Send Orders for Reprints to reprints@benthamscience.net

Current Genomics, 2020, *21*, 531-535

REVIEW ARTICLE

Hypothetical Proteins as Predecessors of Long Non-coding RNAs

Girik Malik[1,2,3], Tanu Agarwal[4], Utkarsh Raj[4], Vijayaraghava Seshadri Sundararajan[2], Obul Reddy Bandapalli[5,6,7,*] and Prashanth Suravajhala[2,8,*]

[1]Khoury College of Computer Sciences, Northeastern University, 360 Huntington Ave., Boston, MA 02115, USA; [2]Bioclues.org, Kukatpally, Hyderabad, 500072, India; [3]Labrynthe Pvt. Ltd., New Delhi, India; [4]NIIT University, NH8, Delhi- Jaipur Highway, District Alwar, Neemrana, Rajasthan 301705, India; [5]Hopp Children's Cancer Center [KiTZ], Heidelberg, Germany; [6]Division of Pediatric Neuro Oncology, German Cancer Research Center [DKFZ], German Cancer Consortium [DKTK], Heidelberg, Germany; [7]Heidelberg University, Medical Faculty, Heidelberg, Germany; [8]Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Statue Circle, Jaipur 302021, RJ, India

ARTICLE HISTORY

Received: March 09, 2020
Revised: April 28, 2020
Accepted: May 16, 2020

DOI:
10.2174/1389202921999200611155418

Abstract: Hypothetical Proteins [HP] are the transcripts predicted to be expressed in an organism, but no evidence of it exists in gene banks. On the other hand, long non-coding RNAs [lncRNAs] are the transcripts that might be present in the 5' UTR or intergenic regions of the genes whose lengths are above 200 bases. With the known unknown [KU] regions in the genomes rapidly existing in gene banks, there is a need to understand the role of open reading frames in the context of annotation. In this commentary, we emphasize that HPs could indeed be the predecessors of lncRNAs.

Keywords: Hypothetical proteins, lncRNA, aptamers, annotation, functional genomics, transcripts.

1. INTRODUCTION

There are known knowns [KK], known unknowns [KU] and unknown knowns [UK] in the genome, as aptly put by David Logan [1]. Fully sequenced genomes have the KUs that encode Hypothetical Proteins [HPs] or domains of unknown function [DUF]. These proteins are predicted through *in silico* approaches and their biological activities are not substantiated by experimental evidence, and hence referred to as HPs [2, 3]. Despite their lack of functional characterization, HPs play an important role in understanding biochemical and physiological pathways [4], for instance, to find new structures and functions [5], biomarkers and relevant targets [6], and early detection for proteomic and genomic research [2]. Importantly, structural and functional characterization of HPs revealed crucial roles in microorganisms, particularly in pathogens associated with human diseases [7, 8]. In the recent past, many robust *in silico* approaches were developed and these tools are publicly available to predict the putative function of the HPs [9]. As the HPs are implicated in a myriad of biological functions, it is important to study the Protein-Protein Interactions [PPIs] as this would allow us to see how many of them are the products of pseudogenes [10]. Further, the HP function could as well be attributed to motif similarity data and homology reference [11] With the advancement of Next-Generation Sequencing [NGS], there is an inherent need to further annotate, classify and curate HPs which will help in not only understanding their functions but also allow us to use them as biomarkers [12]. Across the whole genome, ca. 2% genome encodes for proteins, while the remaining are non-coding or still functionally unknown [13]. In our recent study, to examine the role of HPs and the long non-coding RNAs [lncRNA] in metabolic diseases and cancers, we deciphered the function of HPs using a nine-point classification-scoring schema [14]. Most HPs from GenBank lack protein-coding capacity; therefore, we argue that they could essentially be a part of non-coding RNA [ncRNAs] transcripts. We earlier proposed an interactome of HPs [Hypothome] to define a network of Protein-Protein Interactions [PPI] wherein a connotation to the interactomes for predicted proteins would devalue the integrity of the interactome [15]. In this commentary, we describe a gist of what 'hypothome' is all about and argue that they also serve as predecessors of lncRNAs. This HP protein annotation, which we describe, could largely be inferred for eukaryotes as lncRNAs are considered as subjects.

