THE HUMAN PROTEOME SIZE AS A TECHNOLOGICAL DEVELOPMENT FUNCTION

E.V. Sarygina*, A.S. Kozlova, E.A. Ponomarenko, E.V. Ilgisonis

Institute of Biomedical Chemistry,
10 Pogodinskaya str., Moscow, 119121 Russia; *e-mail: lizalesa@gmail.com

Changes in information on the number of human proteoforms, post-translational modification (PTM) events, alternative splicing (AS), single-amino acid polymorphisms (SAP) associated with protein-coding genes in the neXtProt database have been retrospectively analyzed. In 2016, our group proposed three mathematical models for predicting the number of different proteins (proteoforms) in the human proteome. Eight years later, we compared the original data of the information resources and their contribution to the prediction results, correlating the differences with new approaches to experimental and bioinformatic analysis of protein modifications. The aim of this work is to update information on the status of records in the databases of identified proteoforms since 2016, as well as to identify trends in changes in the quantities of these records. According to various information models, modern experimental methods may identify from 5 to 125 million different proteoforms: the proteins formed due to alternative splicing, the implementation of single nucleotide polymorphisms at the proteomic level, and post-translational modifications in various combinations. This result reflects an increase in the size of the human proteome by 20 or more times over the past 8 years.

Key words: proteomics; proteoforms; post-translational modifications; single-amino acid substitutions; alternative splicing; neXtProt

DOI: 10.18097/PBMC20247005364

INTRODUCTION

The actual number of different types of proteins significantly exceeds the number of genes encoding them [1]. Protein diversity arises due to various reasons: alternative splicing (AS), single nucleotide substitutions in the genome, realized at the proteomic level in the form of single-amino acid polymorphisms (SAP), as well as post-translational modifications (PTM) of proteins [1].

The diversity of proteins encoded by one gene is called by the authors as “protein forms”, “protein isoforms” [2], “modified variants” [3], “protein types” [4]. In 2013, a group led by Professor Kelleher introduced the term “proteoforms” [5] for a group of proteins encoded by one gene. This term designates a set of protein forms that arise due to modifications at the genomic, transcriptomic, and proteomic levels, and also due to endogenous proteolysis [5]. The human proteome is a collection of all proteins in tissues and organs, and therefore the diversity of proteoforms can be referred to as the “width” of the proteome and calculated using models that take into consideration the frequencies of modifications at the genomic, transcriptomic, and proteomic levels [6].

Analysis of the proteome of biological objects is hampered by a wide dynamic range of analyte concentrations [7], due to which the mass spectrometric signal from a specific protein can be lost among the “noise” from other molecules, as well as the limited sensitivity of the analytical method [8]: the higher the sensitivity of the method, the greater the number of proteins (both canonical and proteoforms) that can be determined experimentally. Existing methodological solutions do not allow for experimental analysis of the full spectrum of proteoforms in a high-throughput mode [9, 10]. It has been experimentally shown that the limited analytical sensitivity of modern mass spectrometry is a limiting factor in the use of proteomic technologies in medicine.

In 2016, our group proposed three mathematical models for predicting the possible number of human proteoforms [11]. The use of these models requires information on the number of protein-coding genes for which the encoded proteoform has been experimentally shown, and the total number of splice variants, SAPs, and PTMs. In the 2016 paper, predictions were based on data from the neXtProt version (ver. 2015_06) and predicted the existence of 0.62, 0.88, or 6.13 million human protein species (excluding combinatorial variants, [11]). In the present study, we have updated these estimates taking into account the update of information in the neXtProt database and the development of experimental and bioinformatic approaches for protein detection.

A similar bioinformatics estimate was made by Aebersold et al. [1] in 2018. The researchers estimated the theoretically possible number of proteoforms as the product of the probabilities of the occurrence of all existing (supported by data with experimental confirmation in information resources) PTMs in each amino acid residue of the protein-coding gene.

Taking into account only binary modifications, the number of theoretical proteoforms was estimated as astronomically large (1×10^{27}). At the same time, the diversity of proteoforms that we actually observe in biological systems, is significantly lower than theoretically predicted due to the limited capabilities of existing detection technologies. In this regard, to assess the diversity of proteoforms, it is advisable to use information resources that consolidate the results of experimental studies of the proteome by various scientific groups around the world. The most complete resource on the human proteome is the neXtProt system [12].

neXtProt is an information platform created in 2011 to consolidate data obtained in the study of the human proteome [12]. The basis for neXtProt is provided by data in the UniProtKB/Swiss-Prot (Reviewed) records for *Homo sapiens* (TaxID: 9606) with the keyword "Complete proteome" (KW-0181). The resource is also enriched with external sources of biological data, the number of which has been increasingly integrated since 2016. In addition to Uniprot, which provides the bulk of knowledge about transcript isoforms, as well as the localization of genes on the chromosome and protein products in the cell, as of 2023, neXtProt also includes databases of single nucleotide substitutions (UniProtKB, COSMIC, dbSNP, neXtProt, gnomAD), and PTMs (UniProtKB, neXtProt, PeptideAtlas, GlyConnect), as well as others (Ensembl, PeptideAtlas, PDB via UniProtKB).

Following the completion of the Human Genome Project [13], the technological prerequisites for high-throughput gene product research at the transcriptomic (next-generation sequencing) and proteomic (mass spectrometry, e.g. Orbitrap) levels were created, thus contributing to a significant increase in the amount of experimental data.

This work is a retrospective analysis of changes in information on the number of proteoforms in the neXtProt database 8 years after the work of our group in 2016 [11]. Retrospective assessment of the number of proteoforms helps to identify trends in proteome research, as well as to identify areas requiring additional research or updates. Performing such analysis we can assess, what changes have occurred in the database due to the development of proteome analysis methods and technological advances, as well as to assess the quality of the data contained in the neXtProt database. Previously, a similar approach resulted in identification of the most frequently used experimental and bioinformatic methods for analyzing protein functions [14].

MATERIALS AND METHODS

We used the neXtProt annotation section describing the amino acid sequences encoded by the gene for a healthy person.

For each year, an xml file containing information about all proteins was downloaded from the neXtProt FTP server [15]. NeXtProt provides updated versions several times a year, so we selected one version from each year in the interval from 2016 to 2023, namely: [nextprot_release_]2016-12-02, 2017-08-01, 2018-09-03, 2019-08-22, 2020-11-26, 2021-11-19, 2022-10-31, 2023_09.

Using our own python scripts, we extracted information From each xml file from the following fields: (i) <entry accession>, (ii) <annotation-category> with the category attribute equal to "modified-residue", "cross-link", "disulfide-bond", and "glycosylation-site", "variant", (iii) <property-list>, (iv) <chromosomal-location-list>, (v) <transcript-mapping>.

