Combining mass spectrometry with database interrogation strategies in proteomics

Adam J. Liska, Andrej Shevchenko

Protein identification by mass spectrometry (MS) and sequence-database searching was established in 1993; since then, the proteomics community has witnessed a proliferation of analytical strategies for protein identification. Analytical strategies comprise three components: MS platforms; spectra-database sequence correlation methods; and, sequence databases. Multiple strategies are now applied simultaneously to increase sensitivity, throughput, and reliability of the characterization of proteomes. Now, by assessing the complexity of the interplay of MS, bioinformatics and sequence databases, we can begin to predict future approaches and challenges in the development of proteomics.

Keywords: Bioinformatics; Databases; Mass spectrometry; MS/MS; Proteomics; Spectra interpretation

Abbreviations: 2-D, two-dimensional (gel electrophoresis); cDNA, complementary DNA; DB, database; ESI, electrospray ionization; EST, expressed sequence tag; FT-ICR-MS, Fourier transform ion cyclotron mass spectrometry; FWHM, full width at half maximum; LC, liquid chromatography; MALDI, matrix-assisted laser-desorption/ionization; MS, mass spectrometry; MS BLAST, MS driven BLAST; MSP, MS platform; MS/MS, tandem MS; NanoESI, nanoelectrospray; PMF, peptide mass fingerprinting; PSD, post-source decay; Q(q)TOF, quadrupole time-of-flight; TOF, time-of-flight; TQ, triple quadrupole

Adam J. Liska, Andrej Shevchenko*
Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstrasse 108, D-01307 Dresden, Germany

*Corresponding author.
Tel.: +49 351 210 2615;
Fax: +49 351 210 2000;
E-mail: shevchenko@mpi-cbg.de

1. Introduction

The first problem in conducting proteomics is the development of accurate, versatile protein-identification strategies. Although it has been possible to purify proteins using established methods of biochemistry, the crucial step in the characterization of any proteome is high-throughput protein identification. Protein identification comprises associating an isolated protein with a specific gene or a specific biochemical function based upon sequence identity (reviewed in [1]).

Advances in genomic sequencing, MS, and DB searching now provide a method for high-throughput protein identification in organisms with sequenced genomes (reviewed in [2,3]). To extend these capabilities, as well as to advance the efficacy of protein identification in organisms with unsequenced genomes (reviewed in [4]), a number of recent developments point to new analytical strategies for proteomics. Analytical strategies are built upon three major components:

- *MS platforms* (coupled analytical methods with mass spectrometers);
- *spectra-DB sequence correlation methods*; and,
- *sequence DBs*.

Specific types of mass spectrometers produce spectra of varying quality, and different interpretation methods are suitable for specific types of spectra. When a specific MS platform is combined with a specific correlation method and a specific type of DB, this combination may be more or less effective for protein identification than a different combination.

Here, to discuss these relationships, we will designate specific analytical strategies by the annotation ''MS platform'' (where ''MS/MS'' means any tandem MS method, unless otherwise named) – ''spectra-DB sequence correlation method'' – ''sequence DB.''

To enhance the ability to identify proteins, a number of combinations must be developed, compared, and employed simultaneously in proteome analysis. Here, we attempt to systematically address this complexity and point out potential new strategies for protein identification.

0165-9936/03/$ - see front matter © 2003 Published by Elsevier Science B.V. doi:10.1016/S0165-9936(03)00507-7

2. MS platforms

To characterize the interactions between MS platforms (MSPs), spectra-DB sequence correlation methods, and sequence DBs, the qualities of mass spectra produced by different MSPs will first be briefly considered.

In proteome analysis by MS, proteins are extracted from cells or tissues, digested with proteases (typically trypsin), and peptide fragments are analyzed. Historically, the first MSP for protein identification included two-dimensional (2-D) polyacrylamide gel electrophoresis and matrix-assisted laser desorption/ionization (MALDI) time-of-flight (TOF) MS for peptide mass fingerprinting (PMF) (reviewed in [5]).

MALDI-TOF is most commonly used in proteomics in ion-reflection mode because of its low-femtomole (even attomole) sensitivity and high resolution ($>10{,}000$ FWHM). Peptide-mass fingerprints are routinely acquired with better than 50 ppm mass accuracy with external calibration, and recently reported automated re-calibration methods lower the error of mass measurement below 10 ppm [6,7]. MALDI spectra can be acquired very rapidly, and the entire routine, starting from digestion of proteins, preparation of the MALDI probes and acquiring the spectra, is now automated and optimized for a very high throughput [8,9].

The capabilities of MALDI in protein identification can be additionally strengthened by the acquisition of PSD spectra for a few selected peptide ions [10]. However, acquisition of PSD spectra is rather slow and it is much less sensitive compared to peptide fingerprinting, fragmentation is poorly controlled, and the spectra suffer from low resolution and low mass accuracy, so the technology has not had a significant impact in proteomics. However the recently introduced LIFT method [11] speeds up acquisition of MALDI-PSD spectra considerably.

MALDI-tandem MS (MS/MS) capabilities are mainly explored through the development of instruments with combined mass analyzers, such as MALDI-Q(q)TOF mass spectrometers [12,13] and MALDI-TOF/TOF [14]. Both instruments can acquire peptide fingerprints of high mass accuracy, and enable the control of collision energy in MS/MS mode. However, because of the orthogonal configuration of the ion path, MALDI-Q(q)TOF machines can acquire MS/MS spectra at relatively low collision energy, with the same resolution ($>12{,}000$) and mass accuracy (<20 ppm) as in the MS mode [15].

