Macedonian Journal of Chemistry and Chemical Engineering, Vol. **27**, No. 2, pp. 99–106 (2008)
MJCCA9 – 520 ISSN 1857 – 5552
Received: October 1, 2008 UDC: 612.124:543.51
Accepted: November 11, 2008

Review

DISCOVERING HUMAN PROTEIN DIVERSITY

Dobrin Nedelkov

Institute for Population Proteomics and Intrinsic Bioprobes Inc.
2155 E. Conference Dr. Suite 104, Tempe, AZ 85284
dnedelkov@populationproteomics.org // dnedelkov@intrinsicbio.com

Current emphasis on discovering and correlating human genetic variations lays the foundation for future studies of human protein diversity. Protein posttranslational processing, along with the translation of genetic variations, results in a complex, variable human proteome. Analyzing these protein variations on a grandeur scale has become feasible with the advent of mass spectrometry. Mass spectrometry is the only detection method today that can universally provide information about specific protein structural modifications, without *a priori* knowledge of the modification. However, high-throughput separation approaches are needed to effectively prepare the proteins for mass spectrometric interrogation. Such are the immunoaffinity separations that target single proteins by using highly specific antibodies for their affinity retrieval from the biological fluids. The resulting combination of immunoaffinity separation with MALDI-TOF mass spectrometry, termed Mass Spectrometric Immunoassay (MSIA), has been recently applied in two large studies of protein diversity. The results of these studies reveal a human protein diversity that is far more complex than the variations observed at the genetic level. Assessing the human proteome variations among and within populations will be an important future undertaking with significant clinical and diagnostic implications.

Key words: protein diversity; population proteomics; protein isoforms; plasma; serum; mass spectrometry; immunoassay

ИСПИТУВАЊЕ НА ПРОТЕИНСКИОТ ДИВЕРЗИТЕТ КАЈ ЧОВЕКОВАТА ПОПУЛАЦИЈА

Откривањето и воспоставувањето корелација помеѓу човековите генетски варијации претставува добра основа за идни испитувања на протеинскиот диверзитет кај човековата популација. Протеинскиот диверзитет е резултат на посттранслациските процеси кај протеините, како и варијациите на генетско ниво. Опсежното испитување на овие протеински варијации станува возможно со примена на масена спектрометрија. Масената спектрометрија е единствениот метод за детекција кој може универзално да даде информација за одредени протеински структурни модификации, без претходно знаење за идентитетот на модификацијата. Меѓутоа, брзи и ефикасни сепарациски процеси се потребни за протеините соодветно да се подготват за масена анализа. Такви се имуноафинитетните раздвојувања кои се применуваат за изолација на протеини од биолошки примероци со помош на високо специфични антитела кои имаат афинитет кон анализираните протеини. Комбинацијата на имуноафинитетни раздвојувања со масената спектрометрија MALDI-TOF се нарекува масеноспектрометриска имуноанализа (MSIA) и неодамна беше применета во две опсежни студии на протеинскиот диверзитет кај човековата популација. Резултатите од овие студии сугерираат голем степен на протеинска разновидност која е доста покомплексна од генетскиот диверзитет. Откривањето и студирањето на протеинските варијации во човековата популација ќе биде значаен дел од идните протеински испитувања во клиничките лаборатории и во дијагностиката.

Клучни зборови: протеински диверзитет; протеински изоформи; плазма; серум; масена спектрометрија; имуноанализа

INTRODUCTION

The human genome contains little over 3 billion base pairs, and an estimated 20,000–25,000 protein-coding genes. Immediately after the intial publication of this genome in 2001 [1, 2], a multi-country collaborative project termed HapMap was initiated to generate a haplotype map of the human genome that describes the common patterns of