Each protein entry annotation has a data reliability level [16]: (i) "Gold" (experimentally proven data of the highest quality, corresponding to an error level of <1% by immunohistochemistry, tandem mass spectrometry, etc.); (ii) "Silver" (data predicted by bioinformatics tools, but not confirmed by experimental methods, corresponding to an error level of <5%); (iii) "Bronze" (low-quality data). In this work, only Gold and Silver data obtained from the <annotation quality> tag were considered. Thus, the number of PTM, SAP, and AS corresponding to each protein-coding gene was determined. After that, the obtained information was summarized and the human genome of each of the versions selected for analysis (from 2016 to 2023) was characterized using the following set of descriptors:

[N] — the number of protein-coding genes;

[AS], ([SAP], [PTM]) — the number of amino acid sequences predicted from mRNA alternative splicing data (containing SAP or PTM);

[ASav], ([SAPav], [PTMav]) — the number of amino acid sequence variants encoded by one gene, formed due to alternative splicing (the presence of SAP or PTM); the descriptor is calculated as the ratio of [AS] ([SAP], [PTM]) to [N].

To estimate the potential number of proteoforms, three different cases of combination of PTM, SAP, and AS events were considered according to the following information model equations [11] (1)–(3):

$$Nps = N\times(ASav+SAPav+PTMav) \quad (1);$$

$$Nps = (N+AS)\times(SAPav+PTMav) \quad (2);$$

$$Nps = N\times Asav\times SAPav\times PTMav \quad (3).$$

Equation (1) assumes that PTMs occur exclusively in canonical protein sequences, but not in splice variants. Equation (2) assumes that PTMs and SAPs occur both in proteins encoded by canonical sequences and in splice variants. Equation (3) assumes that all types of modifications (PTMs, SAPs and AS) occur independently of each other.

The results were visualized using our own scripts in python (v3.8), including the pandas (v2.2.2), matplotlib (v3.8.4), and seaborn (v 0.13.2) libraries.

RESULTS AND DISCUSSION

We have analyzed the dynamics of changes in the period 2016–2023 of the key parameters of the models previously proposed for estimating the number of proteoforms: the number of protein-coding genes (PCGs), PTMs, ASs, and SAPs.

The main parameter, the number of PCGs (20300), has remained unchanged since 2018 (Fig. 1). This indicates that the bioinformatics algorithms for genome annotation, which help to identify new genes and their products, as well as clarify information about already known genes, have not changed significantly. The main sources of information about PCGs used by neXtProt curators are databases of genome sequences and their annotations, such as EMBL-EBI (EMBL's European Bioinformatics Institute), in particular Ensembl, KeGG, and others [12].

The analysis of the dynamics of changes in the number of PTMs (Fig. 2) has shown that the number of experimentally confirmed PTMs actually reached a plateau in 2017, while *in silico* approaches are improving and predict more and more possible variants (over 1000 PTMs) [17]. On the one hand, reaching a plateau may indicate reaching the limits of currently available PTM detection

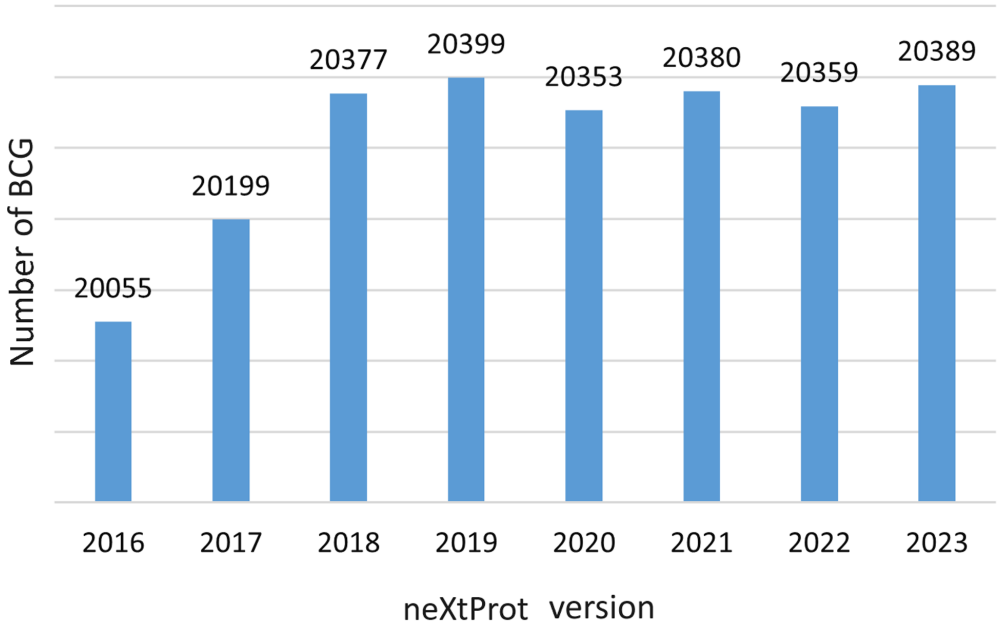

Figure 1. Number of PCGs in the human genome, according to neXtProt data.

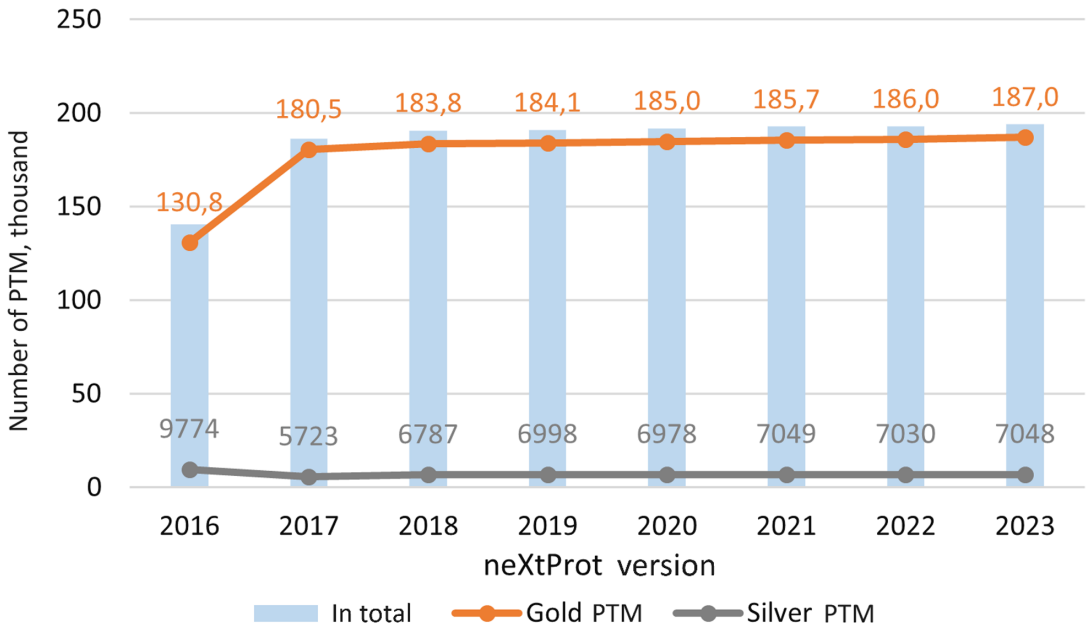

Figure 2. Number of PTMs associated with human PCGs, according to neXtProt data.

technologies (mainly, mass spectrometric methods are used for this purpose) [18]. On the other hand, many PTMs depend on cellular regulatory signals that influence the activity of enzymes responsible for PTMs of amino acid residues (kinases, phosphatases, acetyltransferases, etc.), as well as on external or internal factors, for example, oxidation of aromatic amino acids due to oxidative stress [19]. The presence or absence of these factors at a particular time point determines the possibility of modification and should be taken into account in the study on PTM profiling [20].