2. CHALLENGES

2.1. Essential HPs

We developed a database of HPs a.k.a Hypo2 [12] and a "quick search" enabled us to retrieve the list of HPs, with the

*Address correspondence to these authors at the Hopp Children's Cancer Center [KiTZ], Heidelberg, Germany; Division of Pediatric Neuro Oncology, German Cancer Research Center [DKFZ], German Cancer Consortium [DKTK], Heidelberg, Germany; and Heidelberg University, Medical Faculty, Heidelberg, Germany; Tel: +4962214241809;
E-mail: o.bandapalli@kitz-heidelberg.de and Bioclues.org, Hyderabad, Telangana, India; Department of Biotechnology and Bioinformatics, Birla Institute of Scientific Research, Statue Circle, Jaipur 302021, RJ, India;
Tel: +91-141-2385094; E-mail: prash@bisr.res.in

1389-2029/20 $65.00+.00

new annotation records from NCBI. Similarly, the searches using the keywords "Hypothetical protein" AND "*Homo sapiens*" at NCBI, using the Boolean logic, enabled us to filter 8653 accessions. After a careful validation in checking "M" or "L" [Methionine/Leucine] start sites through an in-house python script, the sequences were mapped to Uniprot Ids using the "ID mapping" tool wherein 4192 out of 6129 linked Gene Indices [GI] were successfully represented to 3212 Uniprot KB identities (Supplementary Table **1**: *labelled as* All proteins). Other 787 (Supplementary Table **1**: *labelled as* 787) were mapped to the UniParc sequence archive. Keeping an epilogue of the HPs turned out to be non-coding RNA or pseudogenes, we blasted our dataset of 3212 proteins with the human Noncode dataset [http://www.noncode.org last accessed, September 11, 2019] and found two hits with E value 0 (Supplementary Table **1**; *highlighted in yellow and labelled as*: Noncode). These two-noncoding RNAs eventually turned out to be pseudogenes [marked in yellow in the LncRNAs sheet]. From further search using filters set with keywords like 'ncRNA' and 'LINC' for the genes, we obtained 12 ncRNA (Supplementary Table **1**: *labelled as* LncRNAs) and lncRNA which were further validated using the Noncode [16] and Lncipedia databases [17]. A total of 809 matched interactions formed the primary basis of 'hypothome' and among them, 73 (Supplementary Table **1**:, *labelled as* 73) were available as Noncode Blast results, which were further used for downstream annotation to check for consolidated pathways and drug interactions. A representation of how an HP could prospectively be annotated as lncRNA is shown in Fig. (**1**). Further, a contextual hub analysis tool [CHAT] [18] network was created using these matched entries and in order to reach a consensus, the same protein interactions were visualized in Osprey (Fig. **2**) as well [19] as well. Interestingly, these were associated with diseases like Alzheimer's, metabolic pathways like Arginine and proline metabolism, DNA repair, *etc*. Thus, we believe that the HPs if annotated could be lncRNAs and they can be used as putative biomarkers for various diseases. Nevertheless, they also could be used for identifying specific antibodies/aptamers and novel targets for drugs with further validation, and finally could serve as 'essential HPs'.

