Trends in Analytical Chemistry ■■ (2015) ■■–■■

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.

© 2015 Elsevier B.V. All rights reserved.

Contents

1. Introduction 2
   1.1. A brief introduction to proteomics 2
   1.2. The special meaning of identification in proteomics 4
2. Discussion 4
   2.1. Pre-identification analytical steps 4
   2.2. Protein identification 4
      2.2.1. Bottom-up MS and general methods 4
      2.2.2. Peptide-mass fingerprinting (PMF) 5
      2.2.3. MS/MS peptide identification 5
      2.2.4. Informatics methods for MS-based protein identification 7
      2.2.5. Special methods for identification of proteoforms 9
      2.2.6. Middle-down proteomics 9
      2.2.7. Top-down proteomics 9
   2.3. Post-identification data analysis 10
3. Conclusions and future prospects 10
   References 11

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/© 2015 Elsevier B.V. All rights reserved.

Please cite this article in press as: Zoltan Szabo, Tamas Janaky, Challenges and developments in protein identification using mass spectrometry, Trends in Analytical Chemistry (2015), doi: 10.1016/j.trac.2015.03.007

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].

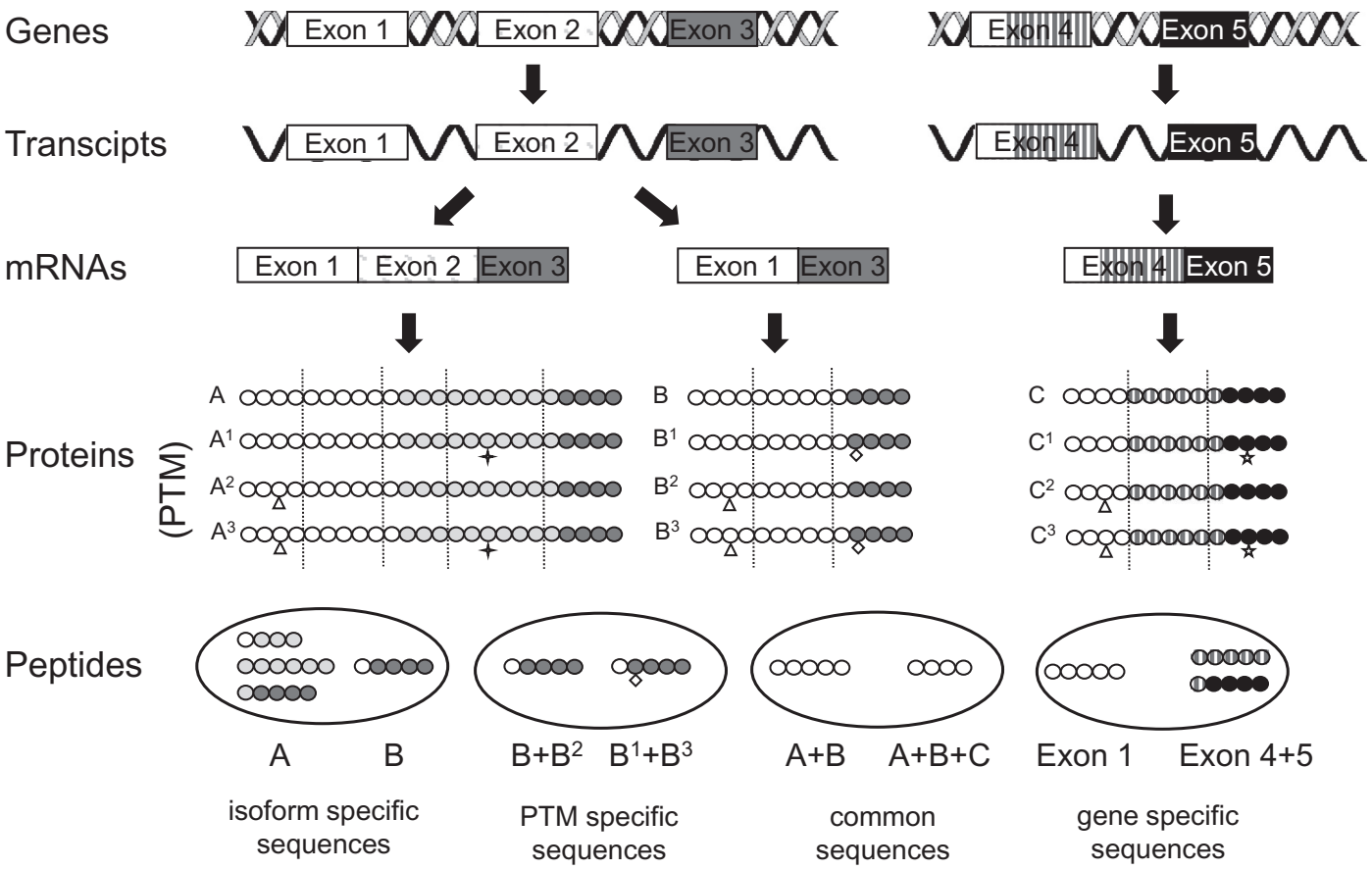


Fig. 1. Origin of proteome complexity (top) and challenges in fully resolving complexity by "bottom-up" proteomics. Presence of proteoforms can be proved by identifying specific peptides. Based on the lack of identification of specific peptides, no proteoform can be excluded; instead, protein groups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

Please cite this article in press as: Zoltan Szabo, Tamas Janaky, Challenges and developments in protein identification using mass spectrometry, Trends in Analytical Chemistry (2015), doi: 10.1016/j.trac.2015.03.007

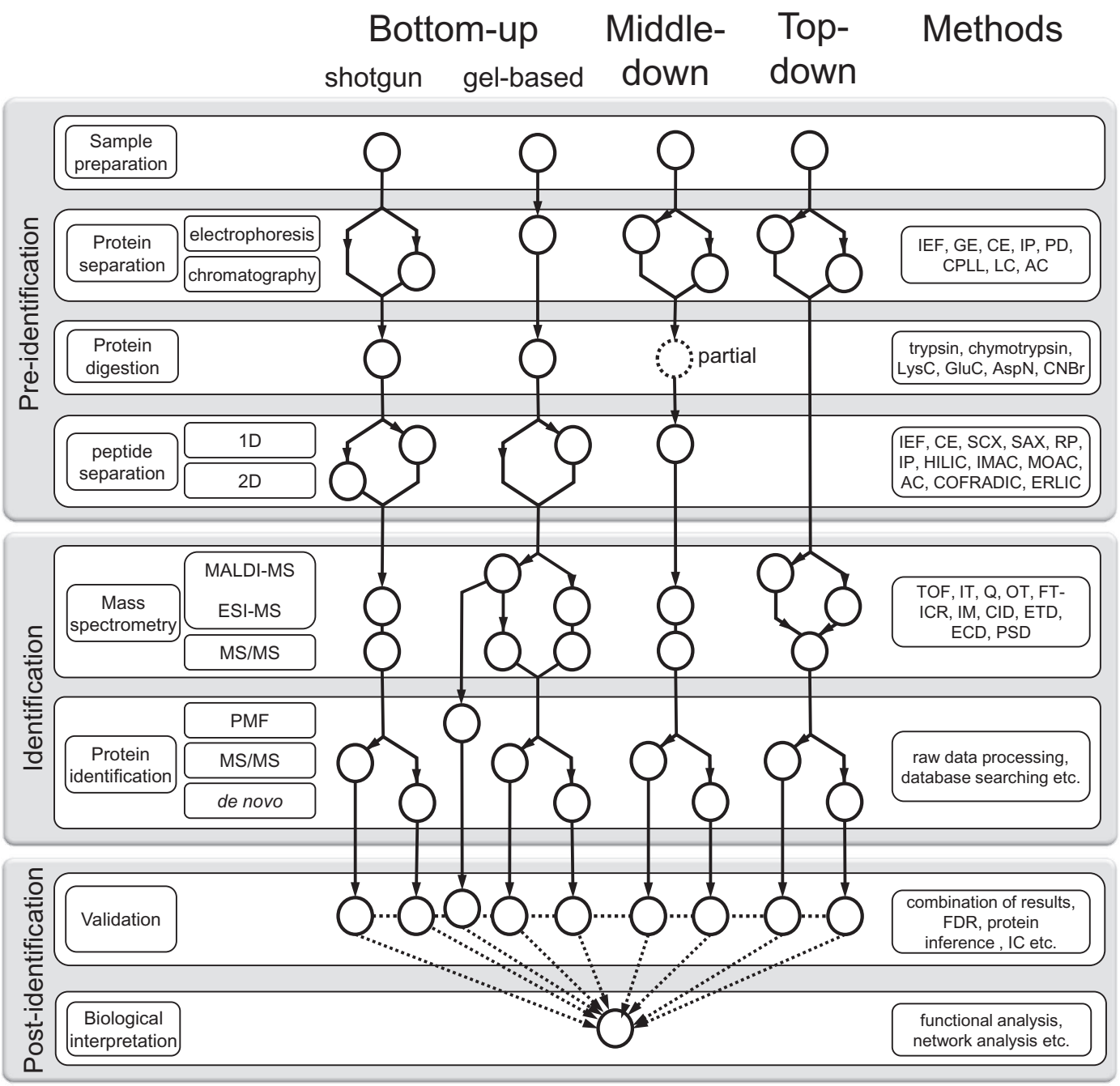


Fig. 2. Common proteomics identification workflows.

