Chapter 03

Zipf's Law in Proteomics

Stanislav Naryzhny[1,2]*, Maria Maynskova1, Victor Zgoda[1] and Alexander Archakov[1]

[1]Orekhovich Institute of Biomedical Chemistry of Russian Academy of Medical Sciences, Russia
[2]B.P. Konstantinov Petersburg Nuclear Physics Institute, National Research Center "Kurchatov Institute", Russia

***Corresponding Author:** Stanislav Naryzhny, B.P. Konstantinov Petersburg Nuclear Physics Institute, National Research Center "Kurchatov Institute", Orlova roscha, Gatchina, Leningrad region, 188300, Russia, Fax: (+7) 8137132303; E-mail: snaryzhny@mail.ru

First Published **February 26, 2018**

This Book Chapter is an excerpt from an article published by Stanislav Naryzhny, et al. at Journal of Proteomics & Bioinformatics in March 2017. (Naryzhny S, Maynskova M, Zgoda V, Archakov A (2017) Zipf's Law in Proteomics. J Proteomics Bioinform 10: 79-84. doi: 10.4172/jpb.1000427)

Abstract

Human cells contain many thousands of protein components, protein species/proteoforms, whose cooperation provides the complicated functional mechanisms of the cellular proteome. Though recent methods still do not allow us to obtain the whole picture of this cooperation, they at least provide an opportunity to develop a representation of the proteome size and quantitative distribution of protein species inside the proteome. Using 2DE analysis followed by both protein staining and ESI LC-MS/MS analysis, we performed an analysis of the quantitative distribution of different protein species in human cells. We have analyzed several human cancer cell lines (HepG2, glioblastoma, MCF7) along with the primary liver cells from tissue samples and found that the dependence of the number of protein species on their abundance is described by Zipf's law:

$y = ax^{-1}$ (1),

where y stands for the number of protein species (N), x stands for the abundance. In the case where the abundance is expressed as %V, and a = 14, the final equation is:

$N=14/\%V$ (2).

It is very likely that this type of distribution reflects the fundamental functional organization of the human cellular proteome since it is the same in all types of cells analyzed.

Keywords

Protein Species/Proteoform; Abundance; 2DE; Proteome; ESI LC-MS/MS

Abbreviations

HCD-Higher Energy Collisional Dissociation; ABC-Ammonium Bicarbonate; ACN-Acetonitrile; PTM-Post-Translation Modification; emPAI- Exponential Modified Form of Protein Abundance Index; C-HPP-Chromosome-Centric Human Proteome Project; FASP-Filter-Aided Sample Preparation; CBB-Coomassie Brilliant Blue

Introduction

After completion of the human genome sequencing and determination of its size, there is a great demand for similar information about the human proteome as proteins mediate almost all processes in a cell. To better understand the functionality of proteins, we need the information about their activity that is directly linked to their abundance. However, the situation is not simple here because of the complexity of proteins themselves. This complexity may arise from allelic variations, alternative splicing of RNA transcripts, and post-translational modifications. All these cellular events create distinct protein molecules, proteoforms/protein species, that modulate a wide variety of biological processes [1,2]. Apparently, by using standard technologies, it has been impossible so far to identify and calculate all protein species/proteoforms present in a single human cell or in human plasma [1,3]. The main problem is a huge dynamic range of concentrations, where the number of copies of different protein species in an object lies in the range from one to a billion molecules. One of quantitative proteomic approaches, a proteomic technique that is mainly performed using 2DE or liquid chromatography-tandem mass spectrometry (LC-MS/MS) is expected to offer an alternative solution this problem [4–6]. Recently, using a shotgun approach, a large amount of information about protein abundance was produced [7–9]. This information is still not enough as we still need to know how many specific molecules (protein species/proteoforms) are present in a cell. In earlier time, to estimate the number of protein species in the human proteome, we have developed and applied a method of extrapolation using 2DE gel-staining method with protein dyes of different sensitivities [10,11]. As we have discussed in these papers, this extrapolation was possible because the abundance distribution of proteoforms inside the cell follows a special formula [10,11]. At present, we have progressed further and performed multiple calculations of the data produced from several types of human cells using protein staining and mass spectrometry analysis, which allowed us to develop a formula. According to these data, the dependency of the number of protein species/proteoforms on their abundances in a cell is not normal but follows Zipf's law (1). This law is a popular member of a family of related discrete power law probability distributions, which approximates many types of data collected from very different areas of study on scaling behavior. In the present paper, we once more confirm the universality of Zipf's law in human proteome.