2.2. Large-Scale Protein Interactions for Structural 'mer' Studies

Given the aforementioned reason for calling them as 'essential HPs", we believe, these protein interactions can be processed as graph structures. The HPs could be ideal candidates for representing graphs wherein one can easily find metrics like cliques, communities, small worlds, *etc*., that lead to an abstract level of understanding of interactions, drug delivery, *etc*. With growing numbers of pseudogenes, the hypothome problem could be scaled towards the existing algorithms in parallel using methods like MapReduce [20]. Although for the initial steps of using BLAST for finding sequence matches, there are other approaches exploiting the use of parallel infrastructure [21-23], however, these approaches are often underutilised because of the complex set-up, compared with the sequential processing, and are limited. Lately, there has been a sudden increase in tools and techniques for data analysis that has helped solve some interesting problems in the bioinformatics domain. Such a technique has also incited the interest of large community projects in using novel techniques for crunching data. Google DeepMind's AlphaFold [https://deepmind.com/blog/article/alphafold last accessed, December 16, 2019] is one such example, which outperformed other techniques at the completion of popular Critical Assessment of Structural Prediction [CASP]. Deep Learning, in general, is picking up rapidly because of the gargantuan amounts of data now accessible that has a bit of leeway of conceivably giving an answer for addressing the data analysis and learning issues found in enormous volumes of input data. Even more explicitly, it helps in automatically extracting complex data representations from a massive volume of unsupervised data. This makes it a profitable means for big data analytics, which encompasses data analysis from huge accumulations of raw data, which is usually unsupervised and un-categorized. With the increased number of HPs turning out to be pseudogenes, there is an accurate need to develop novel techniques of processing that would be immensely useful. Very recently, these techniques have not only started to make their way into functional prediction [24-26] but also proved to be useful for what is described by Logan DC [1], for finding KUs, given the other one, and their combinations. Such techniques are generally rate limited by the processing pipeline, wherein the difference between the amounts of data to be taken as input and given as output is far from insignificant. Eliminating this bottleneck, generally results in faster data processing and a better performance system, in general [27, 28].

2.3. A Case Study on Targeting Domains of Unknown Function

Moving forward to structures, the 3D swap database has a couple of DUFs, which we would like to consider as a case study to infer the role of aptamers for effective specificity. Assuming that the functions of DUFs and HPs can be better used as targets for diagnostics, the most common entity used are antibodies that could possibly circumvent the effect/targets. While the experimental characterization of antibodies is cumbersome, it is assumed that aptamer-protein prediction methods may serve as a benchmark besides providing cost-effective measures [29, 30]. What remains to be elucidated is whether the aptamer is bound? If bound, whether the 3D domains are swapped? If swapped, could it be applied for domains caused due to extensive multimerization as well? To answer this, we have analysed it and found that there is a dimer interface communicated to the catalytic domain of 2A9U. Therefore, we assume that the aptamers specific to this variable fragment could be used. With this, we expect that through the antigen-binding capacity of aptamer with the molecule, a wide number of HPs or DUFs can be targeted, which could be associated with diseases. Thus, we hope active conformation and aptamers as small molecules for therapies could prove to be very useful in the development of medical technology. Therefore, we hypothesize that the role of aptamers over antibody isotypes can be inferred and based on the affinity of aptamers bound to swapped domains particular to HPs or DUFs.