Regardless of the mass analyzers employed, MALDI sources predominantly ionize tryptic peptides as singly charged ions. To fragment singly charged ions, higher collision energy is required and therefore cleavage of amide bonds in the peptide backbone occurs less consistently. Usually, MALDI-MS/MS spectra do not contain continuous ion series that might facilitate the confident determination of long peptide sequences. A number of chemical derivatization methods, localizing the charged groups at the N- or C-terminus of the peptide molecule, have been developed to improve peptide fragmentation patterns [16,17].

MALDI sources were also coupled with an ion trap, allowing the acquisition of MS/MS spectra very rapidly, albeit the mass accuracy of the ion trap is much lower than that of TOF analyzers [18]. However, a large number of acquired MS/MS spectra increases the specificity of DB searching, and compensates for the lack of specificity of DB searching with peptide mass fingerprints, which is heavily dependent on mass accuracy.

ESI methods form another major cluster of MS platforms. In ESI-MS, tryptic peptides are typically ionized as doubly or triply charged ions. Multiply charged ions can be efficiently fragmented at lower collision energy, and their MS/MS spectra are usually dominated by intense y- and b-ions (see [19] for the nomenclature), which facilitates DB searching and also makes the spectra more amenable for *de novo* interpretation [20].

Peptides can be separated by liquid chromatography on-line with the mass spectrometer (LC-MS/MS); alternatively, unseparated peptide mixtures may be directly analyzed by NanoESI [21]. The absence of separation in NanoESI is compensated by much longer spraying time per spectrum, which is made possible by low flow injection rates (typically, 20–30 nL/ min), and many precursor ions can be fragmented successively [22].

However, NanoESI-MS/MS experiments are limited by the ability of the operator to recognize low-abundance precursors masked by chemical noise. The specificity of detection of precursor ions can be increased *via* precursor ion scanning for abundant ammonium ions of amino acid residues (typically, of Leu and Ile) [23]. Precursor-ion scanning is a routine operation mode of TQ mass spectrometers, and recently it has also been set up on Q(q)TOF machines [24,25].

Although fewer peptide precursor ions are typically fragmented in the course of NanoESI-MS/MS analysis, the quality and signal-to-noise ratio in their fragment spectra are routinely better than by LC-MS/MS, because the data accumulation time and collision energy can be precisely tuned by an operator during the acquisition process. However, NanoESI-MS/MS is relatively difficult to automate [26] and it has a limited ability to identify proteins in complex mixtures.

By pre-separating peptides in front of the on-line mass spectrometer, analytical methods gain higher dynamic range and the ability to identify proteins in very complex mixtures [27]. By applying this method, a substantial part of the proteomes of prokaryotic and low eukaryotic organisms can be characterized [27–29]. Further coupling of multidimensional LC-MS/MS analysis enables relative quantification that utilizes peptides enriched with stable isotopes as internal standards, and promises a global survey of quantitative

changes in the proteomes [29–31]. However, there is less control in the process of spectra acquisition, and information content of MS/MS spectra might be compromised. This might not be particularly important for protein identification by pattern searches (see later section) because, with high-resolution instruments, the ion statistics in the peak does not strongly affect mass accuracy, and full representation of fragments in the spectrum is not required [15]. However, poor ion statistics affects the accuracy of *de novo* sequencing, which benefits from recognizing complementary pairs of fragment ions and the full representation of low molecular weight peaks is often critical.

Because of differences in ionization mechanisms, MALDI and ESI produce different data sets when the same protein digest is analyzed [32,33]. Parallel analysis of digests by two methods increases the sequence coverage of peptide maps, but usually requires the employment of different instrumentation. Rapidly switchable combined MALDI/ESI sources [34,35] allow changing between ionization modes within minutes without venting the mass spectrometer, and might provide an effective alternative to expanding costly instrumentation.

Other MSPs, such as MALDI- and ESI-FT-ICR-MS [36,37], linear ion trap [38], and ion trap-TOF [39] mass spectrometers are employed in proteomics, but are currently less prevalent.

3. Sequence DBs

Protein sequence DBs are continually updated with submissions produced from the cloning of genes, from which amino acid sequences are generated by translation of nucleotide sequences in their correct reading frames (www.ncbi.nlm.nih.gov, www.expasy.org/sprot/, www.ebi.ac.uk/).

Whole genome shotgun sequencing also produces large sets of nucleotide sequences that are arranged in contiguous regions (i.e. whole chromosomes). As these genomic sequences are evaluated by gene prediction methods and open reading frames are designated, protein sequence is generated on the basis of nucleotide sequence and contributed to growing protein sequence DBs (see [40] for review).

Apart from the sequencing of individual genes or genomic DNA, mRNA is isolated to generate complementary DNA (cDNA) libraries. cDNAs are then partially sequenced to produce expressed sequence tag (EST) nucleotide sequence DBs [41].

In the hands of the mass spectrometrist, all three types of DB (protein, EST, and genomic) may be interrogated with mass spectra. Depending on the type of mass spectra (discussed above), specific types of DBs can be interrogated with more or less efficiency. We will discuss those relationships in the following section.

4. Spectra-sequence correlation methods and analytical strategies

Mass spectra are correlated with DB sequences primarily in three ways: the mass pattern; the amino-acid sequence; and, the sequence tag (Fig. 1). These three methods derive information of different qualities from peptide MS/MS, and they each suit the interpretation of spectra (more or less effectively) depending on the spectra signal-to-noise, mass accuracy and resolution. Furthermore, each of the different methods has distinct capabilities in identifying analyzed peptides the sequences if which share only partial identity with DB sequences.