The more common proteomics strategies applied to identify proteins are based on a combination of different technologies of separation sciences, mass spectrometry (MS) and bioinformatics. Nowadays, there are three MS-based approaches used to identify proteins (Fig. 2). In the dominant "peptide-centric" bottom-up strategy [7], proteins are cleaved to smaller peptide fragments using some chemical or enzymatic treatment and these peptides are subjected to MS analysis. The identification of proteins may be based on:

a) the match of the observed peptide masses with predicted values of proteins from a protein-sequence database (peptide mass fingerprint, PMF);
b) the match of peptide sequences (deduced from their MS/MS fragmentation spectrum) against *in-silico* fragmentation of any possible combination of amino acids (de-novo sequencing) or that of cleaved peptides from a protein sequence databank (database search).

An alternative strategy is top-down proteomics [8], where the intact proteins are analyzed in their original form using high-resolution MS.

The middle-down proteomics approach is an effective compromise between the two methods above, employing limited digestion to create peptides that can be analyzed relatively easily by MS but large enough to retain a richer sequence and PTM information.

All of these approaches require sophisticated informatics tools. Novel algorithms are continually being developed, or existing ones modified to improve reliability or to derive benefits from new instrumentation and data-collection techniques.

Please cite this article in press as: Zoltan Szabo, Tamas Janaky, Challenges and developments in protein identification using mass spectrometry, Trends in Analytical Chemistry (2015), doi: 10.1016/j.trac.2015.03.007

1.2. The special meaning of identification in proteomics

In most of the proteomics approaches, MS-based protein identification means finding the best match from a set of simulated spectra of a limited set of protein sequences (including possible amino-acid-sequence variations and possible PTMs) to the experimental MS or MS/MS spectra. To discriminate bad, good and best matches, candidate-protein/peptide sequences are generally ordered in a list according to certain kinds of score, describing the significance of a match based on statistical calculations. Usually, some type of probability-based scoring is applied, in which the calculated score represents the probability that the observed match is a random event. The search space (the amino-acid sequences to be considered for identification) can be any possible permutation of amino acids (*denovo* sequencing using MS/MS) or sequences available in a database (database search using MS or MS/MS). The search space is increased by any variable PTM with sub-stoichiometric occupancy, as unmodified and several different modified forms of the same amino-acid sequence have to be analyzed. Generally, *de novo* sequencing is also followed by some kind of database search to assign the identified amino-acid sequences to a protein using a sequence-similarity-search algorithm, well established in genomics [e.g., Basic Local Alignment Search Tool (BLAST)] and adapted to proteomics (i.e., BLASTp or MS-BLAST) [9].

As can be seen from the above, protein identification heavily depends on bioinformatics (algorithms and protein-sequence databases). For each MS experimental approach, there are several dedicated or generalized identification algorithms. We emphasize that the significance of identification is measured statistically. Best matches may not be significant (e.g., poor-quality spectra) or several hits may be statistically significant (e.g., similar sequences), hence the manual investigation by an expert, parallel application of different algorithms, post-identification validation software, and local or general experience-based criteria of acceptance may be applied to increase confidence. The proper choice of search space (sequence database selected, and taxonomy-based limitations) is also a key factor for successful identification.

Sequence databases are mainly based on translated nucleic-acid sequences, and contain only a small proportion of sequenced protein entries. DNA-protein-sequence translation is not necessarily unambiguous, as identification of protein coding sequences (CDSs), and reading frames (start of nucleic-acid triplets, encoding single amino acids) are not completely free of error. Even the most recent, highly-reliable algorithms may result in erroneous protein sequences. Moreover, genetically-coded amino-acid sequences are further processed to get the final form of a protein (splicing) and cannot be predicted from its nucleic-acid sequence. However, there may be experimental proof of the existence (but not necessarily the full sequence) of predicted proteins, and this information may be present in certain highly-annotated databases (e.g., the manually-curated Swiss-Prot database). Sequence databases are continually growing, but are still far from complete. For a large number of species, complete genomes and also predicted "full" proteomes are available (e.g., human and several model species), but there are taxonomy classes that are very poorly represented.

As opposed to the identification of small molecules, in the prevalent bottom-up approach, MS-based identification of a protein is realized via identification of its peptide fragments. Discrimination among different proteoforms requires identification of unique peptides specific to the proteoforms (see Fig. 1). Generally, a small fraction of all possible fragments is identified, leading to poor sequence coverage, particularly in analysis of very complex protein mixtures (e.g., "shotgun" analysis of whole-cell lysates), hence proteoforms are usually indistinguishable in such experiments. Even in the top-down approach, where the total mass of the analyte molecule (a proteofom) is measured, identification faces difficulties in the case of unexpected/unpredictable PTMs or sequence variations. Possession of precise mass and isotope distribution (providing an estimate of elemental composition) and rules of chemical bonding may aid MS identification of small molecules, but, even though proteins consist of well-defined building blocks, a similar level of protein identification is limited by the lack of such rules.

Protein identification can be defined and carried out at different levels of precision, identifying variably-sized groups of proteoforms, or a single proteoform molecule, depending on the goals of the analysis and the success in identifying unique peptides (see Fig. 1).

2. Discussion

2.1. Pre-identification analytical steps

The successful MS identification of proteins greatly depends on the proper treatment of samples prior to an MS analysis. Pre-identification procedures involve all of those steps that help to remove molecules detrimentally interfering with the MS analysis of the intact protein or its enzymatic/chemical cleavage products and help to reduce the complexity of samples containing mixtures of proteins or peptides. Proteins, in cells, tissues or bodily fluids, are present in complex biological matrices, usually interacting with other biomolecules. Extraction, isolation and further treatment of proteins requires techniques that do not change the composition of the proteome and allow one to identify (and to quantify) a large number of proteins (with high speed and sensitivity) in an unbiased manner. While more powerful mass spectrometers, more sophisticated software and more complete databases are keys to exhaustive characterization of a proteome, proper technologies and strategies for sample pretreatment are critically important in order to obtain reliable downstream proteomic results. The extreme complexity and the large dynamic range of different components (10^{-2}–10^{-12} M) present in biological samples are the greatest challenges that face development of proteomics technology. Efficient solubilization, extraction, reduction of sample complexity by removing high-abundance proteins or selective enrichment of low-abundance proteins and front-end electrophoretic or chromatographic separation (at protein or peptide level, or both) prior to MS identification are especially important. Although all the pre-identification steps influence the quality of the final protein identification, these technologies are outside the scope of this review. For people interested in these areas, we suggest some excellent reviews [6,7,10].

2.2. Protein identification

The bottom-up approach is the most widely-used method for MS-based protein identification, as the peptides resulting from protein cleavage are in the mass range that can be routinely analyzed by most MS and LC systems. We discuss the recently developing top-down and middle-down approaches later.

2.2.1. Bottom-up MS and general methods

The MS analysis of the proteolytic peptides is generally carried out using matrix-assisted laser desorption/ionization (MALDI) or electrospray ionization (ESI).

MALDI analysis is relatively simple, and takes a short time for analysis, but it has limited reproducibility, and sample complexity is restricted by ion-suppression effects. MALDI generally produces singly-charged peptide ions in contrast to ESI, which produces multiply-charged ions., The lower charge in MALDI limits a high mass range (higher charge results in a lower mass-to-charge ratio), and MS/MS fragmentation efficiency.

Please cite this article in press as: Zoltan Szabo, Tamas Janaky, Challenges and developments in protein identification using mass spectrometry, Trends in Analytical Chemistry (2015), doi: 10.1016/j.trac.2015.03.007

ESI can be coupled on-line to liquid chromatography (LC) separation, solving the sample-complexity problem, but it also increases analysis time.

An off-line LC-MALDI analysis is also an option, but not competitive in terms of throughput and effectiveness.

2.2.2. Peptide-mass fingerprinting (PMF)

PMF is the most straightforward, simplest method for protein identification in bottom-up proteomics, and also does not have high requirements on instrumentation. In PMF, the simple MS spectrum of a sample containing the enzymatic digest of separated proteins is collected. In the most common protocol, proteins are separated by 2D polyacrylamide gel-electrophoresis (2D-PAGE), and selected gel spots are digested and then analyzed by MALDI-MS, hence one spectrum for each spot (preferably one abundant protein) is produced. Protein identification is possible through matching the observed peptide masses to theoretical masses from the *in-silico* digest of protein entries from a selected protein-sequence database using different algorithms [11]. This approach identifies a protein directly (like the top-down approach), eliminating the protein-inference problem, but without providing information on the intact mass. The significance of identification, which is generally predicted from the uniqueness of the common occurrence of assigned peptides, highly depends on the search space (database size, enzyme specificity, mass precision, and variable PTMs) and sample complexity [with at most a few (3 or 4) proteins in a sample]. Protein separation is therefore vital, and protein contamination (e.g., human keratin during processing) should be avoided.

PMF is nowadays overshadowed by the more reliable, more robust MS/MS-based identification, but it is still applied, as a complementary fast screening method.

Recent developments focus on exploiting the high mass precision of instruments [12] or improving the significance level via novel statistical algorithms [11,13].

The PMF approach may also be incorporated in MS/MS search algorithms as a supporting tool and scoring factor for protein inference [14,15].

2.2.3. MS/MS peptide identification

2.2.3.1. Fragmentation techniques (CID/HCD, ECD/ETD). The backbone fragmentation of positively-charged peptide ions in the gaseous phase by collision-induced dissociation (CID) is an extensively studied process. There are some open theoretical questions, but it is well described experimentally to be suitable for the identification of peptide sequences [16]. Most commonly, fragmentation occurs at the peptide bonds, producing any of the so-called a/x, b/y or c/z ion pairs, related to the sequence fragments of the precursor peptide.

