Automated protein identification by tandem mass spectrometry: issues and strategies. Mass Spectrom Rev

ARTICLE *in* MASS SPECTROMETRY REVIEWS · MARCH 2006

Impact Factor: 7.71 · DOI: 10.1002/mas.20068 · Source: PubMed

3 AUTHORS:

Patricia Hernandez
Austin Community College
32 PUBLICATIONS **435** CITATIONS
SEE PROFILE

Markus Müller
Swiss Institute of Bioinformatics
56 PUBLICATIONS **2,832** CITATIONS
SEE PROFILE

Ron D Appel
Swiss Institute of Bioinformatics
201 PUBLICATIONS **11,201** CITATIONS
SEE PROFILE

AUTOMATED PROTEIN IDENTIFICATION BY TANDEM MASS SPECTROMETRY: ISSUES AND STRATEGIES

Patricia Hernandez,[1]* Markus Müller,[3] and Ron D. Appel[1,2]
[1]*Swiss Institute of Bioinformatics, Geneva, Switzerland*
[2]*University of Geneva and Geneva University Hospital, Geneva, Switzerland*
[3]*Institute of Molecular Systems Biology, Swiss Federal Institute of Technology, Zürich, Switzerland*

Received 30 November 2004; received (revised) 29 March 2005; accepted 6 April 2005

Published online 11 November 2005 in Wiley InterScience (www.interscience.wiley.com) DOI 10.1002/mas.20068

Protein identification by tandem mass spectrometry (MS/MS) is key to most proteomics projects and has been widely explored in bioinformatics research. Obtaining good and trustful identification results has important implications for biological and clinical work. Although well matured, automated software identification of proteins from MS/MS data still faces a number of obstacles due to the complexity of the proteome or procedural issues of mass spectrometry data acquisition. Expected or unexpected modifications of the peptide sequences, polymorphisms, errors in databases, missed or non-specific cleavages, unusual fragmentation patterns, and single MS/MS spectra of multiple peptides of the same m/z are so many pitfalls for identification algorithms. A lot of research work has been carried out in recent years that yielded new strategies to handle a number of these issues. Multiple MS/MS identification algorithms are now available or have been theoretically described. The difficulty resides in choosing the most adapted method for each type of spectra being identified. This review presents an overview of the state-of-the-art bioinformatics approaches to the identification of proteins by MS/MS to help the reader doing the spadework of finding the right tools among the many possibilities offered. © 2005 Wiley Periodicals, Inc., Mass Spec Rev 25:235–254, 2006

Keywords: *tandem mass spectrometry; protein identification; strategies*

I. INTRODUCTION

Proteomic studies involve analytical techniques, data analyses, and the use of databases for the systematic study of the protein content of a given cell, tissue, or organism, at a given time and under specific conditions. The term proteomics was initially devoted to the identification of proteins displayed on two-dimensional gel electrophoresis, but is now associated with a great variety of analysis techniques covering three main axes: protein identification, characterization, and quantification. The scale of a proteomic experiment varies according to its aim, such as determining the protein content of a whole organism, analyzing proteins in a tissue or targeting proteins in cell subcompartments. Two particularly important issues in clinical proteomics are the study of disease-related molecular changes to discover diagnostic markers, and the discovery of new therapeutic targets. This review focuses on peptide identification from MS/MS data. Under constant development and research for 20 years, MS/MS identification is nowadays well established as a method for protein identification. The review starts with a few historical considerations, then introduces mass spectrometry and MS/MS data, before discussing the various strategies applied to automated MS/MS identification. In many cases, the review shall highlight the numerous pitfalls faced by software due to the complexity of the proteome, or of the MS/MS data and shall describe how those pitfalls are completely or partially accounted for.

**Correspondence to:* Patricia Hernandez, Swiss Institute of Bioinformatics, CMU, rue Michel-Servet 1, 1211 Geneva 4, Switzerland.
E-mail: patricia.hernandez@isb-sib.ch

II. HISTORICAL OVERVIEW

The first uses of tandem mass spectrometry (MS/MS) for protein sequence analyses date back to the 1980s. At that time, the method of choice for amino acid sequencing was Edman degradation, which had been introduced 30 years earlier (Edman, 1950). The process consisted in the repeated one-by-one removal of amino acids from the *N*-terminus of proteins or peptides (each cycle took roughly one hour) and determining which amino acid was cleaved. The procedure became the method of choice for protein identification after 1967, through the development of the automated sequencer (Edman & Begg, 1967). Though extremely successful, limitations such as the speed (one sample per day) of the procedure, the need of highly purified samples and a free amino *N*-terminus, stimulated the development of new protein identification approaches. The requirement of highly purified samples was bypassed early in Shimonishi et al. (1980), who proposed to combine mass spectrometry and Edman degradation to sequence peptides in a mixture. The idea was to combine mass spectrometry and Edman degradation in the following way: the masses of the intact peptides from the mixture were measured with field desorption MS. Then the mixture underwent several cycles of Edman degradation. Two computer programs (PAAS (Matsuo, Matsuda, & Katakuse, 1981) and PROSEQ1 (Kitagishi, Hong, & Shimonishi, 1981)) were proposed to determine the amino acid sequences that were compatible with both the intact peptide masses and the released amino acid compositions. Sakurai et al. (1984) tackled the problem of blocked *N*-terminus and modified the PAAS program to determine probable amino acid sequences of a peptide by using only sequence ion peaks obtained by fast atom bombardment MS (Morris et al., 1981). *De novo* peptide sequencing from tandem MS was born. New

approaches followed during that decade, but their impact remained limited. During the 1990s, the landscape radically changed due to the conjunction of two factors: the rapid increase of the number of available protein sequences in public databases, which were intensively fed by high-throughput DNA sequencing projects, and the development of two soft ionization methods for mass spectrometry: electrospray ionization (ESI) (Fenn et al., 1989) and matrix-assisted laser desorption/ionization (MALDI) (Karas & Hillenkamp, 1988; Tanaka et al., 1988). These methods allowed the gentle ionization of large non-volatile biomolecules, and consequently extended detection limits and mass ranges in MS (Henzel, Watanabe, & Stults, 2003). A new concept, which had been sprouting for a couple of years, was simultaneously described by five groups (Henzel et al., 1993; James et al., 1993; Mann, Hojrup, & Roepstorff, 1993; Pappin, Hojrup, & Bleasby, 1993; Yates et al., 1993): the method, called "peptide mass fingerprinting" (PMF) (Pappin, Hojrup, & Bleasby, 1993) was aimed at identifying proteins by measuring their peptide mass composition by means of MS and then correlating the mass values with theoretical masses computed from protein sequences stored in databases. Numerous software based on PMF identification emerged over the last decade and have been reviewed in (Gras & Muller, 2001). Examples are MOWSE (Pappin, Hojrup, & Bleasby, 1993), Mascot (Perkins et al., 1999), MS-Fit (Clauser et al., 1995), PepIdent (Wilkins et al., 1999), and ProFound (Zhang & Chait, 2000). Recently, a new PMF software, Aldente (Tuloup et al., 2003), was made available by the Swiss Institute of Bioinformatics on the ExPASy website [Gasteiger et al., 2005].

One year after the first publications on PMF, Eng, McCormack, & Yates (1994) applied a similar technique to correlate MS/MS spectra with theoretical peptides, whereas Mann and Wilm (1994) proposed to infer partial sequences from the fragment mass differences observed in MS/MS spectra and to use this information for error-tolerant database searching. From this day on, a large number of algorithms for MS/MS identification have been succeeding one another. Each of them has its own particularities and participates in the refinement and maturation of the technique. The scoring schemes too have evolved and tend to include more and more contextual knowledge with the aim to increase the algorithms' performance. Thanks to these improvements, peptide and protein identification by tandem MS has now become a widespread identification method for proteomic studies.

III. TANDEM MASS SPECTROMETRY

A. Tandem Mass Spectrometry Identification Workflow

The typical workflow for protein identification based on tandem MS is illustrated in Figure 1. The first step is to reduce the complexity of a crude biological sample by applying one or

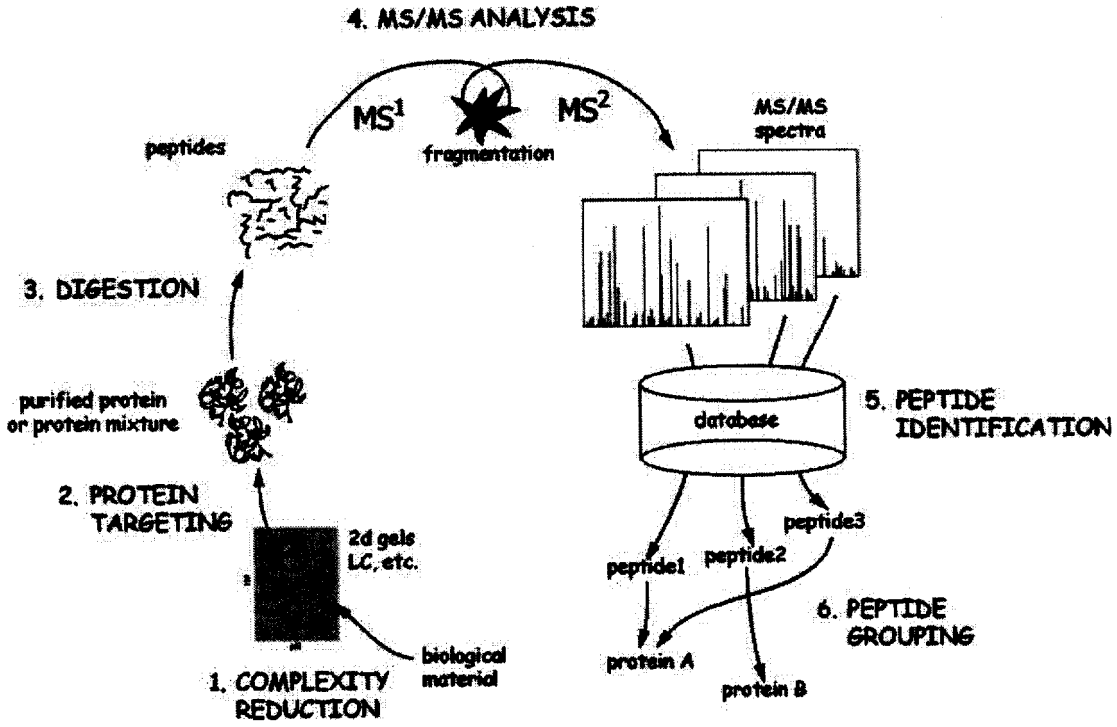


FIGURE 1. Typical workflow for protein identification based on tandem mass spectrometry. In (1) and (2), separation techniques are applied to reduce the protein sample complexity. Then the proteins are digested (3) and the resulting peptides undergo tandem mass spectrometric analysis (4) yielding a collection of MS/MS spectra. The identification procedure consists in correlating the MS/MS spectra with theoretical peptides (5), from which a list of identified proteins is deduced (6).

several protein separation techniques. This step aims at targeting proteins of interest, such as, for example membrane proteins, or proteins that reveal differential expression rates as a response to given stimuli. Depending on the separation techniques applied, the sample to be analyzed by MS/MS is mostly a complex protein mixture (in the simplest case, there is only one purified protein to analyze). The targeted proteins (or protein) are then cleaved into peptides using specific proteolytic enzymes. The most often used enzyme is trypsin, which cleaves peptides at the *C*-terminal side of arginine and lysine. Tandem MS represents a peptide as a readable entity (the MS/MS spectrum), which can then be interpreted and correlated with theoretical peptide sequences from protein or genomic databases. The last step is to combine the peptide identification results into a list of proteins that are most likely present in the sample.

B. Sample Preparation

In proteomic studies, projects, samples, and biology vary enormously, such as whole organisms (Kolker et al., 2003), tissues or physiological fluids (Rose et al., 2004), subcellular compartments (Rappsilber et al., 2002), or protein complexes (MacCoss et al., 2002). Various separation techniques can be used to reduce the protein sample complexity. They are all based on physical or chemical properties of the proteins or peptides. A widely used approach is two-dimensional gel electrophoresis (2-DE). Introduced in the seventies (Kenrick & Margolis, 1970), this technique essentially consists in separating proteins in a rectangular polyacrylamide gel using two orthogonal parameters, typically the isoelectric point and molecular weight of the proteins. The procedure results into a constellation of spots (up to a few thousands), each of them representing one or a few purified protein types. After extraction and proteolysis, proteins of interest can be identified by MS or MS/MS analyses. Unfortunately, 2-DE has some limitations (as have most other separation techniques) (Corthals et al., 2000; Rabilloud, 2002). Luckily, it is not necessary to reach too high a quality of protein purification for MS/MS analysis. An alternative approach, referred to as shotgun proteomics, applies MS/MS directly to large mixture of peptides in solution. In this approach, whose principle dates back to the 1980s (Hunt et al., 1986; Washburn, Wolters, & Yates, 2001), the source (ESI) of the mass spectrometer is coupled in-line with a chromatography system. The most widely used separation techniques include capillary electrophoresis (CE) and high-performance liquid chromatography (HPLC). The dynamic range of protein and polypeptide concentration, which is estimated to more than 10 orders of magnitude in fluids such as plasma, constitutes a challenging problem in proteomics (Anderson & Anderson, 2002). This problem has been targeted using multidimensional separation. The strategy consists in starting with a large amount of sample (e.g., several liters of human plasma) and in performing successive separation steps (e.g., depletion of abundant proteins followed by gel filtration followed by several chromatographies) (Rose et al., 2004).