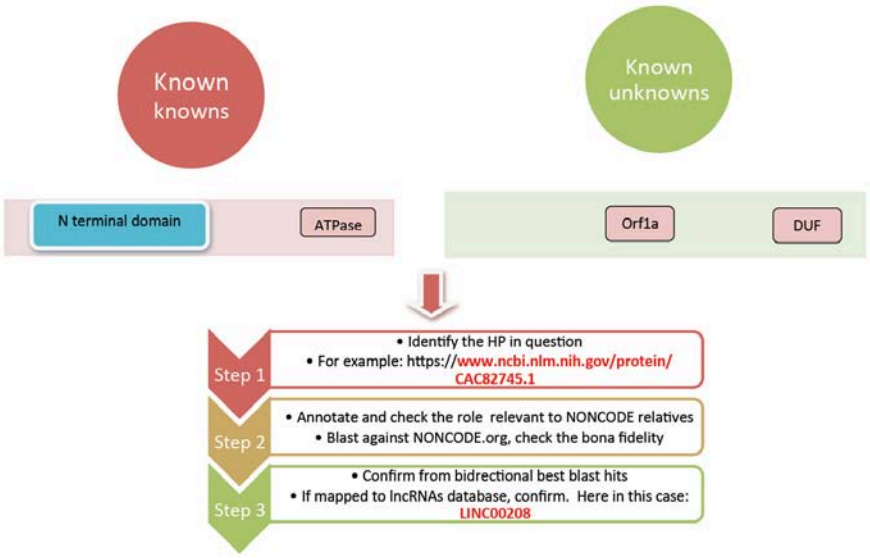

Fig. (1). The figure showing the difference between a known-known and known unknown protein. Precisely, the characteristic domains such as domains of unknown function (DUF) or ORFs unrelated or KIAA domains are associated with hypothetical proteins, which usually are present in the c terminal region of the protein. We show a classic example of how CAC92745, an HP, could be annotated as a lncRNA, *viz*. LINC00208. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

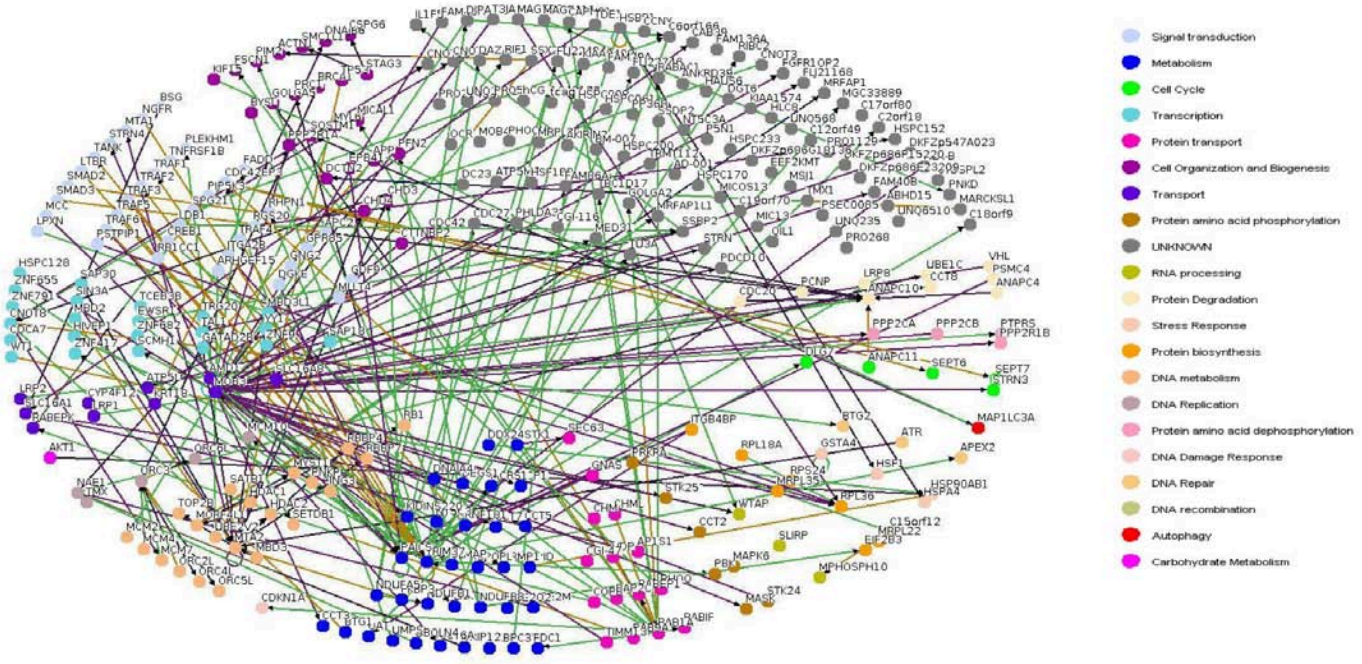

Fig. (2). Osprey visualization showing 314 nodes [genes] and 224 edges [experimental system] with each node colour indicating their GO processes, *viz.* protein transport [pink], cell cycle [green], stress response [radical red], metabolism [royal blue], transport [Egyptian blue], signal transduction [grey], DNA metabolism [tortilla], transcription [azure], biogenesis [purple], RNA processing [lime], protein degradation [white], protein amino acid phosphorylation [brown], DNA damage response [yellow], autophagy [red] and function unknown [black]. (*A higher resolution / colour version of this figure is available in the electronic copy of the article*).