Fragmentation rules for modified amino-acid side chains (producing neutral loss or marker fragment ions) may permit identification and localization of certain PTMs (Fig. 3.). These fragmentation rules have been incorporated into identification software to generate fragment spectra *in silico*, to be matched by the experimental spectra.

The more recent electron-capture dissociation (ECD) and electron-transfer dissociation (ETD) techniques produce different fragment ions; hence, they can complement CID for more reliable peptide identification with higher sequence coverage [18]. Different types of CID cells [ion-trap (IT) or quadrupole (Q)] and techniques (CID or ECD/ETD) produce distinct distributions of fragment ions. Identification algorithms take this into account, and apply scoring profiles or include some machine learning to fine-tune scoring for novel fragmentations [19,20].

The more effective ECD/ETD fragmentation can be utilized in all proteomic approaches [19,20]. The higher tendency to retain the PTM side-chain groups in the fragments produced by these types of fragmentation helps localization of the modification within the peptide. The higher-energy collisional dissociation (HCD) introduced in novel Orbitrap (OT) instruments is a CID technique without the low-mass limitations of other ITs. Peptide identification benefits more from other properties of Orbitraps (e.g., high mass resolution and precision) or the increased duty cycle because of parallel fragmentation and mass analysis [21].

2.2.3.2. Fragmentation modes (DDA, DIA). In the traditional data-dependent MS/MS analysis (DDA), MS and MS/MS cycles are repeated, and ions for MS/MS fragmentation are selected and isolated from each MS scan based on several criteria. The number of ions to be fragmented is chosen by the user and has to be matched to sample complexity. MS scan speed and chromatographic peak width in the case of LC-MS/MS. In MALDI-MS/MS, only the sample amount (spot size and homogeneity) is a limitation. In this way, generally the 2–10 most prevalent ions [from those that meet the other criteria (e.g., charge state, m/z limits, and presence or absence on inclusion or exclusion lists)] are fragmented in each cycle. This method produces a limited depth of identification in the case of the co-elution of more peptide ions, including different charge states of the same peptide, if no charge deconvolution is applied before peak selection.

Depending on the cycle time, quantitation is limited in LC-MS due to its low sampling rate. To overcome these problems, data-independent acquisition (DIA) methods were introduced in the past decade. In these approaches, ions in the whole *m/z* range or in a wide *m/z* range are fragmented simultaneously, and no specific ion-selection is applied. The first "all-ion fragmentation" (the simultaneous acquisition of exact mass at low and high collision energies, MS^E) was introduced by Bateman et al. in 2002 on time-of-flight (TOF) instruments [22]. Dedicated database-search software was soon commercialized [e.g., ProteinLynx Global Server (PLGS), Waters Corp., Milford, MA, USA] for protein identification and quantitation. Later, this approach was adapted to other types of instrument, and now this measurement mode is available in the software of most major MS vendors for variable MS analyzer types (e.g., TOF, IT, and Orbitrap) [23,24].

Another approach devised by Aebersold's group carries out fragmentation on limited *m/z* regions (e.g., 20–25 Da wide) and the fragmentation window is shifted for each fragmentation cycle {SWATH fragmentation [25,26]}. However, this approach generally aims for highly-specific quantification and uses a separate DDA measurement for identification.

In all these cases, the fragmentation of several ions occurs simultaneously, so no direct assignments of MS/MS fragments to precursor ions are possible. The assignment can be performed through common chromatographic features (e.g., identical retention-time profile) of precursor and fragment ions. Fragment assignment can be improved by ultra-performance liquid chromatography (UPLC), applying a SWATH-like approach (limited number of precursor ions) or the combined application of LC and ion-mobility spectrometry (IMS) in high-definition MS analysis ($HDMS^E$) [27,28]. The advantage of DIA methods in protein identification is that fragmentation (and possibly identification) is not limited to ions of highest abundance; hence, a higher dynamic range can be achieved in a single LC-MS run. The disadvantage is that the fragmentation efficiency of peptide ions strongly depends on the mass and the charge state, and it is not possible to apply optimal activation energies for individual ions, as in DDA, resulting in a somewhat poorer MS/MS quality. However, applying "collision-energy ramping", the quality of MS/MS spectra is still satisfactory, and the higher protein-sequence coverage partly compensates for the problem. The processing of DIA data requires special software, and protein identification based on such data is also more reliable with dedicated algorithms, or a modified scoring scheme in existing software packages [15,23,24,29]. A promising step towards generalized pipelines

Please cite this article in press as: Zoltan Szabo, Tamas Janaky, Challenges and developments in protein identification using mass spectrometry, Trends in Analytical Chemistry (2015), doi: 10.1016/j.trac.2015.03.007

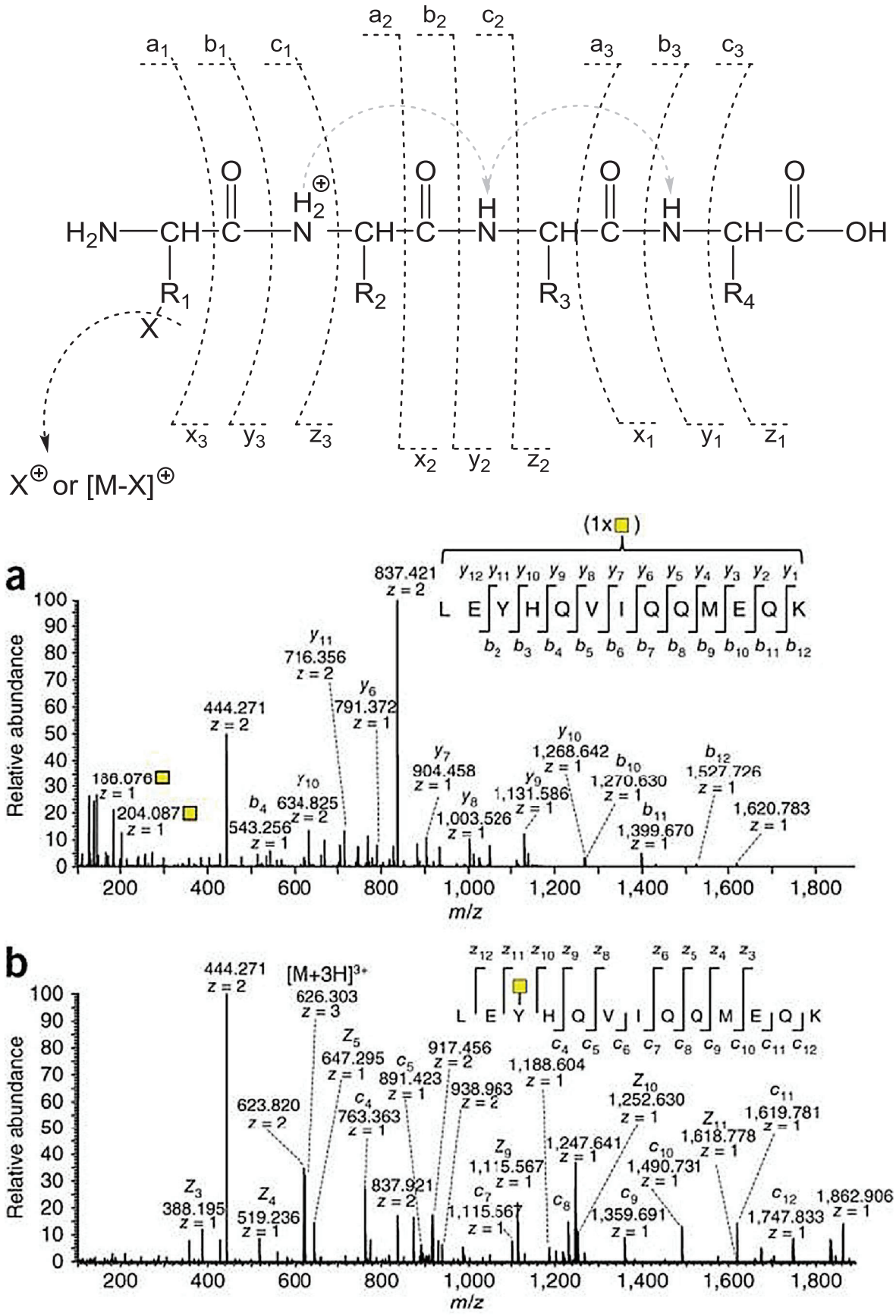


Fig. 3. Peptide fragmentation. Top: Backbone fragmentation patterns. Formation of most important ion types via mobile proton-induced peptide-bond breakage and side-chain fragmentation. Bottom: ESI-MS/MS spectra of an O-glycopeptide on Orbitrap mass spectrometer: (a) HCD-MS/MS spectrum, (b) ETD-MS/MS spectrum. Deduced glycosylated residue is denoted by square. ETD fragmentation allows localization of modification by retained functional group. {Used by permission from [17]}.

Please cite this article in press as: Zoltan Szabo, Tamas Janaky, Challenges and developments in protein identification using mass spectrometry, Trends in Analytical Chemistry (2015), doi: 10.1016/j.trac.2015.03.007

is the recent release of the DIA-Umpire [30] software package, which permits protein identification using traditional database-search tools from MS/MS data collected via any currently-used DIA approach (including SWATH).