Experimental Section

Chemicals and Materials

All reagents used were sourced from Sigma-Aldrich Corp. (St. Louis, MO, USA), unless another manufacturer is specified. The remaining reagents were obtained from the following companies: Thermo Scientific Pierce Protein Research Products, (Rockford, IL, USA): dithiothreitol (DTT), protease inhibitor cocktail; GE Healthcare (Pittsburgh, PA, USA): IPG DryStrip (gel strips), IPG-buffers, DryStrip-coating liquid, Coomassie Brilliant Blue (CBB) R350; Promega Corp., (Madison, WI, USA): Trypsin Gold; Bio-Rad Laboratories, Inc. (Hercules, CA, USA): Precision Plus Protein Dual Color Standards, molecular weight markers for protein electrophoresis; Biolot (St. Petersburg, Russia): RPMI-1640 medium and DMEM for cell growth, fetal calf serum; Orange Scientific (Braine-l'Alleud, Belgium): Carrel culture flasks.

Cell Culture and Culture Conditions

Human cells (hepatocellular carcinoma, HepG2) were cultured in medium (DMEM/F12 or RPMI-1640 supplemented with 10% fetal bovine serum (FBS) and 100 U/ml penicillin) under standard conditions (5% CO_2, 37°C) [10,12]. To prepare cell samples for protein extraction, the cells were detached using 0.25% Trypsin-EDTA solution, washed 3 times with PBS, and treated with lysis buffer [10,13]. Liver tissue samples were provided within the framework of collaboration on the C-HPP. Extraction was performed according to 2DE protocol described in [14].

Sample Preparation and Two-Dimensional Electrophoresis (2DE)

Samples were prepared as described previously [15,16]. Cells (~10^7) containing ~2 mg of protein, were treated with 100 μl of lysis buffer (7 M urea, 2 M thiourea, 4% CHAPS, 1% DTT, 2% ampholytes, pH 3-10, protease inhibitor mixture). Proteins were separated by isoelectric focusing (IEF) using DryStrips pH 3-11, 7 cm and 18 cm ("GE Healthcare") following the manufacturer's protocol. Samples in lysis buffer were mixed with rehydrating buffer (7 M urea, 2 M thiourea, 2 % CHAPS, 0.3% DTT, 0.5% IPG buffer, pH 3-11 NL, 0.001% bromophenol blue) in a final volume of 130 μl (150 μg of protein) for 7-cm strip or 300 μl (800 μg of protein) for 18-cm strips. Strips were passively rehydrated for 6 h at 4°C. IEF was performed on an IPGphor (GE Healthcare,) that was programmed as follows: the first step—500 V 7 h, the second step — gradient to 1000 V, 1 h, the third step — gradient to 10 000 V, 3 h, the fourth step — 10 000 V 4 h, temperature 20°C, and maintained at a voltage 500 V. After IEF, strips were soaked 10 min in the equilibration solution (50 mM Tris, pH 6.8, 6 M urea, 2% SDS and 30 % glycerol) with 1% DTT. This process was followed by 10-min incubation in the equilibration solution containing 5% (w/v) iodoacetamide. The strips were placed on top of the 12 % polyacrylamide gel of the second direction, sealed with a hot solution of 0.5 % agarose prepared in electrode buffer (25mM Tris, pH 8.3, 200 mM glycine, 0.1% SDS), and electrophoresed in the second direction under denaturing conditions using the Hoefer miniVE system (gel size 80 x90x1mm, "GE Healthcare") or Ettan™ DALT six (180x200x1mm, "GE Healthcare"). Electrophoresis was carried out at room temperature at a constant power of 3 W/gel [16,17].