Mass patterns (comprising lists of m/z values of detected peaks along with the corresponding peak intensities) are used in two types of MS analysis – PMF and MS/MS. In PMF, masses of intact peptides are determined and are used for DB searches. Historically, mass patterns derived from peptide mass fingerprints were first used to search protein sequence DBs (PMF-mass pattern-Protein DB) [42–45] (Fig. 2) (Table 1). Observed peptide masses are compared with peptide masses calculated from the *in silico* digestion of protein DB sequences with trypsin, and resulting matches are scored accordingly. In these softwares, mostly mass values have been used, but now there are attempts to incorporate peak intensities to improve the specificity of the identifications [46–48]. When sequences from analyzed peptides deviate from the identity of corresponding sequences in DB entries, either because of amino acid substitution or post-translational modifications, the probability of successful identification by this method diminishes [49], and MS/MS must be employed.

A second common analytical strategy correlates MS/MS spectra through mass patterns with protein DB sequences (MS/MS-mass pattern-Protein DB)[46,50–54]. In these cases, observed masses of peptide precursors and masses and intensities of their fragment ions are compared with theoretical peptide masses and fragments derived from sequence DBs with the application of particular protease specificity and peptide-fragmentation rules. The method of scoring the similarity between MS/MS spectrum and DB sequence employs certain peptide fragmentation models, which are instrument-dependent. To this end, DB searching programs usually allow the specification of instrument type.

MS/MS spectra with higher mass accuracy will be able to interrogate DBs more specifically, increasing the probability of identification with fewer peptides [51]. Furthermore, MS/MS spectra with high signal-to-noise ratio will give the best results, as true peptide fragment ions will not be obscured by background peaks.

Mass pattern methods are currently more diverse and have experienced a greater attention in the proteomics

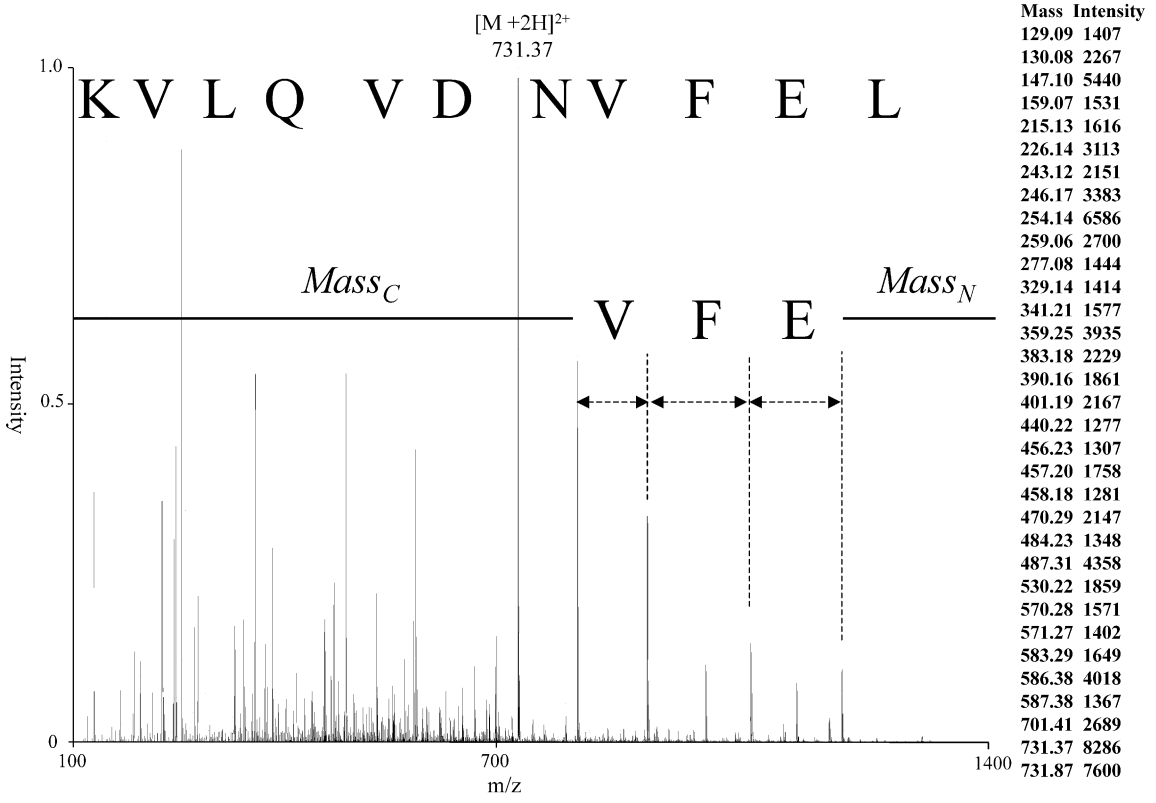

Figure 1. Representative information from the mass spectrum in proteomics. Mass spectra can be represented by primarily three types of information for their correlation with DB sequences. *Mass patterns* are composed of lists of *m/z* values of detected peaks and corresponding peak intensities (partial list shown above/right). *Amino-acid sequences* are derived from spectra considering precise mass differences in ion series, and annotated in the form of series of amino acid symbols (KVLQV. . .). *Sequence tags* consist of partial amino acid sequences combined with two mass values which lock the sequence within the length of a peptide, and a parent mass (the sequence tag for the above spectrum is (815.44)VFE(1190.62), peptide mass 1460.72).

community than other protein identification methods. The correlation of mass patterns with DB sequences also have some ''error-tolerant'' capabilities that withstand amino acid substitutions between the peptides observed and sequences present in a DB [51,55]. In addition to protein DB interrogation, spectra can be correlated through mass patterns with EST DBs (MS/MS-mass pattern-EST DB) [46,50,51,56] and more recently with genomic DBs (MS/MS-mass pattern-Genomic DB) [56,57] (Fig. 2).