Not all human PCGs contain annotations with information on the presence of PTM, AS or SAP: for some genes, such events are either not characteristic or have not yet been detected. The proportion of PCGs for which there is information on encoded proteins with splice variants or SAP, has not changed since 2016 and is 48.3% and 96%, respectively. This means that slightly more than half of the genes in the human genome had not been characterized for the presence of AS variants. At the same time, almost all PCGs are provided with neXtProt records describing variants (at least one) of SAP. Interestingly, the proportion of PCGs, for which the presence of PTM in encoded proteins has been shown, is gradually increasing: in 2016, there were 74.2% of such genes, and in 2023 — 76.4%. This is probably due to the development of mass spectrometric analysis methods that allow characterizing protein PTMs in a high-throughput mode [18]. Most likely, the proportion of such PCG will increase over time, in contrast to the proportion of PCG with annotated AS and SAP — changes in this case should be expected in the event of a breakthrough change in sequencing technologies.

For the number of AS variants (Fig. 3), stabilization of the number of records is also observed. This parameter can change relatively slowly for several reasons. Firstly, some AS events occur in the regulatory regions of transcripts (5′/3′-UTR) and do not lead to diversity of protein structures [21]. Secondly, alternative splicing profiles can change depending on external and internal stimuli, as well as during oncotransformation [22, 23]. Currently, methods and algorithms for searching for new splice forms in RNA sequencing experiments are being improved [24, 25]. Since 2018, the MANE project has been launched to integrate two major human genome annotations (RefSeq and Ensembl/GENCODE) and new splice forms, as well as their validation; however, the project has still not been completed yet [26].

AS events are widely confirmed at the level of mRNA (transcripts). There is significantly less evidence for the existence of protein splice variants due to the low peptide coverage in most panoramic proteomic experiments [27]. This is due to the recognized limitations of proteomics for detecting such events: it is common practice to separate protein isoforms into groups containing the same peptide, since AS does not result in complete polypeptide sequence changes, but only in certain regions [28].

In 2023, Sinitcyn et al. have shown that more than half (about 64%) of the splicing events of highly expressed genes detected by transcriptomics are actually translated and present at the protein level [29]. At the proteomic level, only 22% of splice forms with an intermediate expression level could be detected [29]. Apparently, this evaluation is underestimated due to the highly dynamic nature of protein expression and the problems of detecting differentially expressed splice forms.

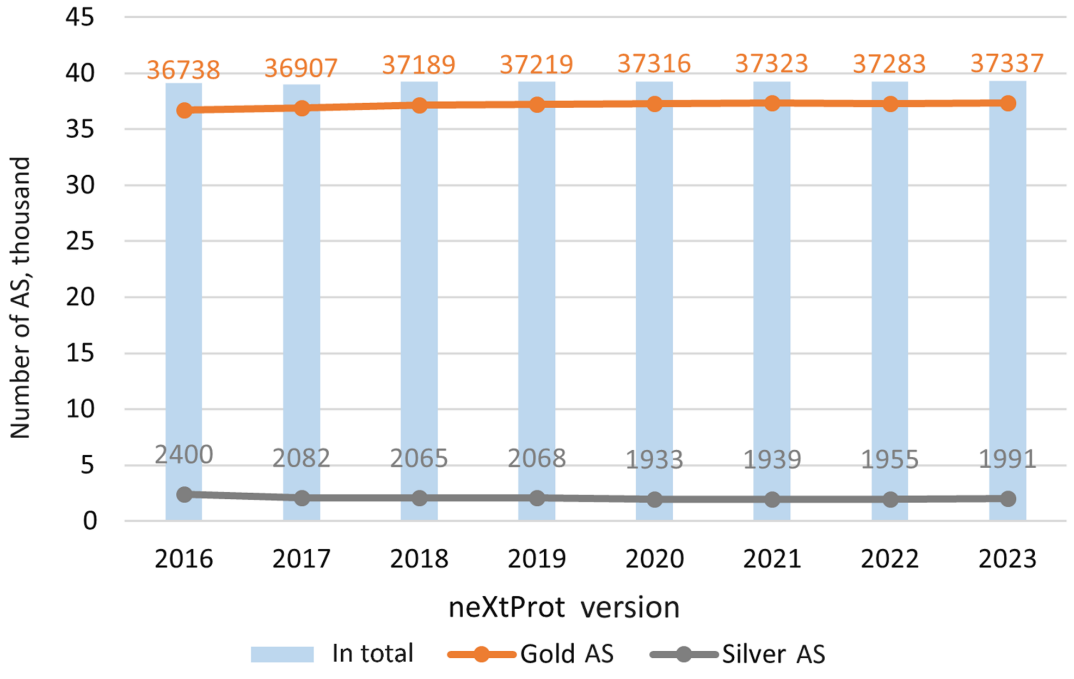

Figure 3. Number of AS associated with human PCGs, according to neXtProt data.

The analysis of changes in the number of known SAPs by year is especially interesting (Fig. 4). In 2019, an explosive growth of SAPs with the "Gold" status was noted, since neXtProt integrated variant frequency data from the Genome Aggregation Database (gnomAD) [30], expanding the information on sequence variations at the protein level.

A similar increase in the number of experimentally detected SAPs in 2023 is likely related to the study by Sinitcyn et al. [29]. This is the deepest proteogenomic study to date, indicating that approximately 73% of non-synonymous SNPs (i.e. SAPs) are translated and present in the proteome.

The chromosome-centric analysis (information on the chromosome number on which the PCG is located was selected from the annotations) of the PCG SAP based on the 2023 version of neXtProt showed a median presence of 81 SAPs for disease-associated variants and 300 for other SAPs (Fig. 5).

Figure 5 confirms the same SAP distribution density for all chromosomes. The outliers reach more than 2000 SAP variants on the PCGs (for example, the TTN gene, SYNE1, and PCLO) for SAPs associated with the development of pathological conditions, and more than eight thousand (e.g., the MUC4 gene, OBSCN, and AHNAK2) among other variants.