Mass Spectrometry

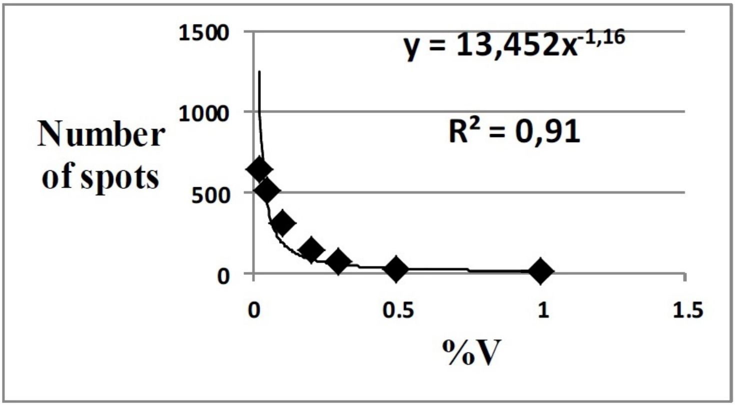
All procedures with gel plugs were performed according to the protocol described previously [12,18,19]. Gel-free sample treatment was performed according to FASP assay [20]. Proteolysis was performed by incubation with trypsin ("Trypsin Gold", 10 μg/ml) at least 4 h at 37°C. Tryptic peptides were dissolved in 5% (v/v) formic acid. Using an Agilent HPLC system 1100 Series (Agilent Technologies), 4 μg of peptides were injected onto a trap column Zorbax 300SB–C18, 5 ×0.3 mm (Agilent Technologies). After washing with 5% ACN containing 0.1% formic acid, peptides were resolved on a 150 mm x 75 μm Zorbax 300SB-C18 reverse phase analytical column (Agilent Technologies) using a 30-min organic gradient of 5-60% ACN, 0.1% formic acid with a flow rate of 300 nL/min. Peptides were then ionized by nano-electrospray at 2.0 kV using a fused silica emitter with an internal diameter of 8 μm (New Objective). MS/MS analysis was carried out in duplicate on an Orbitrap Q-Exactive Plus (Thermo Scientific). Mass spectra were acquired in the positive ion mode. High resolution data was acquired with a resolution of 30 000 (m/z 400) for MS and 7500 (m/z 400) for MS/MS scans. Survey MS scan was followed by MS/MS spectra of five the most abundant precursors. For peptide fragmentation, higher energy collisional dissociation (HCD) was set to 35 eV, the signal threshold was set to 5000 for an isolation window of 2 m/z, and the first mass of HCD spectra was set to 100 m/z. Fragmented precursors were dynamically excluded from targeting for 90 s. Singly charged ions and ions with unassigned charge state were excluded from triggering MS/MS scans. The automatic gain control target value was regulated at 1×10^6 with a maximum injection time of 100 ms and at 1×10^7 with a maximum injection time of 250 ms for MS and MS/MS scans, respectively. The data were searched by Mascot "2.4.1" search engine (www.matrixscience.com) using the following parameters: enzyme = trypsin (allowing for cleavage before proline); maximum missed cleavages = 2; fixed modifications = carbamidomethylation of cysteine; variable modifications = oxidation of methionine; phosphorylation of serine and threonine, acetylation of lysine; precursor mass tolerance = 20 ppm; product mass tolerance = 0.01 Da. NeXtProt database (October 2014) was used as a protein sequence database. For FDR assessment, a separate decoy database was generated from the protein sequence database. False-positive rate of 1% was allowed for protein identification. These parameters have previously been shown to be adequate to identify true positive matches [21]. Exponentially modified PAI (emPAI) defined as the number of identified peptides divided by the number of theoretically observable tryptic peptides for each protein was used to estimate protein abundance [22,23].