Apart from using complete mass patterns for protein identification, the recently developed ''peptide end sequencing'' makes use of N- and C-terminal peptide-fragment ions detected in the low *m/z* region, along with a parent mass, for the identification of low-abundance proteins [58,59]. However, if multiple non-isobaric amino-acid substitutions occur within individual peptides or if unknown multiple post-translational modifications exist, then the probability decreases that the protein will be identified by these methods. In these cases, a different method is employed for interpreting spectra.

Amino acid sequences (deduced from MS/MS spectra) enable spectra interpretation for the identification of proteins that are homologous to DB sequences despite having no peptides of identical precursor mass with those theoretically predicted from DB entries (as required by mass pattern searches). In this analytical strategy, amino-acid sequences produced from MS/MS spectra can be correlated with protein-DB sequences (MS/MS-amino acids-Protein DB) [60–62]. Amino-acid sequences can be produced *de novo* from MS/MS spectra of peptides primarily by two methods: chemical modification of peptide N- or C-termini; or, direct computer-assisted interpretation of spectra by sequence-prediction algorithms (reviewed in [4]).

Chemical modification methods demand relatively large sample quantities, are manually laborious, and ultimately obscure spectra for parallel mass pattern interpretation. However, sequence prediction algorithms rapidly generate putative amino acid sequences, although often multiple degenerate sequences are predicted with similar statistical confidence. To utilize this information, numerous peptide sequences from multiple fragmented peptides must be compiled in a query for a sequence-similarity DB search. This has been accomplished in dedicated softwares based on FASTA [63] and BLAST [64] homology searching algorithms

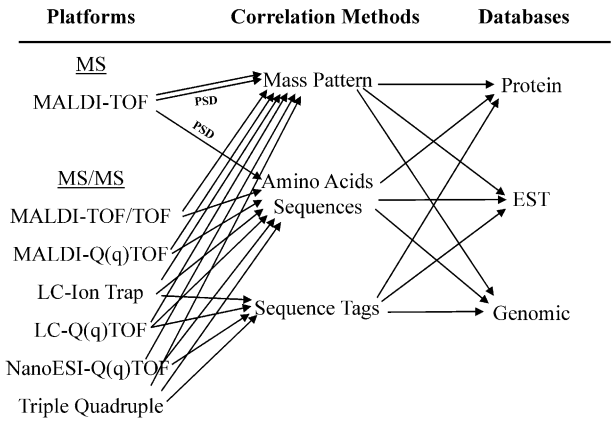

Figure 2. Strategy network. Analytical strategies are composed of three components: MS *platforms*, spectra-sequence *correlation methods*, and sequence *DBs*, which can be read left to right, in the scheme above. Arrows show the interaction of mass spectra with representation methods (mass pattern, amino acid sequences, sequence tags) (see Fig. 1) and nucleotide or amino acid DBs (protein, EST, genomic). Arrows represent robust and less established interactions equally, despite practical limitations to some.

Table 1. Analytical strategies

Analytical strategy	Ref.
MS	
PMF-*mass pattern*-Protein DB	[42]
PMF-*mass pattern*-EST DB	[46]
PSD-*mass pattern*-Protein DB	[10]
PSD-*amino acids*-Protein DB	[81]
MS/MS	
MS/MS-*mass pattern*-Protein DB	[50]
MS/MS-*mass pattern*-EST DB	[50]
MS/MS-*mass pattern*-Genomic DB	[57]
MALDI-Q(q)TOF-*amino acids*-Protein DB	[61]
ESI-Q(q)TOF-*amino acids*-Protein DB	[65]
ESI-TQ-*amino acids*-Protein DB	[65]
LC-Q(q)TOF-*amino acids*-Protein DB	[66]
ESI-Ion Trap-*amino acids*-Protein DB	[68]
MS/MS-*amino acids*-EST DB	*
MS/MS-*amino acids*-Genomic DB	*
MS/MS-*sequence tag*-Protein DB	[69]
LC-Ion Trap-*sequence tag*-Protein DB	[82]
MS/MS-*sequence tag*-EST DB	[71]
MS/MS-*sequence tag*-Genomic DB	[72]

Representative analytical strategies are listed above. Analytical strategies are represented by the annotation *"MS platform"* (where "MS/MS" means any tandem MS method, unless otherwise named) – *"spectra-sequence correlation method"*–*"sequence database."* Researchers who contributed to the development of these strategies are cited in the column on the right.
*These strategies are currently established in-house.

(reviewed in [4]). MS driven BLAST (MS BLAST) is one example of this type of strategy [61].