C. Tandem Mass Spectrometry Analysis

A mass spectrometer is composed of three elements: an ionization source, a mass analyzer, and a detector. The source is the component where the analytes are introduced into the instrument and are ionized. The most widely used types of ionization for protein analysis are MALDI and ESI. With MALDI, the analytes are co-crystallized with a matrix after which pulses of UV laser light vaporize the matrix and liberate the peptides or proteins as gaseous ions. The MALDI generated ions are typically singly charged. In an ESI source, the sample stays in a liquid solution and is sprayed under high-voltage into the source region, where small droplets are formed that subsequently vaporize and give rise to ionized peptides or proteins. In this case, the generated ions are singly and/or multicharged. The mass analyzer separates the ions with respect to their mass-to-charge ratios (m/z), and these ions are then reported by the detector. MS/MS is achieved by performing two mass analyses, either "in space" or "in time." For "in space" configurations, such as triple quadrupole (TQ), quadrupole/time-of-flight (Q-TOF), or time-of-flight/time-of-flight (TOF–TOF), the primary and secondary analyses are performed sequentially as ions travel through the instrument. For "in time" configurations, such as quadrupole ion trap (Q-IT), they are performed consecutively within the same analyzer. In both cases, the first MS analysis step is used to measure ions according to their m/z. Subsequently, ions (precursors) are selected for further processing. The selection can either be manual or automated. In the latter case, the task is often dynamically controlled by the instrument's software (Aebersold & Goodlett, 2001). Redundancy can be moderated by the use of an exclusion list that contains the most recently fragmented m/z values. Then each selected precursor ion undergoes fragmentation. The resulting "product" ions are separated in the second mass analysis step with respect to their m/z values and are recorded as a tandem mass spectrum (MS/MS spectrum). Each instrument has its advantages and disadvantages, including speed, resolution, mass accuracy and price.

Collision induced dissociation (CID) is usually used to generate fragment peptides. If the collision energy is low (10–50 eV), the product ions are primarily formed by cleavage at the peptide bonds. According to Roepstorff's nomenclature (Roepstorff & Fohlman, 1984), ions are denoted as a, b, and c, when the charge is retained on the *N*-terminal side of the fragmented peptide, and x, y, and z when the charge is retained on the *C*-terminal side. As shown in Figure 2, ion types differ by the position of the fragmentation related to the peptide bond.

When using high-energy collision (around 1 kev), additional fragment ions may also be generated, including internal fragments formed by breakage of two peptide bonds, as well as side-chain specific ions (denoted d, v, and w) formed by the loss of all or parts of side chains (Johnson, Martin, & Biemann, 1988). There are a number of studies that have investigated the fragmentation pathway and that support a model called "mobile proton hypothesis" (Dongré et al., 1996; Wysocki et al., 2000). This model states that under low-energy collisional activation conditions, most fragmentation pathways are triggered by protonation of the amide nitrogen or carbonyl oxygen at the cleavage site (Jonsson, 2001). If the precursor ion carries more protons than its number of strongly basic residues, the supplementary protons migrate along the backbone, possibly initiating fragmentation at every amide site, thus giving rise to ionic fragments. When no mobile proton is available, the fragmentation process is affected, leading to spectra with "unusual fragmentation," often qualified

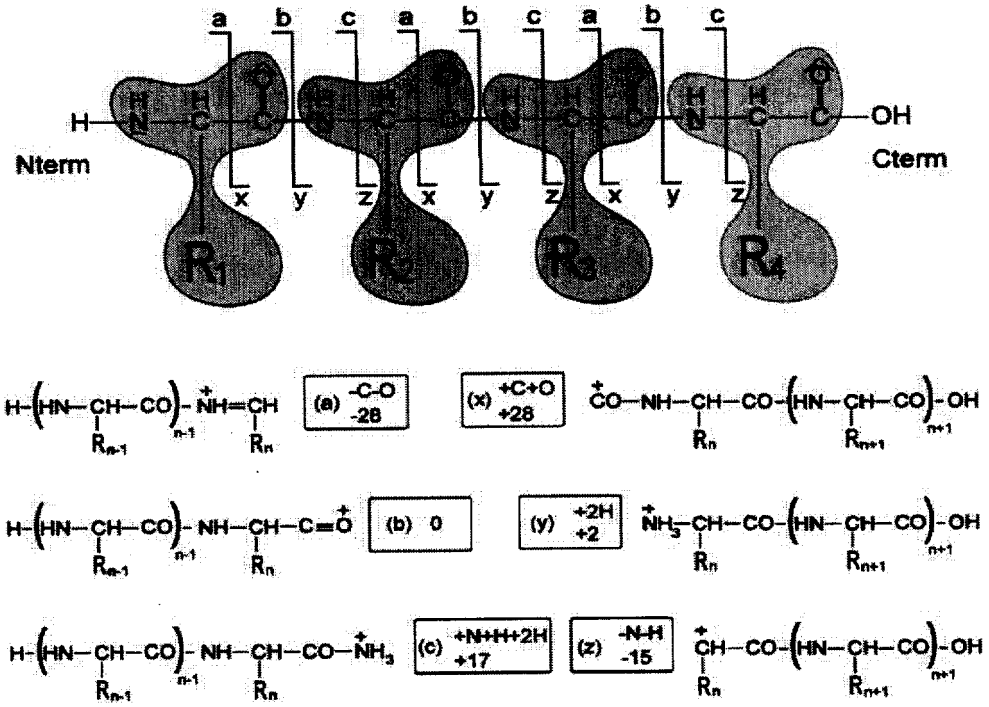


FIGURE 2. This figure shows different possibilities of fragmentation for a tetrapeptide (**top**) and the produced fragment structures (**bottom**). Ideal *N*-terminal fragments are computed by adding the mass of the *N*-terminal group (H) and all amino acid nominal masses before the cleavage position. Ideal *C*-terminal fragments are computed by adding the mass of the *C*-terminal group (OH) and all amino acid nominal masses after the cleavage position. Each ion type is characterized by an offset that represents the mass difference in Daltons between the observed mass and the corresponding *N*- or *C*-terminal ideal fragment. For example, the *b*-ion type offset is 0 (Da), because the mass of a *b*-ion type exactly corresponds to an ideal *N*-terminal fragment, whereas an *a*-ion type offset is −28 (Da), because *a*-ions lose a carbonyl and an oxygen atom compared to an ideal *N*-terminal fragment. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

as "bad" quality spectra. Understanding the rules underlying the gas-phase peptide dissociation is, therefore, important to improve MS/MS identification software.

D. Tandem Mass Spectrometry Spectra

As described above, MS/MS spectra represent peptides that are produced by proteolysis of a protein prior to MS/MS analysis. Most often, the favorite enzyme is trypsin, because of its property to produce peptides with basic *C*-termini that tend to trap a charge, prominently yielding *y*-ion series. In addition, the median length of the peptides produced by trypsin is approximately 15 residues (Tabb et al., 2003c), with a corresponding mass of approximately 1,600 Da that suits well the optimal operating mass range of many mass spectrometers (typically up to 4,000 Da for most instruments). After isolation and fragmentation of one peptide type, the masses of the fragments are measured and reported as a raw spectrum. The latter is composed of a continuous signal and needs to be processed to transform the signal into discrete values. This is called "peak detection." The processed MS/MS spectrum is finally composed of the precursor mass and charge state, as well as of a list of peaks. Each peak is characterized by two values: a measured fragment mass-to-charge ratio (m/z) and an intensity value that represents the number of detected fragments. The number of peaks composing an MS/MS spectrum varies from approximately ten to several hundreds depending on the peptide length, the fragmentation quality, the mass spectrometer type and the parameters used to extract the peaks from the raw spectrum. Interpreting an MS/MS spectrum is not straightforward, as the measured m/z of the peaks depend of many parameters, such as their ionic type, their number of charges and their isotopic distribution. In addition, a fragment can lose specific molecules (like ammonium and water molecules), or carry supplementary molecules (such as post-translational modifications of amino acids). Also, the calibration and internal error of the mass spectrometer affect the position of a peak. Finally, certain peaks may represent noise and disrupt the spectrum interpretation by diluting informative peaks, whereas some fragmentation positions may not be represented at all. Noise peaks include "true" noise peaks—i.e., peaks due to a contaminant—and "false" noise peaks that originate from the peptide but will not be able to be correctly interpreted by the subsequent automated identification procedure. Examples of such false noise peaks are isotopic peaks (peaks that represent the

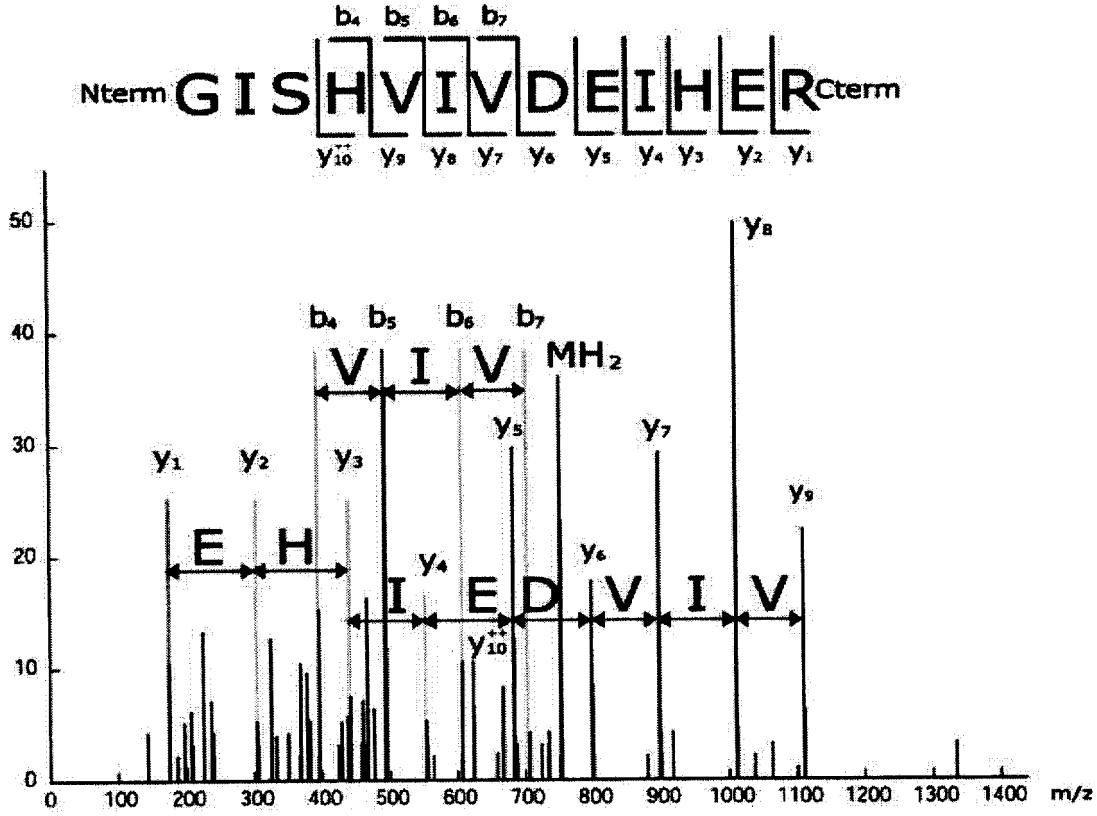


FIGURE 3. An annotated MS/MS spectrum obtained from the peptide GISHVIVDEIHER. Information about the peptide sequence can be inferred from peak differences. Peaks that are not labeled should nevertheless not be systematically associated with noise. Many of them originate from the peptide sequence and could be correctly interpreted by taking into account more ion types.

second or third isotope, but not the first one), internal rearrangements of fragments, internal fragments, or simply peaks whose mass error is greater than the mass error considered by the identification program.

Spectra obtained by tandem MS contain series of peaks that come from successive fragmentation positions in the peptide sequence. This represents a key property of MS/MS spectra, since information about the peptide sequence can be inferred from the mass differences of peaks (Fig. 3). This is the property that makes the MS/MS spectra suitable for *de novo* sequencing.

IV. TANDEM MASS SPECTROMETRY IDENTIFICATION

Tandem mass spectrometry identification consists in correlating peptides present in a sample with their corresponding theoretical amino acid sequences obtained from a protein or genomic database. Obviously, while the basic principle of identification is trivial, the procedure, once applied to real data, becomes quite complex due to the enormous complexity of the proteome. Numerous events may occur before, during, and after the synthesis of proteins, including RNA alternative splicing and post-translational modifications, like signal sequence processing or addition of chemical molecules (Appel & Bairoch, 2004). These events, referred to as pre-, co- and post-translational modifications, form many pitfalls for identification algorithms. As a matter of fact, every situation in which the peptide sequence does not exactly correspond to any candidate peptide from the database will call for special identification strategies. This situation occurs in particular (a) when searching in homologous databases, (b) when the peptide has undergone non-specific proteolytic cleavage, (c) when the peptide contains chemical, pre-, co- or post-translational modifications, (d) in case of sequence polymorphisms, and (e) when there are errors in the database sequences. In addition, other reasons can hamper the identification process: a spectrum may originate from a non-peptide contaminant or may be too noisy; it may originate from multiple peptides with the same *m/z*; the precursor mass may be incorrectly interpreted (e.g., because of incorrect precursor charge assignment); unusual fragmentation patterns can disturb the identification algorithms, for example due to the non-availability of a mobile proton. Luckily, software tools have been

adapted for each of these issues, although none of them are currently able to handle all issues at once. Some are specialized in reducing the number and complexity of MS/MS spectra while increasing their quality; others have been specifically designed to handle unexpected modifications or mutations; some split the identification into several stages and combine different approaches. Finally, identification tools based on similar strategies may use different scoring schemes and then give different results. This diversity also complicates the choice of a given tool. The next sections of this review shall detail different strategies adopted for MS/MS identification.