Results and Discussion

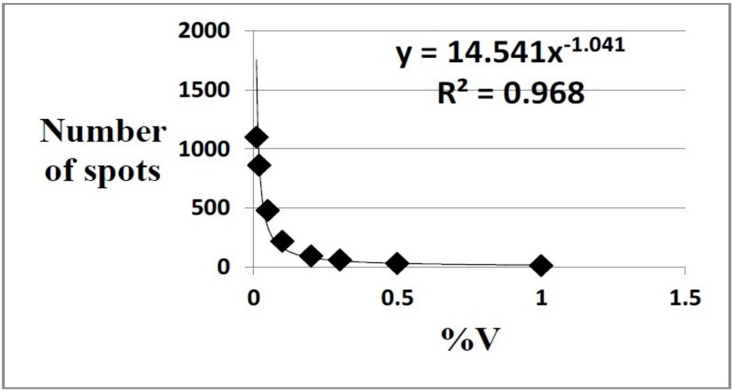
Analysis was performed on protein extracts from HepG2 cells, glioblastoma cells, and the primary hepatocytes (liver). After separation of protein on 2DE, these gels were stained with CBB R350, and each 2DE picture produced was analyzed using ImageMaster 2D Platinum 7.0 (GE Healthcare). All protein spots were quantified and counted. The relative protein abundance (%V) of each spot was calculated according to the staining intensity of the spot. Protein spots were grouped thereafter according to their %V. The first group included those spots with relative abundance of 1% or *greater*, the second — ≥ 0.5%, the third — ≥ 0.3%, the fourth —- ≥ 0.2%, the fifth — ≥ 0.1%, the sixth — ≥ 0.05%, the seventh — ≥ 0.02%, the eighth — ≥ 0.01%. Using Excel, the numbers of spots in each group (N) were plotted against relative abundance (%V), dot graphs were created, and the line of best fit was chosen. In all cases, among available trends, the power function was the most appropriate one with a very high reliability (coefficient of determination R^2 was from 0.92 to 0.98) (Figure 1). It is of interest that the situation was quite similar not only for all types of cells analyzed by our group, but also for the different types of cells (MCF7) studied by another group (24). We have taken the data from this publication [24] and established a curve in the same way. From Figure 1, it is evident that the line of best fit follows the power function or the Zipf's law in particular

$y = ax^{-1}$ (1), where y denotes N (the number of protein spots), x – %V (protein spot abundance), and a = 14. The final equation is

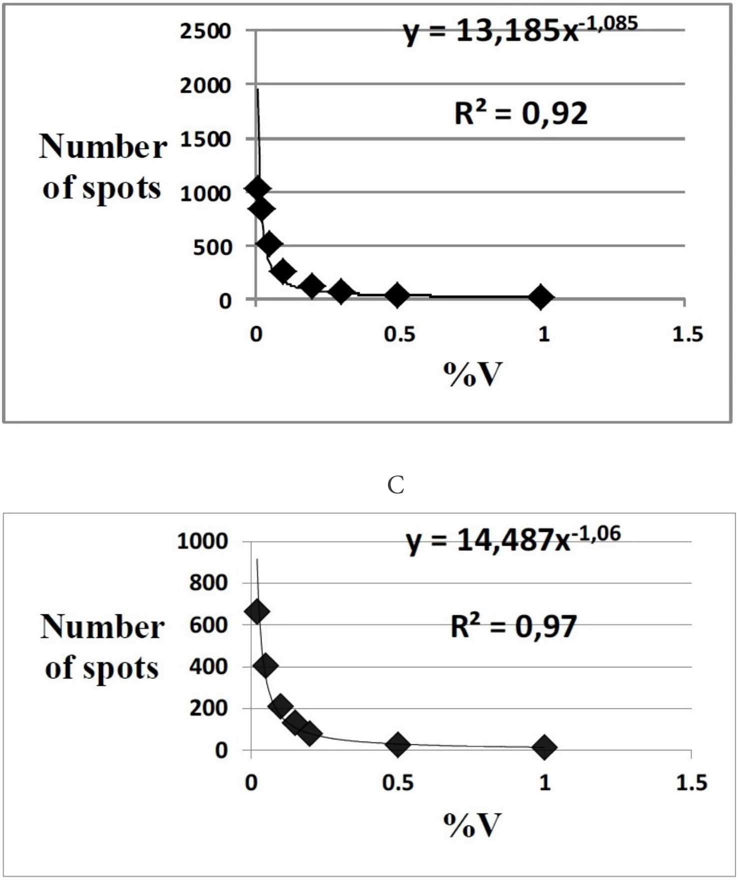
$N=14/\%V$ (2).


