Trends in Analytical Chemistry 69 (2015) 76–87

Contents lists available at ScienceDirect

Trends in Analytical Chemistry

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

Challenges and developments in protein identification using mass spectrometry

Zoltan Szabo *, Tamas Janaky

Department of Medical Chemistry, University of Szeged, Dom sqr. 8, Szeged H-6720, Hungary

ARTICLE INFO

Keywords:
Bioinformatics
Biomarker
Fragmentation
Mass spectrometry
Protein expression
Protein identification
Post-translational modification
Proteoform
Proteome
Proteomics

ABSTRACT

Mass-spectrometry (MS)-based proteomics is the most powerful approach for identifying proteins and determining protein expression in tissues under different conditions to identify post-translational modifications in response to stimuli and to characterize protein interactions. Protein identification is a key step in characterizing proteomes to describe biological processes and to discover disease-related biomarkers, pharmaceutical targets, protein functions or interactions. In all proteomics workflows, whether commonly-applied gel-electrophoresis-based methods or gel-free approaches, MS is an indispensable tool for the identification of protein sequences and modifications. The complexity, high abundance, dynamic range, presence of similar proteins, and several forms of the same protein all raise challenges for analytical instrumentation and data-analysis software. This review provides an introduction to key terms, methods and challenges in protein identification, and summarizes current solutions and trends, including novel data-collection approaches, bioinformatics and instrumentation developments.

Contents

Abbreviations: AC, Affinity chromatography; CE, Capillary electrophoresis; CDS, Protein coding sequence; CID, Collision-induced dissociation; COFRADIC, Combined fractional diagonal chromatography; CPLL, Combinatorial peptide-ligand library; DDA, Data-dependent acquisition; DIA, Data-independent acquisition; ECD, Electron-capture dissociation; ERLIC, Electrostatic repulsion-hydrophilic interaction chromatography; ESI, Electrospray ionization; ETD, Electron-transfer dissociation; FDR, False-discovery rate; FT-ICR, Fourier-transform ion-cyclotron resonance; GE, Gel electrophoresis; HCD, Higher-energy collisional dissociation; HDMS^E, High-definition MS; HILIC, Hydrophilic interaction chromatography; IC, Immunochemistry; IEF, Isoelectric focusing; IM, Ion mobility; IMAC, Immobilized metal-ion chromatography; IMS, Ion-mobility spectroscopy; IP, Immunoprecipitation; IT, Ion trap; LC, Liquid chromatography; MALDI, Matrix-assisted laser desorption/ionization; MOAC, Metal-oxide affinity chromatography; MS, Mass spectrometry; MS/MS, Tandem mass spectrometry; OT, Orbitrap; PAGE, Polyacrylamide gel electrophoresis; PD, Protein depletion; PFF, Peptide-fragment fingerprinting; PMF, Peptide-mass fingerprinting; PSD, Post-source decay; PTM, Post-translational modification; RP, Reversed phase; Q, Quadrupole; SAX, Strong anion exchange; SCX, Strong cation exchange; SELDI, Surface-enhanced laser desorption/ionization; TOF, Time of flight; TPP, TransProteomic Pipeline; UPLC, Ultra-performance liquid chromatography.

* Corresponding author. Tel.: +36 62 545 143.
E-mail address: szabo.zoltan@med.u-szeged.hu (Z. Szabo).

http://dx.doi.org/10.1016/j.trac.2015.03.007
0165-9936/

1. Introduction

1.1. A brief introduction to proteomics

The "proteome" is the complete set of proteins expressed by the genome of a cell, tissue or an organism [1]. While genes determine many of the characteristics of an organism, they do so by providing instructions through mRNA for synthesizing proteins, the building blocks and workhorses of cells – ultimately the functional players that drive different biochemical processes. Within an individual organism, the genome is more or less constant, the transcriptome is more variable, but the proteome is dynamic, complex, and adaptive, varies from cell to cell and reflects the effects of both internal and external environmental stimuli. The first rough draft of human proteome was published recently [2,3], and one of the authors believes "that the human proteome is so extensive and complex that researchers' catalog of it will never be fully complete". The complexity of any proteome is so large that none of the existing technologies can deliver complete detection and quantification of all the proteins that are present. The human genome contains about 20,300 protein-encoding genes, but the total number of proteins in human cells is estimated to be $0.25–1\times10^6$ [4].

The complexity of proteome can be explained by several reasons:

a) each gene may encode several proteins in a process called alternative splicing: one gene may make different mRNA products and, hence, different protein isoforms;
b) one protein may be modified chemically after it is synthesized (PTM, post-translational modification) so that it acquires a different function. The most frequent PTMs are phosphorylation, glycosylation, and acetylation. [5]. Each protein might exist in any one of a multiplicity of chemically-modified proteoforms, resulting in a proteome of even higher complexity (Fig. 1);
c) proteins can interact with each other in complex pathways and networks of pathways often as components of multimolecular complexes, increasing the pool of analytical targets, if the identification of protein complexes is of interest; and,
d) individual variations in the genetic code (e.g., allele variants, or single-nucleotide polymorphism) introduce another level of analytical complexity.