DB searching methods using amino acid sequences are also flexible for the development of alternate analytical strategies (Fig. 2). In one proteome analysis, MS BLAST was employed to correlate MALDI-Q(q)TOF spectra with protein-DB sequences (MALDI-Q(q)TOF-amino acids-Protein DB) [61]. However, this is just one strategy among other possible paths considering available MSPs, DBs, and developments in high-throughput spectra processing and MS BLAST DB searching capabilities. NanoESI-Q(q)TOF-amino acids-Protein DB and NanoESI-TQ-amino acids-Protein DB [65], as well as LC-Q(q)TOF-amino acids-Protein DB [66], are other options for sequence-similarity identification depending on available instrumentation (Table 1).

Furthermore, peptide sequence can be generated *de novo* from LC-ion trap mass spectra and used for protein-DB interrogation (LC-Ion Trap-amino acids-Protein DB) [67,68]. In addition, MS/MS-amino acids-EST DB and MS/MS-amino acids-Genomic DB are other possible DB searching strategies that could also be employed for protein identification with amino-acid sequences (currently both approaches are operating in-house) (Table 1). All of these strategies enable new capabilities for protein identification; however, their efficiency is limited when sequence prediction fails.

Sequence tags enable spectra interpretation for the identification of proteins from low-intensity spectra or where background chemical noise interferes with full-length amino-acid sequence determination. Sequence tags comprise a few (2–4) determined amino acids, along with two mass values, which lock the sequence stretch within the length of the peptide [69] (Fig. 1), effectively combining mass data with sequence data. This requires that only one section of a complete mass spectrum be interpreted correctly by manual inspection. Sequence tags can be employed to identify proteins by correlating MS/MS spectra (of preferably high mass accuracy [51]) with protein (MS/MS-sequence tags-Protein DB) [70], EST (MS/MS-sequence tags-EST DB) [71], or genomic DB sequences (MS/MS-sequence tags-Genomic DB) [72] (Fig. 2). However, it is rather difficult to assemble sequence tags from MALDI-TOF/TOF and MALDI-Q(q)TOF spectra [73], mostly because prominent y- or b-fragment ion series often cannot be determined unambiguously [13].

Error-tolerant sequence tags enable the identification of proteins that are homologous to DB entries [74]. The recently developed MultiTag software correlates multiple (often partial) sequence tags with individual DB entries, whereas previous identification techniques using sequence tags rely upon the correlation of individual spectra to DB entries alone. The MultiTag approach is similar to approaches with mass patterns [46,50,51] or amino acid sequences [60–62] that rely upon information from multiple spectra to increase the confidence of protein identifications. In the future, a

MultiTag approach could greatly facilitate EST and genomic DB searches with sequence tags.

The correlation of mass patterns, amino acid sequences, and sequence tags with DB sequences all have their own unique evaluation schemes to discriminate correct from false positive protein identifications. Mass pattern identification methods primarily score the quality of fit of a tandem mass spectrum to a predicted model spectrum, while taking into consideration other DB search parameters (i.e. SEQUEST [50], Mascot [46], Protein Prospector [51], Scope [52], Sonar MS/MS [53], ProbID [54]).

Protein identifications made using amino acid sequences are evaluated by the significance of the alignment of a sequence query to a DB sequence and the probability of that alignment occurring in a DB of a specific size (i.e. CIDentify [60], MS BLAST [61], and FASTS [62]).

In the evaluation of multiple sequence tags (MultiTag), the probability that such a set of sequence tags align at random to a DB entry (within a DB of a particular size and at a particular mass accuracy) provides a measure of confidence of the identification.

The existence of multiple scoring schemes for the determination of the significance of spectra-sequence alignments raises the question of whether proteins identified by one method are considered positive identifications by another similar method (for instance, two mass pattern matching methods, or two amino acid sequence matching methods), which could ultimately compromise the certainty of proteome characterization. One solution to different scoring systems can be provided with the implementation of algorithms that calculate probabilities that the spectra acquired derive from specific known sequences, rather than that their similarity occurs at random [75]. Another possibility is to develop empirical statistical models based on search results to assess the validity of protein identifications by MS and DB searches [76]. This approach suggests that in the future statistical data interpretation methods can be applied to search results acquired by different MSPs and individual softwares [76].

5. Bridging the gap: a network of strategies

The character of various types of MS/MS spectra directly determines the operation of spectra-DB sequence correlation methods (mass lists, amino acid sequences, and sequence tags) and different methods have different abilities to interrogate DBs (protein, EST, and genomic). Different MSPs, spectra-DB sequence correlation methods, and DBs can be combined to yield varying degrees of effectiveness for protein identification, and new analytical strategies are rapidly creating novel ties between different types of spectra through the three discussed interpretation mechanisms (Fig. 2). Each particular proteome analysis will require a different strategy or strategies to succeed in identifying proteins at a high throughput, depending on the MSPs at hand, the abundance of the proteins, and the availability of species-specific DB sequences, as well as the DB type or types employed. This applies both to the proteomics of organisms with sequence resources as well as those species with unsequenced genomes.

Simultaneously applying multiple analytical strategies has enabled the most effective approaches in the analysis of complex protein mixtures by increasing sensitivity, throughput, and reliability of the characterization of proteomes. In 1996, multiple strategies began to be employed simultaneously with the application of PMF-mass pattern-Protein DB and MS/MS-mass pattern-Protein DB together in one study (Table 2) [77].

