Clinical Chemistry 56:11
1775–1776 (2010)

Citation Classic

Counting the Proteins in Plasma

N. Leigh Anderson[1*]

Featured Article: Anderson NL, Polanski M, Pieper R, Gatlin T, Tirumalai RS, Conrads TP, et al. The human plasma proteome: a non-redundant list developed by combination of four separate sources. Mol Cell Proteomics 2004;3:311–26.[2]

The problem of enumerating protein components in plasma has challenged the best analytical technologies for more than 80 years. The first generation of proteomics methodologies (particularly 2-dimensional electrophoresis) had detected about 60 plasma proteins by the early 1990s *(1)*, and these proteins were more or less the same as those previously tabulated in Frank Putnam's reference books (*The Plasma Proteins*) and purified by the Behring Institute and others. Almost all are present at concentrations >1 μmol/L (roughly 50 mg/L). The development of sensitive specific immunoassays during this period, however, clearly demonstrated the presence of at least a few proteins at concentrations 1000- to 100 000-fold lower, raising the important question of how many proteins lay below the tip of the iceberg then visible to systematic proteome mapping.

When systematic protein methods began to improve about a decade ago, with mass spectrometry and genomic data combining to enable efficient identification of proteins on the basis of sequence, another expansion of the plasma proteome began. Numerous methodologies were being developed, and I thought it would be interesting to assemble a more comprehensive proteome by combining data produced with the new liquid chromatography–tandem mass spectrometry shotgun methods (provided by T.P. Conrads and T.D. Veenstra at the National Cancer Institute, and by J.N. Adkins and J.G. Pounds at Pacific Northwest National Laboratory), obtained by 2-dimensional electrophoresis with added fractionation (from R. Pieper and T. Gatlin, my colleagues from Large Scale Biology Corporation), and gleaned from the published literature (which M. Polanski and I searched). This exercise turned out to be a great deal more work than anticipated, primarily because the names assigned to proteins reported by different methodologies (typically database accession numbers in the mass spectrometry world and sometimes ambiguous biochemical names in the literature) did not mesh. Haptoglobin, for example, was reported under 9 different names among the 4 different data sources we combined. These ambiguities were conquered only by computational brute force: R. Fagan and A. Lobley, who were at a private genomics company in the UK, searched by means of the Basic Local Alignment Search Tool (BLAST)[3] the sequences of all the reported accessions against one another, and we lumped together anything with >95% sequence identity over 15 or more amino acid residues (collapsing all the immunoglobulins into a single cluster, for example). This effort boiled 1735 entries down to 1175 distinct proteins, or about 20-fold more than the previous era.

The real surprise came when we compared the data sets. Only 46 proteins occurred in all 4 (the typical "plasma proteins" from 2-dimensional gel days), whereas 980 proteins appeared in only 1 data set. This result was not the one we had hoped for. What it revealed, correctly as it turned out, was that different methods that had been regarded as fairly comprehensive in fact detect different sets of proteins in plasma. Subsequently, the Human Proteome Organization, better known as HUPO, carried out a larger study of plasma with more proteomics platforms and arrived at a very similar result *(2)*. This "sampling" effect, now a well-understood limitation of shotgun proteomics methods, limits the completeness of a plasma proteome observed by one approach.

Recent work with substantially improved analytical platforms capable of sampling down to approximately 100 pmol/L (approximately 5 μg/L) suggests that 1000–2000 different proteins can now be reliably detected in a single laboratory. This expansion is making a large difference with respect to biomarker discovery. Despite this progress, 3 frustrating limitations remain. First, the effort required to sample the plasma proteome to this depth (dividing it into many fractions analyzed separately) currently precludes running more than a handful of clinical samples. Although this approach helps biomarker discovery, it cannot satisfy the needs of clinical validation, which requires >1000 samples. Second, although many proteins are detected, their concentrations are measured only ap-

[1] The Plasma Proteome Institute, Washington, DC.
* Address correspondence to the author at: The Plasma Proteome Institute, P.O. Box 53450, Washington, DC 20009. E-mail leighanderson@plasmaproteome.org.
Received May 6, 2010; accepted May 25, 2010.
Previously published online at DOI: 10.1373/clinchem.2010.146167
[2] This article has been cited nearly 500 times since publication, according to Google Scholar.

[3] BLAST, Basic Local Alignment Search Tool; HUPO, Human Proteome Organisation.

proximately and not on any absolute scale. The true distribution of proteins as a function of concentration "depth" thus remains obscure. Third, despite intense efforts at collecting proteome data in giant digital repositories, a widely available curated reference plasma proteome has yet to emerge.

Fortunately, these limitations can be addressed through the development of an inverse brute force approach to the proteome: creation of specific assays for at least one form of each of the 20 000+ protein-coding human genes *(3)*. Completion of such a platform will finally provide a clear definition of what is in plasma and permit clinically useful measurement of the disease and population variation necessary to define a new generation of clinical diagnostics.

Author Contributions: *All authors confirmed they have contributed to the intellectual content of this paper and have met the following 3 requirements: (a) significant contributions to the conception and design, acquisition of data, or analysis and interpretation of data; (b) drafting or revising the article for intellectual content; and (c) final approval of the published article.*

Authors' Disclosures of Potential Conflicts of Interest: *No authors declared any potential conflicts of interest.*

Role of Sponsor: The funding organizations played no role in the design of study, choice of enrolled patients, review and interpretation of data, or preparation or approval of manuscript.

References

1. Anderson NL, Anderson NG. A two-dimensional gel database of human plasma proteins. Electrophoresis 1991;12:883–906.
2. Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, et al. Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. Proteomics 2005;5:3226–45.
3. Anderson NL, Anderson NG, Pearson TW, Borchers CH, Paulovich AG, Patterson SD, et al. A human proteome detection and quantitation project. Mol Cell Proteomics 2009;8:883–6.