A. Data Preprocessing

The MS/MS spectra used for identification are the result of automated signal processing algorithms that transform the continuous signal of raw spectra into generic lists of peaks. Peak detection software include, with more or less success, peak centroiding, noise filtering, calibration, deisotoping, and deconvolution. This low-level preprocessing is a key step that can noticeably influence the outcome of the identification (Gentzel et al., 2003). Any proteomic project involving MS/MS data should, therefore, have the opportunity to access the raw data and to process them with most appropriate algorithms. Unfortunately, this is not systematically possible. Raw data are in most cases stored in a proprietary format and processed directly by the instrument's software, even though the information contained in the raw data is of considerable interest. It would thus be best if manufacturers made them available, not only for preprocessing purposes, but also for the wealth of information they hold. For example, raw LC-MS data may be visualized as a two-dimensional image, where the first dimension represents the retention time and the second dimension the *m/z* values. Mass spectrometry related imaging applications such as MSight (Palagi et al., 2005) may then be applied on these images, providing the scientists with new powerful means of data analysis such as: visual data navigation, assessment of data quality and experiment design, discrimination of peptides or proteins from noise, detection of artifacts or post-translational modifications, and automatic or semi-automatic differential proteome analysis.

Higher preprocessing procedures are often performed on the peak lists. While they are not all systematically applied by identification software, they generally tend to enhance the identification quality. Such procedures include removing non mono-isotopic peaks, filtering background noise, and deleting the precursor ion from the peak list, if present. In addition, other preprocessing procedures allow reducing the number of spectra. This step can be more or less important depending on the experiment's scale. Differential proteome analysis using 2-DE (Dowsey, Dunn, & Yang, 2003) or MS-imaging (Palagi et al., 2005) approaches, usually focus on a small subset of proteins that are up- or down-regulated. But this number can reach several thousands per day when chromatographic separation is directly coupled to MS/MS. It becomes, therefore, important to develop strategies that avoid wasting time and computer resources when analyzing large sets of MS/MS spectra. For example, several algorithms have been proposed to remove spectra with incorrect charge-state assignment using information from the fragment ions (Sadygov et al., 2002; Colinge et al., 2003a). Such a situation occurs in particular when the MS/MS data are obtained by low-resolution three-dimensional ion traps operating in full scan mode. Another way to reduce the number of spectra is to remove those spectra that are not likely to yield positive identification results, for example, because they come from background noise or because the spectrum presents unusual peak patterns. The algorithm 2 to 3 from Sadygov et al. (2002) uses as a quality measure the number of "matching pairs" found in the spectrum. A "matching pair" is composed of two fragment ions whose added mass is equal to the precursor mass. Spectra with less than three matching pairs are automatically discarded. FullStat, a perl script from Moore, Young, & Lee (2000) spots "bad quality" spectra by using a set of statistical descriptors—like the number of peaks or the intensity distribution—assembled into a quality scoring function that was optimized by an evolutionary algorithm. A third way to reduce the number of spectra is to group similar ones. Such a method has been described in NoDupe (Tabb et al., 2003a). The algorithm is based on the observation that spectra representing the same peptide show significant similarity. The idea is to group spectra using a similarity criterion, and to build a "consensus" spectrum for each cluster. This allows avoiding multiple analyses of a same peptide, while improving signal-to-noise ratio and mass accuracy. Nevertheless, information is lost during the clustering procedure. The similarity measure used in NoDupe is the spectral contrast angle. The procedure consists in representing each spectrum as a vector in a multidimensional space. Then, the contrast angle is computed for every pair of spectra with similar precursor masses. Angles near zero degree are found for similar spectra. Another example of similarity score applied to spectra clustering was proposed by Pevzner, Dancik, & Tang (2000). This method applies the Needleman and Wunsch sequence alignment algorithm to spectral masses. These similarity-based methods will be described in more details later, as they are also used as scoring functions in identification.

B. *De Novo* Versus Peptide Fragment Fingerprinting

Two main approaches are taken to identify an MS/MS spectrum (Fig. 4). The first one, named *de novo* sequencing, consists in inferring knowledge about the peptide sequence independently of any information extracted from a pre-existing protein or DNA database. Then, the inferred complete or partial sequences are compared to theoretical sequences using specifically developed sequence similarity search algorithms. The second major way to identify a spectrum has recently been designated as "peptide fragment fingerprinting" (PFF) (Blueggel, Chamrad, & Meyer, 2004), by analogy to PMF. In the PFF approach, spectrum analysis is performed specifically for candidate peptides extracted from a database by building theoretical model spectra from theoretical peptides and measuring the degree of similarity between the experimental spectra and the modeled ones. For a given scoring function, the highest scoring theoretical peptide is then taken as the one amongst all candidates that best represents the experimental spectrum.

The various methods described in literature fit either the first category (*de novo* sequencing), or the second one (PFF).

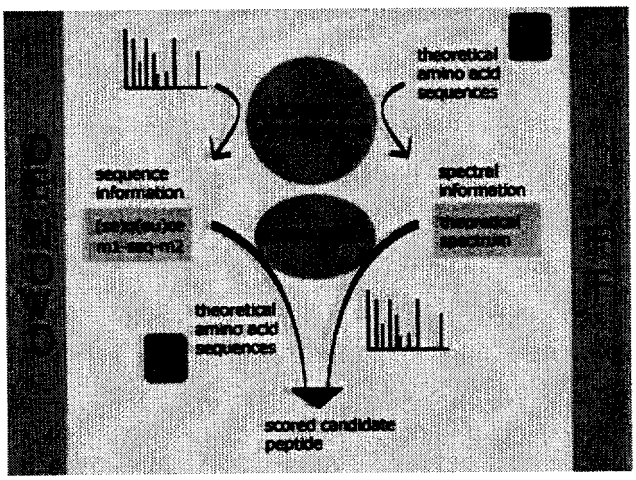


FIGURE 4. Conceptual representation of *de novo* and PFF approaches. Both methods contain a knowledge extraction phase and a comparison phase. In the *de novo* approach, knowledge is directly extracted from the spectrum as *de novo* complete or partial sequences that are then compared to theoretical sequences using a sequence similarity search algorithm. The PFF approach extracts knowledge from the database by building model spectra from the theoretical sequences and then compares these models to the experimental spectra.

Sometimes they combine both techniques. The next sections discuss the two approaches.

C. Methods Based on *De Novo* Sequencing

De novo sequencing-based identification starts by inferring sequence information from the experimental MS/MS spectrum. Before the 1990s, when protein and genomic databases were still at an embryonic stage, the obtained *de novo* sequences were used to design oligonucleotide probes to clone genes of interest. But with the growing number of sequenced genomes, software tools have been conceived to correlate *de novo* sequences with theoretical sequences.

Since *de novo* methods do not use database information during the spectrum interpretation, the search space is the set of all possible sequences that can be represented by the spectrum without any other restriction than the peak disposition. Due to the size of this search space, *de novo* sequencing methods are disadvantaged compared to PFF methods. They require spectra of higher quality with smaller fragment errors and a more or less continuous signal. Despite these disadvantages, *de novo* methods overcome PFF methods in certain cases: for example, when identifying a peptide from an organism whose genome is still not or only partially sequenced, by looking for related proteins in other species; or when using non-curated databases (such as EST databases), because such databases may more easily contain errors than curated ones. More generally, a *de novo* approach should be attempted each time a PFF approach fails in identifying a spectrum with high confidence. Also, *de novo* sequencing algorithms are not hindered by the presence of mutations. The extracted sequences "naturally" contain the mutations, which are then handled by the similarity search algorithm in allowing mismatches between the *de novo* sequences and the database sequences. But *de novo* sequencing algorithms are not well adapted to deal with the presence of modifications on the peptide. One solution is to add, for each putative modification, the corresponding modified amino acid mass to the pool of existing amino acids. But this results into an expansion of the search space; therefore, this solution should be used only for a couple of modifications of interest, or for chemical modifications deliberately introduced during sample preparation, such as cystein carbamidomethylation. Another solution to deal with modified peptides is to extract partial sequences that do not include the modified amino acid(s).

De novo sequencing algorithms can be separated into two classes: the first one applies a "pseudo" PFF approach using a database of random sequences, while the second one exploits the principle of peak succession to extract sequence information from the spectrum.

1. The "Pseudo" Peptide Fragment Fingerprinting Approach

Early *de novo* sequencing algorithms (Sakurai et al., 1984; Hamm, Wilson, & Harvan, 1986) consist of building a "pseudo" sequence database on-the-fly: the sequences are generated by determining all possible amino acid compositions with a total mass matching the experimental precursor mass, and then, for each composition, by determining all possible amino acid permutations. Subsequently, as for a PFF-type approach, theoretical spectra are computed from the "pseudo" sequences and common peaks between the experimental and theoretical spectrum are "counted." The theoretical sequences with the highest scores are the most likely to represent the original peptide. The main drawback of this approach is combinatorial complexity as the number of possible sequences increases exponentially with the precursor mass. This issue can be handled by using additional information, such as the kind of amino acids expected in the peptide, or the minimum and maximum expected number of each amino acid. Heredia-Langner et al. (2004) proposed to build candidate sequences using a genetic algorithm rather than systematically enumerating all amino acid combinations. In another approach by Spengler (2004), a drastic reduction in the number of "pseudo" sequences was achieved by using more accurate precursor masses and the presence of immonium ions in the spectrum. Finally, PEAKS (Ma et al., 2003) used dynamic programming to efficiently select 10000 candidate "pseudo" sequences for further scoring.

2. The Peak Succession Approach

Since the mid-1980s, the tendency has been to use an incremental approach: candidate sequences are built in an iterative way, amino acid by amino acid, until complete sequences that account for the precursor mass are obtained. During sequence building, only partial sequences whose extensions are validated by fragment ions in the spectrum are retained for further extension. In this way, large subsets of permutations are discarded from

analysis, contrary to the previous approach in which every possible sequence is systematically compared to the spectrum. This method is, therefore, much more sensitive to spectrum quality. Thus, a two amino acid gap in the spectrum (or in other words two successive non-fragmented positions on the peptide) causes the correct sequence to be discarded. Early implementations differ from each other by small variations. For example, Zidarov et al. (1990) use the precursor mass as well as information about the amino acid composition from observed immonium ions to limit the search space; Ishikawa's approach (Ishikawa & Niwa, 1986) generates a pool comprising all possible permutations of three amino acids to initiate the extension process and allow multiple amino acid extensions; in PepMatch (Yates et al., 1991), the extension is performed by sequentially subtracting the amino acids from the parent mass; SEQPEP (Johnson & Biemann, 1989) includes information from side-chain losses to differentiate leucine from isoleucine; and Scarberry, Zhang, & Knapp (1995) use a neural network to assign specific ion-types to the observed fragment ions before starting the iterative sequencing.

Bartels (1990) coined the term "spectrum graph," which clearly illustrates the principle of peak succession in MS/MS spectra. The graph structure is widely adopted and refined in several *de novo* methods (Hines et al., 1991; Fernandez-de-Cossio, Gonzalez, & Besada, 1995; Taylor & Johnson, 1997; Dancik et al., 1999). Because it is based on the comparison of peak masses, its construction first requires re-expressing the peaks as single ion type nodes (e.g., *b*-ion type). The drawback of the interpretation process is that each peak is at the origin of as many nodes as the number of ion types taken into account. Among these nodes, at most one is correctly interpreted and the remaining ones are false positives (Fig. 5). Then, nodes that differ by the mass value of one or more amino acids within a given error margin are linked by an edge which is labeled with the corresponding amino acid(s). Each path in the graph defines a sequence that is consistent with the spectrum. The graph is traversed from low mass nodes to high mass nodes following available edges. As a path exploration progresses, a score representing the adequacy between the parsed sub-sequence and the spectrum is computed. Various algorithms have been proposed for parsing these graphs. Hines et al. (1991) "recursively" explore graphs. The authors of SeqMS (Fernandez-de-Cossio, Gonzalez, & Besada, 1995) proposed to use the Dijkstra algorithm, and Chen et al. (2001) reported a dynamic programming algorithm to extract the highest scoring sequences from the spectrum graph, as well as sub-optimal ones (Lu & Chen, 2003a).

