Journal of
Biomolecular Research & Therapeutics

Lisitsa et al., J Biomol Res Ther 2015, 4:4
DOI: 10.4172/2167-7956.1000132

Research Article Open Access

The Width of the Human Plasma Proteome Compared With a Cancer Cell Line and Bacteria

Andrey V Lisitsa*, Ekaterina V Poverennaya, Elena A Ponomarenko and Alexander I Archakov

Institute of Biomedical Chemistry, 119121, Pogodinskaya Street, 10, Moscow, Russian Federation, Russia

Abstract

Whole genome sequencing has revealed the number of protein-encoding genes in a given organism, which can be considered a first approximation of molecular complexity. Due to post-transcriptional and post-translational modifications such as RNA splicing, polymorphisms, covalent modifications and degradation, the total number of different protein species (the proteome) can be much larger than the number of protein-encoding genes. 2-D gel electrophoresis can be used to estimate the width of the human proteome. The number of spots obtained with different stains (dyes) under different protein loading conditions can give a rough idea of the number of different proteins in the sample. Data on human plasma and cell lines and on bacterial cells have been investigated to determine the dependence of the number of spots on the dye sensitivity. Assuming that each spot represents a different protein species, the spots-to-sensitivity dependence was applied as an estimate of the width of the proteome. In theory, there are 1.75 million proteoforms in 1 L of blood plasma, 18 thousand species per individual HepG2 cell, and 6700 species per bacterium.

Keywords: 2-DE; Sensitivity; Proteoforms; Human proteome project

Abbreviations

AB: Amido Black; Cy5min: Cy5 Minimal Labeling; Cy5sat: Cy5 Saturated Labeling; PTM: Post-Translation Modification

Introduction

Since completion of the Human Genome Project determined the number of protein-coding genes, the quest is on to derive similar information regarding the human proteome [1–4]. In this study, we consider the entire complement of protein species (proteoforms) in a given cell or organism as the proteome width.

Assuming the 'one gene one protein" mantra, there should be >20,000 human proteins [5,6]. However, in reality the situation is far more complex, and some estimates suggest there may be 100 protein variants from a single protein-coding gene [7,8]. Variations can include single amino acid substitutions (SAPs) derived from non-synonymous SNPs, translation of the alternatively spliced transcripts (AS), and post-translationally modified forms (PTM) [7]. Inventory of protein diversity by mass-spectrometry was coined as population proteomics over 5 years ago [9].

Proteome width is currently investigated using shotgun mass-spectrometry, and this identified up to 8000 different species in one cell line [10]. For blood plasma the benchmark of 2000 identified proteins was recently achieved [11] but this is evidently just the tip of the iceberg, and will certainly be expanded when identification of multiple proteoforms becomes commonplace [12]. The term "protein species" has been used traditionally [13,14], but this is being replaced by "proteoforms" in the top-down mass-spectrometry community [15,16]. It is still not known exactly how many proteoforms are present in a given biological sample, but estimations based on theory range from 10^4 to 10^6 species (Supplementary Figure S1). We tried to address this question experimentally by investigating the ability of 2-D gel electrophoresis to evaluate the proteome width. The number of spots should be proportional to the proteome width, and in our approach. We exploited this proportionality to assign the width of the proteome of essentially different specimens.

For estimation of proteome width, 2-D electrophoresis (2-DE) is more appropriate method. In contrast to conventional gel-free proteomic approaches. 2-DE allows detection of AS, SAPs and PTMs, which can all affect the protein properties. Existing bottom-up mass spectrometry methods cannot achieve this aim, while a top-down approach is still challenging [12].

Materials and Method

In order to investigate proteome width, we used data from previous studies on blood plasma [3], tissues and cells [17]. From these studies the sensitivity for different staining methods was assigned and a number of 2-D gels were produced using different dyes and varying amounts of protein. The proteome width was considered in two steps: (1) determination of the number of protein spots on a 2-DE image to assess the sensitivity of different staining dyes, and (2) extrapolation of the spots-to-sensitivity function to estimate number of spots the highest theoretical sensitivity.