human DNA sequence variation [3]. In recent years, the focus has moved to individual human genomes, as advances in sequencing technology and a decrease in costs have made sequencing individual genomes feasible. In 2007, the complete (diploid) genome sequences of two individuals were published, those of James Watson (the co-discoverer of the DNA structure) [4] and Craig Venter, the head of the privately funded human genome project [5]. In 2008, a consortium of scientists sequenced the genomes of eight people from diverse ethnic backgrounds, and completed the first sequence-based map of structural variations in the human genomes, such as insertions, deletions, inversions, single nucleotide polymorphisms (SNPs), and copy number differences [6]. And in early 2008, an international research consortium (the "1000 Genomes Project") was formed to create the most detailed and medically useful picture to date of human genetic variation, by sequencing the genomes of at least a thousand people from around the world [7]. In all, the human genetic variation is such a hot topic that it was named the breakthrough of the year 2007 by the magazine Science [8]. Therefore, it is not surprising that population genetics and personal genomics are being rapidly commercialized, with companies like 23andMe (www.23andme.com), deCode Genetics (www.decode.com), and Navigenics (www.navigenics.com) offering personalized genetic analysis to reveal one's risks to specific diseases, traits, and other conditions, based on specific regions and changes in the genetic sequence. The Genographic project, with aims to map the migratory history of the human species, is another example of how genetic data can be used to track one's ancestry [9]. With these concerted efforts and outcomes from the investigations of the human genetic variation, we are only starting to apprehend and appreciate the extent to which our genomes differ from person to person. More importantly, we are learning about the implication that these variations have in disease predisposition and development.

But the genes only control heredity and provide the codes for the life's building blocks – the proteins. If there is so much variation at the gene level, what kind of variations can be expected at the protein level? The answer to that question leads us to the much more complex world of human protein diversity. The protein diversity stems from several dimensions. To start with, a single nucleotide change in the DNA sequence (i.e., SNP) can give rise to a different amino acid in the protein sequence, which in turn can influence protein folding, processing, function, etc. Next, alternative splicing of the pre-mRNA can produce sets of protein isoforms, each with distinct characteristics and functionalities. Then, following the protein sequence translation from the mRNA, a large number of modifications are induced at specific sites of the protein sequence, and can include phosphorylation, glycosylation, disulfides formation, side chain oxidation, various enzymatic processing, N- and C-terminal sequence truncations, etc. These posttranslational processes, as well as the overall protein expression in specific tissues, are heavily influenced by various cell processes, environmental factors, and cell and body cycles, resulting in an overall fluidic state of the human proteome. Analyzing these variations on a grandeur scale has been very challenging – there is no PCR-equivalent technique for protein analysis. Classical biochemical approaches have been used in the past to painstakingly unveil specific protein function and structure, in a one-protein-at-a-time approach. The invention of enzyme immunoassays in the early 1970s [10] enabled researchers to readily determine concentration ranges for a large number of proteins in biological fluids such as human serum, plasma, and urine [11, 12]. These methods have been the cornerstone of the diagnostic world for over three decades, and are still considered the gold standard in the clinical and reference laboratories. However, there has been no easy and rapid method for detecting protein structural modifications – until the advent of mass spectrometry.

Mass spectrometry has been synonymous with the field of proteomics from the very beginning, opening the realms of high-throughput and high-content proteome analysis [13, 14]. Mass spectrometry is the only detection method today that can universally provide information about specific protein structural modifications, without *a priori* knowledge of the modification. Mass spectrometry interrogates the protein mass, which is an intrinsic property of each fully expressed and functional protein. The mass contains information about the gene that encodes the protein, and the post-expression processing that the protein undergoes. Any changes in the gene sequence and/or post-expression protein processing will be reflected in the mass of the whole protein. Mass spectrometry can essentially detect all modifications that result in a mass shift of the wild-type

protein, including sequence truncations, side-chain residue modifications (phosphorylation, sulfonation, oxidation, etc.), deglycosylations, chemical adducts, etc. Many of those modifications have been reported for numerous proteins, yet, to date there is virtually no data on their distribution across the general population, even for the most abundant proteins. Point-mutations at the gene level can also be detected and catalogued as those mutations oftentimes result in detectable mass shifts, even on intact proteins.