CONCLUSION

Automated genome sequence analysis and annotation may provide ways to understand genomes, although with limited precision. Given the challenge of determining protein functions, bioinformatics algorithms have not only allowed us to predict near functions to these HPs but also provided us to benchmark these methods for developing efficient tools. From the experimentally determined partners or interologs [orthologous interacting partner pairs], in principle, it is possible to suggest a role for HPs in a biological context. In recent years, lncRNAs have emerged as key regulators of cellular processes, including transcription, splicing, translation,

DNA repair, and their role in regulation and relation with HPs have not only provided a big hope for characterising the KUs but also revealed new insights in modern biology. The experimental characterization of these and other 'conserved hypothetical' proteins is expected to reveal new directions, for example, understanding crucial aspects of microbial biology as in the case of HPs and, in addition, could also lead to the better functional prediction for characterizing medically relevant human homologs.

CONSENT FOR PUBLICATION

Not applicable.

FUNDING

None.

CONFLICT OF INTEREST

The authors declare no conflict of interest, financial or otherwise.

ACKNOWLEDGEMENTS

Declared none.

SUPPLEMENTARY MATERIAL

Supplementary material is available on the publisher's website along with the published article.

REFERENCES

[1] Logan, D.C. Known knowns, known unknowns, unknown unknowns and the propagation of scientific enquiry. *J. Exp. Bot.,* **2009**, *60*(3), 712-714. http://dx.doi.org/10.1093/jxb/erp043 PMID: 19269994

[2] Galperin, M.Y.; Nikolskaya, A.N.; Koonin, E.V. Novel domains of the prokaryotic two-component signal transduction systems. *FEMS Microbiol. Lett.,* **2001**, *203*(1), 11-21. http://dx.doi.org/10.1111/j.1574-6968.2001.tb10814.x

[3] Eisenstein, E.; Gilliland, G.L.; Herzberg, O.; Moult, J.; Orban, J.; Poljak, R.J.; Banerjei, L.; Richardson, D.; Howard, A.J. Biological function made crystal clear-annotation of hypothetical proteins *via* structural genomics. *Curr. Opin. Biotechnol.,* **2000**, *11*(1), 25-30. http://dx.doi.org/10.1016/S0958-1669(99)00063-4

[4] Sharma, M.; Vedithi, S.C.; Das, M.; Roy, A.; Ebenezer, M. Sequence homology and expression profile of genes associated with DNA repair pathways in *Mycobacterium leprae. Int. J. Mycobacteriol.,* **2017**, *6*(4), 365-378. http://dx.doi.org/10.4103/ijmy.ijmy_111_17 PMID: 29171451

[5] Nimrod, G.; Schushan, M.; Steinberg, D.M.; Ben-Tal, N. Detection of functionally important regions in "hypothetical proteins" of known structure. *Structure,* **2008**, *16*(12), 1755-1763. http://dx.doi.org/10.1016/j.str.2008.10.017 PMID: 19081051

[6] Shahbaaz, M.; Hassan, M.I.; Ahmad, F. Functional annotation of conserved hypothetical proteins from *Haemophilus influenzae* Rd KW20. *PLoS One,* **2013**, *8*(12), e84263. http://dx.doi.org/10.1371/journal.pone.0084263 PMID: 24391926

