Journal of Proteomics 134 (2016) 1–4

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

Towards deciphering proteomes via the proteoform, protein speciation, moonlighting and protein code concepts

New concepts have been presented recently, which will help to improve our knowledge about proteomes in the future, including the proteoform, protein speciation, moonlighting, and the protein code concept. All these concepts have in common that they focus on the relationship of function to the exact chemical formula of proteins. This common discourse is in contrast to the former and still existing protein expression discourse.

The very popular protein expression discourse is based on the one gene – one protein – one function hypothesis. Consequently the DNA decides about the function. It is transcribed in RNA, which is translated in protein. This translation process is mostly described as protein expression. This reductive view includes that the protein fulfills the function fixed in its gene. To reach optimal sensitivity the protein is digested (bottom-up or shotgun approach) and commonly very few of the resulting peptides are used for identification and quantification. If a difference in "protein expression" between two biological situations can be confirmed at the transcriptome level by RT-PCR, it is postulated that the function of this gene/protein is up or down-regulated. The technology for this discourse, bottom-up LC-MS/MS can be run highly automated in a large scale with high sensitivity. It gives us access to a large part of the proteome and to amino acids, which are post- translationally modified.

However, the shotgun based quantification approach produces false negative and false positive results [1]. As soon as there is more than one protein species present, originating from the same gene, single tryptic peptide analysis represents quantification of a mixture of different protein species and therefore of a mixture of different functions. In shotgun results no information about a single function of a protein can be revealed. Protein speciation occurs presumably for most of the mammalian genes – Neil Kelleher has estimated the number of human protein species (proteoforms) to be in the range of 1 billion [2] coded by approximately 20.000 genes. All combinations of the about 30 modifications containing histone H3 alone lead to more than 1 billion theoretically existing H3 protein species [1].

The separation of cellular proteins of Hela cells by two-dimensional gel electrophoresis (2-DE) resulted in the finding of several spots per protein for about 50% of the proteins identified [3]. These proteins are diversified in several forms of the protein, which differ in molecular weight or isoelectric point. A new world arises below the level of proteins. In analogy to the classification of organisms the "protein" represents the genus and the "protein species" the organism species. We named the different forms protein species and defined them by their exact chemical primary structure, their structural formula [4–6]. Protein speciation can be caused by e.g. splicing, truncation, partial degradation and covalent coupling of molecules such as phosphates or glycans. The protein speciation discourse accepts the diversification of a protein into different protein species. These different protein species of one protein differ in their biological functions [5,6]. Consequently protein species have to be correlated with function and not as usual in the protein expression discourse the amount of the mixture of all protein species. To differentiate between the protein species analytically, the protein species have to be separated in their intact form before analysis and this can only be reached by top-down procedures, such as 2-DE-MS or LC-MS. Several other concepts such as proteoforms [7], moonlighting [8], and the protein code [9], face in the same direction. Therefore, we tried to comprise these concepts in a Special Issue about protein species with the aim to open the mind in proteomics for the real molecular level and its connection to biological function. We have collected 19 manuscripts covering the field of protein species.

Only a few protein diagnostics have been introduced during the last 20 years despite the tremendous progression of mass spectrometry-based proteomics. Steffen et al. [10] reviewed the current knowledge about disease-specific protein species as diagnostic markers. Examples for different diseases because of single amino acid exchanges due to gene mutations, alternative splicing, dysregulation of protein synthesis, proteolytic processing and post-translational modifications are given, indicating the importance of protein species for diagnosis. In addition, antibody- and mass spectrometry-based assays were discussed in the review. In clinics, ELISAs are commonly used for the quantification of proteins, but only few antibodies are available for the detection of distinct protein species. Selection reaction monitoring, data independent quantification, mass spectrometry-based immunoassays, SELDI, and top-down mass spectrometry were critically discussed including their potential usage in routine clinical laboratories in the near future.

The moonlighting effect reviewed by Constance Jeffery [11] describing different functions of a protein is perfectly fitting with the protein speciation concept: A protein resulting from one gene has different functions. The chemical solution for these different functions is a different chemical primary structure obtained by splicing, truncation or degradation and post-translational modification. The resulting different protein species of a protein are mirrored by different functions. Constance Jeffery showed several examples of moonlighting proteins where only one PTM switches the function, proteins with high homology to a moonlighting protein but sharing only one of its functions, and several "neomorphic moonlighting proteins" in which a single amino acid mutation leads to a new function. An impressive example is the ribosomal protein L13. After phosphorylation it is released from the ribosome and becomes a translation factor as part of IFN-gamma-activated inhibitor of translation (GAIT complex), which binds to mRNAs for ceruloplasmin and other inflammation-related proteins to decrease their synthesis.

http://dx.doi.org/10.1016/j.jprot.2016.01.012
1874-3919/

Another challenge in addition to the protein species - function relationship is the analysis at the protein species level, which needs complete sequence coverage. This is reviewed by Rainer Bischoff [12], under the title "Genomic variability and protein species — Improving sequence coverage for proteogenomics", focusing on the impact of genetic variations in populations and its consequences for proteomics. Bischoff et al. explain in their chapter "Genomics and protein variability" the reasons for the huge diversity and differences of protein species caused by variants in the genomes of individuals as well as varying gene transcription, splicing and editing, depending on responses to environmental factors. The authors discuss current transcriptome sequencing approaches and explain how sample-specific protein sequence databases can be obtained from transcriptome sequencing data of samples for improving the yield of sequence coverage in protein identifications of individuals. In addition Bischoff et al. thoroughly discuss the application of peptide fragmentation prediction, diverse protein digestion methods and derivatization methods for improving protein sequence coverage in proteomics.