In 2016, the estimated number of human protein species was 0.62, 0.88, or 6.13 million (with 20,043 protein-coding genes), depending on the model used [11].

As of 2023, the estimated number of proteins estimated by the same equations reached 5.8, 16.3, and 124.3 million proteins, respectively (Fig. 6).

Over the past eight years, the estimated size of the human proteome width has demonstrated more than 20-fold growth. These indicators have been calculated only for experimentally confirmed, "Gold" annotations of PTM, SAP (excluding disease-associated substitution variants) and AS and for the genes encoding them, and also excluding substitution variants. Taking into account bioinformatically predicted data (having the "Silver" status), the number of potential protein types in the human body exceeds the values obtained in 2016 by more than 40 times.

The greatest contribution to protein diversity is made by the presence of SNPs in the genome, which are then implemented at the proteomic level in the form of amino acid residue substitutions in the protein product. The contribution to protein diversity of AS and PTMs is much more modest (Fig. 7). This is especially evident for the model calculated using equation (3) (Fig. 6), which is based on the assumption that PTMs arise in any amino acid sequence, and the co-occurrence of PTMs and SAPs occurs in all splice variants and canonical sequences.

During the analyzed period, significant improvements occurred in proteome research methods such as mass spectrometry and bioinformatics, which made it possible to detect and identify proteins that were previously inaccessible to detection. In many ways,

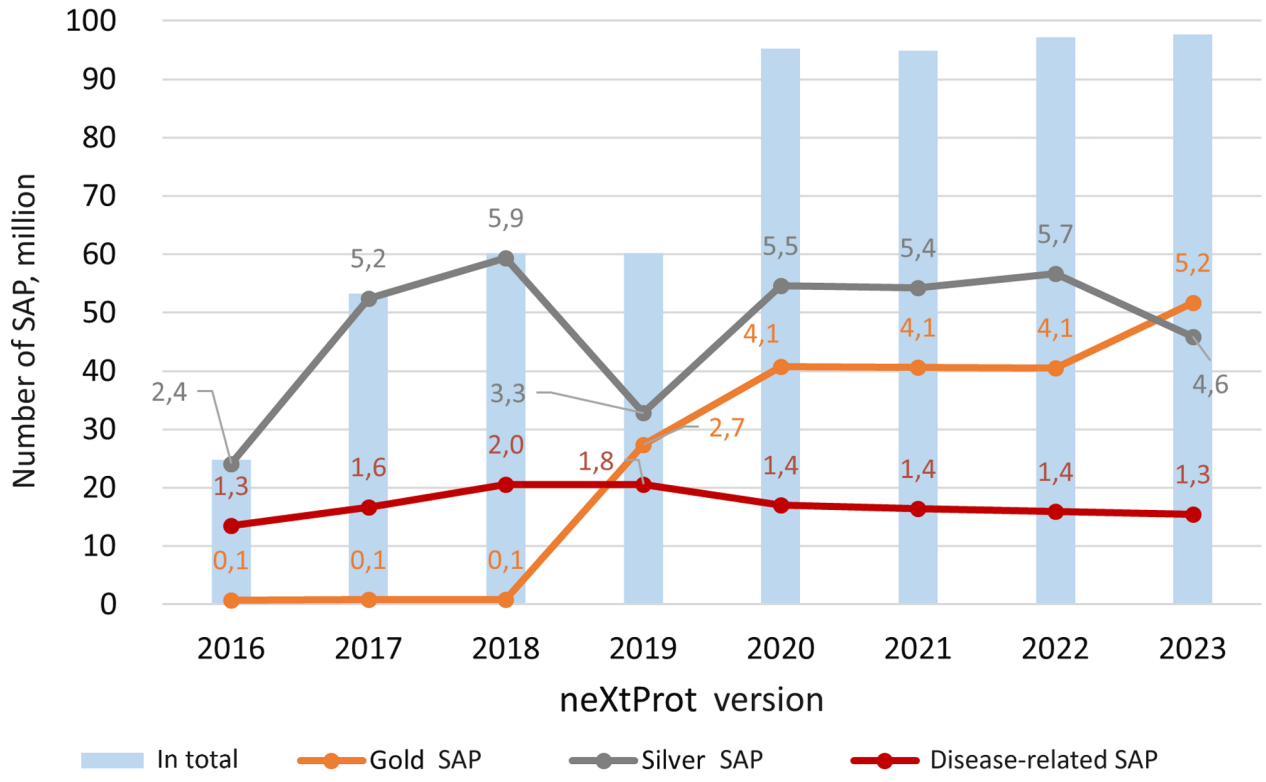

Figure 4. Number of SAPs associated with human PCGs, according to neXtProt data.