3. *Complete Sequences Versus Partial Sequences: The Sequence Tag Approach*

De novo algorithms typically try to infer the whole peptide sequence. Thus, methods based on peak succession build the inferred sequence in an iterative way until the precursor mass is reached. When a spectrum graph is used, it is typically parsed from the first node (the empty sequence) to the last node (the complete sequence). Missing fragmentation positions are handled by using combinations of two or three amino acids. But when several consecutive fragmentation positions are missing in the spectrum, or in case of unexpected modifications, the correct path is split into two (or more) sections, which are, or are not, connected by alternative paths. The extraction process may, therefore, stop before completing the sequence, or produce a sequence that contains sections made of wrong amino acids. Mann & Wilm (1994) proposed to limit the *de novo* sequencing to "islands" of consecutive ions that can generally be observed in the high mass region of spectra. They called these inferred partial sequences "sequence tags" and defined them as short stretches of amino acid sequences flanked by two "docking" masses representing the start mass (or prefix region mass) and the end mass (or suffix region mass) of the tag (Fig. 6). The idea of Mann and Wilm was to use the tags to extract potential candidate sequences from a database using a pattern matching-type algorithm and the flanking masses, and then to rescore the candidate peptides using a PFF approach. Their method deals with modifications—even unexpected ones—by allowing one of the regions to mismatch during the candidate extraction procedure. At that time, tags were extracted manually. A recent program, GutenTag (Tabb, Saraf, & Yates, 2003b) uses a similar algorithm with an enhanced scoring scheme, and extracts the tags by recursively parsing a spectrum graph. It reduces the search space by carefully preprocessing the peaks, assuming all peaks as *y*-ion types, and by limiting the sequence tag lengths. It then extracts candidate sequences from the database that match the tags and at least one of their flanking masses, independently of the precursor mass. If both flanking masses match, the experimental peptide is supposed to be unmodified and the complete theoretical sequence can be evaluated against the spectrum using a PFF approach. If only one flanking mass matches, the comparison focuses on the matching part of the sequences, without attempting to include additional information by taking into account shifts between peaks. The Popitam algorithm, a first version of which is described in 2003 (Hernandez et al., 2003), is also based on a tag approach. It has been specifically designed to identify spectra from modified and/or mutated peptides without any a priori knowledge about the expected type of modifications. Popitam borrows the spectrum graph from *de novo* methods, but it differs from *de novo* sequencing algorithms in that the graph is specifically parsed for each candidate peptide extracted from a database. Popitam's algorithm consists in searching the graph for all tags of the longest possible length that match a subsequence of the current candidate peptide. Then, the tags are combined according to compatibility rules, to build plausible spectrum interpretation scenarios. Typically, a scenario is composed of one or several tags, separated by "gaps." Using the flanking regions of the tags, Popitam evaluates if a gap contains a modification or a mutation, or if it arises from a lack of information in the spectrum. Finally, each scenario is scored with a function previously learned on a set of identified spectra using Genetic Programming (Hernandez Ph.D. thesis, University of Geneva).

4. *Database-Search Algorithms for Data Obtained by De Novo Sequencing*

Once sequence information has been extracted, the *de novo* sequence is correlated with theoretical sequences using database-search algorithms. Without surprise, there are numerous programs that propose the use of sequence information to extract theoretical peptides or proteins from a database.

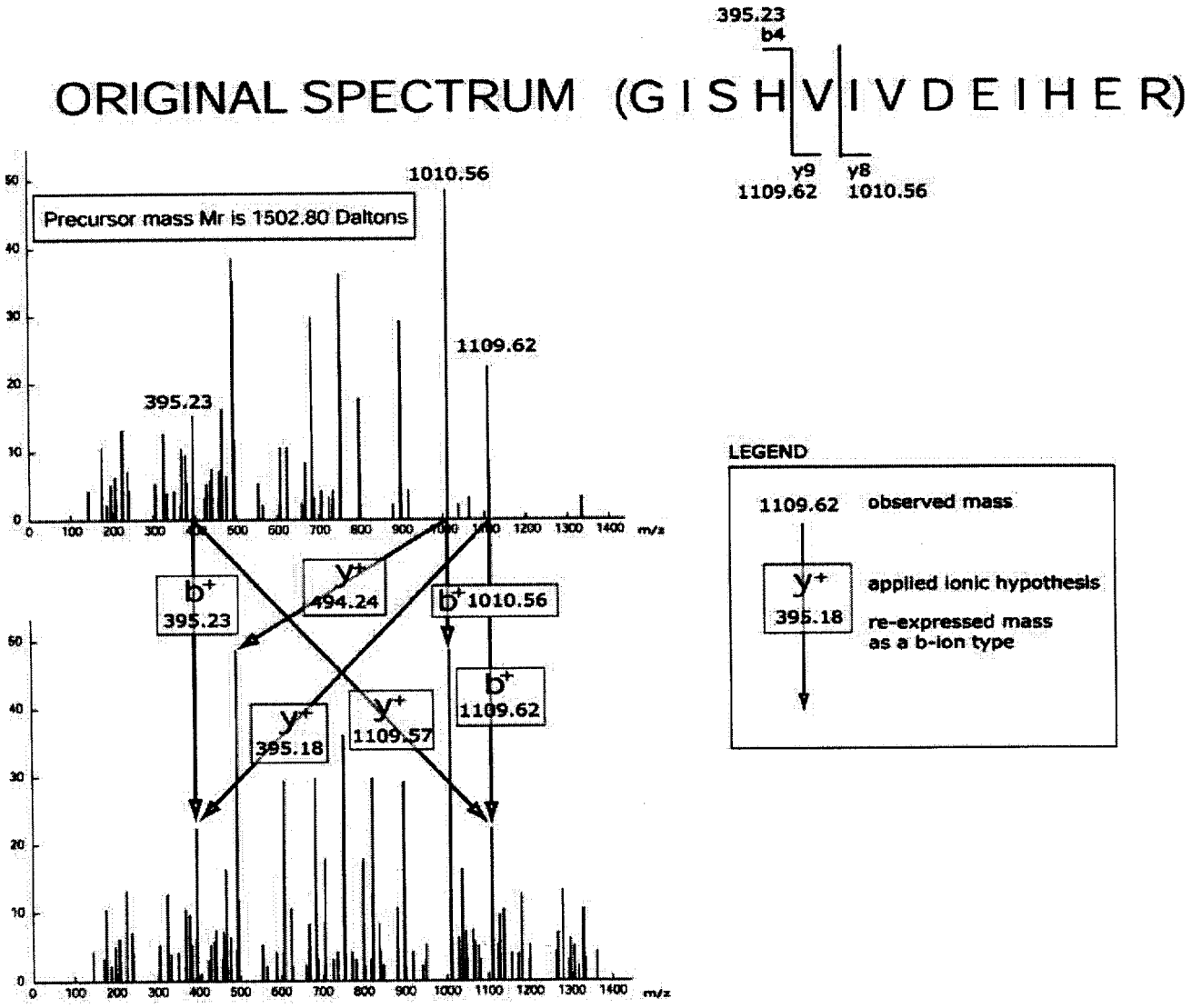


FIGURE 5. MS/MS spectrum before (**top**) and after (**bottom**) re-expression of the peaks as a single ion type (*b*-ion type). Below are formulae that allow to transform a given peak mass (denoted as obs_{Mass}) into a potential *b*-ion type (denoted as b_{Mass}), with a precursor mass M_r and an ionic hypothesis composed of an offset value o (e.g., -28 for *a*-ion types, $+2$ for *y*-ion types), a terminus side t (*N*-term or *C*-term) and a number of charges c:

$$\text{if } (t = N\text{-term}) \quad b_{\text{Mass}} \leftarrow c \times \text{obs}_{\text{Mass}} - (c-1) - o; \tag{a}$$

$$\text{if } (t = C\text{-term}) \quad b_{\text{Mass}} \leftarrow M_r - [c \times \text{obs}_{\text{Mass}} - (c-1) - o]; \tag{b}$$

In this example, only two ion hypotheses are made. The first one states that the peak actually is of b-ion type ($o = 0$, $c = 1$, $t = N$-term), and, according to Equation (a), reports its m/z as it is in the interpreted spectrum. The second one states that the peak is of *y*-ion type ($o = +2$, $c = 1$, $t = C$-term), and then computes the m/z of the potential corresponding *b*-ion type using Equation (b). The six arrows allow one to follow the interpretation process for three particular peaks. As two ion hypotheses are considered, each peak in the original spectrum is translated into two peaks in the interpreted spectrum. Between these two interpreted peaks, at least one originates from a false interpretation. For example, since peak 395.23 is the *b4*-ion of peptide GISHVIVDEIHER, the interpreted *b*-ion type peak 395.23 is correct, whereas the interpreted *b*-ion type peak 1109.57, which is based on the assumption that the original peak is of *y*-ion type, is a false positive. Similarly, peak 1109.62 being the *y9* ion, the interpreted *b*-ion type 1109.62 is a false positive, whereas the interpreted *b*-ion type 395.18 is correct. When two different peaks originate from the same fragmentation position, their correct interpretation converges to a unique m/z value. The symmetry axis on the interpreted spectrum is due to the fact that *b*- and *y*-ion types complement each other to form the complete peptide.

MS-Seq (Clauser, Baker, & Burlingame, 1999) works on a list of masses corresponding to given ion type series (the masses do not have to be all contiguous). MS-Pattern, from the same authors, performs a text-based search with a regular expression syntax. It accounts for mutations and database errors by allowing mismatches between the input sequence and the theoretical ones. PeptideSearch (Mann & Wilm, 1994) also accepts regular expressions, but in addition it can search the database with sequence tags, allowing either one of the flanking masses, or the sequence to mismatch. Several tools are based on the well known sequence similarity search algorithms BLAST (Altschul et al., 1990) and FASTA (Pearson & Lipman, 1988); MS-Blast (Shevchenko et al., 2001), for example is a BLAST-based protocol for using BLAST with *de novo* sequences, whereas FASTS (Mackey, Haystead, & Pearson, 2002) and CIDentify (Taylor & Johnson, 1997) are based on the FASTA algorithm; FASTS allows searching the database with *de novo* sequences of unknown order, and CIDentify was specifically written for database searching with *de novo* sequences obtained with Lutefisk97 (Taylor & Johnson, 1997). Since Lutefisk97 often reports several similar high-scoring *de novo* sequences, CIDentify performs the search by using a set of sequences (extracted from one spectrum) instead of only one sequence or a regular expression. OpenSea (Searle et al., 2004) is designed to align sequences reported by PEAKS (Ma et al., 2003). It initiates the alignments by matching tags composed of unambiguous amino acids to theoretical sequences, and then extends the alignments using a "breadth-first-search" approach based on mass correspondence between matching amino acids or groups of amino acids. Mascot (Perkins et al., 1999) requires each submission to start with the precursor mass of the experimental spectrum, to which additional high-confident information is associated, such as partial sequence(s), amino acid composition or ionic fragment(s). Finally, MS-Shotgun (Huang et al., 2001) and MultiTag (Sunyaev et al., 2003) have been designed to analyze the output of multiple sequence database searches with the aim to identify homologous proteins.

D. Methods Based on the Peptide Fragment Fingerprinting Approach

Peptide fragment fingerprinting methods are based on a direct comparison between the experimental spectrum and theoretical peptide sequences. By exploiting information from the database during the process of spectrum interpretation, PFF methods typically restrain the search to a subset of the spectrum space. This results into better exploration capacities and consequently, the method is generally more efficient than the *de novo* approach, as it allows interpreting the spectrum specifically (then optimally) for each candidate peptide. PFF methods rely strongly on a scoring function that evaluates the correlation between the experimental spectrum and the theoretical peptides. Many identification algorithms based on a PFF approach have been developed. Variations can be found at every step: in the way the candidate peptides are chosen from the database or the virtual spectra are modeled from theoretical amino acid sequences, as well as the way to score the similarity between the experimental spectrum and the virtual spectra or to validate the confidence in the resulting identifications.

1. Choosing Candidate Peptides from the Database

PFF algorithms typically produce candidate peptides by *in silico* digestion of theoretical protein sequences. Cleavage rules depend on the type of enzyme used for proteolysis. For example, if the enzyme was trypsin, the algorithm cleaves the protein sequence after each arginine and lysine, unless the next amino acid in the sequence is proline. Generally, an option allows skipping one or two cleavage sites to account for peptides with one or two missed-cleavages. However, it is not uncommon that MS/MS spectra originate from unspecifically cleaved peptides. In this case, the identification fails because the correct peptide will not be considered as a candidate. A solution is to present as candidates all possible peptides that match the spectrum precursor mass within a given range, without taking into account any specific cleavage rule, although this results into a significant increase of computing time. Recently, Lu & Chen (2003b) proposed a method for database searching without restrictions for the cleavage sites. They used a generalized suffix tree structure to index the whole set of database sequences, thus allowing direct access to all possible subsequence of any protein without enzyme specificity restriction.

Generally, PFF algorithms use filtering criteria such as the species or the precursor mass to reduce the number of candidate peptides. Such criteria reduce the computing time, but one should be aware that they also might prevent the correct identification of peptides; For example, when the precursor mass is incorrect due to a false assignment of the precursor charge state; when the precursor mass error is higher than the selected threshold; or when the precursor mass does not match the mass of the corresponding database sequence (e.g., due to the presence of a modification or a mutation on the peptide or to an error in the database sequence).