The actual sensitivity was assigned to different staining methods by preparing a dilution series (BSA was used to estimate sensitivity; Figure 1a) and determining the lowest detectable concentration (Supplementary Table S1). The response (Z) is then estimated for each staining method (Figure 2b). The number of protein spots is plotted as a function of the amount of protein loaded on the gel; more protein loaded equates to more spots present, up to saturation levels after which gels become overloaded. The tilt angle is different depending on the dye; therefore the dyes were characterized using the following formula:

***Corresponding author:** Andrey V Lisitsa, Institute of Biomedical Chemistry, 119121, Pogodinskaya street, 10, Moscow, Russian Federation, Russia,Tel: (+7)499 2463731; E-mail: lisitsa063@gmail.com

Received July 17, 2015; **Accepted** October 31, 2015; **Published** November 11, 2015

Citation: Lisitsa AV, Poverennaya EV, Ponomarenko EA, Archakov AI (2015) The Width of the Human Plasma Proteome Compared With a Cancer Cell Line and Bacteria. J Biomol Res Ther 4: 132. doi:10.4172/2167-7956.1000132

Z = number of spots (#) / amount of protein (ng)

Data from different staining methods were then placed on a single plot (Figure 1c). Experimental data were used to derive the dependency of response (Z) to the sensitivity of the staining method. The dependency can then be extrapolated to hypothetical detection limits (e.g., one molecule in 1 L of blood, or one molecule per cell) by multiplying the response to the total amount of protein in a given volume of blood plasma or cell. As exemplified in Figure 1c, given a particular value of detection limit (DL) from the sensitivity axis, if DL_1 = one molecule per cell, then the following formula can be applied:

proteins = Z (DL_1)*Q,

where Q is the total amount of protein in the cell. To calculate Z, a dynamic range of five orders of magnitude was covered using selected gel staining methods (Supplementary Table S1). Assuming that a dye with comparatively higher sensitivity develops more protein spots, we explored the spot-to-sensitivity dependency of different biomaterials.

Results and Discussion

In previous work we demonstrated how to calculate the number of protein species in human blood plasma [3]. The number of protein spots can be shown as a function of the total amount of protein applied to the 2-DE gel (Figure 2a; [3]). This same function was applied for each dye, and regions corresponding to optimal gel loading were approximated from the linear trend. This method is now generalized and was applied to data on plasma, as well as data for HepG2 and bacterial cells [17].

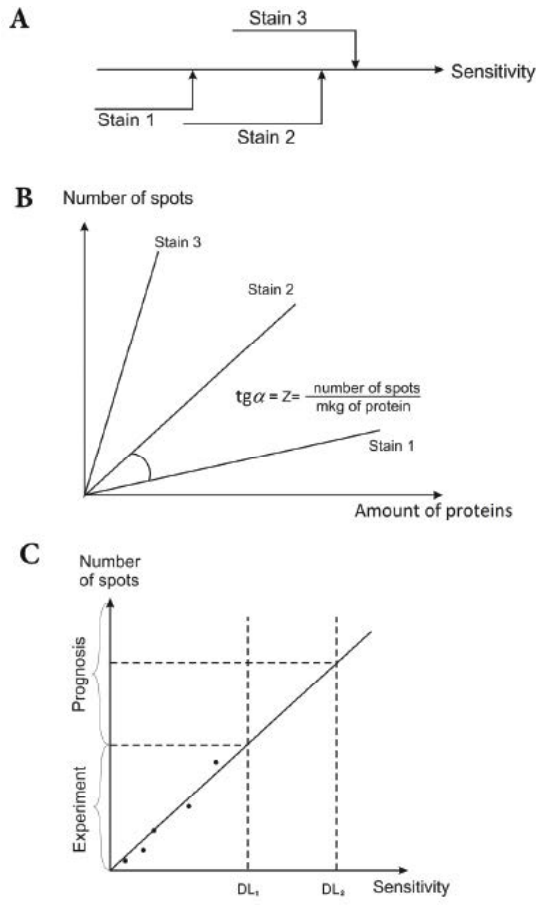

Figure 1 (a,b,c): Estimating the proteome width: (a) determination of the analytical sensitivity for the stains; (b) the number of spots as a function of protein load on gel and (c) spots-to-sensitivity function, DL – detection limit.