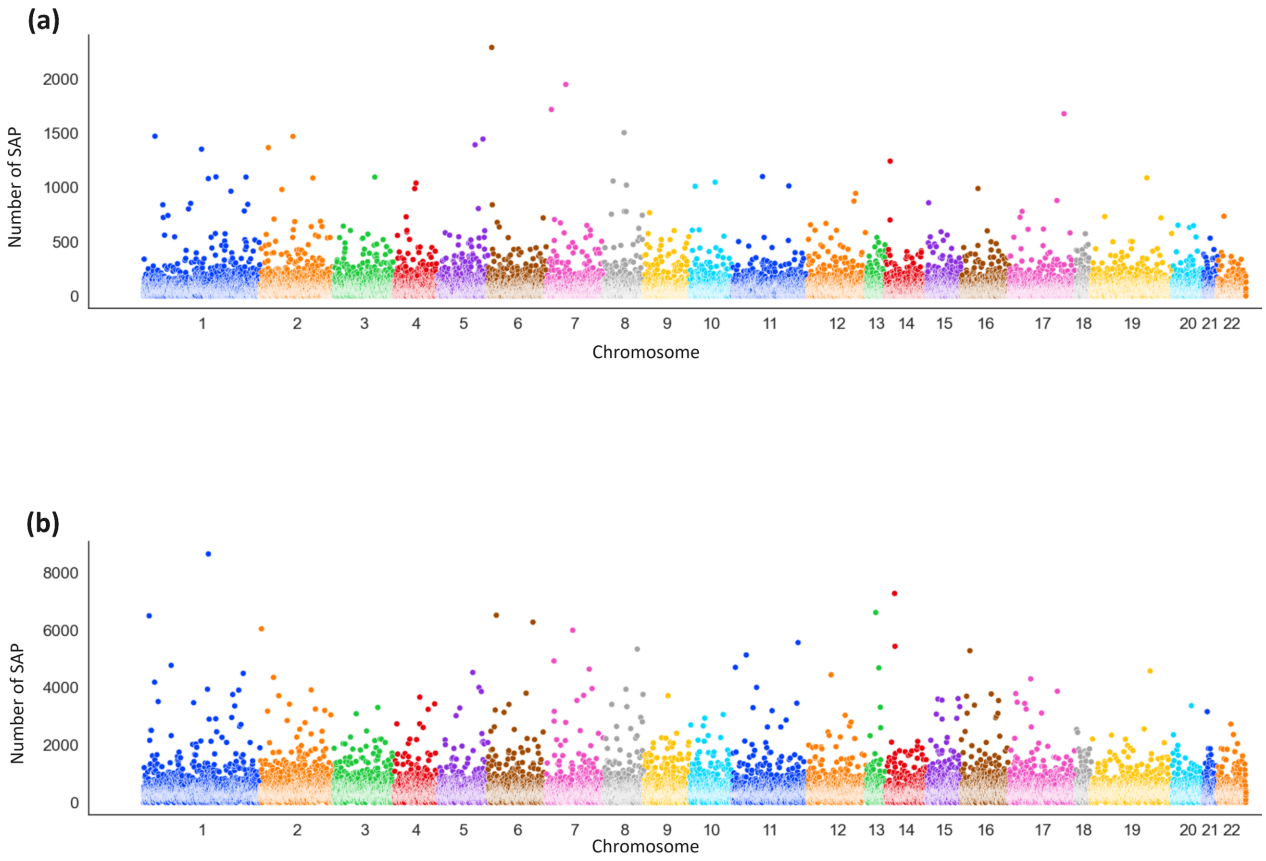

Figure 5. Chromosome-centric distribution of SAPs associated with human PCGs, according to the latest neXtProt 2023 version for SAP variants (**a**) associated with the disease; (**b**) not associated with the pathological process. The dot shows PCGs, the color of the dot encodes the chromosome number.

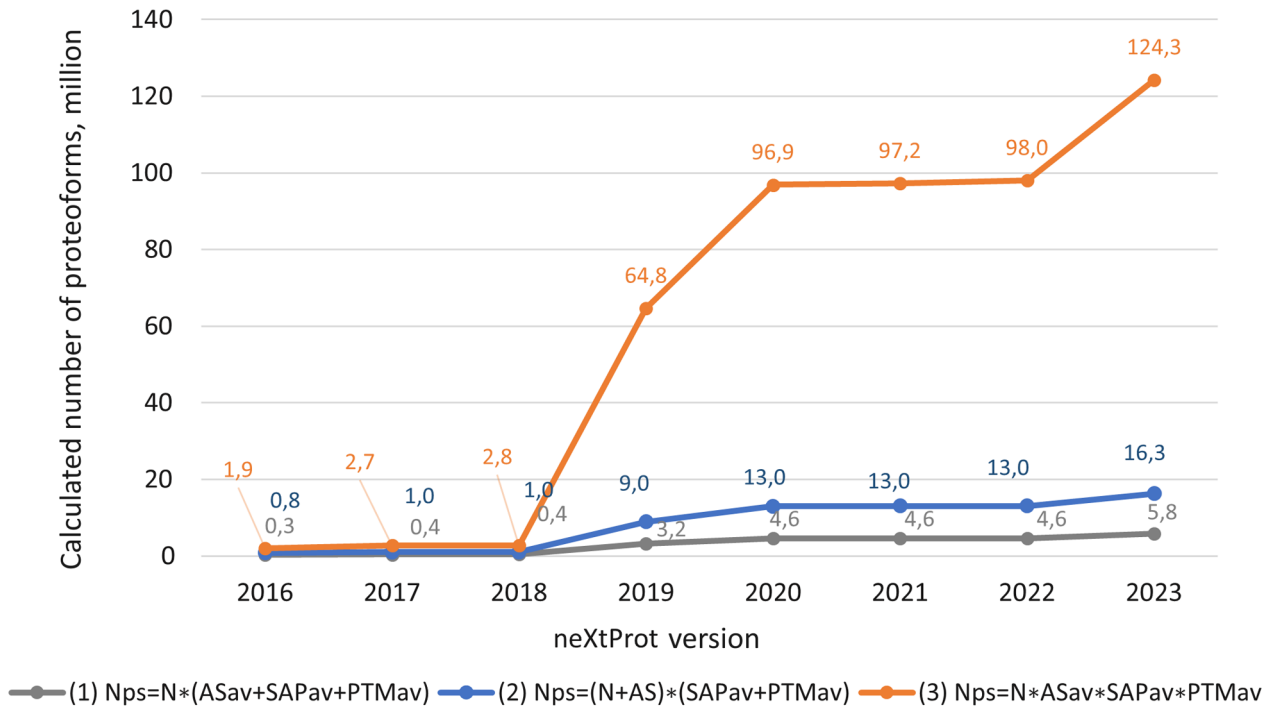

Figure 6. Estimated number of human proteoforms (according to neXtProt data from 2016 to 2023).

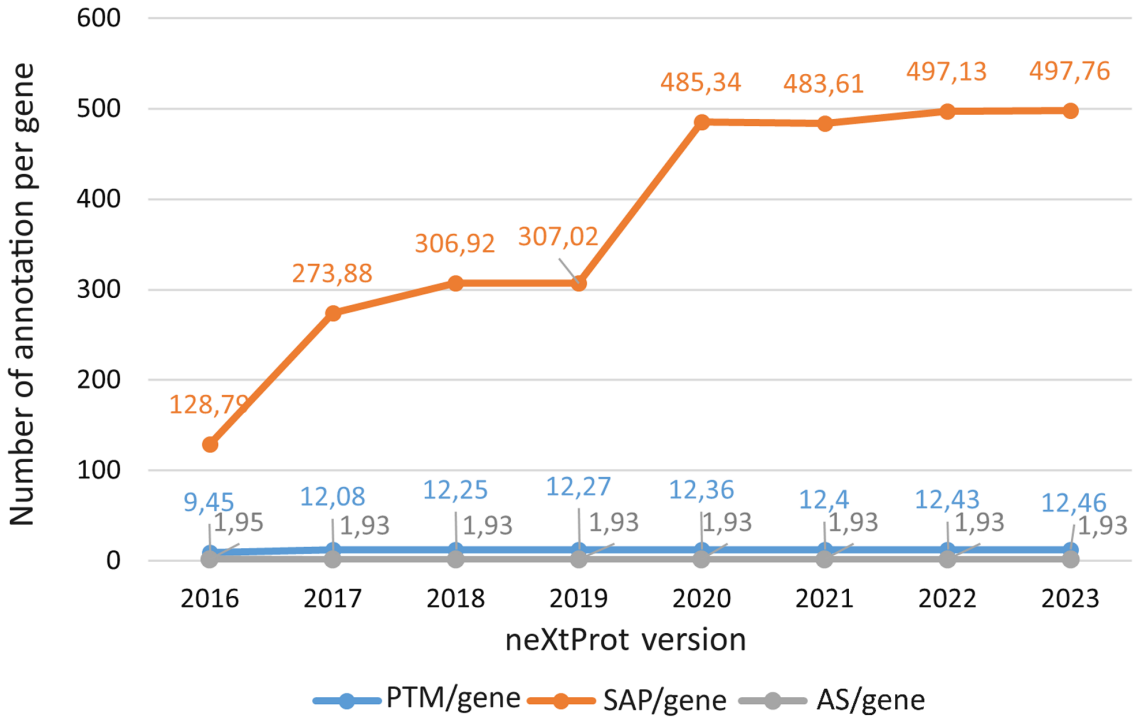

Figure 7. Change in the number of human PCG annotations in neXtProt from 2016 to 2023.