1500
1000
500
0
0
0.5
1
1.5
y = 13,452x-1,16
R² = 0,91
Number of spots
%V

A


2000
1500
1000
500
0
0
0.5
1
1.5
y = 14.541x-1.041
R² = 0.968
Number of spots
%V

B

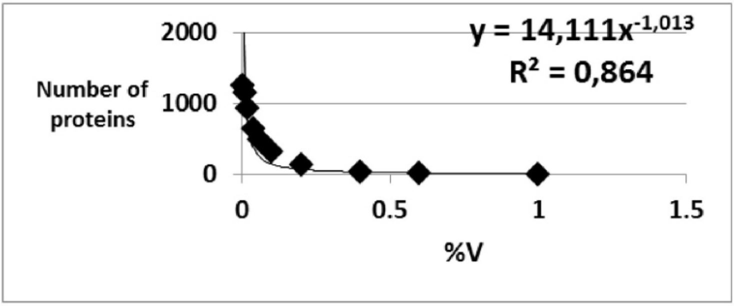


C

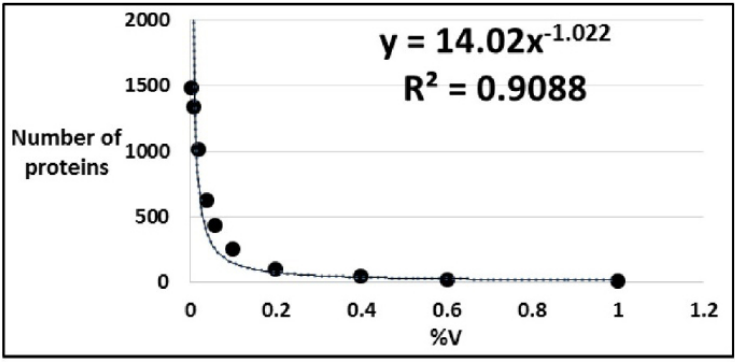
D

Figure 1: Dependency of the number of 2DE protein spots on their abundances (normalization in %V). After 2DE separation, the gels were stained with CBB R350 and analyzed by ImageMaster 2D Platinum software (GE Healthcare). Spots were counted and quantitated. A – liver cells (see Figure 1 in Ref [34]), B – HepG2 (the data adopted from [19]), C – glioblastoma cells (the data adopted from [12]), D – MCF7 breast cancer cells (the data adopted from [24].

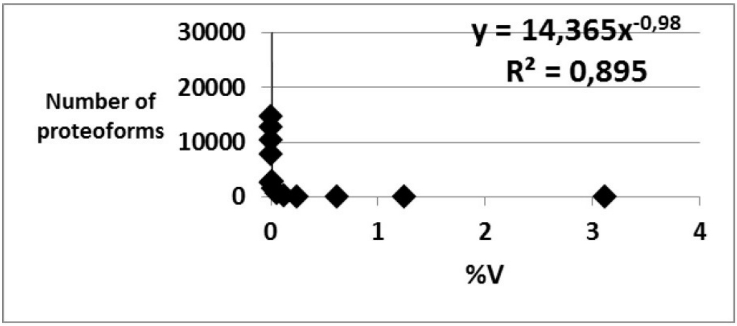
It needs to be mentioned that according to 2DE protein separation principles, each protein spot should represent a specific protein species/proteoform ideally. In practice, the situation is more complicated because a single spot may accommodate numerous different proteins and proteoforms [18,25]. To circumvent this problem in our curve fitting, we assumed a single spot to contain a single protein species, especially if one protein species is dominant and represents the major volume of the spot (at least 70%). Actually, a case like this is frequently observed in our study [18,25]. An alternative and reliable way to evaluate proteoforms is to combine 2DE with ESI LC-MS/MS which makes precise evaluation of proteoforms possible. Thus, we have applied this technique in our current curve fitting analysis [18,25,26], in which we first performed a typical shotgun proteomics experiment, where the cellular extract was trypsinized, the peptides obtained were analyzed by ESI LC-MS/MS (Figure 2A,B), and proteins instead of proteoforms were quantitated using emPAI [18]. Having performed these experiments, this parameter is not necessarily very precise for measuring the abundance of individual proteins in our opinion, it should be quite reliable in estimations among large scale proteomics projects instead [23]. Since the quantitation was done using emPAI, we normalized emPAI to %V units. To accomplish that, the sum of all emPAIs was divided by 100%. We estimated that 5 emPAI corresponds here to 1%V. We established another curve and found that the line of the best fit again follows the Zipfian distribution (2) (Figure 2A, B). A possible weakness of the spot analysis might be the number of detected proteins used in the calculations and curve building. Thus far, this number has been slightly higher than 1000 (Figure 1, Figure 2A, B). Bearing in mind that the cellular proteome may possibly contain at least 70 000 protein species/proteoforms [11,18], it would be more accurate to have quantitative proteomic data for a larger number of proteoforms using label-free and isotope labeling-based approaches. To achieve this objective, we separated proteins by 2DE again and then performed the mass spectrometry analysis of the whole gel by cutting it into small sections as described in [12,19]. In this case, if the same protein was identified in different sections, it was considered to be different protein species/proteoforms. As a result of 2DE separation, we identified nearly 20 000 proteoforms. Once again, we normalized emPAI to %V units. Interestingly enough, the estimated emPAI/%V ratio was much bigger in this case. For instance, 400 emPAI corresponds to 1%V in the HepG2 or glioblastoma analysis, and 100 emPAI — in the analysis of liver proteins. Following the same way as we did with the spots, the protein species were grouped step by step according to their normalized %V, and the curves were established using Excel (Figure 2C-E). Amazingly, the line of the best fit in all cases follows the Zipfian distribution closely (2). Accordingly, the equation (2) is named "*the first equation of the human proteome*", as it appears to give us a shared and common overview of the human proteome across several cell types that we have analyzed. Given that, we still need more data about protein species/proteoforms of low abundance that escaped detection and analysis. Complete information about all protein species/proteoforms in a cell would allow us to build the final "first equation".