2. Modeling Virtual Tandem Mass Spectrometry Spectra from Theoretical Sequences

PFF methods try to score the similarity of a given theoretical sequence with an experimental spectrum. Since comparison can only be applied to similar entities, PFF methods build model spectra from amino acid sequences. This procedure, referred to as spectrum modeling or spectrum prediction, consists in predicting the fragmentation of a peptide given its amino acid sequence, its charge state, as well as experimental conditions. The aim is to build a spectrum model that approaches the corresponding experimental one. Most of the identification tools use basic rules for spectrum prediction: each cleavage position in the theoretical sequence is translated into several expected peaks according to a list of possible ion types, charges and molecule losses. So far little attention has been given to modeling the intensity of the peaks, or to measuring the influence of neighboring amino acids. Recently, several authors carried out work aimed at characterizing sequence-dependent fragmentation patterns using statistical observations from identified MS/MS spectra. For example, Kapp et al. (2003) measured to what extent specific amino acid residues promote or dampen the cleavage on their *N*- or *C*-terminal side, and they analyzed the influence of basic residues on fragment ion peak intensities. Elias et al. (2004) built from a set of identified

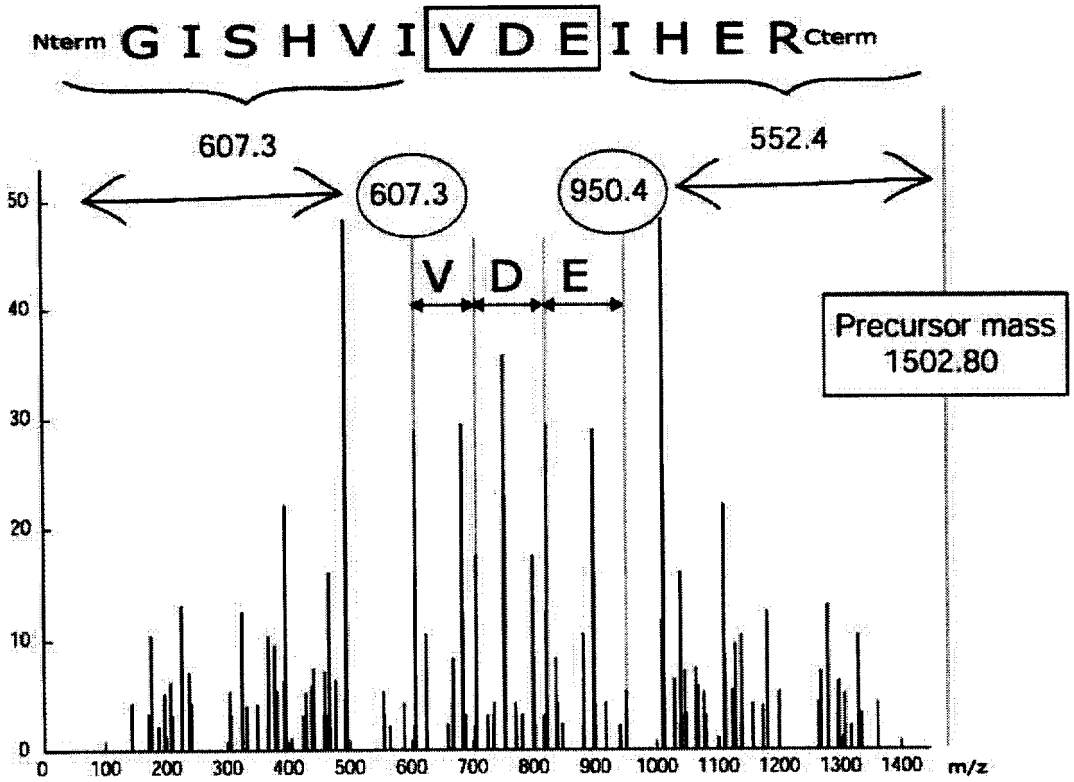


FIGURE 6. A sequence tag inferred from an interpreted spectrum (all peaks are considered as *b*-ions). The tag is composed of two flanking masses (circled) and a sequence. Masses of the prefix and suffix regions are easily computed from the flanking masses and from the precursor mass.

MS/MS spectra, a probabilistic decision tree and use it to estimate the intensity distribution of a fragment ion given an extended list of peptide and fragment attributes. As a last example, Zhang (2004) developed a mathematical model based on classical kinetic rules and on the mobile proton hypothesis. There is little doubt that such studies will be of great help to improve the performance of spectral interpretation methods by increasing the precision of spectrum modeling. They will also allow scoring functions to be more efficient by incorporation of additional knowledge on the fragmentation process.

3. Scoring the Similarity between Experimental and Theoretical Spectra

In this section, we shall discuss how PFF algorithms score the similarity between an experimental and a theoretical spectrum. The scoring is principally based on the mass correspondence between the two spectra. For example, PEP_PROBE (Sadygov & Yates, 2003) derives its scoring from a basic "shared peak count" (SPC). Sequest (Eng, McCormack, & Yates, 1994) applies a cross-correlation function based on the Fourier transforms of the spectra. Cross-correlation functions are typically applied to measure similarities between two signals. They amount to compute the sum of the products of corresponding pairs of points for multiple phase differences τ. If the two spectra are the same, the correlation function maximizes at $\tau = 0$. Sequest reports as similarity score the value of the cross-correlation when $\tau = 0$ minus the mean of the cross-correlation when τ varies from -75 to $+75$. A third type of score that accounts for mass correspondence between two spectra is the spectral angle contrast method, which was implemented in Sonar (Field, Fenyo, & Beavis, 2002) and GutenTag (Tabb, Saraf, & Yates, 2003b). In this method, both spectra are represented as vectors in a N-dimensional space, N being a number of mass intervals. The similarity measure is the cosine value of the angle between the vectors. Actually, all these similarity measures are very close to each other, as Fu et al. (2004) recently pointed out. Additional information, like probabilities of ion series or the number of consecutive fragment matches, complement the "peak matching" similarity measure. Thus, Fu et al. (2004) extended the spectral angle contrast method to make use of the correlative information among fragments.

Most scoring schemes are cast into a probabilistic framework. Mascot (Perkins et al., 1999) evaluates the probability P that the observed similarity score occurs by chance and reports as final score $-10\log(P)$. PEP_PROBE (Sadygov & Yates, 2003) follows a similar approach using a hypergeometric distribution. SCOPE (Bafna & Edwards, 2001), ProbId (Zhang, Aebersold, &

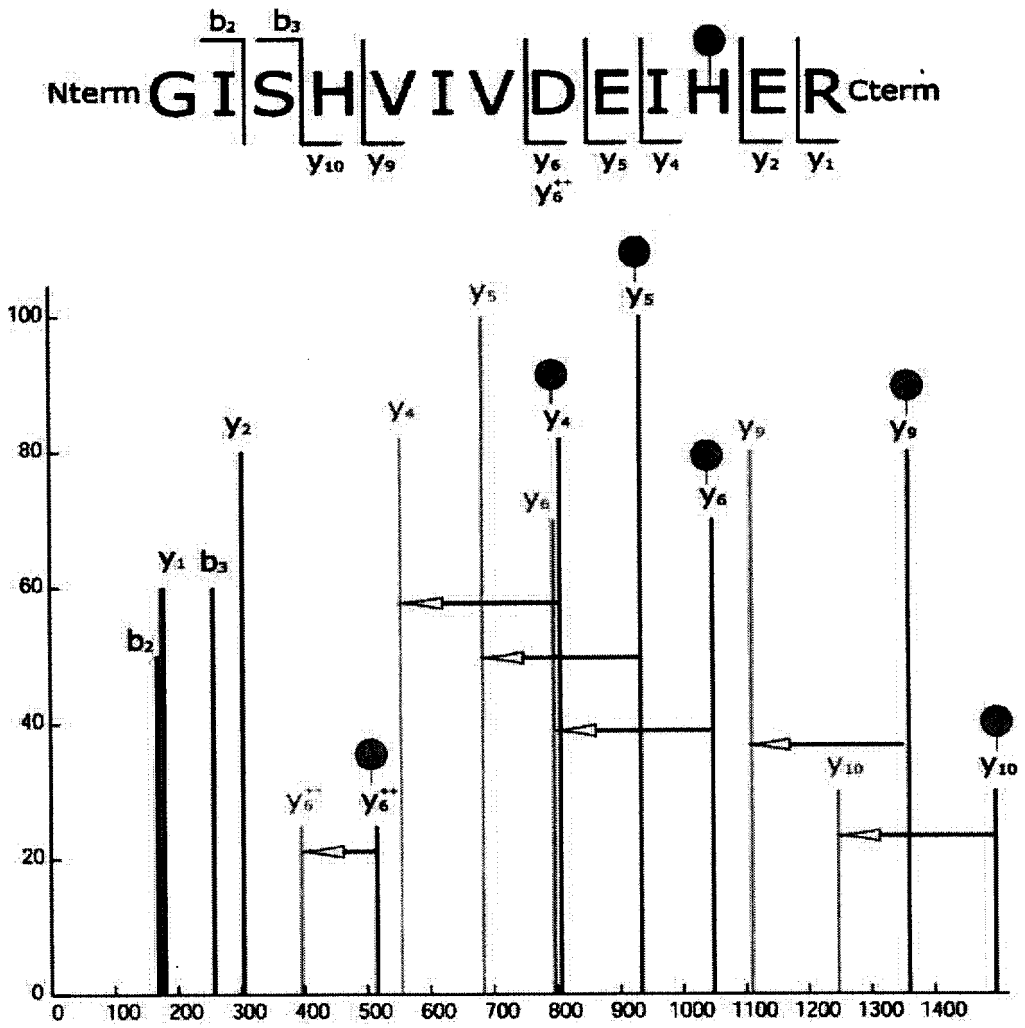


FIGURE 7. A theoretical example of a spectrum originating from a covalently modified peptide. A 250 Da modification was simulated on the second histidine. Peaks marked with a circle represent fragments that carry the modification and are thus shifted by a delta value from their expected position (gray peaks) computed from the unmodified peptide sequence. It should be noted that the shifted peaks are not necessarily grouped in the spectrum and that the delta mass is not constant, since it depends on the number of charges in the fragment. Moreover, according to the type of modification or mutation, peaks may be shifted either to the left or to the right.

Schwikowski, 2002), and Phenyx (based on the Olav scoring system) (Colinge et al., 2003b) also use probabilistic models incorporating multiple components. Thus, for example, Phenyx includes information about matching fragments, mass errors, ion series, peptide amino acid composition, presence of modifications and number of missed cleavages. In this way, information present in the experimental spectrum is more extensively exploited, thus reducing the number of false positive identifications.

A crucial question is how PFF methods handle modifications or mutations on peptides. Figure 7 illustrates how a given post-translational modification may appear in a spectrum. We simulated the presence of a hypothetical 250 Da modification on the histidine of peptide GISHVIVDEIHER. The spectrum was obtained using an in-house utility tool designed to build spectra with specified modifications given an entry peptide sequence and arbitrarily chosen ion-series and intensities. When comparing such a spectrum with its corresponding non-modified theoretical peptide, a certain number of peak matches are lost (on the average 50% of the masses are shifted), reducing the confidence in the identification score. PFF methods have two possibilities to deal with this problem: either they include the modifications and mutations into the modeled spectra, or they use a scoring function able to account for shifts in peaks. The different score functions discussed above have not been designed to handle shifts in peaks. Therefore, software tools that use these kind of functions have to virtually modify the theoretical peptides before the comparison

procedure, so that the modeled spectrum of the correct peptide can closely reflect the experimental one. These methods require the user to specify a list of modifications that one anticipates to appear on the peptide. Certain tools allow the user to define his/her own types of modifications. Unfortunately, because of combinatorial complexity, such an approach limits the number of modifications and mutations that can be considered. In these cases, a solution is to reduce the number of candidates. For example, Phenyx (Colinge et al., 2003b), Mascot (Creasy & Cottrell, 2002), TANDEM (Craig & Beavis, 2004), and VEMS (Matthiesen et al., 2004) propose a two-pass analysis strategy. The first pass is used as a filter and allows the isolation of proteins that are likely to be represented in the experimental mixture. Then a second pass is performed with loose parameters on a limited protein database built from the potentially identified proteins. During the second pass, more modifications can be considered, as well as possible mutations and unspecific cleavages.

Taking into account unexpected modifications with PFF approaches is a challenge. Very few algorithms have been designed to specifically handle unexpected modifications or mutations during the comparison phase. The previously described algorithm Popitam (Hernandez et al., 2003) applies a tag-oriented database-guided spectrum interpretation method. PEDANTA (Pevzner, Dancik, & Tang, 2000) works differently by extending the PFF matching concept. The method consists of aligning the masses of the theoretical and experimental spectra by storing all possible matches (without considering mass similarities) in a matrix, and then searching for the path that best explains the similarity between both spectra. For a better understanding of this algorithm, it is convenient to represent the matrix as a chart delimited by two scaled axes. The first axis represents experimental masses and the second theoretical ones. Each possible intersection between experimental and theoretical masses is marked with a dot. The number of dots that are covered by the diagonal starting from the origin of the two axes is the SPC score. PEDANTA searches the diagonal path in the matrix that covers the largest number of points allowing a small number of shifts between diagonals. In short, the algorithm amounts to split one of the two spectra so as to compare one spectrum with subregions of the other one.