And yet, mass spectrometry is only a method of detection, and as with any other detection technique, the separation processes that prepare the proteins for MS detection are critical. Standard separations such as two-dimensional gel electrophoresis, liquid chromatography, and affinity surfaces, have been used in combination with mass spectrometry extensively in the past 15 years. However, most of them are complex, multifaceted, and often times do not yield reproducible results, even within the same laboratory. To achieve high-throughput reproducible analyses, the fractionation approaches have to be conceptually very simple, highly-reproducible, and yield unambiguous readings and results. Such are the immunoaffinity separations that target single proteins by using highly specific antibodies for their affinity retrieval from the biological fluids. Combining this immunoaffinity separation with MALDI-TOF mass spectrometric detection yields an approach termed Mass Spectrometric Immunoassay (MSIA) (Fig. 1) [15, 16]. MSIA is essentially a rational combination of micro-scale immunoaffinity capture and mass spectrometry. Antibodies are surface-immobilized in small, porous microcolumns that are fitted at the entrance of pipettor tips. Biological samples are repeatedly aspirated and dispensed through these affinity pipettes to expose the immobilized antibody to the protein antigen present in the sample. Once the protein is captured, the affinity pipettes are rinsed to remove any loosely associated and non-specifically bound sample components, and a small volume of MALDI matrix is aspirated into the affinity pipettes. The pH of the matrix solution, and its components, disrupt the antibody-antigen interaction, and the antigen-containing eluate is deposited directly onto a MALDI target for ensuing mass spectrometric analysis. This combination of immunoaffinity capture with mass spectrometry results in a dual specificity assay – the capturing antibody provides the first level of specificity, while the mass spectrometric detection gives the assay another (orthogonal) measure of specificity in that each protein should register in the mass spectrum at a precise m/z value that corresponds to its molecular mass. The ability to see deviations from the predicted mass of the protein in the mass spectra enables detection of protein isoforms and other posttranslational modifications that give rise to the protein diversity.

In an initial study of human protein diversity using mass spectrometric methods of detection, twenty-five plasma proteins from a cohort of ninety-six healthy individuals were investigated via mass spectrometric immunoassays [17]. The protocol and an example of the data generated for one of the proteins – transthyretin (TTR), are outlined in Figure 2. The transthyretin MSIA assays were performed in parallel on the 96 human plasma samples using affinity pipettes derivatized with anti-transthyretin antibody. Following mass spectrometric analysis, data matrix containing all tentatively assigned modifications was assembled. Then, peptide-mapping experiments were performed on selected number of samples to identify the specific modifications and finalize the modifications database. The data for all 25 proteins is presented in Figure 3, which lists the modifications observed for 18 of the 25 proteins studied (modifications were not observed for 7 proteins), and shows the frequency of each modification in the 96-samples cohort. A total of 53 protein variants were observed for these 18 proteins, stemming from posttranslational modifications and point mutations. The largest number of posttranslationally modified protein variants was found to be C- or N-terminal truncated protein isoforms. Deglycosylation, oxidation, and cysteinylation were also observed among several of the proteins. Among the point mutations detected for 4 of the proteins, notable was the high incidence of point mutations for apolipoprotein E and transthyretin, which is consistent with genomic studies that have found these proteins to be highly polymorphic. The overall frequency of the modifications in the 96-samples cohort was wide ranged. Fourteen modifications were observed in all 96 samples, suggesting that they must be regarded as wild-type protein forms. Others, such as most of the point mutations, were present in only few of the samples. Overall, 23 of the modifications were observed in more than 65% of the samples, and 20 in less than 15% of the 96 samples analyzed. Upon further data analysis, and taking into the considera-

tion the gender, age, and ethnicity of the individuals who provided the samples, it was determined that the Gly6Ser mutation in transthyretin was detected only in individuals of Caucasian origin, which is consistent with existing knowledge about the occurrence of this common non-amyloidogenic population polymorphism in Caucasians [18]. Another correlation was observed in regards to interprotein variations in specific individuals: all seven individuals for which carbohydrate deficient transferrin was detected were also characterized with deglycosylated antithrombin III.