this may be the result of the Human Proteome Project in terms of searching for missing proteins [31]. Mass spectrometric approaches such as high-resolution tandem mass spectrometry have become widespread and this has a significant impact on the number of detected proteins. From the bioinformatics view point, algorithms based on machine learning have begun to develop to solve various problems, such as protein structure prediction [32], gene classification [33], DNA sequence analysis etc.

New methods of genome sequencing and analysis of genetic variation have allowed more precise identification of protein variants that may be formed due to genetic differences between individuals and alternative splicing. The creation of long-read RNA sequencing method developed by Oxford Nanopore Technologies (ONT), spanning several exons, opened new possibilities for studying AS by direct identification and quantification of the isoform transcripts [34].

Integration of data from more diverse sources and the application of interdisciplinary approaches such as systems biology and network analysis have also helped to expand the estimated size of the proteome.

The increase in the estimated size of the human proteome width by 20-fold or more over the past 8 years is the result of a combination of technological advances, improved data analysis methods, and a deeper understanding of the biological complexity of the proteome. At the same time, the fact that the number of detected PTMs and AS variants has actually reached a plateau indicates that genomic and transcriptomic technologies are developing significantly faster than proteomic ones. This difference is caused by the scale and complexity of the research objects. Genomes and transcriptomes are more static structures compared to proteins. In addition, the cost of equipment for protein detection still significantly exceeds the cost of sequencing devices. Proteome research requires overcoming many technical and methodological obstacles, including the complexity of protein structure analysis, identification and quantitative assessment of protein components, as well as analysis of their functions and interactions within the cell [35].

CONCLUSIONS

Eight years after the estimation of the proteome size [11], we conducted a retrospective analysis of the change in the trend of the number of human proteoforms. We have used the experimental results of various scientific groups aggregated in the neXtProt resource, the most complete repository of human proteome data. According to various information models, modern experimental methods make it possible to identify from 5 to 125 million different proteoforms, the proteins formed due to AS, implementation of SNPs at the proteome level, and PTMs in various combinations. This result reflects an increase in the size of the human proteome by 20 or more times over the past 8 years. The dynamics of data accumulation indicate the development of methodological approaches: we do not observe an increase in the number of protein-coding genes or genes, for which AS or single-amino acid substitutions have been shown

for the first time over the period from 2016 to 2023. This indicates saturation in terms of such genomic characteristics; unlike genes encoding proteins with variants of the SAP, their number increases, but they are studied not by sequencing methods, but by proteomic methods based on mass spectrometry. An important achievement of recent years is the work [29], which has shown that the theoretically predicted diversity of proteoforms is confirmed experimentally at the proteome level. The limitations of sensitivity methods seriously complicate experimental registration of the proteoforms present in relatively small concentrations (compared to canonical, master forms). One of the possible options for overcoming this problem is the use of different types of biomaterial, different conditions in the hope that somewhere the concentration of a specific proteoform will be sufficient for the operating detector, because of the proteome dynamics.

Our work opens prospects for studying the proteome taking into account the diversity of protein types (proteoforms). For the first time, a proteome of proteoforms was obtained based on a consolidated array of experimental data, reaching up to 125 million by 2023. In the future, we can expect an increase in the number of variants found due to the identification of single nucleotide substitutions at the proteomic level in various types of biomaterial (in norm and pathologies) and PTMs. Probably, the aggregate of such minor structural changes as a whole that is a biomarker of the body's response to pathological processes.

FUNDING

The work was carried out within the framework of the Program of Fundamental Research in the Russian Federation for the long-term period (2021–2030) (No. 122030100168-2).

COMPLIANCE WITH ETHICAL STANDARDS

This article does not contain any research involving humans or the use of animals as objects.