2.2.3.3. Ion-mobility MS (IM-MS). Ion mobility is treated as an orthogonal gas-phase peptide-separation technique in protein MS [31]. It is an additional dimension of separation based on gas-phase collisional cross section and charge, extending the resolving power of the available separation techniques. Although the concept of IMS and initial applications date from the 1970s, bioanalytical applications (e.g., in proteomics) expanded after the commercial availability of hybrid TOF, IT and Orbitrap instruments capable of IMS-MS.

The additional level of separation is exploited in a bottom-up approach in resolving the chimera MS/MS spectra of co-eluting peptides in DDA (if within the mass-selection window) or in DIA (e.g., HDMS^E on Waters Q-TOF instruments) [6,27,32,33].

The processing of IM-MS data requires special software, which is generally provided by the instrument vendor, but free or open source tools are also being developed.

IM-MS is also a powerful tool in top-down proteomics [34], as we show in sub-section 2.2.7 (below).

The combination of MALDI with IM-MS enhances the reliability of identification in imaging applications [35] and allows utilization of DIA fragmentation with this ionization method.

2.2.4. Informatics methods for MS-based protein identification

MS-based proteomics uses different software for data acquisition, processing, analysis, and representation. Table 1 gives a comprehensive, but by no means complete, list of websites of databases, database searching, *de novo* sequencing, spectral library-searching algorithms, and tools for post-identification processing, MS data management and other software.

2.2.4.1. Raw data processing. Most protein-identification platforms match MS peaks to mass or charge values of some theoretical peaks. These software tools work with peak-lists (*m/z*–intensity pairs) generated from raw MS or MS/MS data. The form and the size of the raw data and the required processing steps vary with MS methods. Detailed discussion of these steps (e.g., smoothing, centroiding, charge deconvolution, and peak integration) is beyond the scope of this review. However, the importance of carefully-selected parameters has to be pointed out, as a lot of information may be lost and identification results may change significantly [36]. The novel data-independent MS/MS acquisition methods brought new peak-list-generation challenges by the assignment of fragment ions to precursors via their chromatographic similarity. Insufficient chromatographic separation causes difficulties for this step. One solution is to postpone this decision until later, to the peptide identification step, as in the Apex3D algorithm in PLGS (Waters Corp., Milford, MA, USA) [15]. With this method, fragments are not assigned exclusively to any precursor ion, but, during multi-round searches, fragment and precursor ions that have been identified are removed from the pool of the next round. Several other solutions can be found for processing DIA data to generate input peak-lists for database search [24,37,38]. Because only a small fraction of fragmentation events provide MS/MS spectra of good enough quality for peptide identification, filtering low-quality spectra speeds up the identification search [39].

2.2.4.2. De novo *sequencing and finding short sequence tags. De novo* sequencing is the identification of a protein or peptide sequence without any *a priori* knowledge of or limitation to sequences found in a database. Complete sequence identification relies on finding sequential fragment ion-pair ladders for experimental precursor masses. The search space for the identification of an MS/MS spectrum is limited only by mass precision determined by the instrument applied, hence its size can increase to dimensions that are difficult to handle by most computer algorithms, especially if variable PTMs are also assumed.

For successful automatic sequencing, strict enzyme specificity, high mass precision and no unknown modifications are preferential parameters. Even in a limited search space, the quality of the fragmentation spectrum is the main factor in successful identification. However, quality means a good fit to the theoretically-predicted MS/MS peaks. Improvements in statistical analysis and development of more sophisticated scoring algorithms (e.g., dealing with theoretical or experimental prediction of ion intensities, fragmentation patterns or PTMs) are the focus of recent developments [40].

Table 1
Examples of available software tools and databases for MS-based proteomics

Tool	WWW site
Websites collecting tools, databases, organizations	
Expasy tools	www.expasy.org
EBI tools, databases	www.ebi.ac.uk
Uniprot, Swissprot, neXtProt	www.uniprot.org, www.nextprot.org
NCBI database, tools	www.ncbi.nlm.nih.gov
HUPO, c-HPP	www.hupo.org, www.c-hpp.org
De novo sequencing	
Lutefisk	www.hairyfatguy.com/Lutefisk
PepNovo	proteomics.ucsd.edu/Software/PepNovo/
PEAKS	www.bioinformaticssolutions.com
pNovo+	pfind.ict.ac.cn/software/pNovo/index.html
Database search	
SEQUEST	thermo.com
SpectrumMill	www.chem.agilent.com
Proteinlynx GlobalServer	www.waters.com
ProteinPilot	www.absciex.com
ProteinProspector	prospector.ucsf.edu
MASCOT	matrixscience.com
ProbID	tools.proteomecenter.org/wiki/index.php?title=Software:ProbID
X! Tandem (+the GPMdb database)	www.thegpm.org
MS-GF+	proteomics.ucsd.edu/software-tools/ms-gf/
Morpheus	morpheus-ms.sourceforge.net/
MS Amanda	ms.imp.ac.at/?goto=msamanda
Sequence tag and combined approaches	
InsPecT	proteomics.ucsd.edu/Software/Inspect.html
Popitam	code.google.com/p/popitam
TagRecon, DirectTag	fenchurch.mc.vanderbilt.edu/software.php
ByOnic	www.proteinmetrics.com/products/byonic/
Spectral library search	
SpectraST	www.peptideatlas.org/spectrast/
X! P3	p3.thegpm.org
Bibliospec	skyline.gs.washington.edu
Post-identification processing	
PeptideProphet/ProteinProphet	www.proteomecenter.org/software.php
Percolator	noble.gs.washington.edu/proj/percolator/
Scaffold	www.proteomesoftware.com/
MassSieve	www.ncbi.nlm.nih.gov/staff/slottad/MassSieve/
PeptideClassifier	http://www.mop.uzh.ch/software.html
MS data management, spectral libraries	
PeptideAtlas	www.peptideatlas.org
Proteios	www.proteios.org
SBEAMS	sbeams.org
CPAS	www.labkey.org/
PRIDE	www.ebi.ac.uk/pride/
MASPECTRAS 2	genome.tugraz.at/maspectras
ProteomXchange	www.proteomexchange.org
Multifunctional frameworks and pipelines	
MaxQuant (Andromeda search)	www.biochem.mpg.de/en/rd/maxquant/
Easyprot (Phenyx search)	easyprot.unige.ch/
VEMS 5.0	portugene.com/vems.html
Trans-proteomic pipeline	www.proteomecenter.org/software.php

Please cite this article in press as: Zoltan Szabo, Tamas Janaky, Challenges and developments in protein identification using mass spectrometry, Trends in Analytical Chemistry (2015), doi: 10.1016/j.trac.2015.03.007

Another approach is combining complementary data from several experiments. One excellent example is the pipeline of Bandeira et al. [41], who combined MS/MS spectra of overlapping peptides produced by cleavage with different enzymes, and complementary fragmentation spectra of each peptide using different fragmentation techniques (CID/HCD/ETD). This approach yielded a high-throughput, high-coverage shotgun sequencing of unknown proteins [42].

A further approach is the combination of complementary bottom-up and top-down MS/MS fragmentation [43,44], which aids the alignment of peptide fragments from a bottom-up analysis to the intact protein sequence.

The identification of PTMs by *de novo* sequencing is still a challenge, even though there have been some developments in this direction [40,45]. The traditional manual *de novo* sequencing is also helped by most current proteomics software packages (e.g., calculation and identification of specific mass differences) and is a powerful, but low-throughput, tool that requires great expertise [46].

Even with the combination of results from the most sophisticated algorithms or manual investigation, there may be some gaps in the sequences identified, for which no significant assignment is possible because of the experimental data being imperfect. In such cases, short pieces of amino-acid sequences from sequential MS/MS fragment ladders may be identified, even if the precursor peptide-ion mass does not match any combination allowed. The MS-tag approach utilizes these short sequences (tags) and the missing mass gaps for identification of proteins from a sequence database. This approach can be extremely useful in the case of unknowns or a high rate of peptide modifications, as the unexpectedly altered precursor peptide mass does not completely hinder identification as in other approaches (*de novo* or MS/MS database search), whose methods generally first limit the search space to the precursor-ion mass [47].

2.2.4.3. MS/MS database search (peptide fragment fingerprinting, PFF). From the search algorithm point of view, an MS/MS database search is similar to PMF, but each spectrum search identifies a peptide as a peptide spectrum match (PSM) to an enzymatic cleavage product of a protein taken from the sequence database. However, the majority of PSMs are not proteotypic, and can be assigned to more than one protein, so the protein-inference problem also has to be solved. Peptide identification by MS/MS fragmentation spectra applies on the same fragmentation rules as in automatic *de novo* sequencing, but the search space is highly limited, as only a small fraction of the theoretically-possible amino-acid combinations is realized in living systems. This method can therefore better handle decreased enzyme specificity, or unknown PTMs. In addition to the continuous emergence of novel search algorithms promising higher identification confidence or robustness on traditional DDA CID data, current progress is driven by novel fragmentation techniques (ETD/HCD) and DIA methods.