Today, multiple strategies are employed on a regular basis for high-throughput proteomics. In the recent characterization of the human nucleolus, five strategies (PMF-mass pattern-Protein DB, MS/MS-mass pattern-Protein DB, NanoESI-Q(q)TOF-sequence tags-Protein DB, NanoESI-Q(q)TOF-sequence tags-EST DB, and NanoESI-Q(q)TOF-sequence tags-Genomic DB) were all employed simultaneously [70]. In addition, in this study, multiple protein (nrdb and IPI) and genomic (phases 0–3 of the uncompleted human genome

Table 2. Application of analytical strategies in parallel

Year	Proteomics	Ref.
1993	**Bacteria Proteomics** PMF-*mass pattern*-Protein DB	[42]
1996	**Yeast Proteomics** PMF-*mass pattern*-Protein DB MS/MS-*mass pattern*-Protein DB	[77]
2001	**Maize Proteomics** PMF-*mass pattern*-Protein DB PMF-*mass pattern*-EST DB	[78]
2002	**Pea Symbiosome Proteomics** NanoESI-Ion Trap-*mass pattern*-Protein DB NanoESI-Ion Trap-*mass pattern*-EST DB	[79]
2002	**Human Nucleolus Proteomics** PMF-*mass pattern*-Protein DB MS/MS-*mass pattern*-Protein DB MS/MS-*sequence tag*-Protein DB MS/MS-*sequence tag*-EST DB MS/MS-*sequence tag*-Genomic DB	[70]
2002	**African Clawed Frog Proteomics** PMF-*mass pattern*-Protein DB MS/MS-*mass pattern*-Protein DB MS/MS-*amino acids*-Protein DB MS/MS-*sequence tags*-Protein DB MS/MS-*sequence tag*-EST DB	[80]

Representative proteomic studies are shown above. The name of the organism is in bold. The analytical strategies that were employed in those particular proteome studies are listed below the name. Researchers who conducted these studies are cited in the column on the right and the year of the study in the left hand column. The legend of Table 1 describes strategy annotation.

sequence) DBs were interrogated, adding additional reference sequences to facilitate protein identification.

The effective characterization of proteomes from organisms with unsequenced genomes relies upon the use of multiple analytical strategies as well. The proteome of maize leaves was analyzed using two strategies, PMF-mass pattern-Protein DB and PMF-mass pattern-EST DB [78]. However, using only these two strategies enabled the identification of 216 spots out of 300 analyzed from 2-D gels. The authors recognized the importance of MS/MS methods (perhaps combined with homology-based DB searching) for further studies to be more comprehensive.

Similarly, in the proteomics of the pea symbiosome, NanoESI-Ion Trap-mass pattern-Protein DB and NanoESI-Ion Trap-mass pattern-EST DB methods were applied, but failed to identify almost half the proteins, with 46 identifications out of 89 spots analyzed from 2-D gels, despite the application of MS/MS methods [79].

In another example, five strategies (PMF-mass pattern-Protein DB, NanoESI-Q(q)TOF-mass pattern-Protein DB, NanoESI-Q(q)TOF-amino acids-Protein DB, NanoESI-Q(q)TOF-sequence tags-Protein DB, and NanoESI-Q(q)TOF-sequence tags-EST DB) were applied simultaneously to characterize the African Clawed frog *Xenopus laevis* microtubule-associated proteome, successfully identifying 62 proteins from 55 protein bands from one-dimensional gels [80].

Considering the recent bioinformatics developments presented here, mass spectrometrists can begin systematic application and comparison of the effectiveness of these analytical strategies (Fig. 2). In the future, we can expect new MSPs, new spectra-sequence correlation methods, as well as perhaps new types of DBs to contribute to the proliferation of protein identification strategies. We can also begin to predict future strategies that provide new potential for the MS community, by developing the approaches discussed above. Whereas developments in protein identification have been presented here, the principles underlying these strategies may be employed in the future to nucleic acid characterization and DB searches (perhaps single nucleotide polymorphisms, SNPs).

Sensitive and confident protein identification by MS is a never-ending problem. All of the world's species will not have their genomes sequenced, sequence DBs will always be at various stages of development, and the biological sciences will find new specimens to be analyzed at the level of the proteome. In order to thoroughly and sensitively characterize these proteomes, the application of multiple analytical strategies provides a successful approach. In the future, we can expect that the network between various types of mass spectra and different types of DB sequences will become more and more integrated with the development of novel analytical strategies.

Acknowledgements

We were able to cite only a selection of representative references, and we sincerely apologize to researchers for the work that we were not able to cite and discuss because of space considerations.