Following this small scale protein diversity study, a second study of human protein diversity was recently carried out wherein the number of samples was greatly expanded in order to get an accurate view of the distribution of some of the protein modifications in the general population [19]. One thousand individuals from 4 geographical regions in the United States (California, Florida, Tennessee, and Texas) were selected and the protein modifications for beta-2-microglobulin (b2m), cystatin C (cysC), retinol binding protein (RBP), transferrin (TRFE), and transthyretin (TTR) were delineated (in the 96-samples study, these five proteins accounted for 19 of the 53 protein variants observed). The results of the study are summarized in Figure 4, which lists the protein modifications observed and the frequency of each in the 1,000 samples cohort. A total of 27 protein modifications (20 posttranslational modifications and 7 point mutations) were detected, with various frequencies in the cohort of samples. Variants resulting from oxidation were observed most frequently, along with single amino acid truncations. Least frequent were variants arising from point mutations and extensive sequence truncations. In total, six modifications were observed with high frequency (present in >80% of the samples), 5 were of medium frequency (20-50% of the samples), and 16 were low frequency modifications observed in <7% of the samples. Nine of the low frequency modifications were not observed in the 96 individuals study. Thus, by increasing the size of the population it became possible to detect these low-occurrence protein modifications.

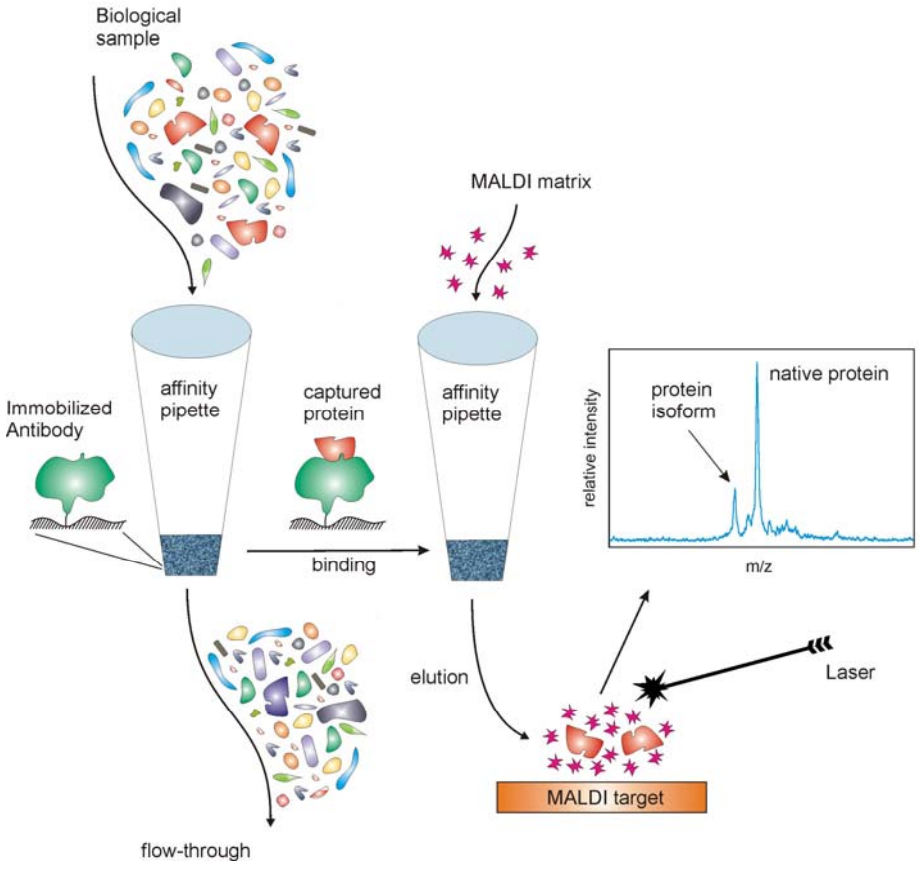

Fig. 1. Mass Spectrometric Immunoassay

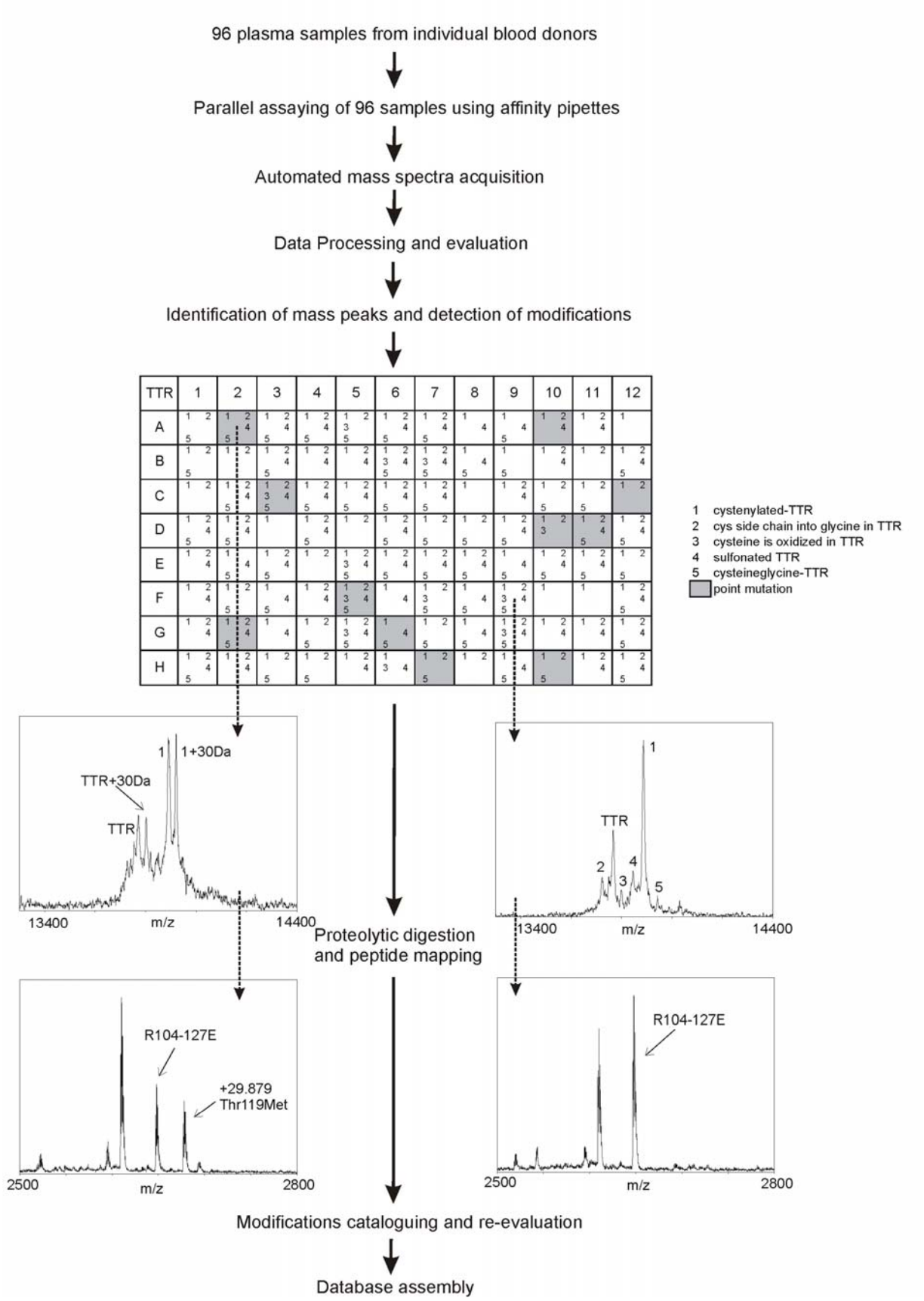

Fig. 2. An outline of the high-throughput mass spectrometric immunoassay using transthyretin (TTR) as an example