Confidence in identification can be increased by probabilistic approaches, applying better statistical algorithms, or incorporating some methods to estimate false-discovery rates (FDRs). A non-probabilistic approach applies additional peptide information other than amino-acid sequence to confirm identification. These can be some predicted physicochemical characteristics (e.g., LC retention time) or specific fragmentation or ionization rules {e.g., the peptide3D algorithm integrated in the ProteinLynx Global Server (Waters Corp., Milford, MA, USA) software system to process data collected by the MS^E/$HDMS^E$, where the DIA mode evaluates 14 different parameters including match to predicted retention time and charge state [14]}. The Percolator algorithm rescores peptide-identification results of several different database search engines using semi-supervised machine learning. The algorithm has already been included as an option in some of the search engines and post-processing tools [48]. Current search engines calculate the FDR, which is essential in the case of shotgun proteomics, identifying up to several thousand of proteins in one analysis, but there are several post-identification tools for this as well [49]. The most common way of FDR estimation is a target-decoy approach, running a search on a decoy database consisting of definitely false-protein entries, as a separate database or attached to the target database. The number of identifications with the acceptance criteria in the decoy database estimates the amount of false positives from the target search. Such a decoy database might be an independent database with random sequences, but the best FDR estimation is achieved if it is created from the applied target database by randomization or reversing sequences, thus having the same database size and protein-size distribution [50]. FDR can be calculated at peptide or protein level, using one of numerous algorithms. The recently-released ProteoStats open-source package tries to synthesize several of these algorithms [51], and there are solutions for FDR estimation without a decoy search [52]. We discuss integration and further application of FDR evaluation in proteomics pipelines later.

Although peptide and protein identification benefits from novel instrumentation, real improvement also needs novel algorithms. For example, Morpheus [53] and MS Amanda [54] are new database-search algorithms for DDA data, which seek to enhance protein identification by exploiting high-resolution MS. Also, MS Amanda evaluates the intensity of fragment ions in most of the currently available fragmentation methods (i.e., CID, HCD, and ETD).

However, novel search engines generally face the problem of being accepted and used by the community for reasons of confidence (from users and reviewers) and convenience (fitting into existing protocols and pipelines). For easier evaluation and comparison of novel software tools, several standard data sets can be found in proteomics data repositories [55]. The "ISB standard protein mix" (mixture of 18 proteins analyzed on eight different MS systems) [56] and standard protein samples are also commercially available from chemical suppliers.

2.2.4.4. Spectral library search. Spectral library search, an alternative way for peptide identification using MS/MS spectra, relies on reproducible fragmentation of peptides. Millions of high-quality peptide-fragmentation spectra can be produced in a series of shotgun proteomics experiments. The ever-growing deposition of identified and annotated MS/MS spectra in peptide spectral libraries {e.g., PeptideAtlas [57]} creates a tremendous amount of data that can be utilized for proteomics identification. The spectral library can also be used to find experimentally-confirmed peaks and MS/MS transitions for targeted protein quantification using text search and spectrum browsing. Protein identification using spectral library matching is around two orders of magnitude faster than MS/MS database search (PFF) due to the limited search space, but may produce higher specificity, as all spectral features of a real spectrum (i.e., not theoretical) can be used in scoring [58].

2.2.4.5. Sequence-similarity searches. In the case of *de novo* sequencing or Edman degradation, when no database is applied for peptide sequencing, the relation of the identified sequences to any known protein should be determined. Sequence-similarity search may be required, if the identification of the protein or peptide is statistically significant but protein assignment is questionable due to low sequence coverage or incorrect taxonomy. Running a sequence alignment is generally faster than a new search of the MS or MS/MS database, and it also provides additional information; and, most protein-identification software packages offer a simple, convenient way to run such a search on the identified sequences.

Sequence-similarity searches are widely used in genomics. The same methodology has been transferred to protein analysis. However, scoring matrices are based on biological similarity and genetic

Please cite this article in press as: Zoltan Szabo, Tamas Janaky, Challenges and developments in protein identification using mass spectrometry, Trends in Analytical Chemistry (2015), doi: 10.1016/j.trac.2015.03.007

relations of amino acids, and changes in the total mass of the peptides are not considered, as they should be for MS analysis.

The MS-blast approach introduces a more MS-related substitution matrix, and is more relevant in this way [9].

2.2.5. Special methods for identification of proteoforms

Modification-specific proteomic identification is one of the fastest growing areas promoted by novel technological developments and accumulated knowledge about their importance (e.g., in protein function, activation or deactivation of enzymes, and role in disease). Protein modifications to be identified can be site specific, modifying particular amino-acid side chains, or location specific, found on specific regions of the polypeptide chain (e.g., C-terminal or N-terminal). Proteoforms can have different amino-acid sequences due to genetic polymorphism [in deoxyribonucleic acid, (DNA)], alternative splicing [in ribonucleic acid (RNA)] or protein processing and degradation. Pre-, co- and post-translational enzymatic modifications, chemical artificial or environmental modifications can introduce side-chain or terminal modifications. In the MS spectrum, these amino-acid modifications appear as mass shifts from the unmodified amino-acid mass, and as the sum of all mass shifts in the mass of a peptide/protein in the case of multiple modifications.

The most complete database of amino-acid mass shifts (UNIMOD, www.unimod.org) contains more than 7000 entries, although about 200 of these have biological relevance. The rest relate to chemical artifacts, derivatization or labeling, or multiple occurrences at different amino acids. During a general database search, a certain number of these mass-shift values can be considered to limit the search space. One solution for identification of unexpected modifications is by running a two-round search. First, proteins are identified using strict modification freedom, then, in a second round, a search run is carried out on the unidentified spectra considering a higher level of variability in modifications (e.g., all possible modifications plus single amino-acid substitutions), but it is limited to proteins identified in the first round. This approach allows the identification of any modification available in the modification database, but it generally provides only a mathematical solution for the observed spectra. The chemical or biological relevance of a suggested combination of modifications needs to be considered by the user. The identification of amino-acid modifications is generally confirmed by modification-specific fragmentation patterns (neutral losses or marker fragment ions depending on fragmentation technique). Confident localization within the peptide requires high sequence coverage with fragments retaining the modification group, which is more probable using electron-based fragmentation (ECD/ETD [17],) and a longer amino-acid sequence (e.g., using the middle-down or top-down method). Special cases are those PTMs in which large, highly variable groups are attached to specific amino acids, as in glycation, glycosylation and ubiquitination. In such cases, an oligosaccharide or a polypeptide chain with very large mass shift (several hundred or thousands of Da) is the intact modifier group, and a structural study of the modifier is then worthwhile [59,60]. However, in most bottom-up protocols and MS ionization/fragmentation methods, modifiers can degrade and small marker groups remain as the signature of modification.

Most biologically-relevant chemical modifications are present with sub-stoichiometric occupancy and only a small fraction of the genetically-expressed protein exists at any moment in a specific protein form. Also, some modifications decrease ionization efficiency (MS sensitivity), so, if one uses the general approaches to identification above, the discovery rate of modification will be low. Experiments targeting the identification of protein modifications generally require specific purification and enrichment procedures. Several immunochemical and chemical enrichment methods have been applied successfully to the study of individual proteins or to large-scale analysis [e.g., glycopeptides can be enriched by hydrophilic interaction chromatography (HILIC) or a hydrazide-derivatization method]. The enrichment of phosphorylated peptides is generally based on the affinity of the phosphate group to some metal oxides (e.g., TiO_2, or ZrO_2) or to a higher charged cation (e.g., Fe^{3+}, Al^{3+}, Cu^{2+}, or Ti^{4+}) in metal-oxide affinity chromatography (MOAC) or immobilized metal-ion affinity chromatography (IMAC), respectively. However, the selectivity of enrichment is made worse by the non-specific binding of amino acids (Asp, Glu) to the affinity matrices in both cases. For more details, dedicated reviews on the very diverse topic of PTM enrichment are available [59,61–63].

The large-scale identification of C- and N-terminals using the bottom-up approach also requires derivatization before protein digestion and direct or indirect enrichment to differentiate between these and termini produced during protein cleavage [64]. Without these steps, searching for the occasional occurrence of predicted terminal or internal peptides unique to processed or spliced forms of individual proteins among the identified peptides is the solution, but its unpredictable success greatly depends on the specificity, the abundance and the detectability of these peptides.

Large-scale, modification-specific bottom-up proteomics methods using state-of-the-art instrumentation and software allow the identification of thousands of modification sites proteome wide in one experiment [65]. Mining the resulting huge databases [5,66,67] provides information on a biologically-realized PTM pool, though with a little or no information on biochemical function.

Most recent studies on PTM identification were conducted with the idea of PTM cross-talk, as activation and functions of proteins are controlled by combinations and patterns of several different PTMs on the same polypeptide chain [68,69]. Information on these patterns and protein processing (e.g., truncation) is lost or hardly observable in most shotgun bottom-up methods. In top-down proteomics, it is much easier to find the right place for each part of the puzzle, as the total size and the origins of pieces are known. The recent progress in top-down proteomics was driven by recognition of these possibilities and promoted by novel instrumentation [8,70].

The results of the first pilot study by the top-down consortium on histons (one of the most extensively modified protein groups) were published recently [71]; 74 proteoforms of human histon H4 were identified by the laboratories involved in the project.

The determination of disulfide bridges within or between protein chains is also very important. The question of whether it is an identification or structural investigation is really academic, but the methodologies applied are closer to structural analysis [72]. Active structure, enzyme resistance and interactions of biologically-active polypeptides (e.g., snake venoms, or antifungal agents) are also controlled by disulfide-bridge patterns [73].

2.2.6. Middle-down proteomics

Middle-down proteomics uses methods similar to the bottom-up approach [74], but the enzymes of choice for protein cleavage (e.g., AspC) have specificity to amino acids with lower sequence frequency [75], so they produce larger peptide fragments (e.g., 3–10 kDa). This allows mapping of PTMs on highly modified proteins {e.g., histons [76]} and the study of modifications that are generally damaged in a bottom-up protocol {e.g., ubiquitin-like peptidyl modifications [77]}.