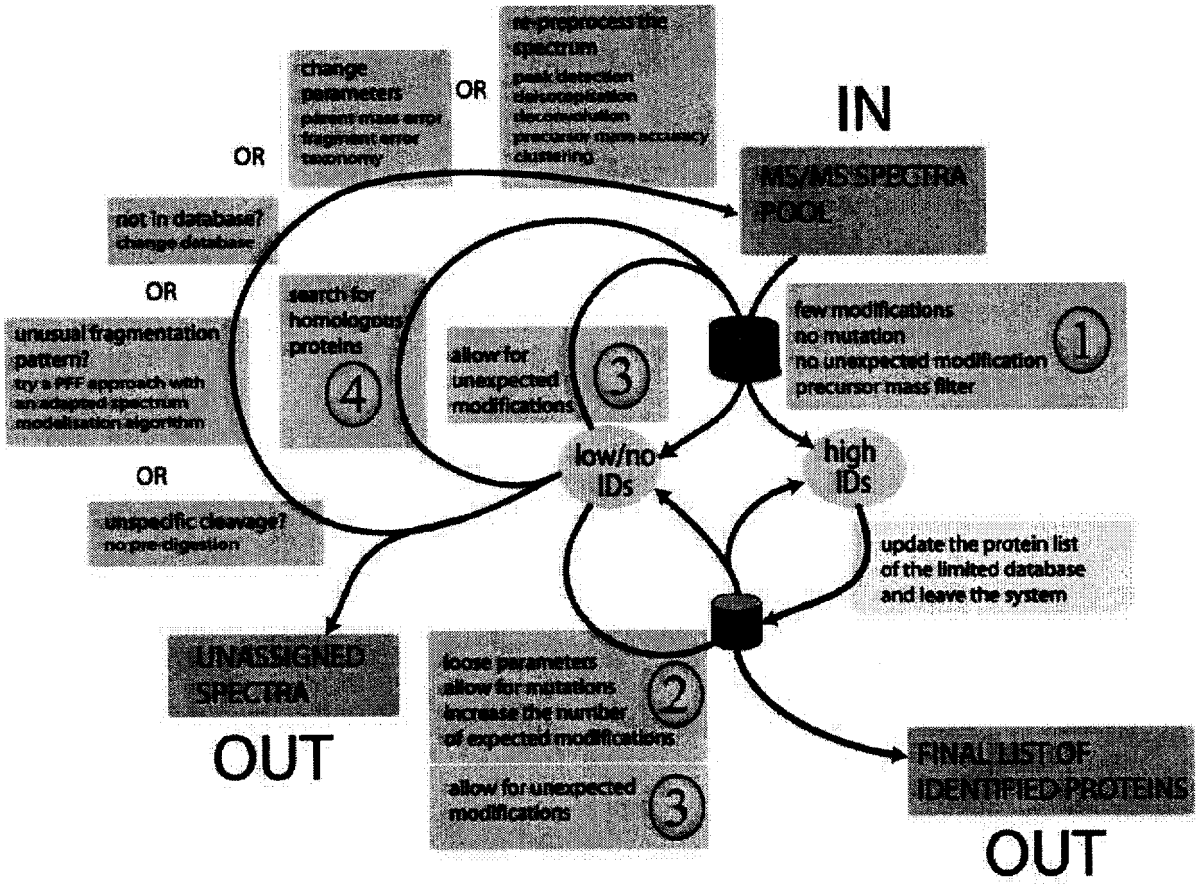
E. Statistical Validation of Results

An MS/MS scoring schema should provide good discrimination between true and random matches, i.e., a low number of both false positives (high specificity) and false negatives (high sensitivity). Many commonly used identification scores, however, are only interpretable by experts and their threshold values often depend on experimental conditions under which the spectra were acquired, the searched sequence database, as well as physico-chemical properties of the peptides or amino acids neighboring the fragmentation point (Kapp et al., 2003). Also, different search engines use different scoring functions, which, makes the comparison between their results cumbersome. Statistics can alleviate these problems substantially: it transforms the various scores into interpretable and unified probability scores, it allows considering database features and experimental conditions, and it provides a framework for incorporating supplementary information in a consistent way.

The task to decide whether a match is a random or a true one is a classification task, and the quality of the score as well as the classification method will define the quality of its outcome. If the score is one-dimensional and if the probabilities for true and random matches are known, the Neyman–Pearson lemma (Ewans & Grant, 2001) can be applied by simply calculating the ratio of the probability of a true match versus the probability of a random match. A threshold is defined either empirically or based on the prior class probabilities (percentage of identifications expected in a database) and on the costs of false positive and false negative identifications. If the ratio is higher than the threshold, then a match is classified as true. The probabilities can be determined in various ways. In the Olav scoring that is implemented in Phenyx, Colinge et al. (2003b, 2004) use parametric models, where the parameters for the true match model are learned from a training set of identified and manually validated spectra, and those for the random model are learned from a set of random sequences. Keller et al. (2002) took an unsupervised learning approach, which works for large-scale datasets. They used the fact that the random and true probability distributions of their score were close to gamma and Gaussian distributions respectively, and applied an expectation-maximization (EM) algorithm to fit these distributions to the data. Sadygov & Yates (2003) explicitly calculated the random match distribution for a simple score: the number of matching fragment masses. However, for more powerful scoring functions, exact analytical expressions for the score distribution have not yet been obtained.

If the score consists of several sub-scores, such as the combination of scores from several search engines, multidimensional classification theory (Hastie, Tibshirani, & Friedman, 2004) must be used. Keller et al. (2002) used discriminant function analysis to find the linear combination of SEQUEST scores that gives the best discrimination between random and true matches. A Naïve Bayes approach, where a set of scores is considered independent, was chosen by Colinge et al. (2003b). Their scores evaluate the number of matching and non-matching peptides, peak intensities, subsequent fragment matches, and other features. Other groups used support vector machines (Anderson et al., 2003) or neural networks (Lokhov et al., 2004). To be consistent with the one-dimensional approach, it is preferable to use classifiers, which also yield the probabilities of a classification being true or false.

Expressing the confidence in a match in terms of probabilities presents several advantages. First, it provides a unified approach, which does not formally depend on the scoring method. Various methods exist to calculate the confidence: probability ratios (Zhang & McElvain, 2000; Colinge et al., 2003b; Havilio, Haddad, & Smilansky, 2003), probabilities of true matches applying a Bayesian formula (Zhang, Aebersold, & Schwikowski, 2002; Nesvizhskii et al., 2003), and random match probabilities or *P* values (Colinge et al., 2003b; Fenyo & Beavis, 2003) (Yates, Eng, & McCormak, 1995; Perkins et al., 1999; Sadygov & Yates, 2003). *P* values are useful if only the probability distribution of the random matches is known, which is the case when too small a training data set is available to obtain the true match distribution. If one works with a large number of candidate sequences, the vast majority of these sequences will produce random matches, which form the random match distribution. Otherwise, a smaller database of random sequences



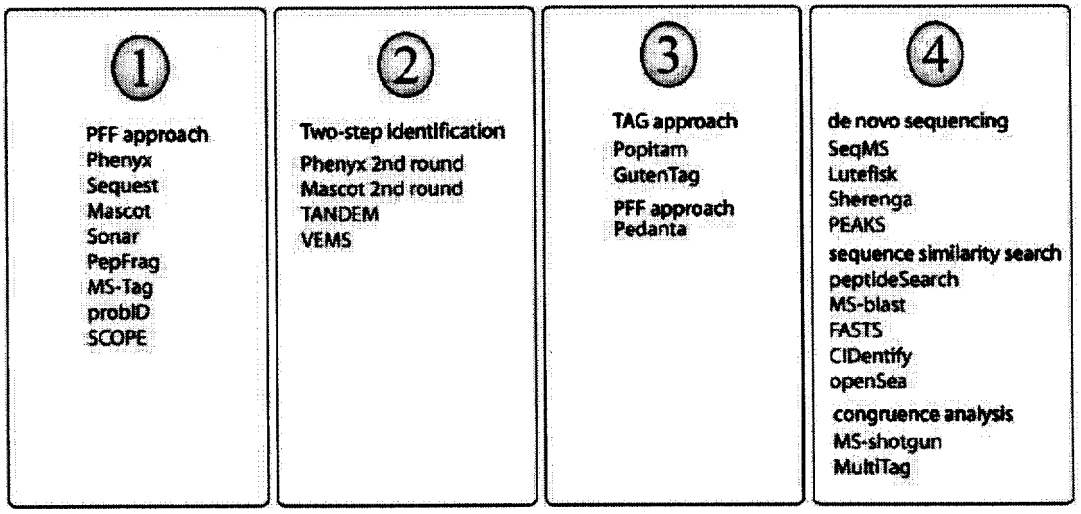


FIGURE 8. A cyclic routine integrating several approaches for MS/MS identification. Possible types of algorithms are given for the different strategies. Note that some of them might not be available or not be adapted to automation. [Color figure can be viewed in the online issue, which is available at www.interscience.wiley.com.]

can be used for this purpose (Colinge et al., 2003b). The calculation of *P* values is subject to various discussions in the statistical literature. An easy approach is to fit a known probability distribution to the data and to calculate the *P* value by means of standard formulas or tables. Other non-parametric approaches, which circumvent fitting a standard distribution and calculate the *P* value directly from the data using extreme value theory, have also been used (Fenyo & Beavis, 2003). The second advantage of using statistical scores is that they are normalized with regard to random matches. This makes the score more objective than just considering an isolated value, since it tests whether this score is significantly different from random matches. Third, the Bayesian framework allows, at least formally, the incorporation of supplementary information that may improve identification (Keller et al., 2002; Nesvizhskii et al., 2003). to control the number of false positives, the database size can be incorporated into probabilistic scores. *P* value thresholds can, for example, be adjusted to the database size by the Bonferroni correction, which simply divides the threshold by the database size. In more complex situations, where smaller databases with many homologous sequences are used, more refined techniques from multiple testing theory have to be considered (Ewans & Grant, 2001).

Probability distributions can vary significantly across experiments and a robust scoring schema should adapt to these varying conditions. Keller et al. (2002) used the EM algorithm for this purpose, assuming that the shape of the distributions remains the same, and only their parameters such as mean and standard deviation shift from experiment to experiment. Colinge et al. (2003b) realized that their probability ratio score distribution, which was calculated from a set of random sequences, was significantly different from MS/MS spectrum to spectrum. Calculating the random score distribution and corresponding *P* value threshold for every spectrum, largely improved the robustness of their method.

V. OUTLOOK

In this review, we depicted the issues that can hamper a confident correlation of a spectrum with a theoretical peptide sequence. We analyzed and classified different approaches according to their aptitude to handle these issues, highlighting their main advantages and drawbacks, but we did not supply a global identification strategy for large-scale analyses. Proteomic experiments often collect large amounts of data and consequently ask for adapted processing strategies. For experiments aimed at cataloging as many as possible proteins present in a sample, PFF approaches are undoubtedly the most adapted, providing the sample belongs to an organism with a sequenced genome. The use of additional algorithms to preprocess the spectra (Section "Data Preprocessing") or to post-validate the results (Section "Statistical Validation of Results") should of course not be neglected. In case where the search has to be applied in a database of homologous sequences, *de novo* approaches followed by error-tolerant alignment should be chosen. If the aim is to highlight the presence of specific modifications in the identified proteins, a PFF approach with expected modification search is the correct choice. If the goal is to characterize a mutation or a modification that could explain an observed dysfunction, or to direct the analysis towards the discovery of unknown post-translational modifications, an "open search modification" approach should be applied together with a PFF approach in a two-step analysis procedure. Ideally, different algorithm types should be triggered according to spectrum quality criteria and scores obtained by different algorithms should be combined. For example, the Proteinscape platform (Chamrad et al., 2003) combines the results of several PFF engines as a meta identification score and can trigger a PTM explorer module for unexplained high-quality MS/MS spectra. We are convinced that improvement in MS/MS identification can arise from a judicious use and combination of complementary algorithms involving tactics mixing, task sharing, search space splitting, and result compilation. Why not imagine a cyclic identification system in which several identification strategies would be coherently used according to criteria like the number of spectra to process, their quality, the size of the searched database, the type of analysis to carry as well as scores obtained by previous identification attempts. Figure 8 shows an example of a cyclic global MS/MS identification strategy. The spectra enter the system and remain there until a given criterion is met (typically, a confident identification or on the contrary, a definitively non-identifiable spectrum). As they circulate inside the system, the spectra are annotated to keep trace of the various paths they followed. The annotations, associated to other algorithms, will help the system decide what way a spectrum should follow. As identified spectra leave the circuit, more and more time consuming methods may be applied to the remaining ones. In Figure 8, each strategy has been labeled with a number. For each of them, a choice of existing software or algorithms is indicated.