[7] Ansell, B.R.E.; Pope, B.J.; Georgeson, P.; Emery-Corbin, S.J.; Jex, A.R. Annotation of the Giardia proteome through structure-based homology and machine learning. *Gigascience,* **2019**, *8*(1), 8. http://dx.doi.org/10.1093/gigascience/giy150 PMID: 30520990

[8] Yang, Z.; Tsui, S.K. Functional annotation of proteins encoded by the minimal bacterial genome based on secondary structure element alignment. *J. Proteome Res.,* **2018**, *17*(7), 2511-2520. http://dx.doi.org/10.1021/acs.jproteome.8b00262 PMID: 29757649

[9] Murakami, M.; Nakagawa, M.; Olson, E.N.; Nakagawa, O. A WW domain protein TAZ is a critical coactivator for TBX5, a transcription factor implicated in Holt-Oram syndrome. *Proc. Natl. Acad. Sci. USA,* **2005**, *102*(50), 18034-18039. http://dx.doi.org/10.1073/pnas.0509109102 PMID: 16332960

[10] Shidhi, P.R.; Nair, A.S.; Suravajhala, P. Identifying pseudogenes from hypothetical proteins for making synthetic proteins. *Syst. Synth. Biol.,* **2014**, *8*(2), 169-171. http://dx.doi.org/10.1007/s11693-014-9148-4 PMID: 24799963

[11] Rehman, H.U.; Benso, A.; Di Carlo, S.; Politane, G.; Savino, A.; Suravajhala, P. Combining homolog and motif similarity data with Gene Ontology relationships for protein function prediction. **2012**. *IEEE International Conference on Bioinformatics and Biomedicine,* pp. 1-4. http://dx.doi.org/10.1109/BIBM.2012.6392719

[12] Sundararajan, V.S.; Malik, G.; Ijaq, J.; Kumar, A.; Das, P.S.; Shidhi, P.R.; Nair, A.S.; Dhar, P.K.; Suravajhala, P. Hypo: a database of human hypothetical proteins. *Protein Pept. Lett.,* **2018**, *25*(8), 799-803. http://dx.doi.org/10.2174/0929866525666180828110444 PMID: 30152276

[13] Comfort, N. Genetics: we are the 98%. *Nature,* **2015**, *520*(7549), 615. http://dx.doi.org/10.1038/520615a

[14] Ijaq, J.; Malik, G.; Kumar, A.; Das, P.S.; Meena, N.; Bethi, N.; Sundararajan, V.S.; Suravajhala, P. A model to predict the function of hypothetical proteins through a nine-point classification scoring schema. *BMC Bioinformatics,* **2019**, *20*(1), 14. http://dx.doi.org/10.1186/s12859-018-2554-y PMID: 30621574

[15] Desler, C.; Zambach, S.; Suravajhala, P.; Rasmussen, L.J. Introducing the hypothome: a way to integrate predicted proteins in interactomes. *Int. J. Bioinform. Res. Appl.,* **2014**, *10*(6), 647-652. http://dx.doi.org/10.1504/IJBRA.2014.065247 PMID: 25335568

[16] Liu, C; Bai, B; Skogerbø, G; Cai, L; Deng, W; Zhang, Y; Bu, D; Zhao, Y; Chen, R. NONCODE: an integrated knowledge database of non-coding RNA. *Nucleic Acids Res.,* **2005**, *33*, D112-5.

[17] Volders, PJ; Helsens, K; Wang, X; Menten, B; Martens, L; Gevaert, K; Vandesompele, J; Mestdagh, P LNCipedia: a database for annotated human lncRNA transcript sequences and structures. *Nucleic Acids Res.,* **2013**, *41*(D1), D246-D251. http://dx.doi.org/10.1093/nar/gks915

[18] Muetze, T.; Goenawan, I.H.; Wiencko, H.L.; Bernal-Llinares, M.; Bryan, K.; Lynn, D.J. Contextual Hub Analysis Tool (CHAT): A Cytoscape app for identifying contextually relevant hubs in biological networks. *F1000 Res.,* **2016**, *5*, 1745. http://dx.doi.org/10.12688/f1000research.9118.1 PMID: 27853512