2.2.7. Top-down proteomics

The main advantage of the top-down proteomics workflow (analyzing intact proteins by MS) is the ability to identify proteins directly, so the protein-inference problem is eliminated, and it permits the identification of specific proteoforms based on intact molecular mass. Still, an analysis of proteins with larger molecular mass can be a challenge to the MS instrumentation. Ion detection in the higher mass range (e.g., 5–200 kDa) of intact proteins is generally not a

Please cite this article in press as: Zoltan Szabo, Tamas Janaky, Challenges and developments in protein identification using mass spectrometry, Trends in Analytical Chemistry (2015), doi: 10.1016/j.trac.2015.03.007

major problem as mass spectrometers work with mass/charge (*m/z*) values, and proteins generally acquire a higher number of charges (e.g., more than 10 protons) at the low pH of analysis in an electrospray ion source. However, acceptable sensitivity (generally still lower than in the bottom-up approach) requires tuning of ionization and ion transfer in the instrument. Chemical additives ("supercharging" reagents) may also be applied to enhance ionization and to increase the charge of protein ions. Also, the determination of exact mass with those high charges requires high mass resolution to resolve isotope distribution, and to separate proteins or proteoforms with similar masses.

Recent developments in TOF and Orbitrap mass analyzers make this higher resolution available for a wider range of laboratories. Problems related to mass resolution and sensitivity suggest that protein purification, separation and enrichment are also crucial steps prior to MS analysis to decrease complexity and to increase the signal-to noise ratio of MS spectra. The application of an ion-mobility cell in the MS instrument can also help to resolve protein ions with a similar mass, but different structure.

Several IMS-TOF instruments have been available for years, but the real benefits of these in top-down or bottom-up proteomics were recognized just recently and accepted by the proteomics community [78]. However, even with the highest currently available precision, the measured protein mass is unsatisfactory for the unambiguous identification of a protein species, as numerous amino-acid sequences can have molecular mass within the precision limits of the determined mass. As the combination of possible PTMs increases the number of candidates enormously, MS/MS analysis is essential. The fragmentation of high-mass protein ions has a preference for backbone cleavage, hence retaining PTM groups on amino-acid side chains, as opposed to fragmentation of lower mass peptides, which have a greater tendency to lose these groups, hindering their localization. The poor fragmentation using traditional CID may be improved by applying electron-induced fragmentation techniques (ECD, or ETD). The resulting MS/MS spectra may be hard to analyze, because of high complexity related to the large number of possible fragments with various charge states. An ion-mobility analysis of fragments can be performed to reduce complexity (mainly based on ion charge) [34], and several software solutions have been developed to interpret the spectra and to identify amino-acid sequences and PTMs of proteins [8,70,79].

2.3. Post-identification data analysis

As we have shown, the validation and the confirmation of identification results, especially in the case of large-scale studies, might be necessary. An estimation of FDR is one way of improving sensitivity. A combination of results from different database search engines, or from database search and *de novo* sequencing, bottom-up/top-down methods also increases confidence in common identifications and increases sensitivity by complementary results. The incorporation of these processes into ever more user-friendly modular proteomics pipelines is a tendency that promotes the production of reliable data by the proteomics community [80].

The protein-inference problem is addressed in most of the bottom-up peptide-identification software. However, the solutions available there may be unsatisfactory or too inflexible to satisfy specific needs. Still, in a combination of peptide identifications from several search engines, it has to be solved separately on the combined data. FDR calculations have already been discussed, as it is not necessarily a separate step in the pipeline (it may be done within the identification software), but it also has to be reevaluated in combination of results, at least at the protein level.

The general solution for protein inference is to find the simplest (minimal number of proteins) combination that accounts for the presence of all the unique and common peptides identified with high confidence [81,82]. Changing the database or acceptance criteria will change the set of inferred proteins; hence, protein FDR estimation and protein inference are solved using the same software tool. For example, the PeptideProphet/ProteinProphet algorithms invented in the TransProteomic Pipeline (TPP) estimates score distribution of valid and false identifications [83]. The result of protein inference is generally a list of protein groups, where groups are differentiated by unique peptides, but members of the same group cannot be differentiated using the given set of data [81]. Several pipelines or frameworks are available for complete quantitative and qualitative proteomic analysis [84], and transfer of data to data repositories [85,86] or a further biological investigation [87].

Another problem in the gel-based proteomics method, which remains unpopular, relates to the independent gel-based quantification and MS identification. As the gel-staining intensity is proportional to the total protein content in any gel spot in 2D PAGE, it is impossible to assign quantitative and qualitative information if more than one protein is identified in the spot. This is the case in the majority of spots using high-sensitivity MS/MS identification, due to the limited resolution of 2D PAGE [88]. One solution is to perform an MS identification using a protocol that provides absolute or relative quantitative information, which may be used to identify major components or to correlate MS quantitation with gel staining [89].

Any MS-based proteomics identification should be validated by orthogonal methods (e.g., immunochemistry) or by high-throughput targeted MS quantitation, which is, of course, possible for only a small subset of the results of shotgun analysis, generally selected based on biological significance. In close partnership with the Human Protein Atlas project [www.proteinatlas.org], Atlas Antibodies has developed over 18,000 antibodies covering 15,000 human-gene products and an additional 19,000 advanced reagents for MS-based quantitative proteomics [atlasantibodies.com]. Proteoform identification and validation of data will benefit from community efforts on integration of genomic, transcriptomic and proteomic data, building databases of proteoforms and their relation to diseases (e.g., c-HPP, neXtProt on human proteome, see Table 1.).

Once protein inference has been solved by an analytical scientist, the biological inference problem should be solved: the list of proteins or PTMs identified should be put into the correct biochemical context [90]. As these steps are generally carried out by a different researcher, biological goals and the limitations of the analytical methods should be harmonized. As shown above, biological processes are governed by interacting protein complexes, which may be constructed using a different stoichiometry of the different proteoforms of each member. The detection of quantitative changes or identification of any specific member or sub-group may lead to the wrong conclusions, even if the most up-to-date ontology or network analysis is carried out [87].

3. Conclusions and future prospects

Recent technological developments have led to huge amounts of large-scale bottom-up protein identifications and PTM-analysis data being available in the literature and databases. New developments in instrumentation also initiated the successful application of top-down methods in a large-scale way. In such experiments, after producing several thousand protein identifications in a single analysis, manual validation of data is impossible, so, with the rise of shotgun proteomics methods, publication guidelines were established by proteomics organizations to ensure the quality of published data and to facilitate independent validation [84,91]. Novel techniques and types of data require continual updates of these guidelines and requirements, extending to PTM identification and top-down methods. As the identification of specific proteoforms is of growing importance, methodologies with better capabilities for

Please cite this article in press as: Zoltan Szabo, Tamas Janaky, Challenges and developments in protein identification using mass spectrometry, Trends in Analytical Chemistry (2015), doi: 10.1016/j.trac.2015.03.007

the identification of these (e.g., top-down methods) will be viewed as more important, and will be more widely applied and developed. Combination of different approaches also holds the promise of achieving a greater depth of analysis and making progress in proteoform identification. Another hot topic, label-free MS-based protein quantification, is beyond the scope of this review, but developments in it will also have an impact on protein-identification methods and requirements (e.g., data-independent fragmentation methods, quantitation and identification of proteoforms, and novel solutions for the protein-inference problem).