A


y = 14.02x-1.022
R² = 0.9088
Number of proteins
2000
1500
1000
500
0
0
0.2
0.4
0.6
0.8
1
1.2
%V

B


y = 14,365x-0,98
R² = 0,895
Number of proteoforms
30000
20000
10000
0
0
1
2
3
4
%V

C

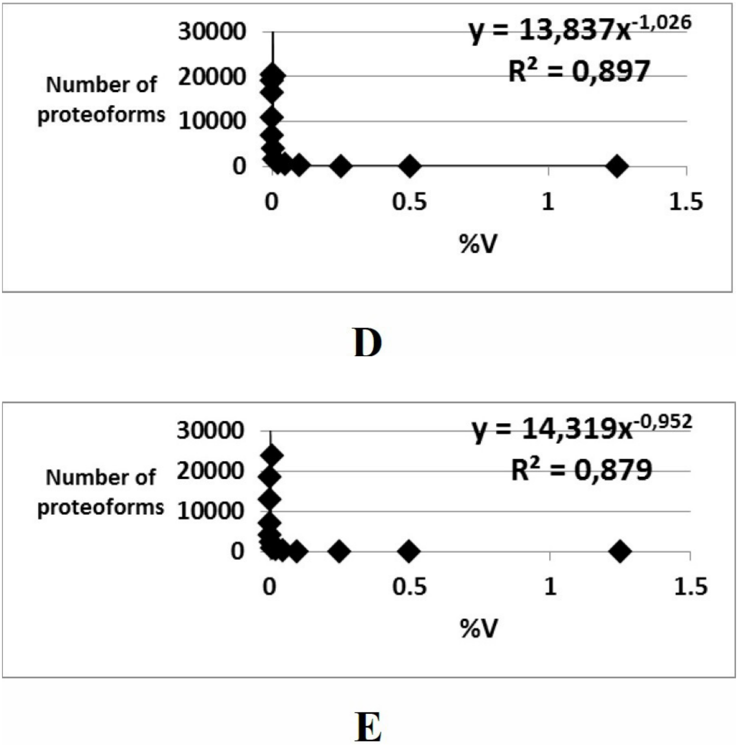


Figure 2: Dependency of the number of proteins (A, B) or protein species/proteoforms (C-E) on their abundances (normalization in %V). After 2DE separation (C-E), the gels were stained with CBB R 350 and cut into 96 sections. Each section was treated and analyzed by ESI LC-MS/MS. All proteoforms were counted and quantitated. A – liver cells (the analysis of the whole extract by ESI LC-MS/MS without 2DE, protein normalization 5 emPAI = 1%V) (See Table 1 in Ref [35]). B – HepG2 cells (the analysis of the whole extract by ESI LC-MS/MS without 2DE, protein normalization 5 emPAI = 1%V) (see Table 1 in Ref [35]). C – liver cells (normalization 100 emPAI = 1%V) (see Table 2 in Ref [34]), D – HepG2 cells (normalization 400 emPAI = 1%V), the data adopted from [19], E – glioblastoma cells (normalization 400 emPAI =1%V), the data adopted from [12].

