proteins we were not able to detect by targeted MS method. The number of undetected proteins increases when we add a matrix, such as human blood plasma or E-coli protein.
*Conclusion*: Taking chromosome 18 as an example, the size of the human proteome was predicted based on NeXtProt data. We found that biological matrix significantly affects the list of detected proteins and obviously affects proteome *width* and *depth*: MS-signals from presenting proteins may be lost or new MS-signals may appear resulting in false-positive results.
*Acknowledgements*: El. P., Ek. P. and Ek. I. acknowledge the Leading Scientific School of Prof. Andrey Lisitsa (No. NSh6313.2018.4).

*References*

1. Ponomarenko E., Poverennaya E., Ilgisonis E. et al. (2016) The Size of the Human Proteome: The Width and Depth. Int J Anal Chem. 2016:7436849.
2. Aebersold R., Agar J.N., Amster I.J. et al. (2018) How many human proteoforms are there? Nat Chem Biol. Feb 14;14(3):206-214.