Ten years ago the complete proteome of the human brain was envisaged and an understanding of neurological diseases and processes such as thinking could be understandable by the elucidation of the human brain proteome. Now it is becoming clear that we did not reach the functional level with listing up the proteins present in an organ and quantification at the protein level is not reliable with the correlation to function. Smaller compartments and components such as the proteasomes (Hirano et al.-) [13] and salivary proline-rich proteins (Manconi et al.) [14] revealed after considering of protein speciation a new universe within this small compartments of a whole organism. Hirano et al. reviewed the knowledge of post-translational modification in 26S proteasome of yeast. For the 35 proteins constituting the 26S proteasome 345 sites were already found modified with 14 kinds of post-translational modifications. Several examples with correlations between the modifications and a function are already known. They announced the introduction of the data into the ModProt database. Manconi et al. in their review give deep insights into the complexity of species of salivary proline-rich proteins which arise from only six genes. The huge number of these species is formed by alternative splicing, diverse proteolytic processes and the addition of several groups of post-translational modifications such as dimerization, N-pyroglutaminylation, N-/O-glycosylation and phosphorylation. Surprisingly, not much is known about the physiological function of this family of protein species.

Xiong et al. [15] reviewed investigations on post-translational modifications in cyanobacteria. Twenty years ago it was believed that PTMs are mainly occurring in eukaryotes, but it became more and more clear that the bacterial proteome is also highly diversified by post-translational modifications.

Gianazza et al. [16] contribute to the protein species concept by a review about the computational prediction of PTM-induced changes in protein conformation and their impact on protein function. In their review Gianazza et al. give helpful information about informatics resources and databases regarding PTMs and discuss how fragment mass spectra from MS/MS experiments can be interpreted with respect to PTMs. Furthermore the authors reviewed how the prediction of protein function is achieved via the computational analysis of changes in affinities between the target protein and interacting molecules such as other proteins (protein-protein interactions) or ligands (receptors) or substrates (enzymes).

Savaryn et al. including Neil Kelleher [17] show their improvements in reaching rare proteins such as transcription factors by top-down proteomics. They focused on a single protein, NF-kappa-B, and enriched it with HaloTag technology and recombinant protein expression. Then enough protein species were obtained for top-down MS and elucidation of N-terminally acetylated NF-kappa-B was the most abundant protein species identified. In a TNFα/LepB treated cell a phosphorylated protein species was postulated from a + 79 Da shift in an MS measurement without MS/MS confirmation. These results with an exogenously enriched protein are a first step toward the ultimate goal of targeted transcription factor characterization from endogenous loci. This article contains also some thoughts about the terms protein species and proteoforms.

The tremendous diversity of potential histone species is mainly caused by phosphorylation, methylation and acetylation. From these three PTMs only methylation on a single site may cause diversity itself by mono-, di-, and trimethylation. This kind of diversity is much more elaborated for glycosylation, which is one of several hundred known PTMs and covered in this Special Issue by five articles because of its importance. At one site of a protein many different monocarbohydrates may be combined not only in a sequence, but even in different branches. Rosenlöcher et al. [18] investigated the different protein species arising from the glyco-diversity in alpha-1-antitrypsin by MALDI-MS, estimated the number of potential protein species and coined a new term: glycoprotein species as a synonym for glycoform [19], describing a chemically clearly defined molecule. Tran et al. with David Goodlett as coauthor [20] applied Orbitrap and FTICR MS with ESI and MALDI sources using ETD for fragmentation to elucidate the protein species of two antibodies with top-down and middle-down proteomics. The emerging real complexity is very high and if antibodies are used as therapeutics this complexity has to be considered for quality control. Important contributions about the impact of differences in post-translational modifications introduced by *in-vitro* processing of protein species are given by Milkovska-Stamenova and Hoffmann, who report on the experimental investigation of hexose-derived glycation sites in different milk products [21] and the identification and quantification of bovine protein lactosylation sites [22]. Treatment of milk with thermal treatment is necessary for getting rid of pathogenic microorganisms, which can be present in milk. Although heat treatment reduces the risk of infections by microbes very significantly, it affects the endogenous milk molecules changing their original composition by diverse chemical reactions such as glycation reactions and oxidation. Also milk proteins can be denatured and derivatized. Changes of the chemical composition of milk proteins can result in a decrease of their nutritional value and an increase of their immunogenicity. Astonishingly, Milkovska-Stamenova and Hoffmann observed that beside the kind of thermal treatment also the technical process independently from thermal treatment has a large effect on the glycation reactions of milk proteins. Therefore the authors concluded that it should be possible by optimization of the technical processes to reduce lactosylation and hexose-associated glycation levels of milk proteins even if thermal treatment is applied to milk. What can we learn from these studies from Milkovska-Stamenova and Hoffmann regarding protein species? Any kind of sample preparation can induce changes in the exact chemical composition of proteins making the analysis of protein species even more challenging because some of them might be generated in vitro while other species at the same time will disappear.

Rosal-Vela et al. [23] detected by a 2-DE-MS approach of spleen in an arthritis model several spots arthritis-dependently changing their intensity, identified the proteins within these spots and found serotransferrin in six spots. Glycopeptide analysis revealed several glycans per spot, showing the incomplete protein species separation using 2-DE. At the glycoform level, differences in amount between fucosylated glycans and branching were found in arthritis compared with a control. They concluded that serotransferrin changed significantly in its glycosylation pattern in arthritic mice. The reader becomes aware of the complexity arising by different glycosylations and the necessity of adding a further separation step before 2-DE to separate the glycoprotein species.

Tremendous progress has been achieved during the last decade to identify phosphorylated peptides by mass spectrometry. However, a key issue in phosphoproteomics is a better understanding of the functional and structural implications. Pan et al. [24] used top-down mass spectrometry with ETD to study different phosphorylated species of