The spot-to-sensitivity function was used to estimate the number of protein species that could be detected, given unlimited sensitivity. The theoretically feasible maximum sensitivity could be either one molecule per 1 L of blood plasma (the reverse Avogadro number [2]), or 1 molecule per average cell for bacteria or HepG2 cells.

This dependence (Figure 2a) can then be used to compare the experiments at fixed amounts of protein in the sample. Generally, the number of spots is a function of two parameters; sensitivity and amount of material, and the dependency on sensitivity can be assigned by attributing a value for the amount of protein. The number of spots produced by different dyes can then be compared for a fixed amount of total protein on the gel. Less sensitive dyes such as Amido Black (AB) and Coomassie (CBB) produce only one or two spots for 1 μg of loaded total protein, whereas more sensitive dyes such as silver anhydride (ST) can give 12 spots, and fluorescent dyes such as Cy5 and Cy5-sat can give even more.

Substituting x=1 μg in the equations in Figure 2, the number of spots were plotted as a function of the dye sensitivity (Figure 2b). Each experimental point in the figure corresponds to a particular dye, and they are well approximated by the exponential dependency, which gives a straight line on a double log plot (R^2=0.93).

Firstly, linear regression was used to approximate the number of protein species in blood plasma (Figure 2b). For a sensitivity of one molecule per 1 μL (10^{-18} M), blood plasma could yield 14,500 spots (the benchmark of 10^{-18} M is used as a lowermost clinically relevant value, enabling the detection of a biomarker shed from a cancerous focus of less than 1 mm in diameter [18]). At the present time, high resolution 2-DE of plasma combined with pre-separation and sample depletion can only resolve 400 spots [19]. This discrepancy means that many potential protein biomarkers are yet to be identified and therefore cannot yet be exploited [20].

The dependency (Figure 2b) was also extended to the lowermost detection limit, known as reverse Avogadro's number [2]. The physiological role of ultra-rare protein species present at only one molecule per 1 L is unknown at present, but may well be physiologically relevant. At a detection limit of 10^{-24} M the dependency shown in Figure 2b could achieve 1.75 million different protein species. This value matches closely to the total number of modified and unmodified protein species annotated in the NextProt database (~ 1.8 million, not including somatic mutations) [21].

The experiments with the eukaryotic HepG2 and bacterial cells were performed following the same protocol as was used for blood plasma [17], and the data on three different types of biomaterial were acquired (Figure 2c; Supplementary Table S2). The dependency was higher for HepG2 cells than for plasma, and even coefficient was 0.75 for plasma, 1 for HepG2, and 1.5 for bacterial higher still for bacterial cells. Therefore spot-to-sensitivity function appeared to be specific for the type of biomaterial: the exponent power cells. Interestingly, this function was indistinguishable between analyzed bacterial species, *E. coli* and *P. furiosus*.

From the trends in Figure 2c the proteome width can be probed. With plasma, at a sensitivity of 10^{-24} M, millions of protein species could be distinguished. However, approximating to the reverse Avogadro number seems meaningless for the cells due to their limited volume. The approximation to one molecule per cell is more meaningful, which

for a HepG2 cell with a 20 nm diameter was 10^{-12} M. From the equation in Figure 2c, the number of protein spots expected for a single HepG2 cell was 368 spots. However, in practice a population of the cells rather than a single cell is used for analysis [22]. Normalizing the response Z (Figure 1c) to 1 ng of total protein loaded on the gel corresponded to 10^3 HepG2 cells. Therefore to resolve a typical protein species from an average cell, which is one thousandth of the neighboring single cells, a sensitivity of 10^{-15} M (10^{-12} M diluted 10^3 times) should be approached. From the spot-to-sensitivity equation (Figure 2c), at this sensitivity an average HepG2 cell could generate 18,000 different protein spots.