CONFLICT OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

1. *Aebersold R., Agar J.N., Amster I.J., Baker M.S., Bertozzi C.R., Boja E.S., Costello C.E., Cravatt B.F., Fenselau C., Garcia B.A., Ge Y., Gunawardena J., Hendrickson R.C., Hergenrother P.J., Huber C.G., Ivanov A.R., Jensen O.N., Jewett M.C., Kelleher N.L., Kiessling L.L., Krogan N.J., Larsen M.R., Loo J.A., Ogorzalek Loo R.R., Lundberg E., MacCoss M.J., Mallick P., Mootha V.K., Mrksich M., Muir T.W., Patrie S.M., Pesavento J.J., Pitteri S.J., Rodriguez H., Saghatelian A., Sandoval W., Schlüter H., Sechi S., Slavoff S.A., Smith L.M., Snyder M.P., Thomas P.M., Uhlén M., van Eyk J.E., Vidal M., Walt D.R., White F.M., Williams E.R., Wohlschlager T., Wysocki V.H., Yates N.A., Young N.L., Zhang B.* (2018) How many human proteoforms are there? Nat. Chem. Biol., **14**(3), 206–214. DOI: 10.1038/nchembio.2576
2. *Zhang F., Chen J.Y.* (2016) A method for identifying discriminative isoform-specific peptides for clinical proteomics application. BMC Genomics, **17**(Suppl 7), 522. DOI: 10.1186/s12864-016-2907-8
3. *Prabakaran S., Lippens G., Steen H., Gunawardena J.* (2012) Post-translational modification: Nature's escape from genetic imprisonment and the basis for dynamic information encoding. Wiley Interdiscip. Rev. Syst. Biol. Med., **4**(6), 565–583. DOI: 10.1002/wsbm.1185
4. *Schlüter H., Apweiler R., Holzhьtter H.G., Jungblut P.R.* (2009) Finding one's way in proteomics: A protein species nomenclature. Chem. Cent. J., **3**, 11. DOI: 10.1186/1752-153X-3-11
5. *Smith L.M., Kelleher N.L., Consortium for Top Down Proteomics* (2013) Proteoform: A single term describing protein complexity. Nat. Methods, **10**(3), 186–187. DOI: 10.1038/nmeth.2369
6. *Semba R.D., Enghild J.J., Venkatraman V., Dyrlund T.F., van Eyk J.E.* (2013) The human eye proteome project: Perspectives on an emerging proteome. Proteomics, **13**(16), 2500–2511. DOI: 10.1002/pmic.201300075
7. *Wasinger V.C., Locke V.L., Raftery M.J., Larance M., Rothemund D., Liew A., Bate I., Guilhaus M.* (2005) Two-dimensional liquid chromatography/tandem mass spectrometry analysis of Gradiflow fractionated native human plasma. Proteomics, **5**(13), 3397–3401. DOI: 10.1002/pmic.200401160
8. *Vavilov N., Ilgisonis E., Lisitsa A., Ponomarenko E., Farafonova T., Tikhonova O., Zgoda V., Archakov A.* (2022) Number of detected proteins as the function of the sensitivity of proteomic technology in human liver cells. Curr. Protein Pept. Sci., **23**(4), 290–298. DOI: 10.2174/1389203723666220526092941
9. *Po A., Eyers C.E.* (2023) Top-down proteomics and the challenges of true proteoform characterization. J. Proteome Res., **22**(12), 3663–3675. DOI: 10.1021/acs.jproteome.3c00416
10. *Carvalho A.S., Penque D., Matthiesen R.* (2015) Bottom up proteomics data analysis strategies to explore protein modifications and genomic variants. Proteomics, **15**(11), 1789–1792. DOI: 10.1002/pmic.201400186
11. *Ponomarenko E.A., Poverennaya E.V., Ilgisonis E.V., Pyatnitskiy M.A., Kopylov A.T., Zgoda V.G., Lisitsa A.V., Archakov A.I.* (2016) The size of the human proteome: The width and depth. Int. J. Anal. Chem., **2016**, 7436849. DOI: 10.1155/2016/7436849
12. *Lane L., Argoud-Puy G., Britan A., Cusin I., Duek P.D., Evalet O., Gateau A., Gaudet P., Gleizes A., Masselot A., Zwahlen C., Bairoch A.* (2012) neXtProt: A knowledge platform for human proteins. Nucleic Acids Res., **40**(Database issue), D76-D83. DOI: 10.1093/nar/gkr1179
13. *Lander E.S., Linton L.M., Birren B., Nusbaum C., Zody M.C., Baldwin J., Devon K., Dewar K., Doyle M., FitzHugh W., Funke R., Gage D., Harris K., Heaford A., Howland J., Kann L., Lehoczky J., LeVine R., McEwan P., McKernan K., Meldrim J., Mesirov J.P., Miranda C., Morris W., Naylor J., Raymond C., Rosetti M., Santos R., Sheridan A., Sougnez C., Stange-Thomann Y., Stojanovic N., Subramanian A., Wyman D.,*

Rogers J., Sulston J., Ainscough R., Beck S., Bentley D., Burton J., Clee C., Carter N., Chen Y.J., Szustakowki J., International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. Nature, **409**(6822), 860–921. DOI: 10.1038/35057062