"Proteomics" is the large-scale comprehensive study of a proteome, including information on the abundances of proteins, their variations and modifications, and their interacting partners and networks. Proteomics technologies can perform the qualitative and quantitative comparison of proteomes under different conditions (e.g., normal and pathological) to further unravel complex biological processes, to discover biomarkers and to provide information for systems biology to build integrated network of cells.

To be able to characterize the large diversity of proteins in biological samples, the technologies and the chemistries need to be diverse and complex. Nevertheless, technological progress and new instrumentation has advanced to where this characterization can be realized on a large scale [6,7].

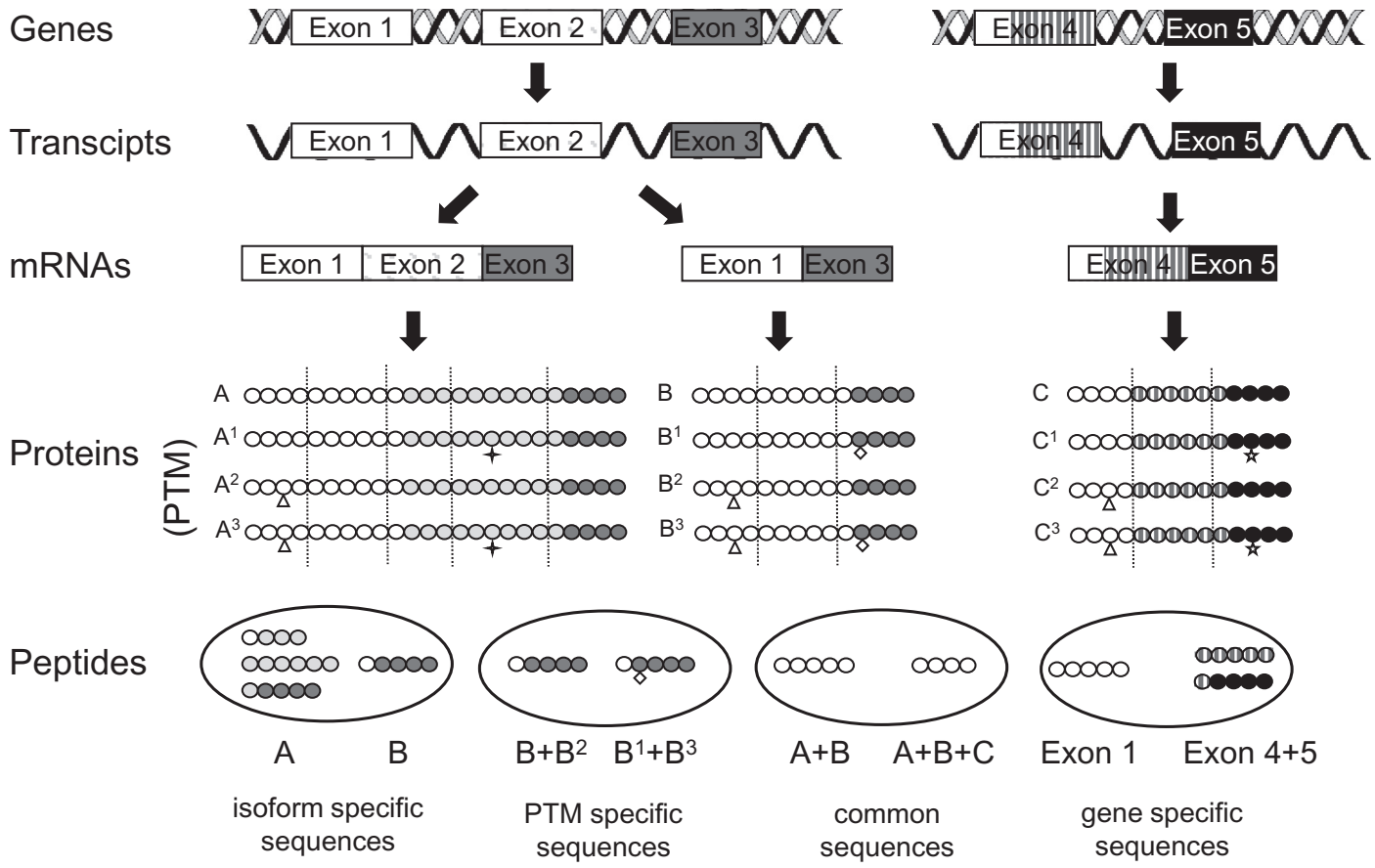

Fig. 1. Origin of proteome complexity (top) and challenges in fully resolving complexity by "bottom-up" proteomics. Presence of proteoforms can be proved by identifying specific peptides. Based on the lack of identification of specific peptides, no proteoform can be excluded; instead, protein groups are identified via common peptides (bottom). (Vertical dotted line (⋮): enzymatic cleavage places; △, ✦, ◊, ★: post-translational modifications).