The first type of algorithms to be applied in this workflow belongs to a PFF approach (Strategy 1). The aim is to carry out a first sorting of the peptides, without considering any potential polymorphisms or post-translational modifications. According to the obtained scores, the spectra can either follow the "low/no IDs" or the "high IDs" paths. The "high IDs" are the spectra that obtain a high enough identification score to be assigned without doubt to a protein or peptide. They are used to feed a limited list of proteins that are potentially present in the sample. Once a spectrum has been classified as "high ID's," it definitively leaves the system and participates to the final list of identifications. The "low/no IDs" remain in the cycle and follow one of the subsequent paths. A first possibility consists in using a similar PFF algorithm with looser parameters and on the limited database only (Strategy 2). Algorithms based on a tag approach could possibly be used to explore the spectrum for unexpected modifications. This path, therefore, allows increasing the coverage of potentially identified proteins present in the limited database, but does not allow one to "catch" additional proteins. It is thus important to perform such an approach on the initial database too (Strategy 3). Annotations collected by a spectrum as well as decision algorithms can help in setting hypotheses and selecting appropriate algorithms, or changing the identification parameters for the spectrum. Finally, other modules, such as a module for result compilation and another one for post-validation, may complete the system.

Such integrated systems using a combination of identification algorithms and applying different strategies as a response to specific hypotheses, as well as searching several databases, are

coming to the fore as a promising solution for the identification of MS/MS spectra.

VI. CONCLUDING REMARKS

MS/MS identification is becoming a mature procedure in proteomics. Advancements are all the more impressive since research progresses on several fronts: separation techniques, mass spectrometry, computing sciences, and database development. Different groups tackle the identification issue from different angles, resulting in a bundle of specialized techniques that more efficiently cover the numerous difficulties arising during the identification process. Nowadays, particular attention is given to the identification of modified peptides, to the correlation of *de novo* peptide sequences with homologous proteins, and to spectrum modeling. Recently, new tools have combined several strategies, and multi-step identification procedures are starting to appear. This is undoubtedly a good direction to be taken for achieving better identification performances. Future work has to focus on improving scoring schemes even more, to reduce the number of false negative and false positive identifications. Moreover, the availability of newly developed tools, the emergence of open source projects, and the unification of MS/MS spectrum and database formats will hopefully boost the development of global identification systems.

LIST OF IDENTIFICATION ALGORITHMS

MS identification algorithms and URLs	
PMF	
Aldente	http://www.expasy.org/tools/aldente/
Mascot	http://www.matrixscience.com/search_form_select.html
MOWSE	http://srs.hgmp.mrc.ac.uk/cgi-bin/mowse
MS-Fit	http://prospector.ucsf.edu/ucsfhtml4.0/msfit.htm
PeptIdent	http://www.expasy.org/tools/peptident.html
ProFound	http://65.219.84.5/service/prowl/profound.html
MS/MS identification algorithms and URLs	
PFF	
Phenyx	http://www.phenyx-ms.com/
Sequest	http://fields.scripps.edu/sequest/index.html
Mascot	http://www.matrixscience.com/search_form_select.html
PepFrag	http://prowl.rockefeller.edu/prowl/pepfragch.html
MS-Tag	http://prospector.ucsf.edu/ucsfhtml4.0/mstagfd.htm
ProbID	http://projects.systemsbiology.net/probid/
Sonar	http://65.219.84.5/service/prowl/sonar.html
TANDEM	http://www.proteome.ca/opensource.html
SCOPE	N/A
PEP_PROBE	N/A
VEMS	http://www.bio.aau.dk/en/biotechnology/vems.htm
PEDANTA	N/A
De novo sequencing	
SeqMS	http://www.protein.osaka-u.ac.jp/rcsfp/profiling/SeqMS.html
Lutefisk	http://www.hairyfatguy.com/Lutefisk
Sherenga	N/A
PEAKS	http://www.bioinformaticssolutions.com/products/peaksoverview.php
Sequence similarity search	
PeptideSearch	http://www.narrador.embl-heidelberg.de/GroupPages/Homepage.html
PepSea	http://www.unb.br/cbsp/paginiciais/pepseaseqtag.htm
MS-Seq	http://prospector.ucsf.edu/ucsfhtml4.0/msseq.htm
MS-Pattern	http://prospector.ucsf.edu/ucsfhtml4.0/mspattern.htm
Mascot	http://www.matrixscience.com/search_form_select.html
FASTS	http://www.hgmp.mrc.ac.uk/Registered/Webapp/fasts/
MS-Blast	http://dove.embl-heidelberg.de/Blast2/msblast.html
OpenSea	N/A
CIDentify	http://ftp.virginia.edu/pub/fasta/CIDentify/
Congruence analysis	
MS-Shotgun	N/A
MultiTag	N/A
Tag approach	
Popitam	http://www.expasy.org/tools/popitam/
GutenTag	http://fields.scripps.edu/GutenTag/index.html

VII. ABBREVIATIONS

CE	capillary electrophoresis
HPLC	high-performance liquid chromatography
ESI	electrospray ionization
MALDI	matrix-assisted laser desorption/ionization
FAB	by fast atom bombardment
CID	collision induced dissociation
FD	field desorption
TQ	triple quadrupole
Q-TOF	quadrupole/time-of-flight
TOF–TOF	time-of-flight/time-of-flight
Q-IT	quadrupole ion trap
MS	mass spectrometry
MS/MS	tandem mass spectrometry
PMF	peptide mass fingerprinting
PFF	peptide fragment fingerprinting
SPC	shared peak count
EM	expectation-maximization

ACKNOWLEDGMENTS

We thank Luc Otten for its careful reading of the manuscript and Robin Gras for many fruitful discussions.

REFERENCES

Aebersold R, Goodlett DR. 2001. Mass spectrometry in proteomics. Chem Rev 101(2):269–295.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3):403–410.

Anderson NL, Anderson NG. 2002. The human plasma proteome: History, character, and diagnostic prospects. Mol Cell Proteomics 1(11):845–867.

Anderson DC, Li W, Payan DG, Noble WS. 2003. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: Support vector machine classification of peptide MS/MS spectra and SEQUEST scores. J Proteome Res 2(2):137–146.

Appel RD, Bairoch A. 2004. Post-translational modifications: A challenge for proteomics and bioinformatics. Proteomics 4(6):1525–1526.

Bafna V, Edwards N. 2001. SCOPE: A probabilistic model for scoring tandem mass spectra against a peptide database. Bioinformatics 17(Suppl 1): S13–S21.

Bartels C. 1990. Fast algorithm for peptide sequencing by mass spectrometry. Biomed Environ Mass Spectrom 19:363–368.

Blueggel M, Chamrad D, Meyer HE. 2004. Bioinformatics in proteomics. Curr Pharm Biotechnol 5(1):79–88.

Chamrad DC, Koerting G, Gobom J, Thiele H, Klose J, Meyer HE, Blueggel M. 2003. Interpretation of mass spectrometry data for high-throughput proteomics. Anal Bioanal Chem 376(7):1014–1022.

Chen T, Kao MY, Tepel M, Rush J, Church GM. 2001. A dynamic programming approach to *de novo* peptide sequencing via tandem mass spectrometry. J Comput Biol 8(3):325–337.

Clauser KR, Baker P, Burlingame AL. 1999. Role of accurate mass measurement (+/−10 ppm) in protein identification strategies employing MS or MS/MS and database searching. Anal Chem 71(14):2871–2882.

Clauser KR, Hall SC, Smith DM, Webb JW, Andrews LE, Tran HM, Epstein LB, Burlingame AL. 1995. Rapid mass spectrometric peptide sequencing and mass matching for characterization of human melanoma proteins isolated by two-dimensional PAGE. Proc Natl Acad Sci USA 92(11):5072–5076.

Colinge J, Magnin J, Dessingy T, Giron M, Masselot A. 2003a. Improved peptide charge state assignment. Proteomics 3(8):1434–1440.

Colinge J, Masselot A, Giron M, Dessingy T, Magnin J. 2003b. OLAV: Towards high-throughput tandem mass spectrometry data identification. Proteomics 3(8):1454–1463.

Colinge J, Masselot A, Cusin I, Mahe E, Niknejad A, rgoud-Puy G, Reffas S, Bederr N, Gleizes A, Rey PA, Bougueleret L. 2004. High-performance peptide identification by tandem mass spectrometry allows reliable automatic data processing in proteomics. Proteomics 4(7):1977–1984.

Corthals GL, Wasinger VC, Hochstrasser DF, Sanchez JC. 2000. The dynamic range of protein expression: A challenge for proteomic research. Electrophoresis 21(6):1104–1115.

Craig R, Beavis RC. 2004. TANDEM: Matching proteins with tandem mass spectra. Bioinformatics 20(9):1466–1467.

Creasy DM, Cottrell JS. 2002. Error tolerant searching of uninterpreted tandem mass spectrometry data. Proteomics 2(10):1426–1434.

Dancik V, Addona TA, Clauser KR, Vath JE, Pevzner PA. 1999. *De novo* peptide sequencing via tandem mass spectrometry. J Comput Biol 6(3–4):327–342.

Dongré AR, Jones JL, Somogyi A, Wysocki WH. 1996. Influence of peptide composition, gas-phase basicity, and chemical modification on fragmentation efficiency: Evidence for the mobile proton model. J Am Chem Soc 118:8365–8374.

Dowsey AW, Dunn MJ, Yang GZ. 2003. The role of bioinformatics in two-dimensional gel electrophoresis. Proteomics 3(8):1567–1596.

Edman P. 1950. Method for determination of the amino acid sequence in peptides. Acta Chem Scand 4:283–293.

Edman P, Begg G. 1967. A protein sequenator. Eur J Biochem 1(1):80–91.

Elias JE, Gibbons FD, King OD, Roth FP, Gygi SP. 2004. Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat Biotechnol 22(2):214–219.

Eng JK, McCormack AL, Yates J III. 1994. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. J Am Soc Mass Spectrom 5:976–989.

Ewans WJ, Grant GR. 2001. Statistical methods in bioinformatics. New York: Springer-Verlag.

Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM. 1989. Electrospray ionization for mass spectrometry of large biomolecules. Science 246(4926):64–71.

Fenyo D, Beavis RC. 2003. A method for assessing the statistical significance of mass spectrometry-based protein identifications using general scoring schemes. Anal Chem 75(4):768–774.

Fernandez-de-Cossio J, Gonzalez J, Besada V. 1995. A computer program to aid the sequencing of peptides in collision- activated decomposition experiments. Comput Appl Biosci 11(4):427–434.

Field HI, Fenyo D, Beavis RC. 2002. RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database. Proteomics 2(1):36–47.

Fu Y, Yang Q, Sun R, Li D, Zeng R, Ling CX, Gao W. 2004. Exploiting the kernel trick to correlate fragment ions for peptide identification via tandem mass spectrometry. Bioinformatics 20(12):1948–1954.

Gasteiger E, Hoogland C, Gattiger A, Duvaud S, Wilkins MR, Appel RD, Bairoch A. 2005. Protein identification and analysis tools on the ExPASy server. In: The proteomics protocols handbook. New Jersey: Humana Press. pp 571–607.

Gentzel M, Kocher T, Ponnusamy S, Wilm M. 2003. Preprocessing of tandem mass spectrometric data to support automatic protein identification. Proteomics 3(8):1597–1610.

Gras R, Muller M. 2001. Computational aspects of protein identification by mass spectrometry. Curr Opin Mol Ther 3(6):526–532.

Hamm CW, Wilson WE, Harvan DJ. 1986. Peptide sequencing program. Comput Appl Biosci 2(2):115–118.

Hastie T, Tibshirani R, Friedman J. 2004. The elements of statistical learning. New York: Springer-Verlag.

Havilio M, Haddad Y, Smilansky Z. 2003. Intensity-based statistical scorer for tandem mass spectrometry. Anal Chem 75(3):435–444.

Henzel WJ, Watanabe C, Stults JT. 2003. Protein identification: The origins of peptide mass fingerprinting. J Am Soc Mass Spectrom 14(9):931–942.

Henzel WJ, Billeci TM, Stults JT, Wong SC, Grimley C, Watanabe C. 1993. Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases. Proc Natl Acad Sci USA 90(11):5011–5015.

Heredia-Langner A, Cannon WR, Jarman KD, Jarman KH. 2004. Sequence optimization as an alternative to *de novo* analysis of tandem mass spectrometry data. Bioinformatics 20(14):2296–2304.

Hernandez P, Gras R, Frey J, Appel RD. 2003. Popitam: Towards new heuristic strategies to improve protein identification from tandem mass spectrometry data. Proteomics 3(6):870–878.

Hines WM, Falick AM, Burlingame AL, Gibson BW. 1991. Pattern-based algorithm for peptide sequencing from tandem high-energy collision-induced dissociation mass spectra. J Am Soc Mass Spectrom 3:326–336.

Huang L, Jacob RJ, Pegg SC, Baldwin MA, Wang CC, Burlingame AL, Babbitt PC. 2001. Functional assignment of the 20 S proteasome from Trypanosoma brucei using mass spectrometry and new bioinformatics approaches. J Biol Chem 276(30):28327–28339.

Hunt DF, Yates JR III, Shabanowitz J, Winston S, Hauer CR. 1986. Protein sequencing by tandem mass spectrometry. Proc Natl Acad Sci USA 83(17):6233–6237.

Ishikawa K, Niwa Y. 1986. Computer-aided peptide sequencing by fast atom bombardment mass spectrometry. Biomed Environ Mass Spectrom 13:373–380.

James P, Quadroni M, Carafoli E, Gonnet G. 1993. Protein identification by mass profile fingerprinting. Biochem Biophys Res Commun 195(1):58–64.

Johnson RS, Martin SA, Biemann K. 1988. Collision-induced fragmentation of $(M+H)^+$ ions of peptides. Side chain specific sequence ions. Int J Mass Spectrom Ion Processes 86:137–154.