Concluding Remarks

The power-law distributions have been identified in physics, biology, and the social sciences [27]. One of a family of related discrete power law probability distributions is a Zipfian distribution or Zipf's law. This law states, in particular, that the frequency of any word in a language is inversely proportional to its rank in the frequency table. For example, in the Brown Corpus of American English texts, consisting of over one million words, only 135 words represent half of the word volume [28,29]. The most frequent word "the" occurs here approximately twice as often as the second most frequent word "and", three times as often as the third most frequent word "to", etc. The same relationship occurs in many other rankings unrelated to language, such as the size of cities in various countries, corporation sizes, income rankings [30,31]. The formula (1) may be shown as:

$S_R = S_1/R$ (3),

where S – size of an organization (city, corporation, income, etc.), R – rank of organization. As we have already shown, when we are talking about the distribution of protein species, we practically deal with the same formula:

$N=14/\%V$ (2). Only instead of size (S_R) we have here a number (N), R is not a rank but protein species abundance (%V), S_1 is equal to 14 (14 is the number of most abundant protein species with $\%V \geq 1$). As Zipf's law is so popular, it is reasonable to think that there is a universal origin for such a distribution in nature. Importantly, gene expression data was previously found to also obey Zipf's law [32,33]. So far, people can only hypothesize about the general ubiquity of Zipfian distribution [34], but in the case of the human proteome, we can say that this kind of distribution is a reflection or a result of functionality of different protein species and their abundance inside the proteome. On the one hand, a human cell needs a high copy number (millions) of only a few protein species (like actin or tubulin) for its structural organization, *e.g.* in the cytoskeleton. But on the other hand, only a few copies each of many thousands of protein species are involved in such processes as signaling or protein turnover. In summary, our analysis provides the first quantitative overview of protein species/proteoforms in the human cellular proteome. Therefore, we think that the equation of the Zipfian distribution that we identified reflects a fundamental functional organization of the human cell proteome.

Acknowledgements

The study was performed in frames of Program of Fundamental Research of State Academies of Sciences for 2013-2020. We acknowledge the IBMC "Human Proteome" Core Facility for assistance with the generation of mass-spectrometry data. Carita Lanner is acknowledged for the editing assistance.

References

1. Smith LM, Kelleher NL. Proteoform: a single term describing protein complexity. Nat Methods. 2013; 10: 186–187.
2. Schlüter H, Apweiler R, Holzhütter H-G, Jungblut PR. Finding one's way in proteomics: a protein species nomenclature. Chem Cent J. 2009; 3: 11.
3. Toby TK, Fornelli LKN. Progress in Top-Down Proteomics and the Analysis of Proteoforms. Annu Rev Anal Chem (Palo Alto Calif). 2016; 9: 499–519.
4. Ong S-E, Mann M. Mass spectrometry-based proteomics turns quantitative. Nat Chem Biol. 2005; 1: 252–262.
5. Bantscheff M, Schirle M, Sweetman G, Rick J, Kuster B. Quantitative mass spectrometry in proteomics: a critical review. Anal Bioanal Chem. 2007; 389: 1017–1031.
6. Nikolov M, Schmidt C, Urlaub H. Quantitative mass spectrometry-based proteomics: an overview. Methods Mol Biol. 2012; 893: 85–100.
7. Weiss M, Schrimpf S, Hengartner MO, Lercher MJ, Von Me-

ring C. Shotgun proteomics data from multiple organisms reveals remarkable quantitative conservation of the eukaryotic core proteome. Proteomics. 2010; 10: 1297–1306.

8. Wang M, Herrmann CJ, Simonovic M, Szklarczyk D, von Mering C. Version 4.0 of PaxDb: Protein abundance data, integrated across model organisms, tissues, and cell-lines. Proteomics. 2015; 15: 3163–3168.
9. Schwanhausser B, Busse D, Li N, Dittmar G, Schuchhardt J, et al. Global quantification of mammalian gene expression control. Nature. 2011; 473: 337–342.
10. Naryzhny SN, Lisitsa AV, Zgoda VG, Ponomarenko EA, Archakov AI. 2DE-based approach for estimation of number of protein species in a cell. Electrophoresis. 2014; 35: 895–900.
11. Naryzhny SN, Zgoda VG, Maynskova MA, Ronzhina NL, Belyakova NV, et al. Experimental estimation of proteome size for cells and human plasma. Biomed Khim. 2015; 61: 279–285.
12. Naryzhny SN, Maynskova MA, Zgoda VG, Ronzhina NL, Novikova SE, et al. Proteomic profiling of high-grade glioblastoma using virtual-experimental 2DE. J Proteomics Bioinform. 2016; 9: 158–165.
13. Shtam TA, Naryzhny SN, Landa SB, Burdakov VS, Artamonova TO, et al. Purification and in vitro analysis of exosomes secreted by malignantly transformed human cells. Cell Tissue Biol. 2012; 6: 317–325.
14. Zabel C Klose J. Protein Extraction for 2DE. Methods Mol Biol. 2009; 519: 171-196.
15. Naryzhny SN, Lee H. Proliferating cell nuclear antigen in the cytoplasm interacts with components of glycolysis and cancer. FEBS Lett. 2010; 584: 4292–4298.