The smaller volume of a bacterial cell means that a concentration of one molecule per bacterium is 10^{-9} M rather than 10^{-12} M for a HepG2 cell. However, many more bacterial cells are used to generate the same amount of protein sample. A sensitivity of 10^{-12} M is therefore appropriate for detecting one protein species in 1,000 bacterial cells. Applying this function (Figure 2c), at a sensitivity of 10^{-12} M, 6,900 spots could be generated from a typical bacterial cell.

Our estimate of 7,000–18,000 protein species per cell may be an underestimate, due to problems associated with 2-DE [23-26]. For instance, 2-D gels have limited resolution, and a single protein spot can contain up to 20 protein species [26]. Despite the drawbacks, our speculative estimates for the number of proteoforms appear to be reasonable due to choosing two straightforward correlations: the number of spots vs. amount of protein loaded and the number of spots vs. the sensitivity of the staining method. It should be emphasized that the proteome width can vary between different cells. Although confocal microscopy and other observational methods can investigate single cells [22], high-throughput proteomic approaches cannot operate at this level at the present time. The average cell is therefore used in proteomics, and this may be the average of thousands or even millions of cells [27]. The problem of proteome heterogenic is relevant to blood plasma as well. The proteome width of blood plasma is dependent on the minimal sample volume, which could represent the whole diversity of proteoforms.

There are evident objections to the approach presented herein. First, the total amount of proteins was simply estimated by some calculation and linear regression, while estimation is highly dependent on the process of different dye staining. The described method also cannot account for proteome dependence of growth stage or culture condition in mammalian or bacterial cells. However, the overall trend is captured in the experiments, compliant with the difference in the dynamic range of plasma, cell and bacteria [10,28,29].

To emphasize the problem let's look into the figure 3 borrowed from the publication of [30]. We see the normal distribution of the molecules versus their concentrations as a result of proteomic experiments. The figure looks quite comfortable, as observations of the molecules are compliant with the biochemistry view-style of measurements.

However, from the 2DE data presented in this article it is concluded that such distribution is false if we observe the individual molecules. Previously 2-DE was the main method of proteomics and then undeservedly forgotten so deprived of one shortage. It allows to observe proteins as separate proteoforms rather than a result of identification of the peptide mixture. That is pointed out by a thin line drawn over the picture at Figure 1.

Experiments reported herein just show that number of biomolecules is infinitely increasing, when we increase either sensitivity or selectivity or any other analytical parameter [31]. It is comparable to

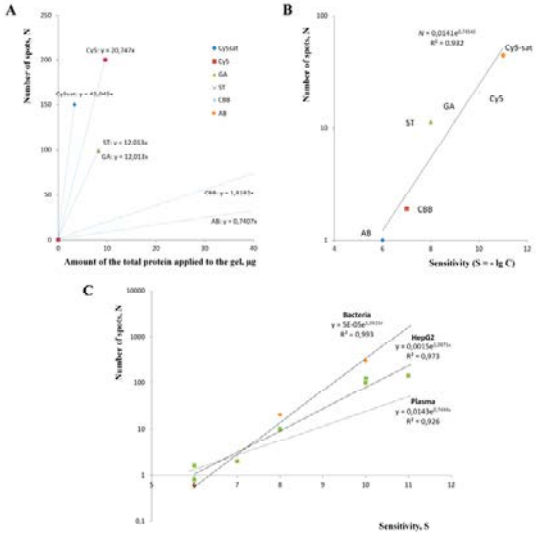

Figure 2 (a,b,c): Number of protein spots on 2-DE gel of blood plasma as a function of: (a) amount of the total protein applied to the gel: AB – Amido Black, CBB - Coomassie Brilliant Blue, ST – Silver Thiosulfate, GA - Glutaric Aldehyde, Cy5 – fluorescent labelling by Cyanines, Cy5-sat – Cyanines saturated and (B) sensitivity of the dye (double-log scale). S = - lgC, where C is the lowermost detectable concentration, M. (C) Number of spots as a function of the sensitivity of the staining dyes for HepG2, bacterial cells (datapoints were taken from [17]), and for human blood plasma (datapoints were taken from [3] S = -lgC, where C is a limit of detection of analytical method.