References

[1] J. Rappsilber, M. Mann, Trends Biochem. Sci. 27 (2002) 74.
[2] R. Aebersold, D.R. Goodlett, Chem. Rev. 101 (2001) 269.
[3] M. Mann, R.C. Hendrickson, A. Pandey, Ann. Rev. Biochem. 70 (2001) 437.
[4] A.J. Liska, A. Shevchenko, Proteomics 3 (2003) 19.
[5] D.J.C. Pappin, Methods Mol. Biol. 211 (2002) 211.
[6] J. Gobom, M. Mueller, V. Egelhofer, D. Theiss, H. Lehrach, E. Nordhoff, Anal. Chem. 74 (2002) 3915.
[7] M. Bantscheff, B. Duempelfeld, B. Kuster, Rapid Commun. Mass Spectrom. 16 (2002) 1892.
[8] M. Traini, A.A. Gooley, K. Ou, M.R. Wilkins, L. Tonella, J.C. Sanchez, D.F. Hochstrasser, K.L. Williams, Electrophoresis 19 (1998) 1941.
[9] V. Egelhofer, J. Gobom, H. Seitz, P. Giavalisco, H. Lehrach, E. Nordhoff, Anal. Chem. 74 (2002) 1760.
[10] B. Spengler, J. Mass Spectrom. 32 (1997) 1019.
[11] V. Schnaible, S. Wefing, A. Resemann, D. Suckau, A. Bucker, S. Wolf-Kummeth, D. Hoffmann, Anal. Chem. 74 (2002) 4980.
[12] A.N. Krutchinsky, A.V. Loboda, V.L. Spicer, R. Dworschak, W. Ens, K.G. Standing, Rapid Commun. Mass Spectrom. 12 (1998) 508.
[13] A.V. Loboda, A.N. Krutchinsky, M. Bromirski, W. Ens, K.G. Standing, Rapid Commun. Mass Spectrom. 14 (2000) 1047.
[14] K.F. Medzihradszky, J.M. Campbell, M.A. Baldwin, A.M. Falick, P. Juhasz, M.L. Vestal, A.L. Burlingame, Anal. Chem. 72 (2000) 552.
[15] A. Shevchenko, A. Loboda, W. Ens, K.G. Standing, Anal. Chem. 72 (2000) 2132.
[16] T. Keough, R.S. Youngquist, M.P. Lacey, Proc. Natl. Acad. Sci. USA 96 (1999) 7131.
[17] M. Bartet-Jones, W.A. Jeffery, H.F. Hansen, D.J.C. Pappin, Rapid Commun. Mass Spectrom. 8 (1994) 737.
[18] A.N. Krutchinsky, M. Kalkum, B.T. Chait, Anal. Chem. 73 (2001) 5066.
[19] K. Biemann, Biomed. Environ. Mass Spectrom. 16 (1988) 99.
[20] A. Shevchenko, I. Chernushevich, M. Wilm, M. Mann, Mol. Biotechnol. 20 (2002) 107.
[21] M. Wilm, M. Mann, Anal. Chem. 66 (1996) 1.
[22] M. Wilm, A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer, T. Fotsis, M. Mann, Nature 379 (1996) 466.
[23] M. Wilm, G. Neubauer, M. Mann, Anal. Chem. 68 (1996) 527.
[24] H. Steen, B. Kuster, M. Fernandez, A. Pandey, M. Mann, Anal. Chem. 73 (2001) 1440.
[25] K. Ekroos, I.V. Chernushevich, K. Simons, A. Shevchenko, Anal. Chem. 74 (2002) 941.
[26] S. Geromanos, J. Philip, G. Freckleton, P. Tempst, Rapid Commun. Mass Spectrom. 12 (1998) 551.
[27] M.P. Washburn, D. Wolters, J.R. YatesIII, Nat. Biotechnol. 19 (2001) 242.
[28] F. Xiang, G.A. Anderson, T.D. Veenstra, M.S. Lipton, R.D. Smith, Anal. Chem. 72 (2000) 2475.
[29] T.P. Conrads, K. Alving, T.D. Veenstra, M.E. Belov,