References

[1] M.R. Wilkins, J.C. Sanchez, A.A. Gooley, R.D. Appel, I. Humphery-Smith, D.F. Hochstrasser, et al., Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it, Biotechnol. Genet. Eng. Rev. 13 (1996) 19–50.
[2] M.S. Kim, S.M. Pinto, D. Getnet, R.S. Nirujogi, S.S. Manda, R. Chaerkady, et al., A draft map of the human proteome, Nature 509 (2014) 575–581.
[3] M. Wilhelm, J. Schlegl, H. Hahne, A. Moghaddas Gholami, M. Lieberenz, M.M. Savitski, et al., Mass-spectrometry-based draft of the human proteome, Nature 509 (2014) 582–587.
[4] G.B. Smejkal, Genomics and proteomics: of hares, tortoises and the complexity of tortoises, Expert Rev. Proteomics 9 (2012) 469–472.
[5] G.A. Khoury, R.C. Baliban, C.A. Floudas, Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database, Sci. Rep. 1 (2011) doi:10.1038/srep00090.
[6] Z. Zhang, S. Wu, D.L. Stenoien, L. Pasa-Tolic, High-throughput proteomics, Annu. Rev. Anal. Chem. 7 (2014) 427–454.
[7] Y. Zhang, B.R. Fonslow, B. Shan, M.C. Baek, J.R. Yates 3rd, Protein analysis by shotgun/bottom-up proteomics, Chem. Rev. 113 (2013) 2343–2394.
[8] H. Zhang, Y. Ge, Comprehensive analysis of protein modifications by top-down mass spectrometry, Circ. Cardiovasc. Genet. 4 (2011) 711–721.
[9] A. Shevchenko, S. Sunyaev, A. Loboda, A. Shevchenko, P. Bork, W. Ens, et al., Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching, Anal. Chem. 73 (2001) 1917–1926.
[10] T.E. Angel, U.K. Aryal, S.M. Hengel, E.S. Baker, R.T. Kelly, E.W. Robinson, et al., Mass spectrometry-based proteomics: existing capabilities and future directions, Chem. Soc. Rev. 41 (2012) 3912–3928.
[11] Z. He, C. Yang, W. Yu, Peak bagging for peptide mass fingerprinting, Bioinformatics 24 (2008) 1293–1299.
[12] E.D. Dodds, B.H. Clowers, P.J. Hagerman, C.B. Lebrilla, Systematic characterization of high mass accuracy influence on false discovery and probability scoring in peptide mass fingerprinting, Anal. Biochem. 372 (2008) 156–166.
[13] Y. Li, P. Hao, S. Zhang, Y. Li, Feature-matching pattern-based support vector machines for robust peptide mass fingerprinting, Mol. Cell. Proteomics 10 (2011) M110.005785, doi:10.1074/mcp.M110.005785.
[14] G.Z. Li, J.P. Vissers, J.C. Silva, D. Golick, M.V. Gorenstein, S.J. Geromanos, Database searching and accounting of multiplexed precursor and product ion spectra from the data independent analysis of simple and complex peptide mixtures, Proteomics 9 (2009) 1696–1719.
[15] S.J. Geromanos, J.P. Vissers, J.C. Silva, C.A. Dorschel, G.Z. Li, M.V. Gorenstein, et al., The detection, correlation, and comparison of peptide precursor and product ions from data independent LC-MS with data dependant LC-MS/MS, Proteomics 9 (2009) 1683–1695.
[16] I.A. Papayannopoulos, The interpretation of collision-induced dissociation tandem mass spectra of peptides, Mass Spectrom. Rev. 14 (1995) 49–73.
[17] C. Steentoft, S.Y. Vakhrushev, M. Vester-Christensen, K.T.-B.G. Schjoldager, Y. Kong, E.P. Bennett, et al., Mining the O-glycoproteome using zinc-finger nuclease-glycoengineered SimpleCell lines, Nat. Methods 8 (2011) 977–982.
[18] A.J. Creese, H.J. Cooper, Liquid chromatography electron capture dissociation tandem mass spectrometry (LC-ECD-MS/MS) versus liquid chromatography collision-induced dissociation tandem mass spectrometry (LC-CID-MS/MS) for the identification of proteins, J. Am. Soc. Mass Spectrom. 18 (2007) 891–897.
[19] M.S. Kim, A. Pandey, Electron transfer dissociation mass spectrometry in proteomics, Proteomics 12 (2012) 530–542.
[20] M. Sarbu, R.M. Ghiulai, A.D. Zamfir, Recent developments and applications of electron transfer dissociation mass spectrometry in proteomics, Amino Acids 46 (2014) 1625–1634.
[21] T. Geiger, J. Cox, M. Mann, Proteomics on an orbitrap benchtop mass spectrometer using all-ion fragmentation, Mol. Cell. Proteomics 9 (2010) 2252–2261.
[22] R.H. Bateman, R. Carruthers, J.B. Hoyes, C. Jones, J.I. Langridge, A. Millar, et al., A novel precursor ion discovery method on a hybrid quadrupole orthogonal acceleration time-of-flight (Q-TOF) mass spectrometer for studying protein phosphorylation, J. Am. Soc. Mass Spectrom. 13 (2002) 792–803.
[23] J.D. Chapman, D.R. Goodlett, C.D. Masselon, Multiplexed and data-independent tandem mass spectrometry for global proteome profiling, Mass Spectrom. Rev. 33 (2014) 452–470.
[24] K.P. Law, Y.P. Lim, Recent advances in mass spectrometry: data independent analysis and hyper reaction monitoring, Expert Rev. Proteomics 10 (2013) 551–566.
[25] J.D. Venable, M.Q. Dong, J. Wohlschlegel, A. Dillin, J.R. Yates, Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra, Nat. Methods 1 (2004) 39–45.
[26] L.C. Gillet, P. Navarro, S. Tate, H. Rost, N. Selevsek, L. Reiter, et al., Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis, Mol. Cell. Proteomics 11 (2012) 1–17.
[27] P.V. Shliaha, N.J. Bond, L. Gatto, K.S. Lilley, Effects of traveling wave ion mobility separation on data independent acquisition in proteomics studies, J. Proteome Res. 12 (2013) 2323–2339.
[28] S. Helm, D. Dobritzsch, A. Rödiger, B. Agne, S. Baginsky, Protein identification and quantification by data-independent acquisition and multi-parallel collision-induced dissociation mass spectrometry (MSE) in the chloroplast stroma proteome, J. Proteomics 98 (2014) 79–89.
[29] K. Buts, S. Michielssens, M.L.A.T.M. Hertog, E. Hayakawa, J. Cordewener, A.H.P. America, et al., Improving the identification rate of data independent label-free quantitative proteomics experiments on non-model crops: a case study on apple fruit, J. Proteomics 105 (2014) 31–45.
[30] T. Avtonomov, L. Tucholska, C. Gingras, et al., DIA-Umpire: comprehensive computational framework for data-independent acquisition proteomics, Nat. Methods 12 (2015) 258–264.
[31] J.A. McLean, B.T. Ruotolo, K.J. Gillig, D.H. Russell, Ion mobility–mass spectrometry: a new paradigm for proteomics, Int. J. Mass Spectrom. 240 (2005) 301–315.
[32] K.L. Crowell, E.S. Baker, S.H. Payne, Y.M. Ibrahim, M.E. Monroe, G.W. Slysz, et al., Increasing confidence of LC-MS identifications by utilizing ion mobility spectrometry, Int. J. Mass Spectrom. 354–355 (2013) 312–317.
[33] S.J. Valentine, M.A. Ewing, J.M. Dilger, M.S. Glover, S. Geromanos, C. Hughes, et al., Using ion mobility data to improve peptide identification: intrinsic amino acid size parameters, J. Proteome Res. 10 (2011) 2318–2329.
[34] N.F. Zinnel, P.J. Pai, D.H. Russell, Ion mobility-mass spectrometry (IM-MS) for top-down proteomics: increased dynamic range affords increased sequence coverage, Anal. Chem. 84 (2012) 3390–3397.
[35] N.E. Mascini, R.M.A. Heeren, Protein identification in mass-spectrometry imaging, TrAC-Trend. Anal. Chem. 40 (2012) 28–37.
[36] F. Mancuso, J. Bunkenborg, M. Wierer, H. Molina, Data extraction from proteomics raw data: an evaluation of nine tandem MS tools using a large orbitrap data set, J. Proteomics 75 (2012) 5293–5303.
[37] C.R. Weisbrod, J.K. Eng, M.R. Hoopmann, T. Baker, J.E. Bruce, Accurate peptide fragment mass analysis: multiplexed peptide identification and quantification, J. Proteome Res. 11 (2012) 1621–1632.
[38] M. Bern, G. Finney, M.R. Hoopmann, G. Merrihew, M.J. Toth, M.J. MacCoss, Deconvolution of mixture spectra from ion-trap data-independent-acquisition tandem mass spectrometry, Anal. Chem. 82 (2010) 833–841.
[39] J. Salmi, T.A. Nyman, O.S. Nevalainen, T. Aittokallio, Filtering strategies for improving protein identification in high-throughput MS/MS studies, Proteomics 9 (2009) 848–860.
[40] N. Bandeira, K.R. Clauser, P.A. Pevzner, Shotgun protein sequencing: assembly of peptide tandem mass spectra from mixtures of modified proteins, Mol. Cell. Proteomics 6 (2007) 1123–1134.
[41] A. Guthals, N. Bandeira, Peptide identification by tandem mass spectrometry with alternate fragmentation modes, Mol. Cell. Proteomics 11 (2012) 550–557.
[42] H. Chi, H. Chen, K. He, L. Wu, B. Yang, R.X. Sun, et al., pNovo+: de novo peptide sequencing using complementary HCD and ETD tandem mass spectra, J. Proteome Res. 12 (2013) 615–625.
[43] X. Liu, L.J.M. Dekker, S. Wu, M.M. Vanduijn, T.M. Luider, N. Tolic, et al., De novo protein sequencing by combining top-down and bottom-up tandem mass spectra, J. Proteome Res. 13 (2014) 3241–3248.
[44] A. Resemann, D. Wunderlich, U. Rothbauer, B. Warscheid, H. Leonhardt, J. Fuchser, et al., Top-down de novo protein sequencing of a 13.6 kDa camelid single heavy chain antibody by matrix-assisted laser desorption ionization-time-of-flight/time-of-flight mass spectrometry, Anal. Chem. 82 (2010) 3283–3292.
[45] A. Guthals, J.D. Watrous, P.C. Dorrestein, N. Bandeira, The spectral networks paradigm in high throughput mass spectrometry, Mol. Biosyst. 8 (2012) 2535–2544.
[46] K. FMedzihradszky, R.J. Chalkley, Lessons in de novo peptide sequencing by tandem mass spectrometry, Mass Spectrom. Rev. 34 (2015) 43–63.
[47] L. McHugh, J.W. Arthur, Computational methods for protein identification from mass spectrometry data, PLoS Comput. Biol. 4 (2008) e12.
[48] L. Kall, J.D. Canterbury, J. Weston, W.S. Noble, M.J. MacCoss, Semi-supervised learning for peptide identification from shotgun proteomics datasets, Nat. Methods 4 (2007) 923–925.
[49] A.I. Nesvizhskii, A survey of computational methods and error rate estimation procedures for peptide and protein identification in shotgun proteomics, J. Proteomics 73 (2010) 2092–2123.
[50] A.K. Yadav, D. Kumar, D. Dash, Learning from decoys to improve the sensitivity and specificity of proteomics database search results, PLoS ONE 7 (2012) e50651, doi:10.1371/journal.pone.0050651.
[51] A.K. Yadav, P.K. Kadimi, D. Kumar, D. Dash, ProteoStats – a library for estimating false discovery rates in proteomics pipelines, Bioinformatics 29 (2013) 2799–2800.