[19] Hallen, M.A.; Martin, J.W.; Ojewole, A.; Jou, J.D.; Lowegard, A.U.; Frenkel, M.S.; Gainza, P.; Nisonoff, H.M.; Mukund, A.; Wang, S.; Holt, G.T.; Zhou, D.; Dowd, E.; Donald, B.R. OSPREY 3.0: Open-source protein redesign for you, with powerful new features. *J. Comput. Chem.,* **2018**, *39*(30), 2494-2507. http://dx.doi.org/10.1002/jcc.25522 PMID: 30368845

[20] Finocchi, I.; Finocchi, M.; Fusco, E.G. Clique counting in MapReduce: theory and experiments. *J. Exp. Algorithmics,* **2014**, *20*.

[21] de Castro, M.R.; Tostes, C.D.S.; Dávila, A.M.R.; Senger, H.; da Silva, F.A.B. SparkBLAST: scalable BLAST processing using in-memory operations. *BMC Bioinformatics,* **2017**, *18*(1), 318. http://dx.doi.org/10.1186/s12859-017-1723-8 PMID: 28655296

[22] Meng, Z.; Li, J.; Zhou, Y.; Liu, Q.; Liu, Y.; Cao, W. bCloudBLAST: An efficient mapreduce program for bioinformatics applications. Proceedings of *4th International Conference on Biomedical Engineering and Informatics [BMEI], Shanghai,* **2011**, pp. 2072-2076.

[23] Yang, X.; Liu, Y.; Yuan, C.; Huang, Y. Parallelization of BLAST with MapReduce for long sequence alignment. *2011 Fourth International Symposium on Parallel Architectures, Algorithms and Programming,* Tianjin, **2011**, pp. 241-246. http://dx.doi.org/10.1109/PAAP.2011.36

[24] Gao, R.; Wang, M.; Zhou, J.; Fu, Y.; Liang, M.; Guo, D.; Nie, J. Prediction of enzyme function based on three parallel deep CNN and amino acid mutation. *Int. J. Mol. Sci.,* **2019**, *20*(11), 2845. http://dx.doi.org/10.3390/ijms20112845 PMID: 31212665

[25] Sureyya Rifaioglu, A.; Doğan, T.; Jesus Martin, M.; Cetin-Atalay, R.; Atalay, V. DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks. *Sci. Rep.,* **2019**, *9*(1), 7344. http://dx.doi.org/10.1038/s41598-019-43708-3 PMID: 31089211

[26] Liu, X. Deep recurrent neural network for protein function prediction from sequence. *arXiv preprint. arXiv:1701.08318,* **2017**.

[27] Lavallée-Adam, M.; Park, S.K.; Martínez-Bartolomé, S.; He, L.; Yates, J.R., III From raw data to biological discoveries: a computational analysis pipeline for mass spectrometry-based proteomics. *J. Am. Soc. Mass Spectrom.,* **2015**, *26*(11), 1820-1826. http://dx.doi.org/10.1007/s13361-015-1161-7 PMID: 26002791

[28] Keller, A.; Shteynberg, D. Software pipeline and data analysis for MS/MS proteomics: the trans-proteomic pipeline. *Bioinformatics for Comparative Proteomics*; Humana Press, **2011**. pp. 169-189 http://dx.doi.org/10.1007/978-1-60761-977-2_12

[29] Song, K-M.; Lee, S.; Ban, C. Aptamers and their biological applications. *Sensors (Basel),* **2012**, *12*(1), 612-631. http://dx.doi.org/10.3390/s120100612 PMID: 22368488

[30] Suravajhala, P; Burri, HVR; Heiskanen, a combining aptamers and *in silico* interaction studies to decipher the function of hypothetical proteins **2014**, *3*(8), 809-810.