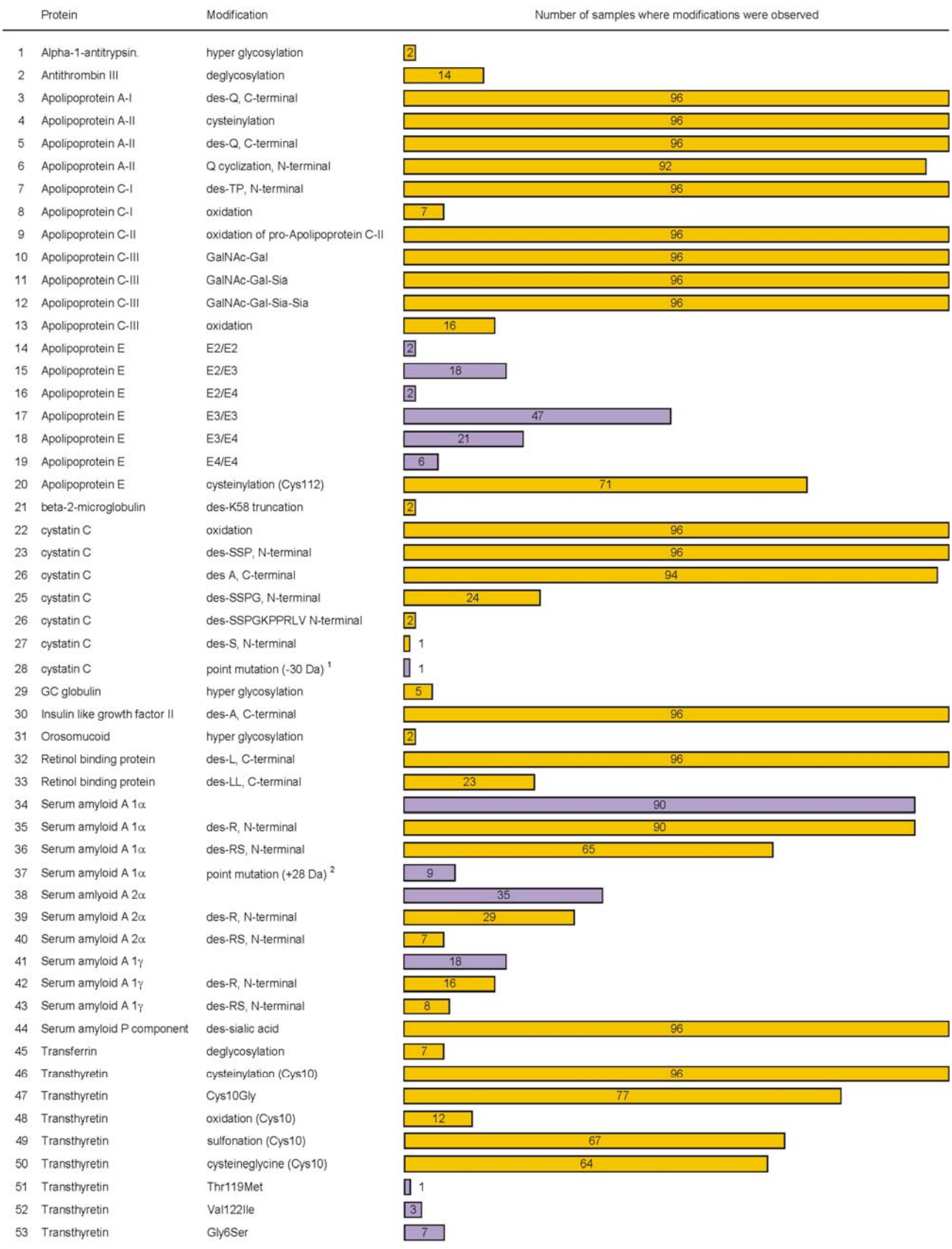

Fig. 3. Modifications observed in 18 of the 25 proteins analyzed from 96 human plasma samples (modifications were not detected for albumin, cerruloplasmin, C-reactive protein, insulin like growth factor I, lysozyme, plasminogen, and urine protein 1. Reprinted with permission from Nedelkov *et al*. (17).