Please cite this article in press as: Zoltan Szabo, Tamas Janaky, Challenges and developments in protein identification using mass spectrometry, Trends in Analytical Chemistry (2015), doi: 10.1016/j.trac.2015.03.007

[52] B. Teng, T. Huang, Z. He, Decoy-free protein-level false discovery rate estimation, Bioinformatics 30 (2013) 675–681.
[53] C.D. Wenger, J.J. Coon, A proteomics search algorithm specifically designed for high-resolution tandem mass spectra, J. Proteome Res. 12 (2013) 1377–1386.
[54] V. Dorfer, P. Pichler, T. Stranzl, J. Stadlmann, T. Taus, S. Winkler, et al., MS Amanda, a universal identification algorithm optimized for high accuracy tandem mass spectra, J. Proteome Res. 13 (2014) 3679–3684.
[55] Y. Perez-Riverol, R. Wang, H. Hermjakob, M. Muller, V. Vesada, J.A. Vizcaino, Open source libraries and frameworks for mass spectrometry based proteomics: a developer's perspective, Biochim. Biophys. Acta 2014 (1844) 63–76.
[56] J. Klimek, J.S. Eddes, L. Hohmann, J. Jackson, A. Peterson, S. Letarte, et al., The standard protein mix database: a diverse dataset to assist in the production of improved peptide and protein identification software tools, J. Proteome Res. 7 (2007) 96–103.
[57] E.W. Deutsch, H. Lam, R. Aebersold, PeptideAtlas: a resource for target selection for emerging targeted proteomics workflows, EMBO Rep. 9 (2008) 429–434.
[58] H. Lam, R. Aebersold, Building and searching tandem mass (MS/MS) spectral libraries for peptide identification in proteomics, Methods 54 (2011) 424–431.
[59] H. Johnson, C.E. Eyers, Analysis of post-translational modifications by LC-MS/MS, Methods Mol. Biol. 658 (2010) 93–108.
[60] A.G. Wetie, A.G. Woods, C.C. Darie, Mass spectrometric analysis of post-translational modifications (PTMs) and protein-protein interactions (PPIs), Adv. Exp. Med. Biol. 806 (2014) 205–235.
[61] M. Cerny, J. Skalak, H. Cerna, B. Brzobohaty, Advances in purification and separation of posttranslationally modified proteins, J. Proteomics 92 (2013) 2–27.
[62] J. Huang, F. Wang, M. Ye, H. Zou, Enrichment and separation techniques for large-scale proteomics analysis of the protein post-translational modifications, J. Chromatogr. A 1372 (2014) 1–17.
[63] T. Liu, W.J. Qian, M.A. Gritsenko, D.G. Camp 2nd, M.E. Monroe, R.J. Moore, et al., Human plasma N-glycoproteome analysis by immunoaffinity subtraction, hydrazide chemistry, and mass spectrometry, J. Proteome Res. 4 (2005) 2070–2080.
[64] Z.W. Lai, A. Petrera, O. Schilling, Protein amino-terminal modifications and proteomic approaches for N-terminal profiling, Curr. Opin. Chem. Biol. 24 (2015) 71–79.
[65] J.V. Olsen, M. Mann, Status of large-scale analysis of post-translational modifications by mass spectrometry, Mol. Cell. Proteomics 12 (2013) 3444–3452.
[66] P. Minguez, I. Letunic, L. Parca, L. Garcia-Alonso, J. Dopazo, J. Huerta-Cepas, et al., PTMcode v2: a resource for functional associations of post-translational modifications within and between proteins, Nucleic Acids Res. 43 (2014) D494–D502, doi:10.1093/nar/gku1081.
[67] A.P. Lothrop, M.P. Torres, S.M. Fuchs, Deciphering post-translational modification codes, FEBS Lett. 587 (2013) 1247–1257.
[68] M. Peng, A. Scholten, A.J. Heck, B. van Breukelen, Identification of enriched PTM crosstalk motifs from large-scale experimental data sets, J. Proteome Res. 13 (2014) 249–259.
[69] A.S. Venne, L. Kollipara, R.P. Zahedi, The next level of complexity: crosstalk of posttranslational modifications, Proteomics 14 (2014) 513–524.
[70] J.D. Tipton, J.C. Tran, A.D. Catherman, D.R. Ahlf, K.R. Durbin, N.L. Kelleher, Analysis of intact protein isoforms by mass spectrometry, J. Biol. Chem. 286 (2011) 25451–25458.
[71] X. Dang, J. Scotcher, S. Wu, R.K. Chu, N. Tolic, I. Ntai, et al., The first pilot project of the consortium for top-down proteomics: a status report, Proteomics 14 (2014) 1130–1140.
[72] M.S. Goyder, F. Rebeaud, M.E. Pfeifer, F. Kalman, Strategies in mass spectrometry for the assignment of Cys-Cys disulfide connectivities in proteins, Expert Rev. Proteomics 10 (2013) 489–501.
[73] C. Hung, S. Jung, J. Grötzinger, C. Gelhaus, M. Leippe, A. Tholey, Determination of disulfide linkages in antimicrobial peptides of the macin family by combination of top-down and bottom-up proteomics, J. Proteomics 103 (2014) 216–226.
[74] J. Cannon, K. Lohnes, C. Wynne, Y. Wang, N. Edwards, C. Fenselau, High-throughput middle-down analysis using an orbitrap, J. Proteome Res. 9 (2010) 3886–3890.
[75] C. Wu, J.C. Tran, L. Zamdborg, K.R. Durbin, M. Li, D.R. Ahlf, et al., A protease for 'middle-down' proteomics, Nat. Methods 9 (2012) 822–824.
[76] S. Sidoli, V. Schwammle, C. Ruminowicz, T.A. Hansen, X. Wu, K. Helin, et al., Middle-down hybrid chromatography/tandem mass spectrometry workflow for characterization of combinatorial post-translational modifications in histones, Proteomics 14 (2014) 2200–2211.
[77] E.M. Valkevich, N.A. Sanchez, Y. Ge, E.R. Strieter, Middle-down mass spectrometry enables characterization of branched ubiquitin chains, Biochemistry 53 (2014) 4979–4989.
[78] A.D. Catherman, O.S. Skinner, N.L. Kelleher, Top Down proteomics: facts and perspectives, Biochem. Biophys. Res. Commun. 445 (2014) 683–693.
[79] A. Drabik, A. Bodzon-Kulakowska, P. Suder, Application of the ETD/PTR reactions in top-down proteomics as a faster alternative to bottom-up nanoLC-MS/MS protein identification, J. Mass Spectrom. 47 (2012) 1347–1352.
[80] S. Carr, R. Aebersold, M. Baldwin, A. Burlingame, K. Clauser, A. Nesvizhskii, et al., The need for guidelines in publication of peptide and protein identification data: working group on publication guidelines for peptide and protein identification data, Mol. Cell. Proteomics 3 (2004) 531–533.
[81] A.I. Nesvizhskii, R. Aebersold, Interpretation of shotgun proteomic data: the protein inference problem, Mol. Cell. Proteomics 4 (2005) 1419–1440.
[82] P.N.J. Lukasse, A.H.P. America, Protein inference using peptide quantification patterns, J. Proteome Res. 13 (2014) 3191–3199.
[83] D. Shteynberg, E.W. Deutsch, H. Lam, J.K. Eng, Z. Sun, N. Tasman, et al., iProphet: multi-level integrative analysis of shotgun proteomic data improves peptide and protein identification rates and error estimates, Mol. Cell. Proteomics 10 (2011) M111.007690, doi:10.1074/mcp.M111.007690.
[84] A.P. Drabovich, E. Martinez-Morillo, E.P. Diamandis, Toward an integrated pipeline for protein biomarker development, Biochim. Biophys. Acta 1854 (2014) 677–686.
[85] R.C. Jimenez, J.A. Vizcaino, Proteomics data exchange and storage: the need for common standards and public repositories, Methods Mol. Biol. 1007 (2013) 317–333.
[86] A. Csordas, D. Ovelleiro, R. Wang, J.M. Foster, D. Rios, J.A. Vizcaino, et al., PRIDE: quality control in a proteomics data repository, Database (Oxford) 2012 (2012) bas004, doi:10.1093/database/bas004.
[87] T. Rabilloud, P. Lescuyer, The proteomic to biology inference, a frequently overlooked concern in the interpretation of proteomic data: a plea for functional validation, Proteomics 14 (2014) 157–161.
[88] N. Campostrini, L.B. Areces, J. Rappsilber, M.C. Pietrogrande, F. Dondi, F. Pastorino, et al., Spot overlapping in two-dimensional maps: a serious problem ignored for much too long, Proteomics 5 (2005) 2385–2395.
[89] Z. Szabo, J.S. Szomor, I. Foeldi, T. Janaky, Mass spectrometry-based label free quantification of gel separated proteins, J. Proteomics 75 (2012) 5544–5553.
[90] A. Schmidt, I. Forne, A. Imhof, Bioinformatic analysis of proteomics data, BMC Syst. Biol. 8 (Suppl. 2) (2014) S3, doi:10.1186/1752-0509-8-S2-S3.
[91] X. Li, A. Pizarro, T. Grosser, Elective affinities – bioinformatic analysis of proteomic mass spectrometry data, Arch. Physiol. Biochem. 115 (2009) 311–319.

Please cite this article in press as: Zoltan Szabo, Tamas Janaky, Challenges and developments in protein identification using mass spectrometry, Trends in Analytical Chemistry (2015), doi: 10.1016/j.trac.2015.03.007