16. Naryzhny SN. Blue Dry Western: Simple, economic, informative, and fast way of immunodetection. Anal Biochem. 2009; 392: 90–95.

17. Naryzhny SN. Upside-down stopped-flow electrofractionation of complex protein mixtures. Anal Biochem. 1996; 238: 50–53.

18. Naryzhny SN, Zgoda VG, Maynskova MA, Novikova SE, Ronzhina NL, et al. Combination of virtual and experimental 2DE together with ESI LC-MS/MS gives a clearer view about proteomes of human cells and plasma. Electrophoresis. 2016; 37: 302–309.

19. Naryzhny SN, Maynskova MA, Zgoda VG, Ronzhina NL, Kleyst OA, et al. Virtual-Experimental 2DE Approach in Chromosome-Centric Human Proteome Project. J Proteome Res. 2016; 15: 525-530.

20. Wiśniewski JR, Zougman A, Nagaraj N, Mann M. Universal sample preparation method for proteome analysis. Nat Methods. 2009; 6: 359–362.

21. Larance M, Ahmad Y, Kirkwood KJ, Ly T, Lamond AI. Global Subcellular Characterization of Protein Degradation Using Quantitative Proteomics. Mol Cell Proteomics. 2013; 12: 638–650.

22. Ishihama Y, Oda Y, Tabata T, Sato T, Nagasu T, et al. Exponentially Modified Protein Abundance Index (emPAI) for Estimation of Absolute Protein Amount in Proteomics by the Number of Sequenced Peptides per Protein. Mol Cell Proteomics. 2005; 4: 1265–1272.

23. Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, et al. Protein abundance profiling of the Escherichia coli cytosol. BMC Genomics. 2008; 9: 102.

24. Hardouin J, Canelle L, Vlieghe C, Lasserre J-P, Caron M, et al. Proteomic analysis of the MCF7 breast cancer cell line.

Cancer Genomics and Proteomics. 2006; 3: 355–368.

25. Thiede B, Koehler CJ, Strozynski M, Treumann A, Stein R, et al. Protein species high resolution quantitative proteomics of HeLa cells using SILAC-2-DE-nanoLC/LTQ-Orbitrap mass spectrometry. Mol Cell Proteomics. 2012; 12: 529–538.
26. Naryzhny SN. Towards the Full Realization of 2DE Power. Review. Proteomes. 2016; 4: 33.
27. Andriani P, McKelvey B. Beyond Gaussian averages: redirecting international business and management research toward extreme events and power laws. J Int Bus Stud. 2007; 38: 1212–1230.
28. Fagan S, Gençay R. An introduction to textual econometrics. In: Handbook of Empirical Economics and Finance. 2010; 133–153.
29. Moreno-Sánchez I, Font-Clos F, Corral Á. Large-scale analysis of Zipf's law in English texts. PLoS One. 2016; 11.
30. Kirby G. ZlPF'S LAW. UK J Nav Sci. 1985; 10: 180–185.
31. Jiang B, Jia T. Zipf's Law for All the Natural Cities in the United States: A Geospatial Perspective. Int J Geogr Inf Sci. 2010; 10.
32. Furusawa C, Kaneko K. Zipf's law in gene expression. Phys Rev Lett. 2003; 90: 88102.
33. Kuznetsov VA, Knott GD, Bonner RF. General statistics of stochastic process of gene expression in eukaryotic cells. Genetics. 2002; 161: 1321–1332.
34. Kawamura K, Hatano N. Universality of Zipf's Law. J Phys Soc Japan. 2002; 71: 1211–1213.
35. Naryzhny S, Maynskova M, Zgoda V, Archakov A. Data set of protein species from human liver. Data Brief. 2017; 12: 584–588.