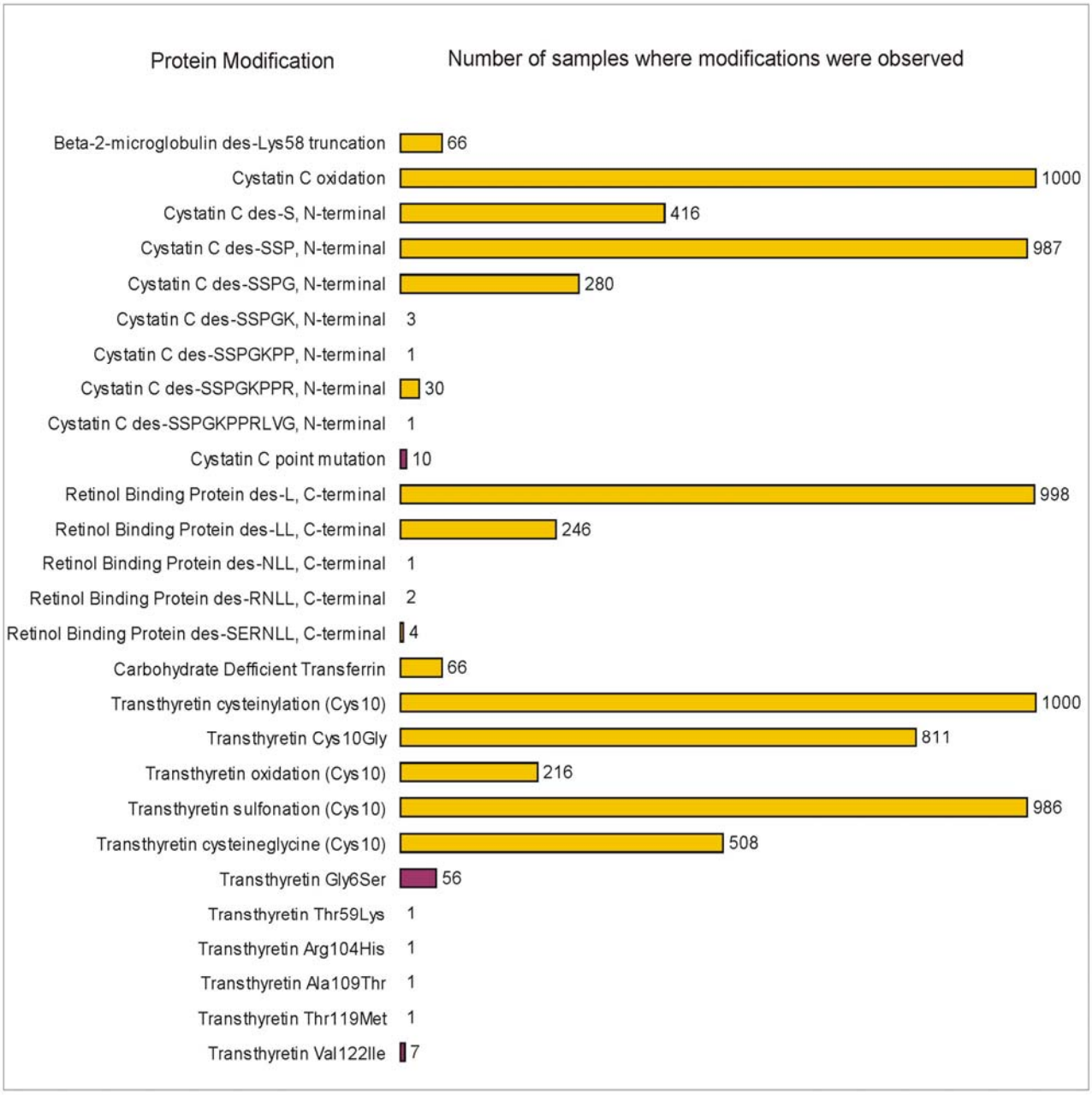

Fig. 4. Modifications observed for five proteins studied from 1,000 human plasma samples. Reprinted with permission from Nedelkov *et al*. (19).

When the frequencies of the modifications in the two studies were compared, an excellent correlation was obtained. For example, in both cohorts ~7% of the individuals were characterized with carbohydrate deficient transferrin. Upon further data analysis based on the gender, age, and geographical origin of the individuals who provided the samples, it was determined that the samples obtained from California contained significantly less protein modifications than the samples obtained from Florida, Tennessee, and Texas, even though the samples from all four states were collected in the same way within a three-month window in the spring of 2005, and stored under identical conditions until analysis. Correlations were also made in regards to the gender distribution of two protein modifications. Carbohydrate deficient transferrin was observed in ~1% of the females and ~10% of the males in the 1,000 cohort. Carbohydrate deficient transferrin is an FDA-approved clinical biomarker for alcoholism, and this gender correlation can partially be explained by the higher prevalence of alcohol dependence in males than in females. The second gender correlation was re-

lated to cystatin C: all 10 of the cystatin C point mutations were found in males.