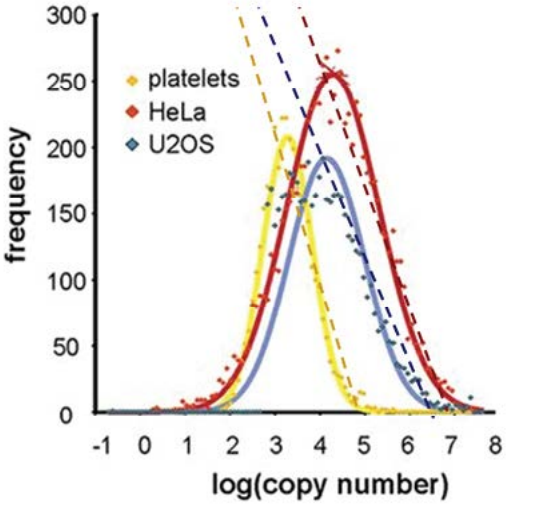

Figure 3: The distribution of copy numbers is similar for 3 proteomes, reproduced from Bukhart et al. [30]. Dashed lines show hypothetical approximation of actual proteome width.

the observation of stars and galaxies – whenever we construct next new telescope we see more objects [32].

That has an important consequence of the curves in Figure 2, but this consequence is not easily accepted from the scratch. We do not see individual molecules not because of the technical reasons of resolution or dynamic range, or whatever. Post genome molecular science should to accept, that we simply - do not know what if all of these molecules in one moment would become visible.

Acknowledgement

The work was done in the framework of the State Academies fundamental research program (2013-2020) and RSF grant (# 15-15-30041).