G.A. Anderson, D.J. Anderson, M.S. Lipton, L. Pasa-Tolic, H.R. Udseth, W.B. Chrisler, et al., Anal. Chem. 73 (2001) 2132.
[30] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, R. Aebersold, Nat. Biotechnol. 17 (1999) 994.
[31] S.E. Ong, B. Blagoev, I. Kratchmarova, D.B. Kristensen, H. Steen, A. Pandey, M. Mann, Mol. Cell. Proteomics 1 (2002) 376.
[32] K.F. Medzihradszky, H. Leffler, M.A. Baldwin, A.L. Burlingame, J. Am. Soc. Mass Spectrom. 12 (2001) 215.
[33] A. Shevchenko, A. Loboda, W. Ens, B. Schraven, K.G. Standing, A. Shevchenko, Electrophoresis 22 (2001) 1194.
[34] A.N. Krutchinsky, W. Zhang, B.T. Chait, J. Am. Soc. Mass Spectrom. 11 (2000) 493.
[35] G. Baykut, J. Fuchser, M. Witt, G. Weiss, C. Gosteli, Rapid Commun. Mass Spectrom. 16 (2002) 1631.
[36] T.W. Chan, L. Duan, T.P. Sze, Anal. Chem. 74 (2002) 5282.
[37] Y. Ge, B.G. Lawhorn, M. El Naggar, E. Strauss, J.H. Park, T.P. Begley, F.W. McLafferty, J. Am. Chem. Soc. 124 (2002) 672.
[38] J.C. Schwartz, M.W. Senko, J.E. Syka, J. Am. Soc. Mass Spectrom. 13 (2002) 659.
[39] B.A. Collings, J.M. Campbell, D. Mao, D.J. Douglas, Rapid Commun. Mass Spectrom. 15 (2001) 1777.
[40] C. Mathe, M.F. Sagot, T. Schiex, P. Rouze, Nucleic Acids Res. 30 (2002) 4103.
[41] M.D. Adams, J.M. Kelley, J.D. Gocayne, M. Dubnick, M.H. Polymeropoulos, H. Xiao, C.R. Merril, A. Wu, B. Olde, R.F. Moreno, et al., Science 252 (1991) 1651.
[42] W.J. Henzel, T.M. Billeci, J.T. Stults, S.C. Wong, C. Grimley, C. Watanabe, Proc. Natl. Acad. Sci. USA 90 (1993) 5011.
[43] M. Mann, P. Hojrup, P. Roepstorff, Biol. Mass Spectrom. 22 (1993) 338.
[44] J.R. YatesIII, S. Speicher, P.R. Griffin, T. Hunkapiller, Anal. Biochem. 214 (1993) 397.
[45] P. James, M. Quadroni, E. Carafoli, G. Gonnet, Biochem. Biophys. Res. Commun. 195 (1993) 58.
[46] D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell, Electrophoresis 20 (1999) 3551.
[47] S. Gay, P.A. Binz, D.F. Hochstrasser, R.D. Appel, Proteomics 2 (2002) 1374.
[48] K.C. Parker, J. Am. Soc. Mass Spectrom. 13 (2002) 22.
[49] M.R. Wilkins, K.L. Williams, J. Theor. Biol. 186 (1997) 7.
[50] J. Eng, A. McCormack, J. Yates, J. Am. Soc. Mass Spectrom. 5 (1994) 976.
[51] K.R. Clauser, P. Baker, A.L. Burlingame, Anal. Chem. 71 (1999) 2871.
[52] V. Bafna, N. Edwards, Bioinformatics 17 (2001) S13.
[53] H.I. Field, D. Fenyo, R.C. Beavis, Proteomics 2 (2002) 36.
[54] N. Zhang, R. Aebersold, B. Schwikowski, Proteomics 2 (2002) 1406.
[55] D.M. Creasy, J.S. Cottrell, Proteomics 2 (2002) 1426.
[56] J.S. Choudhary, W.P. Blackstock, D.M. Creasy, J.S. Cottrell, Trends Biotechnol. 19 (2001) S17.
[57] J.S. Choudhary, W.P. Blackstock, D.M. Creasy, J.S. Cottrell, Proteomics 1 (2001) 651.
[58] M.L. Nielsen, K.L. Bennet, B. Larsen, M. Moniatte, M. Mann, J. Proteome Res. 1 (2002) 63.
[59] A. Schlosser, W.D. Lehmann, Proteomics 2 (2002) 524.
[60] J.A. Taylor, R.S. Johnson, Anal. Chem. 73 (2001) 2594.
[61] A. Shevchenko, S. Sunyaev, A. Loboda, P. Bork, W. Ens, K.G. Standing, Anal. Chem. 73 (2001) 1917.
[62] A.J. Mackey, T.A.J. Haystead, W.R. Pearson, Mol. Cell. Proteomics 1 (2002) 139.
[63] W.R. Pearson, Methods Mol. Biol. 132 (2000) 185.
[64] S.F. Altschul, T.L. Madden, A.A. Schaffer, J. Zhang, Z. Zhang, W. Miller, D.J. Lipman, Nucleic Acids Res. 25 (1997) 3389.
[65] A. Shevchenko, S. Sunyaev, A. Liska, P. Bork, A. Shevchenko, Methods Mol. Biol. 211 (2002) 221.
[66] S. Nimkar, J.A. Loo, Application of a new algorithm for automated database searching of MS sequence data to identify proteins, in: Proc. 50th Am. Soc. Mass Spectrom. Allied Top., abstract 334, 2002.
[67] J. Qin, C.J. Herring, X. Zhang, Rapid Commun. Mass Spectrom. 12 (1998) 209.
[68] A. Marina, M.A. Garcia, J.P. Albar, J. Yague, J.A. Lopez de Castro, J. Vazquez, J. Mass Spectrom. 34 (1999) 17.
[69] M. Mann, M. Wilm, Anal. Chem. 66 (1994) 4390.
[70] J.S. Andersen, C.E. Lyon, A.H. Fox, A.K. Leung, Y.W. Lam, H. Steen, M. Mann, A.I. Lamond, Curr. Biol. 12 (2002) 1.
[71] M. Mann, Trends Biochem. Sci. 21 (1996) 494.
[72] B. Kuster, P. Mortensen, J.S. Andersen, M. Mann, Proteomics 1 (2001) 641.
[73] A. Wattenberg, A.J. Organ, K. Schneider, R. Tyldesley, R. Bordoli, R.H. Bateman, J. Am. Soc. Mass Spectrom. 13 (2002) 772.
[74] S. Sunyaev, A.J. Liska, A. Golod, A. Shevchenko, A. Shevchenko, Anal. Chem. 75 (2003) 1307.
[75] M.J. MacCoss, C.C. Wu, J.R. YatesIII, Anal. Chem. 74 (2002) 5593.
[76] A. Keller, A.I. Nesvizhskii, E. Kolker, R. Aebersold, Anal. Chem. 74 (2002) 5383.
[77] A. Shevchenko, O.N. Jensen, A.V. Podtelejnikov, F. Sagliocco, M. Wilm, O. Vorm, P. Mortensen, H. Boucherie, M. Mann, Proc. Natl. Acad. Sci. U.S.A. 93 (1996) 14440.
[78] L. Porubleva, K. Van der Velden, S. Kothari, D.J. Oliver, P.R. Chitnis, Electrophoresis 22 (2001) 1724.
[79] G. Saalbach, P. Erik, S. Wienkoop, Proteomics 2 (2002) 325.
[80] A.J. Liska, A. Popov, S. Sunyaev, A. Shevchenko, B. Habermann, P. Bork, E. Karenti, A. Shevchenko, *submitted*.
[81] B. Spengler, F. Luetzenkirchen, S. Metzger, P. Chaurand, R. Kaufmann, W. Jeffery, M. Bartlet-Jones, D. Pappin, Int. J. Mass Spectrom. 169 (1997) 127.
[82] P. Huang, D.B. Wall, S. Parus, D.M. Lubman, J. Am. Soc. Mass Spectrom. 11 (2000) 127.