14. *Ilgisonis E.V., Pogodin P.V., Kiseleva O.I., Tarbeeva S.N., Ponomarenko E.A.* (2022) Evolution of protein functional annotation: Text mining study. J. Pers. Med., **12**(3), 479. DOI: 10.3390/jpm12030479
15. neXtProt downloads. FTP-server. Retrieved August 6, 2024, from: https://download.nextprot.org/pub/previous_releases
16. *Gaudet P., Argoud-Puy G., Cusin I., Duek P., Evalet O., Gateau A., Gleizes A., Pereira M., Zahn-Zabal M., Zwahlen C., Bairoch A., Lane L.* (2013) neXtProt: Organizing protein knowledge in the context of human proteome projects. J. Proteome Res., **12**(1), 293–298. DOI: 10.1021/pr300830v
17. *Li Z., Li S., Luo M., Jhong J.H., Li W., Yao L., Pang Y., Wang Z., Wang R.., Ma R., Yu J., Huang Y., Zhu X., Cheng Q., Feng H., Zhang J., Wang C., Hsu J.B., Chang W.C., Wei F.X., Huang H.D., Lee T.Y.* (2022) dbPTM in 2022: An updated database for exploring regulatory networks and functional associations of protein post-translational modifications. Nucleic Acids Res., **50**(D1), D471–D479. DOI: 10.1093/nar/gkab1017
18. *Yang F., Wang C.* (2020) Profiling of post-translational modifications by chemical and computational proteomics. Chem. Commun. (Cambridge), **56**(88), 13506–13519. DOI: 10.1039/d0cc05447j
19. *Santos A.L., Lindner A.B.* (2017) Protein posttranslational modifications: roles in aging and age-related disease. Oxid. Med. Cell. Longev., **2017**, 5716409. DOI: 10.1155/2017/5716409
20. *Basak S., Lu C., Basak A.* (2016) Post-translational protein modifications of rare and unconventional types: Implications in functions and diseases. Curr. Med. Chem., **23**(7), 714–745. DOI: 10.2174/0929867323666160118095620
21. *Lim C.S., Wardell S.J.T., Kleffmann T., Brown C.M.* (2018) The exon-intron gene structure upstream of the initiation codon predicts translation efficiency. Nucleic Acids Res., **46**(9), 4575–4591. DOI: 10.1093/nar/gky282
22. *Sciarrillo R., Wojtuszkiewicz A., Kooi I.E., Gómez V.E., Boggi U., Jansen G., Kaspers G.J., Cloos J., Giovannetti E.* (2016) Using RNA-sequencing to detect novel splice variants related to drug resistance in *in vitro* cancer models. J. Vis. Exp., **9**(118), 54714. DOI: 10.3791/54714
23. *Roy M., Xu Q., Lee C.* (2005) Evidence that public database records for many cancer-associated genes reflect a splice form found in tumors and lack normal splice forms. Nucleic Acids Res., **33**(16), 5026–5033. DOI: 10.1093/nar/gki792
24. *Cmero M., Schmidt B., Majewski I.J., Ekert P.G., Oshlack A., Davidson N.M.* (2021) MINTIE: Identifying novel structural and splice variants in transcriptomes using RNA-seq data. Genome Biol., **22**, 296. DOI: 10.1186/s13059-021-02507-8
25. *Adamopoulos P.G., Kontos C.K., Scorilas A., Sideris D.C.* (2020) Identification of novel alternative transcripts of the human Ribonuclease κ (RNASEK) gene using 3′ RACE and high-throughput sequencing approaches. Genomics, **112**(1), 943–951. DOI: 10.1016/j.ygeno.2019.06.010
26. *Morales J., Pujar S., Loveland J.E., Astashyn A., Bennett R., Berry A., Cox E., Davidson C., Ermolaeva O., Farrell C.M., Fatima R., Gil L., Goldfarb T., Gonzalez J.M., Haddad D., Hardy M., Hunt T., Jackson J., Joardar V.S., Kay M., Kodali V.K., McGarvey K.M., McMahon A., Mudge J.M., Murphy D.N., Murphy M.R., Rajput B., Rangwala S.H., Riddick L.D., Thibaud-Nissen F., Threadgold G., Vatsan A.R., Wallin C., Webb D., Flicek P., Birney E., Pruitt K.D., Frankish A., Cunningham F., Murphy T.D.* (2022) A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature, **604**(7905), 310–315. DOI: 10.1038/s41586-022-04558-8
27. *Reixachs-Solé M., Eyras E.* (2022) Uncovering the impacts of alternative splicing on the proteome with current omics techniques. Wiley Interdiscip. Rev. RNA, **13**(4), e1707. DOI: 10.1002/wrna.1707
28. *Nesvizhskii A.I., Keller A., Kolker E., Aebersold R.* (2003) A statistical model for identifying proteins by tandem mass spectrometry. Anal. Chem., **75**(17), 4646–4658. DOI: 10.1021/ac0341261
29. *Sinitcyn P., Richards A.L., Weatheritt R.J., Brademan D.R., Marx H., Shishkova E., Meyer J.G, Hebert A.S., Westphall M.S., Blencowe B.J., Cox J., Coon J.J.* (2023) Global detection of human variants and isoforms by deep proteome sequencing. Nat. Biotechnol., **41**(12), 1776–1786. DOI: 10.1038/s41587-023-01714-x
30. *Lek M., Karczewski K.J., Minikel E.V., Samocha K.E., Banks E., Fennell T., O'Donnell-Luria A.H., Ware J.S., Hill A.J., Cumming B.B., Tukiainen T., Birnbaum D.P., Kosmicki J.A., Duncan L.E., Estrada K., Zhao F., Zou J., Pierce-Hoffman E., Berghout J., Cooper D.N., Deflaux N., de Pristo M., Do R., Flannick J., Fromer M., Gauthier L., Goldstein J., Gupta N., Howrigan D., Kiezun A., Kurki M.I., Moonshine A.L., Natarajan P., Orozco L., Peloso G.M., Poplin R., Rivas M.A., Ruano-Rubio V., Rose S.A., Ruderfer D.M., Shakir K., Stenson P.D., Stevens C., Thomas B.P., Tiao G, Tusie-Luna M.T., Weisburd B., Won H.H., Yu D., Altshuler D.M., Ardissino D., Boehnke M., Danesh J., Donnelly S., Elosua R., Florez J.C., Gabriel S.B., Getz G, Glatt S.J., Hultman C.M., Kathiresan S., Laakso M., McCarroll S., McCarthy M.I., McGovern D., McPherson R., Neale B.M., Palotie A., Purcell S.M., Saleheen D., Scharf J.M., Sklar P., Sullivan P.F., Tuomilehto J., Tsuang M.T., Watkins H.C., Wilson J.G., Daly M.J., MacArthur D.G., Exome Aggregation Consortium* (2016) Analysis of protein-coding genetic variation in 60,706 humans. Nature, **536**(7616), 285–291. DOI: 10.1038/nature19057
31. *Omenn G.S., Lane L., Overall C.M., Corrales F.J., Schwenk J.M., Paik Y.K., van Eyk J.E., Liu S., Snyder M., Baker M.S., Deutsch E.W.* (2018) Progress on identifying and characterizing the human proteome: 2018 metrics from the HUPO human proteome project. J. Proteome Res., **17**(12), 4031–4041. DOI: 10.1021/acs.jproteome.8b00441
32. *Senior A.W., Evans R., Jumper J., Kirkpatrick J., Sifre L., Green T., Qin C., Žídek A., Nelson A.W.R., Bridgland A., Penedones H., Petersen S., Simonyan K., Crossan S., Kohli P., Jones D.T., Silver D., Kavukcuoglu K., Hassabis D.* (2020) Improved protein structure prediction using potentials from deep learning. Nature, **577**(7792), 706–710. DOI: 10.1038/s41586-019-1923-7
33. *Walker A.S., Clardy J.* (2021) A machine learning bioinformatics method to predict biological activity from biosynthetic gene clusters. J. Chem. Inf. Model., **61**(6), 2560–2571. DOI: 10.1021/acs.jcim.0c01304
34. *Wright C.J., Smith C.W.J., Jiggins C.D.* (2022) Alternative splicing as a source of phenotypic diversity. Nat. Rev. Genet., **23**(11), 697–710. DOI: 10.1038/s41576-022-00514-4
35. *Chandramouli K., Qian P.-Y.* (2009) Proteomics: Challenges, techniques and possibilities to overcome biological sample complexity. Human Genomics Proteomics, **2009**, 239204. DOI: 10.4061/2009/239204

Received: 01. 05. 2024.
Revised: 23. 07. 2024.
Accepted: 08. 08. 2024.

РАЗМЕР ПРОТЕОМА ЧЕЛОВЕКА КАК ФУНКЦИЯ РАЗВИТИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ТЕХНОЛОГИЙ И МЕТОДОВ БИОИНФОРМАТИКИ

Е.В. Сарыгина*, А.С. Козлова, Е.А. Пономаренко, Е.В. Ильгисонис

Научно-исследовательский институт биомедицинской химии имени В.Н. Ореховича,
119121, Москва, ул. Погодинская, 10; *эл. почта: lizalesa@gmail.com

Представлен ретроспективный анализ изменений сведений о количестве протеоформ человека, событий посттрансляционных модификаций (ПТМ), альтернативного сплайсинга (АС), одноаминокислотных полиморфизмов (ОАП), ассоциированных с белок-кодирующими генами в базе данных neXtProt. В 2016 году нашей группой были предложены три математические модели для предсказания количества различных белков (протеоформ) в протеоме человека. Спустя восемь лет мы сравнили исходные данные информационных ресурсов и их вклад в результаты предсказаний, сопоставив различия с новыми подходами экспериментального и биоинформатического анализа модификаций белков. Цель данной работы — актуализировать информацию о статусах записей в базах данных о выявленных протеоформах с 2016 года, а также выявить тренды изменений количеств этих записей. Согласно различным информационным моделям, современные экспериментальные методы позволяют выявить от 5 до 125 млн различных протеоформ — белков, образованных в результате альтернативного сплайсинга, реализации на протеомном уровне однонуклеотидных полиморфизмов и посттрансляционных модификаций в различных комбинациях. Данный результат отражает увеличение размера человеческого протеома на 20 и более раз за последние 8 лет.

Полный текст статьи на русском языке доступен на сайте журнала (http://pbmc.ibmc.msk.ru).

Ключевые слова: протеомика; протеоформы; посттрансляционные модификации; одноаминокислотные замены; альтернативный сплайсинг; neXtProt

Финансирование. Работа выполнена в рамках Программы фундаментальных исследований в Российской Федерации на долгосрочный период 2021–2030 годы (№ 122030100168-2).

Поступила в редакцию: 01.05.2024; после доработки: 23.07.2024; принята к печати: 08.08.2024.