References

1. Wilkins MR, Appel RD, Van Eyk JE, Chung MCM, et al. (2006) Guidelines for the next 10 years of proteomics. Proteomics 6: 4–8.
2. Archakov AI, Ivanov YD, Lisitsa AV, Zgoda VG (2007) AFM fishing nanotechnology is the way to reverse the Avogadro number in proteomics. Proteomics 7: 4–9.
3. Archakov A, Ivanov Y, Lisitsa A, Zgoda V (2009) Biospecific irreversible fishing coupled with atomic force microscopy for detection of extremely low-abundant proteins. Proteomics 9: 1326–1343.
4. Archakov A, Zgoda V, Kopylov A, Naryzhny S, et al. (2012) Chromosome-centric approach to overcoming bottlenecks in the Human Proteome Project. Expert Rev Proteomics 6: 667–676.
5. Casado-Vela J, Lacal JC, Elortza F (2013) Protein chimerism: novel source of protein diversity in humans adds complexity to bottom-up proteomics. Proteomics 13: 5–11.
6. Collins FS, Lander ES, Rogers J, Waterson R (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931–945.
7. Khoury GA, Baliban RC, Floudas CA (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. Sci. Rep 90: srep00090
8. Roth MJ, Forbes AJ, Boyne MT, Kim Y-B, et al. (2005) Precise and parallel characterization of coding polymorphisms, alternative splicing, and modifications in human proteins by mass spectrometry. Mol Cell Proteomics : MCP 4: 1002–1008.
9. Nedelkov D (2008) Population proteomics: investigation of protein diversity in human populations. Proteomics 8: 779–786.
10. Geiger T, Wehner A, Schaab C, Cox J, Mann M (2012) Comparative proteomic analysis of eleven common cell lines reveals ubiquitous but varying expression of most proteins. Mol Cell Proteomics: MCP 11: M111.014050.
11. Tinoco AD, Kim Y-G, Tagore DM, Wiwczar J, et al. (2011) A peptidomics strategy to elucidate the proteolytic pathways that inactivate peptide hormones. Biochemistry 50: 2213–2222.
12. Jungblut PR (2014) The proteomics quantification dilemma. Journal of Proteomics 107: 98-102.
13. Jungblut P, Thiede B, Zimny-Arndt U, Müller E, et al. (1996) Resolution power of two-dimensional electrophoresis and identification of proteins from gels. Electrophoresis 17: 839–847.
14. Jungblut PR, Holzhütter HG, Apweiler R, Schlüter H (2008) The speciation of the proteome. Chemistry Central journal 2: 16.
15. Smith LM, Kelleher NL (2013) Proteoform: a single term describing protein complexity. Nature methods 10: 186–187.
16. Lisitsa A, Moshkovskii S, Chernobrovkin A, Ponomarenko E, Archakov A (2014) Profiling proteoforms: promising follow-up of proteomics for biomarker discovery. Expert review of proteomics 11: 121–129.
17. Naryzhny SN, Lisitsa AV, Zgoda VG, Ponomarenko EA, Archakov AI (2014) 2DE-based approach for estimation of number of protein species in cell. Electrophoresis 35: 895-900.
18. Hori SS, Gambhir SS (2011) Mathematical model identifies blood biomarker-based early cancer detection strategies and limitations. Science translational medicine 3: 109ra116.
19. Trifonova O, Larina I, Grigoriev A, Lisitsa A, et al. (2010) Application of 2-DE for studying the variation of blood proteome. Expert review of proteomics 7: 431–438.
20. Veenstra TD (2011) Where are all the biomarkers? Expert review of proteomics 8: 681–683.
21. Gaudet P, Argoud-Puy G, Cusin I, Duek P, et al. (2013) neXtProt: organizing protein knowledge in the context of human proteome projects. J Proteome Res 12: 293–298.
22. Wang D, Bodovitz S (2010) Single cell analysis: the new frontier in “omics”. Trends in biotechnology 28: 281–290.
23. Corthals GL, Wasinger VC, Hochstrasser DF, Sanchez JC (2000) The dynamic range of protein expression: a challenge for proteomic research. Electrophoresis 21: 1104–1115.
24. Gygi SP, Corthals GL, Zhang Y, Rochon Y, Aebersold R (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. Proc Natl Acad Sci U.S.A 97: 9390–9395.
25. Klose J (1999) Large-gel 2-D electrophoresis. Methods in molecular biology (Clifton, N.J.) 112: 147–172.
26. Thiede B, Koehler C, Strozynski M, Treumann A, et al. (2013) High resolution quantitative proteomics of HeLa cells protein species using stable isotope labeling with amino acids in cell culture(SILAC), two-dimensional gel electrophoresis(2DE) and nano-liquid chromatograpohy coupled to an LTQ-OrbitrapMass spectromet. Mol Cell Proteomics 12: 529–538.
27. Altelaar AFM, Heck AJR (2012) Trends in ultrasensitive proteomics. Current opinion in chemical biology 16: 206–213.
28. Magdeldin S, Yamamoto K, Yoshida Y, et al. (2014) Deep proteome mapping of mouse kidney based on OFFGel prefractionation reveals remarkable protein post- translational modifications. J. Proteome Res 13: 1636–1646.
29. Farrah T, Deutsch EW, Omenn GS, et al. (2014) State of the Human Proteome in 2013 as Viewed through PeptideAtlas: Comparing the Kidney, Urine, and Plasma Proteomes for the Biology- and Disease-Driven Human Proteome Project. J. Proteome Res 13: 60–75.
30. Burkhart JM, Vaudel M, Gambaryan S, Radau S, Walter U, et al. (2012) The first comprehensive and quantitative analysis of human platelet protein composition allows the comparative analysis of structural and functional pathways. Blood 120: e73–82.
31. Naryzhny SN, Zgoda VG, Maynskova MA, Novikova SE, Ronzhina NL, et al. (2015) A combination of virtual and experimental 2DE together with ESI LC-MS/MS gives a clearer view about proteomes of human cells and plasma. Electrophoresis Oct 2015.
32. Azmak O, Bayer H, Caplin A, Chun M, Glimcher P, et al. (2015) Using Big Data to Understand the Human Condition: The Kavli HUMAN Project Big Data 3: 173–188.