Johnson RS, Biemann K. 1989. Computer program (SEQPEP) to aid in the interpretation of high-energy collision tandem mass spectra of peptides. Biomed Environ Mass Spectrom 18(11):945–957.

Jonsson AP. 2001. Mass spectrometry for protein and peptide characterisation. Cell Mol Life Sci 58(7):868–884.

Kapp EA, Schutz F, Reid GE, Eddes JS, Moritz RL, O'Hair RA, Speed TP, Simpson RJ. 2003. Mining a tandem mass spectrometry database to determine the trends and global factors influencing peptide fragmentation. Anal Chem 75(22):6251–6264.

Karas M, Hillenkamp F. 1988. Laser desorption ionization of proteins with molecular masses exceeding 10,000 Da. Anal Chem 60(20):2299–2301.

Keller A, Nesvizhskii AI, Kolker E, Aebersold R. 2002. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 74(20):5383–5392.

Kenrick KG, Margolis J. 1970. Isoelectric focusing and gradient gel electrophoresis: A two-dimensional technique. Anal Biochem 33(1):204–207.

Kitagishi T, Hong YM, Shimonishi Y. 1981. Computer-aided sequencing of a protein from the masses of its constituent peptide fragments. Int J Pept Protein Res 17(4):436–443.

Kolker E, Purvine S, Galperin MY, Stolyar S, Goodlett DR, Nesvizhskii AI, Keller A, Xie T, Eng JK, Yi E, Hood L, Picone AF, Cherny T, Tjaden BC, Siegel AF, Reilly TJ, Makarova KS, Palsson BO, Smith AL. 2003. Initial proteome analysis of model microorganism Haemophilus influenzae strain Rd KW20. J Bacteriol 185(15):4593–4602.

Lokhov PG, Tikhonova OV, Moshkovskii SA, Goufman EI, Serebriakova MV, Maksimov BI, Toropyguine IY, Zgoda VG, Govorun VM, Archakov AI. 2004. Database search post-processing by neural network: Advanced facilities for identification of components in protein mixtures using mass spectrometric peptide mapping. Proteomics 4(3):633–642.

Lu B, Chen T. 2003a. A suboptimal algorithm for *de novo* peptide sequencing via tandem mass spectrometry. J Comput Biol 10(1):1–12.

Lu B, Chen T. 2003b. A suffix tree approach to the interpretation of tandem mass spectra: Applications to peptides of non-specific digestion and post-translational modifications. Bioinformatics 19(Suppl 2):II113–II121.

Ma B, Zhang K, Hendrie C, Liang C, Li M, Doherty-Kirby A, Lajoie G. 2003. PEAKS: Powerful software for peptide *de novo* sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 17(20):2337–2342.

MacCoss MJ, McDonald WH, Saraf A, Sadygov R, Clark JM, Tasto JJ, Gould KL, Wolters D, Washburn M, Weiss A, Clark JI, Yates JR III. 2002. Shotgun identification of protein modifications from protein complexes and lens tissue. Proc Natl Acad Sci USA 99(12):7900–7905.

Mackey AJ, Haystead TA, Pearson WR. 2002. Getting more from less: Algorithms for rapid protein identification with multiple short peptide sequences. Mol Cell Proteomics 1(2):139–147.

Mann M, Hojrup P, Roepstorff P. 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol Mass Spectrom 22(6):338–345.

Mann M, Wilm M. 1994. Error-tolerant identification of peptides in sequence databases by peptide sequence tags. Anal Chem 66(24):4390–4399.

Matsuo T, Matsuda H, Katakuse I. 1981. Computer program PAAS for the estimation of possible amino acid sequence of peptides. Biomed Mass Spectrom 8(4):137–143.

Matthiesen R, Bunkenborg J, Stensballe A, Jensen ON, Welinder KG, Bauw G. 2004. Database-independent, database-dependent, and extended interpretation of peptide mass spectra in VEMS V2.0. Proteomics 4(9):2583–2593.

Moore RE, Young MK, Lee TD. 2000. Method for screening peptide fragment ion mass spectra prior to database searching. J Am Soc Mass Spectrom 11(5):422–426.

Morris HR, Panico M, Barber M, Bordoli RS, Sedgwick RD, Tyler A. 1981. Fast atom bombardment: A new mass spectrometric method for peptide sequence analysis. Biochem Biophys Res Commun 101(2):623–631.

Nesvizhskii AI, Keller A, Kolker E, Aebersold R. 2003. A statistical model for identifying proteins by tandem mass spectrometry. Anal Chem 75(17):4646–4658.

Palagi PM, Walther D, Quadroni M, Catherinet S, Burgess J, Zimmermann-Ivol CG, Sanchez JC, Binz PA, Hochstrasser DF, Appel RD. 2005. MSight: An image analysis software for liquid chromatography-mass spectrometry. Proteomics 5:2381–2384.

Pappin DDJ, Hojrup P, Bleasby AJ. 1993. Rapid identification of proteins by peptide-mass finger printing. Curr Biol 3:327–332.

Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85(8):2444–2448.

Perkins DN, Pappin DDJ, Creasy DM, Cottrell JS. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. Electrophoresis 20:3551–3567.

Pevzner PA, Dancik V, Tang CL. 2000. Mutation-tolerant protein identification by mass spectrometry. J Comput Biol 7(6):777–787.

Rabilloud T. 2002. Two-dimensional gel electrophoresis in proteomics: Old, old fashioned, but it still climbs up the mountains. Proteomics 2(1): 3–10.

Rappsilber J, Ryder U, Lamond AI, Mann M. 2002. Large-scale proteomic analysis of the human spliceosome. Genome Res 12(8):1231–1245.

Roepstorff P, Fohlman J. 1984. Proposal for a common nomenclature for sequence ions in mass spectra of peptides. Biomed Mass Spectrom 11(11):601.

Rose K, Bougueleret L, Baussant T, Bohm G, Botti P, Colinge J, Cusin I, Gaertner H, Gleizes A, Heller M, Jimenez S, Johnson A, Kussmann M, Menin L, Menzel C, Ranno F, Rodriguez-Tome P, Rogers J, Saudrais C, Villain M, Wetmore D, Bairoch A, Hochstrasser D. 2004. Industrial-scale proteomics: From liters of plasma to chemically synthesized proteins. Proteomics 4(7):2125–2150.

Sadygov RG, Yates JR III. 2003. A hypergeometric probability model for protein identification and validation using tandem mass spectral data and protein sequence databases. Anal Chem 75(15):3792–3798.

Sadygov RG, Eng J, Durr E, Saraf A, McDonald H, MacCoss MJ, Yates JR III. 2002. Code developments to improve the efficiency of automated MS/MS spectra interpretation. J Proteome Res 1(3):211–215.

Sakurai T, Matsuo T, Matsuda H, Katakuse I. 1984. Paas 3: A computer program to determine probable sequence of peptides from mass spectrometric data. Biomed Mass Spectrom 11(8):396–399.

Scarberry RE, Zhang Z, Knapp D. 1995. Peptide sequence determination from high-energy collision-induced dissociation spectra using artificial neural networks. J Am Soc Mass Spectrom 6:947–961.

Searle BC, Dasari S, Turner M, Reddy AP, Choi D, Wilmarth PA, McCormack AL, David LL, Nagalla SR. 2004. High-throughput identification of proteins and unanticipated sequence modifications using a mass-based alignment algorithm for MS/MS *de novo* sequencing results. Anal Chem 76(8):2220–2230.

Shevchenko A, Sunyaev S, Loboda A, Shevchenko A, Bork P, Ens W, Standing KG. 2001. Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology. Anal Chem 73(9):1917–1926.

Shimonishi Y, Hong YM, Kitagishi T, Matsuo T, Matsuda H, Katakuse I. 1980. Sequencing of peptide mixtures by Edman degradation and field-desorption mass spectrometry. Eur J Biochem 112(2):251–264.

Spengler B. 2004. *De novo* sequencing, peptide composition analysis, and composition-based sequencing: A new strategy employing accurate mass determination by Fourier transform ion cyclotron resonance mass spectrometry. J Am Soc Mass Spectrom 15(5):703–714.

Sunyaev S, Liska AJ, Golod A, Shevchenko A, Shevchenko A. 2003. MultiTag: Multiple error-tolerant sequence tag search for the sequence-similarity identification of proteins by mass spectrometry. Anal Chem 75(6):1307–1315.

Tabb DL, Saraf A, Yates JR III. 2003b. GutenTag: High-throughput sequence tagging via an empirically derived fragmentation model. Anal Chem 75(23):6415–6421.

Tabb DL, MacCoss MJ, Wu CC, Anderson SD, Yates JR III. 2003a. Similarity among tandem mass spectra from proteomic experiments: Detection, significance, and utility. Anal Chem 75(10):2470–2477.

Tabb DL, Smith LL, Breci LA, Wysocki VH, Lin D, Yates JR III. 2003c. Statistical characterization of ion trap tandem mass spectra from doubly charged tryptic peptides. Anal Chem 75(5):1155–1163.

Tanaka K, Waki H, Ido Y, Akita S, Yoshida T. 1988. Protein and polymer analyses up to *m/z* 100,000 by laser ionization time-of-flight mass spectrometry. Rapid Commun Mass Spectrom 2(8):151–153.

Taylor JA, Johnson RS. 1997. Sequence database searches via *de novo* peptide sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 11(9):1067–1075.

Tuloup M, Hernandez C, Coro I, Hoogland C, Binz PA, Appel RD. 2003. Aldente and BioGraph: An improved peptide mass fingerprinting protein identification environment. Swiss Proteomics Society 2003 Congress [Fontis Media], 174–176, 12th Feb, 2003.

Washburn MP, Wolters D, Yates JR III. 2001. Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19(3):242–247.

Wilkins MR, Gasteiger E, Wheeler CH, Lindskog I, Sanchez JC, Bairoch A, Appel RD, Dunn MJ, Hochstrasser DF. 1999. Multiple parameter cross-species protein identification using Multident—A world-wide web accessible tool. Electrophoresis 19:3199–3206.

Wysocki VH, Tsaprailis G, Smith LL, Breci LA. 2000. Mobile and localized protons: A framework for understanding peptide dissociation. J Mass Spectrom 35(12):1399–1406.

Yates JR III, Eng JK, McCormak AL. 1995. Mining genomes: Correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal Chem 67(18):3202–3210.

Yates JR III, Griffin PR, Hood LE, Zhou JX. 1991. Computer aided interpretation of low energy ms/ms spectra of peptides. In: Techniques in protein chemistry II. San Diego: Academic Press.

Yates JR III, Speicher S, Griffin PR, Hunkapiller T. 1993. Peptide mass maps: A highly informative approach to protein identification. Anal Biochem 214(2):397–408.

Zhang Z. 2004. Prediction of low-energy collision-induced dissociation spectra of peptides. Anal Chem 76(14):3908–3922.

Zhang N, Aebersold R, Schwikowski B. 2002. ProbID: A probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. Proteomics 2(10):1406–1412.

Zhang W, Chait BT. 2000. ProFound: An expert system for protein identification using mass spectrometric peptide mapping information. Anal Chem 72(11):2482–2489.

Zhang Z, McElvain JS. 2000. *De novo* peptide sequencing by two-dimensional fragment correlation mass spectrometry. Anal Chem 72(11):2337–2350.

Zidarov D, Thibault P, Evans MJ, Bertrand MJ. 1990. Determination of the primary structure of peptides using fast atom bombardment mass spectrometry. Biomed Environ Mass Spectrom 19(1):13–26.

Patricia Hernandez, after obtaining a license in Biology at the University of Geneva, was enrolled in the first Swiss edition of the Master in Bioinformatics, organized by the Swiss Institute of Bioinformatics in 1999. Then she started a Ph.D. thesis on peptide identification by tandem mass spectrometry in the Proteome Informatics Group of the SIB. She is now finalizing her Ph.D. work, entitled "Peptide Identification by Tandem Mass Spectrometry: A Tag Oriented Open Modification Search Method." The method is implemented as a peptide identification and characterization tool called Popitam.

Dr. Marcus Müller has been involved for several years in the development of algorithms and software for the analysis of proteomic MS data. He wrote programs for the detection of peptide signals in MALDI and ESI spectra. In many cases, such as the molecular scanner or LC-MS, mass spectra are not measured independently and their correlation was used to enhance the signal to noise ratio and to improve the identification of proteins. Further work included statistical interpretation of a PMF scoring schema and a study about protein expression patterns for biomarker detection. Currently, he has the position of a group leader for bioinformatics at the Institute of Molecular Systems Biology in Zurich, Switzerland. His current research interests are MS/MS identification, LC-MS data processing and the identification of protein complexes by means of mass spectrometry.

Ron Appel is Professor of Bioinformatics at the University of Geneva in the Department of Computer Science of the Faculty of Sciences and the Department of Structural Biology and Bioinformatics of the Faculty of Medicine. In 1998, he co-founded the Swiss Institute of Bioinformatics (SIB), a non-profit academic foundation where he serves as a member of the Council and of the Executive Board. His research group at the SIB, known as the Proteome Informatics Group, is performing research and development work on differential analysis of proteomes and automation of proteome analysis as well as integration of proteome databases. Ron Appel also serves on the editorial board of several scientific journals, notably as senior editor of Proteomics.