Two conclusions can be made from these two systematic studies of protein modifications and variants. First, mass spectrometry is capable of detecting structural protein modifications, and, when coupled to immunoaffinity separations, it can be employed in a high-throughput systematic study of human protein diversity, in a discipline termed population proteomics [20-22]. Second, the human protein diversity is far more complex than the variation observed at the genetic level. While it might be premature to declare the human proteins variation "the next big thing", it is reasonable to predict that assessing human proteome variations among and within populations will be a paramount effort that can facilitate biomarker discovery. Such endeavor would represent a paradigm shift in proteomics with significant clinical and diagnostic implications, as protein variations, quantitative and qualitative, begin to be associated with specific diseases. The time to start these studies has arrived.

REFERENCES

[1] E. S. Lander *et al*., Initial sequencing and analysis of the human genome. *Nature,* **409**, 860–921(2001).

[2] J. C. Venter *et al*., The sequence of the human genome. *Science,* **291**, 1304–1351(2001).

[3] D. Altshuler *et al*., A haplotype map of the human genome. *Nature*, **437**, 1299–1320 (2005).

[4] D. A. Wheeler *et al*., The complete genome of an individual by massively parallel DNA sequencing. *Nature,* **452**, 872–876 (2008).

[5] S. Levy *et al*., The diploid genome sequence of an individual human. *PLoS Biol*, *5*, e254 (2007).

[6] J. M. Kidd *et al*., Mapping and sequencing of structural variation from eight human genomes. *Nature,* **453**, 56-64 (2008).

[7] J. Kaiser, DNA sequencing. A plan to capture human diversity in 1000 genomes. *Science*, **319**, 395 (2008).

[8] E. Pennisi, Breakthrough of the year. Human genetic variation. *Science*, **318**, 1842–1843 (2007).

[9] D. M. Behar, S. Rosset, J. Blue-Smith, O. Balanovsky, S. Tzur, D. Comas, R. J. Mitchell, L. Quintana-Murci, C. Tyler-Smith, R. S. Wells, The Genographic project public participation mitochondrial DNA database. *PLoS Genet* **3**, e104 (2007).

[10] R. M. Lequin, Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA). *Clin. Chem.,* **51**, 2415–2418 (2005).

[11] W. Y. Craig, T. B. Ledue, R. F. Ritchie, *Plasma Proteins: Clinical Utility and Interpretation*, Foundation for Blood Research, Scarborough, ME 2001.

[12] R. F. Ritchie (Ed.), *Serum Proteins in Clinical Medicine*, Foundation for Blood Research, Scarborough, ME 1999.

[13] R. Aebersold, M. Mann, Mass spectrometry-based proteomics. *Nature*, **422**, 198–207 (2003).

[14] B. F. Cravatt, G. M. Simon, J. R.Yates, 3rd, The biological impact of mass-spectrometry-based proteomics. *Nature*, **450**, 991–1000 (2007).

[15] D. Nedelkov, Mass spectrometry-based immunoassays for the next phase of clinical applications. *Expert Rev Proteomics*, **3**, 631–640 (2006).

[16] R. W. Nelson, J. R. Krone, A. L. Bieber, P. Williams, Mass-spectrometric immunoassay. *Anal. Chem.* **67**, 1153 –1158 (1995).

[17] D. Nedelkov, U. A. Kiernan, E. E. Niederkofler, K. A. Tubbs, R. W. Nelson, Investigating diversity in human plasma proteins. *Proc. Natl. Acad. Sci. U S A*, **102**, 10852–10857 (2005).

[18] L. H. Connors, A. Lim, T. Prokaeva, V. A. Roskens, C. E. Costello, Tabulation of human transthyretin (TTR) variants, *Amyloid*, **10,** 160–184 (2003).

[19] D. Nedelkov, D. A. Phillips, K. A. Tubbs, R. W. Nelson, Investigation of human protein variants and their frequency in the general population. *Mol. Cell. Proteomics*, **6**, 1183–1187 (2007).

[20] D. Nedelkov, Population proteomics: addressing protein diversity in humans. *Expert Rev. Proteomics*, **2**, 315–324 (2005).

[21] D. Nedelkov, Population proteomics: investigation of protein diversity in human populations. *Proteomics*, **8**, 779–786 (2008).

[22] D. Nedelkov, U. A. Kiernan, E. E. Niederkofler, K. A. Tubbs, R. W. Nelson, Population proteomics: the cConcept, attributes, and potential for cancer biomarker research. *Mol Cell Proteomics*, **5**, 1811–1818 (2006).